Data-driven personalisation of Television Content: A Survey

Lyndon Nixon[1], Jeremy Foss[2], Konstantinos Apostolidis[3] and Vasileios Mezaris[3]

[1]MODUL Technology, Vienna, Austria.
[2]Birmingham City University, Birmingham, England.
[3]CERTH-ITI, Thessaloniki, Greece.

This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s00530-022-00926-6. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use, https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.

Data-driven personalisation of Television Content: A Survey

Lyndon Nixon[1], Jeremy Foss[2], Konstantinos Apostolidis[3] and Vasileios Mezaris[3]

[1]MODUL Technology, Vienna, Austria.
[2]Birmingham City University, Birmingham, England.
[3]CERTH-ITI, Thessaloniki, Greece.

Contributing authors: lyndon.nixon@modul.ac.at; jeremy.foss@bcu.ac.uk; kapost@iti.gr; bmezaris@iti.gr;

Abstract

This survey considers the vision of TV broadcasting where content is personalised and personalisation is data-driven, looks at the AI and data technologies making this possible and surveys the current uptake and usage of those technologies. We examine the current state-of-the-art in standards and best practices for data-driven technologies and identify remaining limitations and gaps for research and innovation. Our hope is that this survey provides an overview of the current state of AI and data-driven technologies for use within broadcasters and media organisations. It also provides a pathway to the needed research and innovation activities to fulfil the vision of data-driven personalisation of TV content.

Keywords: Broadcasting, Data-driven TV, Deep Learning, Media Analysis, Media Annotation, Personalisation, Recommendation

1 Introduction

The world of television broadcasting has changed fundamentally in the past decades. The Web introduced the most significant change for TV content distribution since the introduction of colour. As soon as broadband connections were available, audiovisual content could be streamed to Internet-connected

devices and TV gradually shifted online. The combination with Web technologies, especially unicast IP, meant "broadcast" did not have to be "multicast", and individual viewers could choose what they want to watch when they want to watch it. We are now entering the next phase of the future of television where not only the choice of program can be personalised to the individual viewer but the content of that program can be adapted to each one. In the broadcasting industry, data-driven personalisation will be critical to compete with the ever-growing number of amateur and semi-professional content producers online.

Data-driven personalisation refers to the adaptation and delivery of TV content to viewers according to their interests and preferences as well as connected to trending and emerging topics of interest among the target audience. It presents the opportunity to both gain and retain an audience in the light of persistent online competition due to the increased relevance and usefulness of the content to the viewer. It is made possible by the digitisation of media resources and the growing production and use of data in describing, understanding, processing and re-using those resources along the whole media value chain (from content creation through to delivery). Despite the presence of this data being created by various software tools as content moves along the media value chain, the full value that is possible in the re-use of this data for later steps in the value chain is not being exploited. Typically data produced at one step is meant only for that same step, or maybe the output from one step acts as the input for the next step. However, e.g. data from the content production process is not readily available for exploitation in the content delivery step. Furthermore, effective personalisation of content needs data beyond the "low level" technical metadata of content processing tools. It needs annotation in terms of higher-level "semantic" concepts which can be understood by computer systems due to their descriptions within knowledge representations (such as ontologies or graphs). The presence of semantic descriptions of resources supports improved resource management, discovery, combination and re-use, especially when combined with state-of-the-art Artificial Intelligence (AI) techniques, such as deep neural networks. In short, data-driven personalisation of TV content can only be a reality when the latest innovations in AI and data-centric technologies can be taken up by broadcasters, applied to their media collections and supported by a new set of software applications that can exploit data (aided by shared standards and specifications) to re-purpose and personalise their audiovisual content according to their audience. In this survey paper, we look at the state-of-the-art in the technological innovations needed for data-driven personalisation of TV content, the software applications that support the uptake of these technologies within media organisations, standards and best practices around these applications and technologies, and identify remaining limitations and gaps for research and innovation. Our hope is that this survey can encourage both more uptake of current solutions within the media industry as well as point to the research and innovation activities still needed to fulfil the vision of data-driven personalisation of TV content.

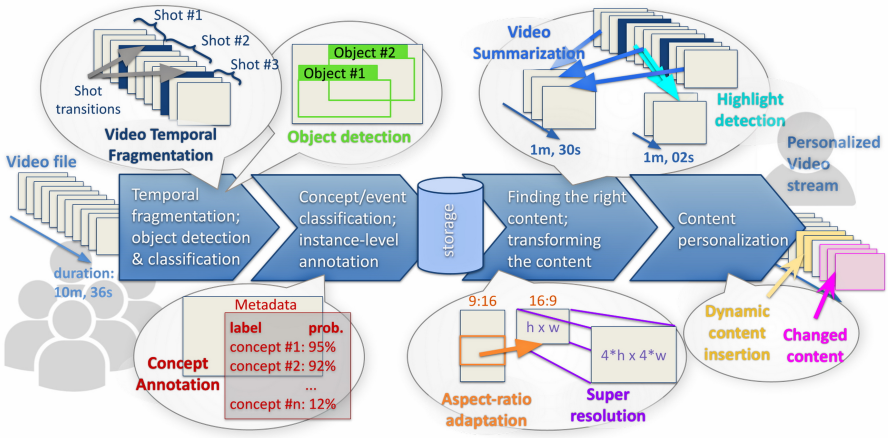

Fig. 1 Illustration of technologies in an example data-driven TV personalisation framework

The rest of the paper is structured as follows: In Section 2 we review various classes of technologies from the scientific literature that can contribute to data-driven TV personalisation. Moving from the purely scientific literature to more applied science, Section 3 surveys on deployable applications which contribute to the vision of data-driven personalised TV experience, i.e. relevant tools and Web services, including such tools and services that integrate technologies discussed in Section 2. Section 4 presents the use of standards defined to support activities in the media trading and value chain, which can be utilised in enabling data-driven personalised TV. Section 5 discusses the current open problems for all technologies and standards, proposing future research and innovation directions. Finally, concluding remarks are given in Section 6.

2 Technologies

In this section, we look at the state-of-the-art in various classes of technologies that are needed for data-driven personalisation of TV content, as illustrated in Fig. 1:

- TV content decomposition: temporal video fragmentation; object detection.
- TV content annotation: classification of media assets; annotation with classes or instances.
- TV content re-purposing: finding the right content (incl. cross-modal representation and retrieval); transforming the content (incl. video summarization, highlight-detection, super-resolution, aspect-ratio adaptation).
- TV content personalisation: in-stream personalisation, incl. changing content, insertion of content inside streams.

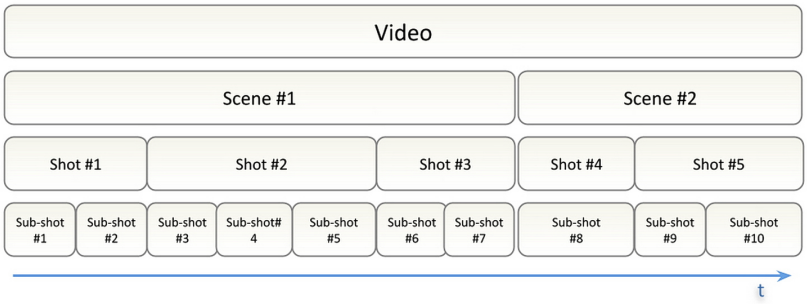

Fig. 2 An example video fragmentation to various levels of granularity

2.1 TV content decomposition

Fine-grained access to video materials is the key to their subsequent personalisation. Similarly to text documents, that can be decomposed into chapters, paragraphs, sentences and words, videos can be decomposed into hierarchically structured temporal segments (Section 2.1.1), as well as be spatially segmented to objects (Section 2.1.2). By discovering the structure of the video, its subsequent understanding and re-purposing is facilitated.

2.1.1 Temporal video fragmentation

Temporal video fragmentation deals with the identification of the underlying temporal structure of the video. It can consider various levels of granularity, but most often starts by detecting the elementary building blocks of an edited video, called shots, which are defined as sequences of frames captured uninterruptedly with the use of a single camera [1]. Temporal video fragmentation is usually the first necessary step in a video analysis pipeline.

Early shot-detection methods used hand-crafted features and rules based on colour characteristics [2–4] and/or local image descriptors [1, 5, 6]. Due to the success of deep learning in various fields of computer vision, more recent efforts are based on the use of deep Convolutional Neural Networks (CNNs). One of the first learning-based approaches is SBD [7], which employs spatio-temporal CNNs. As shot boundary detection training datasets at that time were not adequately large to optimally train a deep CNN, taking into consideration the data-hungry nature [8] of most deep-learning-based methods, [7] aimed to alleviate this by introducing a new dataset containing more than 3.5 million frames of videos that included abrupt and gradual transitions. Another notable work is [9], which introduces two novelties: i) the use of a 3D convolutional architecture, and ii) a technique to artificially create training data which essentially involves taking a raw video, shortening it, and combining the detected video shots with some type of transition. [9] achieves impressive efficiency even to this day, with a reported speed of over 120x *real-time* (*real-time* referring to the duration of the video being processed). Finally, one of the most recent shot-boundary detection methods is TransNetV2 [10], which is based on

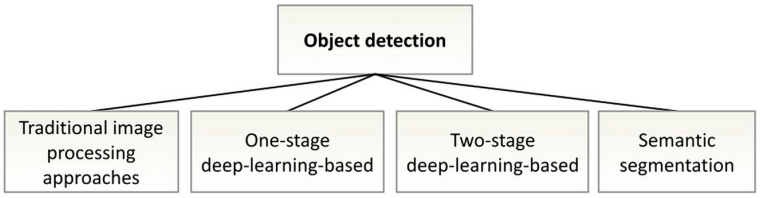

Fig. 3 A taxonomy of object detection methods

the TransNet [11] and enhances it by introducing a pre-processing step where a resized input sequence of frames is initially processed with dilated deep CNN cells [12]. TransNetV2 achieves state-of-the-art results on standard benchmark datasets, i.e. ClipShots [13], BBC Planet Earth [14] and RaiSceneDetection [1].

Apart from shot-boundary detection, a coarser level of fragmentation (see Fig. 2) is the identification of scenes [15]. These are semantically coherent time segments that are formed by grouping consecutive shots. The literature on scene-boundary detection is relatively limited compared to shot-boundary detection. Indicative works include: [15, 16], which utilise multi-modal features to better group shots into scenes; [17, 18], which employ parameter-free deep neural networks eliminating the need for fine-tuning for different types of content. Finally, on the finer side of fragmentation, shots with dynamic and gradually changing visual content can be decomposed into smaller and visually coherent parts, to produce an even more detailed segmentation of the video. These parts are commonly referred to as sub-shots [19]. Shot-boundary detection methods most commonly suffice for the temporal segmentation of a TV program. Contrarily, sub-shot-boundary detection is more appropriate for one-shot user-generated content, while scene-boundary detection results are rather coarse to optimally support fine-grained content annotation and re-use. Additionally, shot segmentation is practically a solved problem, with numerous very well-performing CNN-based methods having been developed, as discussed above. For these reasons, shot detection is both an ideal technology and a typical first step in the analysis of a TV program.

2.1.2 Object detection

Object detection deals with identifying objects in an image. The object detection task can be further broken down to two individual sub-tasks, specifically: i) localising an arbitrary number of candidate objects (detection task), ii) classifying each candidate object, i.e. assigning a label to it (classification task). Object detection can also support the higher-level task of class-based annotation, such as recognising brand logos in a TV program (and thus annotating assets with the identified TV channel), cf. Subsection 2.2.2.

Older iconic works that ignited the development of traditional object detection methods include the Viola–Jones object detection framework based on

[1] https://aimagelab.ing.unimore.it/imagelab/researchActivity.asp?idActivity=19

Haar features [20], the object detection scheme using the Scale Invariant Feature Transform (SIFT) proposed in [21] that inspired multiple works on object detection using local descriptors, and DPM [22] that first introduced bounding box regression. In the last few years, the rapid advances of deep learning techniques have greatly accelerated advances in object detection. Employing deep networks and harnessing the computing power of modern GPUs, the performance and accuracy of object detector frameworks have greatly improved. The deep-learning-based methods can be categorised into two main types (see Fig. 3): i) one-stage, and ii) two-stage. Two-stage algorithms follow a more traditional object detection pipeline, generating region proposals at first and then classifying each proposal into different object categories. In general, methods of this category achieve the highest detection accuracy while one-stage object detectors prioritise inference speed.

Considering the two-stage object detection framework, early methods adopt a region-proposal-based approach; such methods include R-CNN [23] and improved variants of it, e.g. Fast R-CNN [24] and Faster R-CNN [25]. The most recent evolution of the region proposal family of methods is G-RCNN [26], which achieves more accurate extraction of object regions by incorporating the concept of granulation [27] in a deep CNN.

Regarding one-stage methods, their main difference to the two-stage ones is that the region proposal stage is skipped and the detection is carried out directly over a dense sampling of possible locations. An iconic technique in this area is YOLO [28], which only predicts over a limited number of bounding boxes. The whole detection pipeline is designed as a unified single network and is optimised in an end-to-end fashion. Several performance upgrades resulted in enhanced versions, namely the YOLOv2, a.k.a. YOLO9000 [29], and YOLOv3 [30]. SSD [31] is another one-stage detector and one of the first attempts at using CNN's pyramidal feature hierarchy for the efficient detection of objects of various sizes. While YOLO and SSD are amongst the fastest methods [32], there are other works that focus on improving accuracy. In RetinaNet [33] it is argued that the main problem of one-stage detection frameworks, regarding their accuracy, is that many negative examples (i.e. the background class, where no object is detected) are used in the training process. They introduce "focal loss", where the contribution of these easy negative samples to the learning procedure of the model is weighed down so as not to dominate the loss, thus leading the learning to concentrate on the few interesting cases.

Recently, both categories of object detection methods (i.e. one-stage and two-stage) have leveraged deep learning models derived from neural architecture search, a technique that lets machines optimise not only the weights but also the structure of the deep network. Early results of neural architecture search efforts on the task of image annotation, including EfficientNets [34], were found to achieve state-of-the-art accuracy with an order of magnitude fewer parameters. Inspired by the success of EfficientNets, such design approaches were also employed for object detection, proposing several key optimizations

that resulted in EfficientDets [35]. EfficientDets achieve a step-up of 1.5 points from the prior state-of-the-art on the standard object detection benchmark dataset Microsoft COCO [36], while being 4x-9x smaller and 2x-4x faster. Along with EfficientDets, among the most recent and top-performing works are YOLOv4 [37], employing numerous improvements and optimizations over the previous version of the same family, and the most recent YOLOR [38], which proposes a unified network to encode implicit and explicit knowledge together, in an attempt to mimic the human brain. Analysing the related literature, one can observe that the performance gap between one-stage and two-stage methods is closing, therefore the faster one-stage methods are now very promising for use in a computational resource-conscious pipeline for data-driven TV content personalisation.

A sub-domain of object detection that goes one step further in the direction of finer-grained spatial localization of objects is semantic segmentation. Such methods detect for each pixel the object category that it belongs to; thus, instead of generating a bounding box for a detected object, the input video frame is segmented to arbitrary regions, each denoting an object. Image semantic segmentation was first introduced by [39] which employed a fully convolutional network (i.e. not containing any dense layers as in traditional CNNs), with other early works in the field being [40] which adopted the use of deconvolution layers so as to obtain instance-wise segmentation, and [41] which used the pyramid image decomposition method to improve the performance. Another work with widespread usage is Mask R-CNN [42] that extends Faster R-CNN to pixel-level image segmentation by adding a third branch for predicting an object mask, in parallel to Faster R-CNN's existing branches for classification and localization. Among the most recent and top-performing semantic segmentation methods are [43–45]), according to the online available leaderboard[2]). The interested reader is directed to the recent survey of [46] for more details on semantic segmentation.

2.2 TV content annotation

For the purposes of data-driven TV personalisation, new requirements which emerge for TV content annotation are:

1. Describing the decomposed content (as referred to in the previous subsection 2.1.1) in terms of the classes and instances of concepts that occur within that decomposition (text, audio or video segment);
2. Describing the classes and instances which are the targets of the content annotation (previous item) in terms of relationships with each other (e.g. class-instance, subclass, equivalence);
3. Describing the characteristics or properties of those classes and instances such that relationships between them can be construed (e.g. having the same creator, existing at the same time, coming from the same location, etc.).

[2] https://paperswithcode.com/task/semantic-segmentation

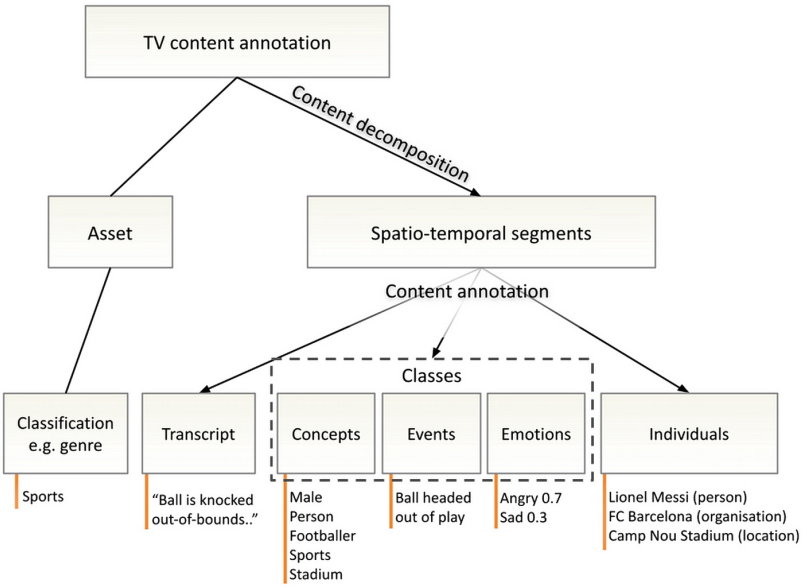

Fig. 4 A taxonomy of TV content annotation methods (orange lines illustrate examples for each type of annotation)

To provide relevant definitions in the domain of annotation: **concepts** are generally anything that might be the subject of discourse, **classes** are the different categories into which concepts may be grouped, and **instances** are the individuals that make up each class. For example, for a general concept such as human beings, a class might be Footballers and an instance of that class would be Lionel Messi.

Research into TV content annotation is not only about how to represent these content descriptions through well-defined metadata models (for more on how current standards align to the above requirements and approaches, see the later chapter 4) but also how to derive them in a (semi-)automated fashion to support organisations with large and growing collections of TV content. Manual annotation suffers from a lack of agreement across annotators as well as a difficulty to scale up, especially as the granularity of the required annotation gets finer (as is needed for later content processing steps like re-purposing and personalisation). Computer systems can automate the annotation task, including determining the right (entity) target of an annotation (a.k.a. disambiguation), at different levels of granularity:

- Asset level: classification (e.g. into genres)
- Spatio-temporal segments: transcription (e.g. speech to text), class-level annotation (e.g. concepts, events, emotions), instance-level annotation (i.e. individuals)

Figure 4 shows the different forms of annotation typically used with media assets such as TV programs along with examples of each type.

2.2.1 Classification of media assets

Looking at the different types of annotation (see Fig. 4), we begin with the classification of multimedia assets. A typical task in the media domain would be labelling a media asset with the genre(s) it belongs to. However, the fundamental approach to classification is not different in the case of media asset classification - a set of example videos labelled with one or more classes from a list is provided to a network for training such that it learns the discriminative features of each class. In [47] beyond state-of-the-art performance is reported; it is achieved by combining two common approaches into one classification model - early fusion combines features before classification while late fusion combines the outputs of classifiers from different features - which they call double fusion. Features such as colour and shot segmentation are the basis for the visual feature-based classification, e.g. [48].

Different machine learning algorithms have proven to be state-of-the-art for classification in general, e.g. deep CNNs such as Subclass Wide ResNet [49], but given the need to process entire (large) media collections, an issue for media classification with deep learning is the computational efficiency. Residual Learning [50] has become a commonly used approach in combination with neural networks to make media classification more efficient for deeper networks, e.g. [51, 52]. Other approaches use pruning, i.e. removing network connections and/or nodes to reduce the complexity, e.g. [53, 54]. Classification in the media domain is almost assuredly multi-label, i.e. each asset may belong to multiple classes, and the classes may differ across the timeline of the audiovisual asset (e.g. a news program probably also has a sports segment and then the weather), both of which are particular challenges that are not tackled in traditional classification frameworks. In the domain of genre classification, where there is a limited list of reasonably well-defined genres, there remain further issues for classification such as the subjectivity of some classifications (e.g. labelling "comedy" might rely on cues like audience laughter whereas some people find quite serious-seeming content as humorous) and the difficulty in training when classes are less clearly distinctive (e.g. "horror comedy"). RAI, the national public broadcasting company of Italy, has presented a novel approach to genre classification [55] in a tool called DENOTER, which can both suggest genres for media assets and reclassify existing assets with new genres, however, based on analysis of the textual metadata. [56] uses a combination of audio and text features to classify 200 hours of BBC content by genre, with a reported accuracy of 98.6%. A combination of genre classification approaches with deep learning-based visual feature analysis would be a subject of future research. Works such as [57] suggest this would be a promising approach.

2.2.2 Class-level annotation

Class-level annotation is the labelling of the spatio-temporal segments of the media asset with concepts drawn from a set of classes. Concepts are a very

general term and annotation tools are trained for a specific set of concepts, which may be, for example, classes (of things), events or emotions.

Class-based annotation methods may be based on the low-level analysis of visual features and/or associated textual metadata (existing titles and descriptions, transcripts, or subtitles) to derive the annotations. Whereas text-based annotation has a longer research history (e.g. [58]) supported by NLP (Natural Language Processing) and NER/NEL (Named Entity Recognition and Linking), basing an annotation on a title/description clearly misses more granular details within the TV programming whereas a transcript only captures the concepts which are spoken about, not those which are visible.

Annotation with classes is closely related to object detection, discussed in 2.1.2. After all, the objects that can be detected are usually classes rather than individuals (e.g. "cat" or "dog" rather than identifying the specific cat or dog individual). While object detection methods as previously discussed may also provide for bounding box detection or even segmentation for specifying the location of the object in a video frame, which also means additional metadata for an annotation, video concept detection methods may focus on accurately labelling video fragments with classes without the detection of object position, for example, are trained for more abstract concepts that do not directly relate to tangible objects such as "Sunny" or "Outdoor". This has been a common task in the TRECVID workshops where different concept detection methods have been benchmarked against a manually labelled video collection [59]. While top-performing methods at that time (the 2000s) were typically machine learning, e.g. support vector machines (SVMs), more recent work in the past decade (2010s) has focused on deep learning with neural networks such as CNNs, e.g. vitrivr [60]. In the last few years, published research on video concept detection has dropped as accuracy scores stabilised for well-known concept sets (e.g. TRECVID) and researchers have focused on more challenging tasks such as scene graphs (describing the content of visual media in terms of both objects and the relations between them [61]), video event detection and video summarization (meaning here the production of natural language descriptions of video content [62]).

The significant advances in the accuracy of annotation methods based on visual features are largely attributable to advances in AI/Machine Learning, i.e. deep neural networks combined with sufficient training data (a.k.a. deep learning for computer vision). The breakthrough is generally seen as the moment AlexNet [63], a CNN, won the ImageNet Large Scale Visual Recognition challenge in 2012 with a significantly better result than the runner-up. The key improvement was seen in the depth of the model and the use of GPUs to handle the computational complexity. [64] review deep learning approaches to computer vision. A state-of-the-art model for visual classification is Fix-EfficientNet [65]. It demonstrated top-performing results on the ImageNet dataset [66] (a standard benchmark for computer vision research) with 480M parameters, a top-1 accuracy of 88.5%, and top-5 accuracy of 98.7%, although

of course new models are repeatedly announced with incremental improvements on state-of-the-art accuracy (a list of the latest benchmarking results for deep learning networks on computer vision using ImageNet can be found at [3]). However, it should be noted that since these benchmarks are measured against a general set of 1000 visual concepts defined by ImageNet, for annotation tasks beyond the most generic cases it is always necessary for organisations to train a chosen network specifically for their content (and the visual features relevant to it). Pre-trained models (usually trained with the generic ImageNet data) can be downloaded (e.g. in the Python library Keras) and then integrated with further network layers which are then trained for the more specific annotation task, an approach known as Transfer Learning.

Event annotation, as in e.g. [67], goes beyond the identification of concepts in the video and detects relevant actions between them that signify certain events - either at the literal description level (e.g. a boy kicks a ball between two sticks) or a higher interpretative level (e.g. a football player scores a goal). The TRECVID series of workshops have also provided a benchmark for event detection methods in the Multimedia Event Detection (MED) task. Deep learning has also become the de facto state-of-the-art here. As each model is typically tuned to a specific task, it has been recognized that ensemble models - a combination of different models whose outputs are combined - can be a solution to broader tasks. For example, in Semantic Event Detection, an ensemble model outperforms state-of-the-art single models in classifying scenes to natural disaster events [68].

Finally, annotation can go beyond the identification of objectively present objects or events in the content to more subjective concepts such as emotions [69], where both the textual [70] or the audio-visual components [71, 72] may be used as input. However, there is still quite some debate in the research community if AI-based approaches can truly detect correctly emotions through visual cues (e.g. facial expressions) and if such results should be relied upon [4].

2.2.3 Instance-level annotation

Instance-level annotation is identifying the specific instance of any class of objects that appears in a video, e.g. a specific person or organisation or museum. The needed semantic annotations might include entities derived from textual metadata created by other algorithms or the (audio) transcript, however, here computer vision techniques are also being increasingly applied to semi-automatically identify relevant details within the visual component of the assets. In the most advanced case, multi-modal annotation uses the combination of different modal inputs, e.g. text, audio, video (such methods are discussed in Sec. 2.3.1) to produce a more accurate or precise annotation, e.g. a visual detector detecting "person" combined with a transcript mentioning the person's name may be combined to annotate the video fragment with that

[3] https://paperswithcode.com/sota/image-classification-on-imagenet, last accessed 1 Feb 2022

[4] https://www.theatlantic.com/technology/archive/2021/04/artificial-intelligence-misreading-human-emotion/618696/

specific person. For example, VRT, the national public-service broadcaster for the Flemish Community of Belgium, combined the transcript-based annotation with visual features extracted by a deep learning framework to refine the entity identification for news videos [73].

With the need for subtitles in production systems for issues of accessibility, transcription of media has been an annotation task done by media organisations since a long time, first manually but now automatically. Speech-to-text methods have become very effective to automatically create transcripts from audio (and can be built from open-source tools such as done by the BBC [74]), but language support is variable (YLE responded to the lack of Finnish support by collecting speech samples from Finnish speakers to train its own model) and modern frameworks do not immediately address issues like use of dialects or background noise [75]. Broadcasters need to be aware of the challenges in automatic transcription for their content, e.g. even the news - which is usually read out by a presenter in a very clear manner - might involve interviews with non-native speakers. Research in addressing such limitations (e.g. [76]) will gradually flow into the latest version of transcription tools.

Such transcripts can contain an association with a time point in the media for display purposes but the same association could be used to support temporal annotation, i.e. we assume that the speaking of some sentence is relevant for the time point in the media when it is spoken; we can use NER to identify references to relevant concepts in the text and annotate the time fragment of the media with those concepts. Such an approach was used in the LinkedTV project[5] to automatically create semantic annotations of TV programming from their timed transcripts [77]. Transcripts and transcript-based annotations can be very useful for search and discovery within media collections but are not as useful for downstream personalisation tasks because (i) they may result in being too fine-grained (we do not need to know every single time a person is mentioned) and (ii) they are limited to the concepts being explicitly talked about (often when things are visible, their presence is not actually explicitly mentioned as it is assumed the viewer is aware of them).

Other annotation tasks can be applied to the asset as a whole but are best used with a spatio-temporal fragmentation of the asset as the relevance of a given concept in a media annotation will usually be restricted to certain parts of the media item, e.g. a news programme is made up of various different news stories, each concerning a distinct set of entities. An asset-level annotation of the news video would not allow the selection of the relevant news item as a search result, but a video fragmentation combined with fragment-level annotations would. The main annotation task arising in this case from the need to enable data-driven personalisation is semantic descriptions of the content of the media asset which could cover everything perceivable and relevant to a viewer, associated with the respective fragment (spatial, temporal or other).

In conclusion, we can say that the classification of TV content can be performed in an accurate manner, especially if the classes to be applied are well

[5] https://www.iti.gr/iti/projects/LinkedTV.html

understood and classification models either have been trained specifically on that set of classes (e.g. high-level genres of TV programming) or can be trained by the user with sufficient labelled audiovisual content (e.g. a broadcaster who has previously manually labelled their content). Instance-level annotation on the other hand is a more challenging area of research since the correct annotation of any decomposition of TV content with an instance of relevance will depend on the explicit occurrence of that instance (in the text, a mention in the audio, or an object visible in the video) and the training of the annotation method to identify occurrences of the instance, where the set of instances is significantly larger than any classification scheme. Here, gaps will have to be accepted by any user (e.g. if detecting persons, there will always be new faces appearing on TV who do not previously exist with a label in the data used for training). Advances in unsupervised learning (approaches that learn from data that is not previously labelled) combined with the use of the Web as a knowledge source will allow future annotation models to best-guess new instances and overcome the limitation of needing to pre-define the finite set of instances to train a model on.

2.3 TV content re-purposing

2.3.1 Finding the right content

Annotations as the ones discussed in the previous section can be used in text-based video retrieval where the textual query is matched to the labels or tags annotated to (segments of) candidate video (where a knowledge model is used behind the interpretation of the textual query, e.g. to expand queries into synonyms or related entities, a technique referred to as "semantic search"). However, video retrieval can also be made possible according to queries of different modalities (visual, audio or audiovisual). In a personalisation scenario, which requires implicitly inferring which media item would be ideal according to the user's interests and preferences, it is beneficial to take into consideration information from sources of different modalities (i.e. while a user may be interested in a particular entity, e.g. a celebrity, personalisation requires that the media items which are annotated with that celebrity also demonstrate the preferred style and substance of the user). Recent advances in representation learning have demonstrated the ability to represent information from different modalities such as video, text, and audio in a joint feature space [78], as can be seen in Fig. 5. The interested reader is directed to [79] and [80] for a survey on cross-modal representation learning.

From the vast field of cross-modal representation methods, of particular interest for TV applications are text-to-video techniques where ad-hoc queries described in natural language can be used to retrieve unlabeled videos. Text-to-video retrieval, differently from the video retrieval task, requires an understanding of both video and language together. A common solution is to utilise Recurrent Neural Networks (RNNs) to learn a dense vector representation for the natural language sentence and CNNs to extract video-segment-level

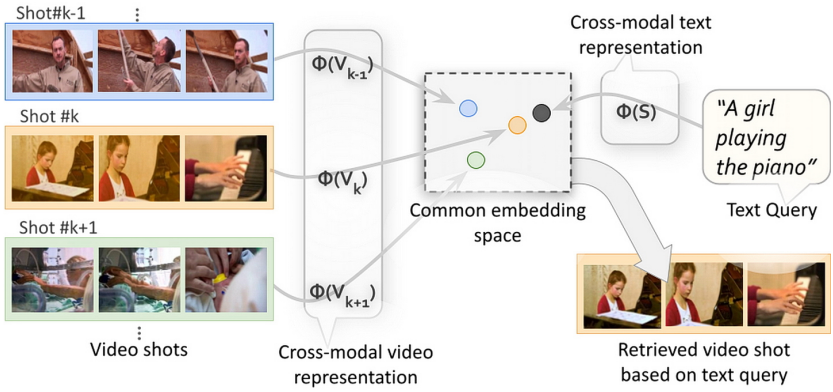

Fig. 5 An example of cross-modal video shot retrieval. V is a video shot and S is a text string. Both V and S are translated into a common embedding space $\Phi(\cdot)$, resulting in two new representations $\Phi(V)$ and $\Phi(S)$ that are directly comparable.

features, then treat the resulting vectors (both video and text) as global representations that are consequently mapped into a joint embedding space [81]. [82] proposed Video2vec, introducing an embedding method to learn the entire representation from freely-available Web videos and their descriptions using an embedding between video features and term vectors. In [83], an extension of Word2VisualVectors [84] is proposed, resulting in W2VV++, a deep learning method for query representation learning which requires no explicit concept modelling, matching and selection. In [85] the problem of unlabeled video retrieval using textual queries is addressed by extending a dual encoding network, introduced in [86], which makes use of more than one encodings of the visual and textual content, as well as two different attention mechanisms. A recent trend is the application of transformers to video processing for cross-modal representation learning (introduced in [87]), inspired by the success of transformer-based models on natural language processing tasks. Transformer-based models can be roughly organised into two categories [88]: i) single-stream transformers (e.g. VideoBERT [87], HERO [89], ClipBERT [90]), where embeddings of different modalities are input into a single transformer to capture their intra- and inter-modality information, and ii) multi-stream transformers (e.g. CBT [91], ActBERT [92], Univl [93]), where each modality is fed into independent transformers to capture information within modalities and then build cross-modal relationships via for example another transformer. Such a multi-stream transformer-based method, Clip2TV [94], which explores where the critical elements lie in transformer-based methods, is one of the best-performing in this domain, achieving state-of-the-art accuracy on the standard benchmark dataset, MSR-VTT [95].

2.3.2 Transforming the content

Content transformation is the process of making existing content more versatile and thus reusable in a different context. When it comes to media items, there

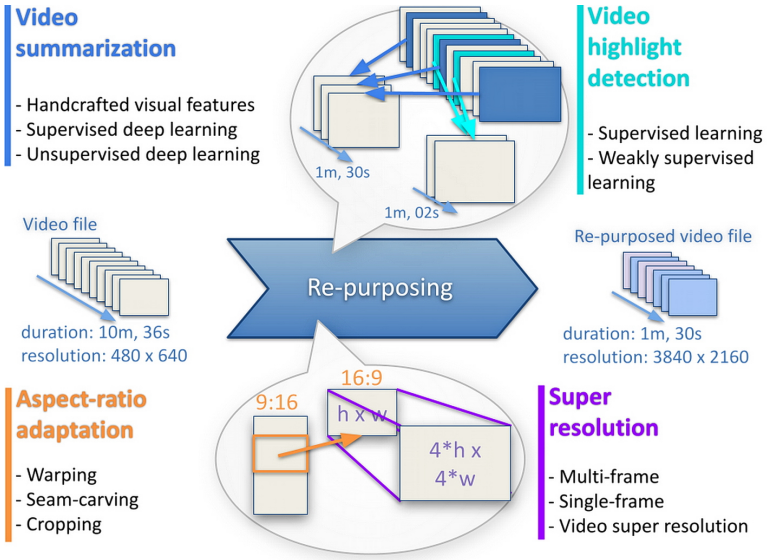

Fig. 6 Overview of technologies and their taxonomy for content transformation

exists a number of different techniques under the umbrella term "content transformation" (see Fig. 6), offering the ability to produce re-purposed versions of an original content item, which are compatible with the constraints of different publishing vectors. Video summarization and highlight detection consider transforming media content in a way that shorter versions are produced, which apart from complying with a vector's constraints, can also optimise a media item for consumption under different conditions or scenarios. For example, we can reduce a 10' news footage to a brief 30" summary, ideal for consumption by a specific target group within an online social platform. A different set of methods aims at improving the quality of the content, to enable for example the conversion of user-generated or old archival content (videos recorded using a less-than-ideal camera) to high-definition television standards. Finally, video aspect-ratio adaptation methods enable transforming the video to facilitate its consumption in different platforms or in devices with different screen sizes.

Video summarization methods aim to provide a short visual summary that encapsulates the flow of the story and the essential parts of the full-length video, by adapting the video content and generating shorter versions of it. A rough categorization of the relevant literature approaches (Fig. 6) includes: i) older methods that utilised hand-crafted low-level visual features, ii) supervised deep-learning-based methods, and finally iii) unsupervised deep-learning-based methods. Regarding older methods, these relied on the extraction and analysis of low-level visual features from the video frames, e.g. [96, 97]; clustering-based techniques that group frames according to their visual similarity and extract key-frames from the clusters' centres, e.g. [98–101]; dictionary learning approaches aiming to approximate the gap between

low-level visual features and high-level visual semantics, e.g. [102–104]; and visual attention modelling that imitates the human attention mechanism that is used to spot the most important parts of the video for generating the summary, e.g. [105–107]. It is commonly accepted that learned features which are extracted automatically to solve a specific task, are more effective at it than handcrafted features. This gave rise to machine learning-based methods, with the early supervised ones aiming to capture the underlying frame selection criterion from human-created summaries to produce video summaries that meet human expectations. Most notable examples include [108–110] which directly optimise multiple objectives for video summarization, such as representativeness, relevance, importance, diversity, uniformity, storyness, and actioness. In order to overcome the need for handpicking of desired characteristics in the final summary, deep-learning video summarization approaches that are trained in a supervised manner have emerged. These are trained using pairs of videos and user-created ground-truth summaries; early examples of such methods include [111–114]. A more recent set of supervised techniques utilises advanced variations of Recurrent Neural Networks (RNN) to capture the temporal dependency over sequential data (Long Short-Term Memory (LSTM) units [115] and Gated Recurrent Units (GRU) [116]); such methods include [117–119]. [120] goes one step further by introducing an architecture with memory augmented networks, which utilises an external memory to record visual information of the whole video, thus tackling video summarization in a more global manner that involves the extraction of knowledge about the temporal inter-dependency across the entire video. [121] proposes a 2-layer LSTM architecture where the first layer extracts and encodes data about the video structure and the second layer uses this data to define the key-fragments of the video. This work is extended in [122] to exploit the shot-level temporal structure of the video and compute shot-level confidence scores for producing a key-shot-based summary of the video. [123] describes a Dilated Temporal Relational (DTR) Generative Adversarial Network (GAN) to exploit long-range dependencies at different temporal windows, with the discriminator being trained via a 3-player loss to distinguish between the learned summary and a summary consisting of randomly selected frames.

Due to the lack of large-scale annotated training data, researchers also began to explore unsupervised training schemes for video summarization. The use of GANs for learning in a fully-unsupervised manner is a current trend, as several well-performing methods ([124–128]) rely on this framework. For example, [124], proposes an architecture that embeds an Actor-Critic model into a GAN and formulates the selection of important video fragments to form the summary as a sequence generation task. On the other hand, the best-performing supervised approaches utilise memory networks [129] or tailored attention mechanisms ([130–133]) to capture variable- and long-range temporal dependencies. One of the recent approaches in the latter direction [134] combines global and local multi-head attention mechanisms to discover different modellings of the frames' dependencies at different levels of granularity,

and also integrates a component that encodes the temporal position of video frames. For a recent and more comprehensive review of the deep-learning-based video summarization literature, the interested reader is directed to [62].

Video highlight detection aims to reduce a video to highlight moments. Video highlights can be defined as the most memorable parts of a video with high emotion intensity [135]. Highlight detectors are typically domain-specific, i.e. they are tailored to a category of video since the definition of what constitutes a highlight often depends on the domain. It is worth noting the core difference between highlight detection and video summarization: whereas summarization aims to provide a complete synopsis of the whole video, highlight detection aims to score individual video segments for their worthiness as highlights [136]. Video highlight detection can provide users with generated media items that aim to quickly revisit important events of a longer TV program, e.g. a sports game or a talk show.

Since the definition of highlight is both subjective and context-dependent, most early approaches focus on highlight detection on specific domains, e.g. sports [137–140], social media videos [141], Formula 1 TV content [142] and first-person camera shooting [143]. In general, literature works on this topic can be generally categorised into two classes: i) supervised learning methods [141, 143–145], which are trained using human-annotated training corpora, and ii) weakly-supervised approaches [136, 146–149], where various weak supervisory signals are exploited to define highlights, including the frequent occurrence of specific segments within a video [146–148], the duration of a video [136] and the information from segment bags [149]. Supervised methods might detect highlights with greater accuracy, yet it is noteworthy that in many tasks it is difficult to get strong supervision information, i.e. ground-truth labels, due to the high cost of the data-labelling process. Thus, there are scenarios where it is more feasible for machine-learning techniques to work with weak supervision.

For training their models, most of the weakly-supervised methods [136, 141, 143–145, 149] followed the principle of pair-based learning, comparing a highlight video segment with a non-highlight one, with the former being expected to rank higher than the latter. An iconic method that follows this paradigm is [144], which proposed a Robust Deep RankNet that, given a video, generates a ranked list of its segments according to their suitability as a highlight. Most importantly, they introduced the Video2GIF dataset, which contains over 100,000 pairs of GIFs, collected from popular GIF websites, and their source videos, collected from YouTube, thus creating a large dataset for training supervised highlight detection methods. In contrast, [136] does not use manually labelled highlights but offers a new way to take advantage of freely available videos from the Internet, based on the insight that video segments from shorter user-generated videos are more likely to be the selected highlights than those from longer videos. Of particular interest are the methods that aim to address multiple domains. [141] proposes a method that, given a search query (domain) such as “surfing”, mines the YouTube database to find pairs of raw and corresponding edited videos. Then, they obtain pair-wise

ranking constraints to train their model, based on the assumption that edited videos are more likely to contain highlights than the trimmed parts of the raw video. Similarly, [150] proposes a framework that learns to adapt highlight detection to a user by exploiting the user's history. In an attempt for true domain-agnostic highlight detection, [135] proposes a hierarchical structure for emotion categories and analyses emotion intensity and type by using arousal- and valence-related features hierarchically.

Most recent works in this domain include [151–153]. In [151] a user-item interaction graph is formulated and TransGRec is proposed; the latter is an inductive graph-based transfer learning framework for personalised video highlight recommendation. [152] explores the cross-category video highlight detection problem through learning two types of knowledge about highlight moments and applying it to the target video category, while [153] utilises multi-modal information by including content-agnostic audio-visual synchrony representations and mel-frequency cepstral coefficients to capture other intrinsic properties of audio.

Super-resolution's objective is to produce an up-scaled and enhanced image or video, either by combining a sequence of low resolution images/frames of a scene or by attempting to reconstruct a high-resolution image from a single low resolution observation. Super-resolution can be used to enhance user-generated content or archival material (i.e. video captured with older cameras) in order to create high-definition content.

Traditionally, the first super-resolution approaches were devised to be applied to still images. More conventional methods can be categorised in two classes: i) multi-frame, e.g. [154–156], where reconstruction produces one high-resolution image from a set of low-resolution images, ii) single-image, e.g. [157, 158] and iii) methods dedicated for video super resolution. The first deep learning methods aimed to create a mapping between low and high-resolution images, e.g. SRCNN [159]. [160] was the first to claim that mean square error is not the ideal way to express the human perception of image fidelity and proposes the use of alternative metrics, such as the structural similarity index [161]. In contrast, the intuition that we do not need to penalize a deep-learning-based model for pixel differences that do not bother a human viewer, gave rise to methods that compute the difference between feature maps of a pre-trained network instead of directly comparing the input images [162]. Additionally, the same principle brought about methods that employ GANs, e.g. SRGAN [163] and Enhance-SRGAN [164]. There are also different ways of addressing the problem of super-resolution besides comparing high-resolution and downscaled images, like methods that employ variational auto-encoders, e.g. [165, 166]. Most recently, the super-resolution problem is being tackled using diffusion models [167]. A very recent method, [168], demonstrated that their diffusion model outperforms GANs on high-fidelity image generation on the ImageNet dataset utilising auxiliary image classifiers to boost sample quality. In [169],

cascaded diffusion models are employed to generate high fidelity images, without any assistance from auxiliary image classifiers. Finally, in [170], SR3 is proposed as an approach to image super-resolution using repeated refinement.

Regarding super-resolution for video, it should be highlighted that applying single-image methods successively to each video frame is feasible but leads to a lack of temporal coherency [171]. Only recently, the focus has shifted to techniques dedicated to video, i.e. video super-resolution (VSR), where additional temporal information from neighbouring frames is exploited for further improving the quality of the result for a given frame [172]. The most recent and among the top-performing VSR methods are: [173], which integrates spatial and temporal contexts from continuous video frames using a recurrent encoder-decoder module; [171], which is a GAN-based spatio-temporal approach to VSR, renders temporally-consistent super-resolution videos; [172], which proposes a Recurrent Residual Network (RRN) network architecture for efficient VSR; and [174], which focuses on properly rendering fast-moving objects. For a recent and more comprehensive review of the deep-learning-based video super-resolution literature, the interested reader is directed to [175].

Aspect-ratio adaptation tackles the problem of transforming a video, originally captured in one aspect ratio, to a different (target) aspect ratio, so that it can be optimally consumed through various devices, e.g. video originally captured for the TV, to be optimally viewed in a mobile phone; or, inversely, video captured in portrait format using a mobile phone, to be used as part of a traditional TV broadcast.

The video aspect-ratio transformation algorithms of the literature can be divided in three main categories: i) warping [176, 177], ii) cropping [178–181], and i) seam carving [182, 183]. Warping methods, instead of resizing the entire video frame uniformly, determine scaling factors in a content-adaptive way: the frame is divided using a grid and important regions are left untouched, while scale factors are applied to other less-important areas. Cropping techniques select a rectangular area in the image/frame and discard visual content outside of it. Seam carving algorithms remove seams of uninteresting pixels, i.e. connected paths of pixels inside the frame are discarded. There are also multi-operation techniques that combine two or more operations, e.g. cropping and warping [184], or seam carving and cropping [185]. It is worth noting that when applying warping or seam carving to the frames of a video, apart from undesirable artefacts introduced [186], the original video's semantic content might be distorted significantly [187]. In [188] it is argued that cropping methods are more suitable for video aspect ratio transformation when the minimization of semantic distortions is a prerequisite, as they select a region of interest in the video frames but do so without introducing any distortion to the visual content. Cropping methods typically extract some kind of feature for assessing the importance of different regions in a frame [188]. Then a crop window is fitted in the frame so as to contain the most important regions. Additional effort is taken to ensure the smooth motion of this crop window throughout the video (e.g. in [178] camera operations are derived by optimising the path of

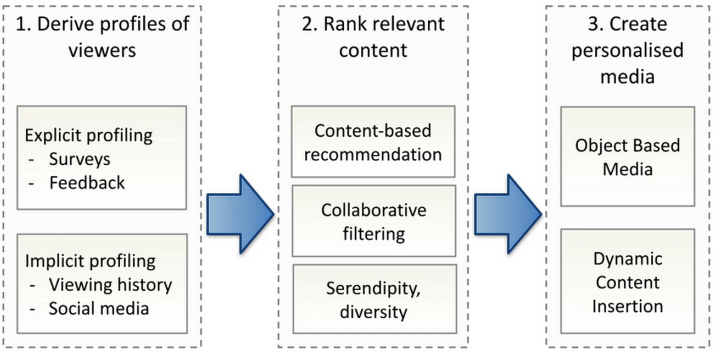

Fig. 7 TV content personalisation approaches

this window, seeking to adhere to the principles of cinematography). In [189] a Structural Similarity feature is proposed based on blur detection to identify whether an image contains a blurred background. In [180] low-level features are employed while in [178, 179] eye-gaze information is utilised. In Google's AutoFlip[6], a solution to smart video reframing (i.e. video aspect-ratio adaptation), face and object detection results are employed.

The most recent methods in this domain include [188, 190]. In [188], apart from proposing a fast smart-cropping technique, a benchmark dataset for video retargeting, RetargetVid, is introduced. Both [190] and [188] find candidate subjects to follow (the first employing object detection, the second utilising saliency detection) and both aim to select the main focus in each sequence of frames (e.g. in [188] by clustering the detected salient blobs and selecting the most appropriate cluster according to the introduced criteria).

2.4 TV content personalisation

Personalisation refers to a degree of identification of the interests of the user, such that content may be offered according to those interests. This typically requires user profiling, whether explicit (asking the user for their interests) or implicit (learning from data, e.g. viewing history), as well as a personalisation platform to derive and manage the user profiles. Finally, a **recommendation engine** matches a set of content items to the interest profile of a user to produce a ranked list where the top items are considered those of most interest to the viewer (see Fig. 7).

In reaction to the growing choice of TV programs across channels and the ability to watch online non-linearly (i.e. catch up on TV programs that were previously broadcast, thus increasing further the choice available), TV content personalisation has been researched extensively, whether for personalised EPGs [191], broadcast (IP)TV [192] or on-demand viewing [193]. Recommendations can be done by comparing the closeness of the user profile with an annotation of each media asset using the same categories (or with a mapping

[6] https://google.github.io/mediapipe/solutions/autoflip

between user and media categories), i.e. **content-based recommendation**. This produces a score that is an indicator of the closeness of the match. Typically, this leads to the problem that users who like certain types of content are recommended more content of the same type (e.g. watching lots of Westerns leads to more recommendations for Westerns). To address this problem, **collaborative filtering** can be used which follows the rule "The user would like to watch the content other users with similar profiles have watched". State of the art methods are typically hybrid combinations of both approaches, e.g. [194, 195].

The NoTube project proposed a Beancounter which analysed the content of a user's social media feeds to build up an interest profile for them and a recommendation engine based on Linked Data so that the profile could be matched to content that is not directly related yet relevant (through semantic links between content, our Western fan might be recommended another film genre because a favourite actor is present or the same period of history is covered) [196].

In [197], the profiling of TV viewers is addressed in a different way making use implicitly of user feedback to online content. The paper proposes a personalised viewer profiling technique that creates individual viewer models dynamically using an incremental learning algorithm to learn from viewer comments, likes and shares on streamed content. The suggested approach reduces prediction errors of previous algorithms and so increases the accuracy of the recommendations.

The need in the TV context to avoid sending viewers into a "filter bubble" where they only see more TV that lies within their modelled interests has led to work on *diversity* and *serendipity* in recommendations [198, 199].

Such personalisation approaches are typically used in the recommendation of whole media assets to the viewer. As a last step in data-driven TV personalisation, we are interested in their application to the appropriate delivery of decomposed, annotated, and re-purposed media content to the viewer in newly and dynamically authored scenes in an individualised manner. Unlike recommendation engines (which are out of the scope for the remainder of this article), we are not considering interfaces where the user can access a set of recommended items (possibly with some explanation of the choice) and choose for themselves which item they want to consume. We consider an alternative future of TV content delivery where the media stream can be automatically adapted to the viewer without their active participation but based on their derived interests and preferences (we assume that explicit consent for personalisation has been given, e.g. at the beginning of access to that media stream).

Based on the previous technologies enabling the media decomposition, annotation and re-purposing, the re-purposed contents can now be repackaged in different ways for individual (IP) delivery to media consumers (personalisation). The key objective of this technology area is to support the media organisation in the (semi-)automatic packaging of different, possibly re-purposed content assets for the delivery to the consumer. It, therefore,

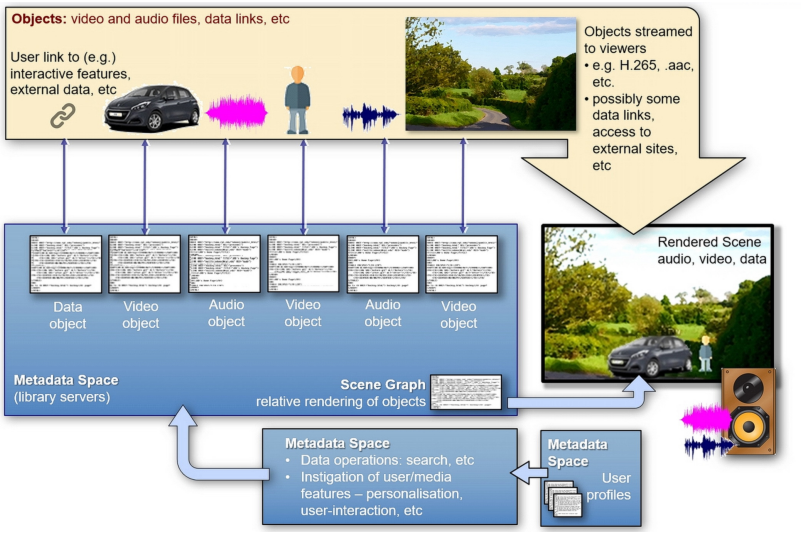

Fig. 8 Example of an Object Based Media application

encompasses search and selection across media collections based on their metadata (annotation) and the packaging of those selected assets into a multi-modal content scene that can be delivered across an IP network to an end device.

A leading example of a possible complete solution of TV content personalisation is the Object Based Media (OBM - see Fig. 8) activity pioneered by the BBC [200]. This is the representation of a media asset as a set of individual assets together with metadata describing their relationships and associations. At the moment of consumption, different individual assets can be packaged together in different ways to provide a personalised content experience to the consumer. OBM has been demonstrated in different ways, e.g. a personalised cookery program [201], interactive storytelling in film [202] or personalised radio broadcasts [203].

In the broadcast industry, this type of technology is only being applied in the case of advertisements (Dynamic Ad Insertion or DAI), where an advertisement break may have multiple slots and different ads are played out in each slot to different viewers based on their profiles. Whereas DAI is part of the current "programmatic television" offer, research work remains on other areas such as insertion of ads as overlays inside the TV program itself, where it is important to choose an appropriate on-screen area [204] or insertion of ads independent of the broadcasters ad slots [205]. The ReTV project looked at the selection and insertion of a choice of program trailers (automatically summarised videos) based on viewer profiles under the name Content sWitch (cf. ReTV project deliverable D6.2 in [7]). Fig. 9 shows four different viewers watching the same media stream but receiving four different content items in parallel such as two different ads, a program trailer or a summarised weather

[7] https://retv-project.eu/wp-content/uploads/2020/01/ReTV_D6.2_final.pdf

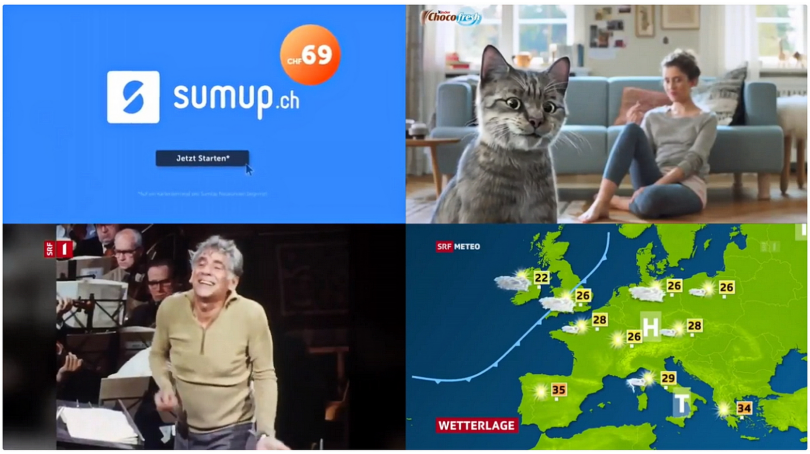

Fig. 9 Demo of the Content sWitch

report. This is presented as part of a future technological approach the project termed "Dynamic Content Insertion" [8].

The brevity of this section reflects that this technology area is the least matured for enabling true personalised TV at the level of the video content. Both Object based Media as well as Dynamic Content Insertion, are firmly in the area of research rather than (TV) production at the time of writing. Besides the further research, common standards and specifications for the scene graph that can represent the recomposed media content are needed so that toolkits can emerge with interoperable software to enable TV content personalisation. The future promise is that TV becomes this composition of contents, dynamically performed for each individual viewer based on the previously described steps of content decomposition, annotation, and re-purposing.

3 Applications

3.1 Applications for TV content decomposition

As already discussed in Section 2.1.1, video fragmentation is most commonly the first step of a video analysis pipeline. In a content-based retrieval scheme video fragmentation enables the specific segment retrieval, i.e. retrieving not only whole media items but multiple video segments from different items that match the search criteria. Additionally, most frameworks that perform some kind of high-level video analysis task, have a video fragmentation technique embedded. For example, Google's MediaPipe is a collection of customizable machine-learning solutions for live and streaming media. The set of solutions offered varies from object detection and face detection to finger tracking and even hair segmentation. MediaPipe is a free and open-source product, so it is

[8] https://retv-project.eu/portfolio-item/dynamiccontentinsertion/

quite easy to ascertain that for most of the offered solutions a shot-boundary-detection technique is applied at an early stage. IBM Video Analytics is a content-indexing platform that uses cognitive analytics to quickly and easily extract key insights, patterns in streaming and archived video. Another example is the free online services for video fragmentation and reverse image search of [9] and the video analysis service of [10] where it is explicitly mentioned that a "shot and scene segmentation" method is employed.

Object detection in videos can produce object-based metadata for each detected video fragment enabling richer indexing capabilities, which in turn allows the more accurate browsing of content thanks to the larger amount of parameters available. Products like Playment[11], an all-in-one data labelling platform, or the object detection solution included in Google's MediaPipe can provide accurate metadata annotation on videos for TV applications. Furthermore, the additional spatial information of the annotations resulting from object detection frameworks, can be exploited to enable various TV personalisation scenarios, such as: i) allow a new media object in an OBM-enabled system (see Section 2.4) to be positioned on a broadcast video sequence according to user's preferences, without occluding any key action (e.g. a personalised object media to be overlayed on a soccer match without occluding any player or the ball), ii) detecting the presence of a channel logo, a technique commonly used in advertisement detection systems [206], which in turn. is a crucial step towards advertisement replacement, or iii) more elaborate systems which would allow a user to receive a personalised notification when a favourite actor or product appears in a broadcast content [207].

3.2 Applications for TV content annotation

The first metadata models for TV content were limited in their descriptive ability - e.g. title, abstract, tags - and not aligned with regard to the vocabulary used, i.e. one annotator might refer differently from another annotator to the same thing. To better support search and discovery, there has been gradually an uptake of semantic technologies (structured metadata based on a schema or ontology) for content annotation [208]. However, as the use case for content metadata expanded from content discovery to supporting new types of content analysis techniques which required content re-purposing, personalisation or recommendation, also the recently-adopted semantic metadata models and tools prove insufficient. For example, they were often used to describe the media asset as a whole, and could not support the use case of finding a fragment of the asset which matches a search request. Another case was descriptive properties still taking natural language for their values which can not be as easily parsed by software as concepts expressed as part of a knowledge model.

Current approaches still vary based on the tools used and often rely on internal representations which work with the legacy IT infrastructure but

[9] http://multimedia3.iti.gr/video_fragmentation/service/start.html
[10] http://multimedia2.iti.gr/onlinevideoanalysis_v5/service/start.html
[11] https://www.playment.io/

would prove limited if any data interchange would take place between organisations. Since external, public knowledge graphs (which could be used as targets for concept annotations) are found to be noisy in cases and incomplete in other cases, organisations that have already started to consider annotation in terms of entities and relationships have tended to build their own internal knowledge graphs and annotate with those as targets. For example, Finnish broadcaster YLE reports that it has developed its own concept vocabulary [209] with a focus on entities missing in public knowledge bases such as Finnish persons and the vocabulary currently encompasses circa 200.000 unique concepts with 20-30 new concepts being added each day. This process can bear fruit for internal data-driven activities but will encounter limitations if ever the annotations are to be re-used in other contexts where access to the knowledge graph is restricted (e.g. by external organisations). An option is to ensure where possible links between internal entities and their equivalents (or similar) in public graphs like DBPedia and WikiData. YLE uses Wikidata[12] both as a source of entities as well as adding new entities to Wikidata when they are found to be missing and are first created internally.

Besides the issue of specifications to support the semantic descriptions of TV content items, applications are needed to extract the descriptions from the existing data. Previously, software applications were installed in the intranets of organisations to process the audiovisual content and provide tags (classes or instances) for each content item. Interestingly, another approach that has emerged as the scale of video has increased has been crowdsourcing, i.e. involving a larger group of human annotators in watching and labelling video. In terms of machine automated applications, off the shelf NLP/NER/NEL tools such as DBPedia Spotlight[13] or AIDA [210] can be used to annotate assets with entities based on textual metadata such as transcripts. It should also be noted that detection of references to creative works in text can be particularly difficult in NER (i.e. titles of books, films, music albums, etc.) and specific training of NER/NEL systems is needed for this case. In the ReTV project, a customisation of the Recogynze tool performed significantly better than the previously mentioned "off the shelf" systems [211]. These annotation tools are generically focused on the identification of the occurrence of entities in text, rather than the specific use case of describing the content of TV programming for personalisation services.

Under the name "Linked Media" [212] a proposal was put forward for semantic annotations which covered the requirements of data-driven TV personalisation by promoting the use of the W3C Media Fragments URI specification and entity references from the Linked Open Data (LOD) cloud. The EBU (European Broadcasting Union) also reported on the use of Semantic Web technologies for annotation of broadcaster assets [209] where the BBC has been a leading advocate of the use of Linked Data, for example publishing the 2010 FIFA World Cup website with the support of an RDF Triple

[12] https://wikidata.org
[13] https://www.dbpedia-spotlight.org/

Store where each World Cup entity (players, coaches, teams, venues, etc.) was described and linked, and each piece of World Cup content tagged with the relevant entities so that entity-centred views could be dynamically created for Website visitors (e.g. see all stats and news around a player or a team). This was expanded into the entire Sports section[14] and now the BBC bases various data-driven content services around its own Linked Data Platform[15]. However, it should be noted that such approaches were based on textual analysis for annotation rather than audiovisual assets.

Many modern annotation solutions are available as SaaS (Software as a Service) i.e. via APIs where the content to be annotated is POSTed to the service and after the analysis process to produce the annotation it is possible to GET the semantic description created for the content. Off-the-shelf services for video annotation are emerging later than the equivalent services for images, with Google claiming to be the first to launch such a service in 2017 (Cloud Video Intelligence[16]). Google's service can label video by segment very generally with concepts and objects that are present; in comparison, Microsoft Azure's Video Analyzer[17] is specialised for particular tasks like recognising persons or object detection (classes of object). AWS launched Amazon Rekognition[18] which similarly focuses on objects, people, text, scenes, and activities in video.

Such services are trained to generically annotate video with objectively identifiable objects visible in the frames, together with time information (generally making their own decomposition of the video and labelling each fragment with the sum of objects identified within that fragment). As with any specific annotation task, media organisations have realised that best results are only possible when they train their own video annotation systems with their content and for the classes and entities they are interested in. In 2019, Amazon Rekognition launched Custom Labels which allows businesses to train the system to detect objects or scenes unique to their business needs. Google similarly enables customised training of their service via the Vertex AI platform which uses AutoML technology[19].

3.3 Applications for TV content re-purposing

Video summarization can automatically generate previews of full-length videos, "teasers" of a program for sharing on social platforms or be used whenever a trimmed-down version of the original video is needed. As discussed in Section 2.3.2, video summarization is an active research area, and a plethora of scientific papers on summarization exist, some of them providing source code. However, these methods are far from what TV broadcasters need, i.e. a stable and customizable solution. A notable example of the few commercial video

[14] https://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html
[15] https://www.infoq.com/presentations/bbc-data-platform-api/
[16] https://cloud.google.com/video-intelligence
[17] https://vi.microsoft.com/
[18] https://aws.amazon.com/rekognition/
[19] https://cloud.google.com/vertex-ai/docs/start/automl-model-types#video

summarization tools available is the Content Wizard[20], a professional-grade Web-based tool that specialises in trans-vector publishing of video content in one seamless, semi-automated workflow which supports video summarization capabilities. On the free side of products, there is the online video summarization service of [21] [213] in which users can submit videos and generate summaries for use in various social media channels. Similarly, video highlight detection technologies are usually embedded in editing tools such as Wedit[22], which can generate video clips from an automatic video highlights search and unify them in a single clip ready for broadcast.

Many super resolution and video aspect ratio transformation methods have been proposed over the years, mostly from the scientific community and academia, as discussed in Section 2.3.2. Examples of commercial applications for super-resolution are PixOp[23], a Web-based application with video denoising and super resolution capabilities which follows a Pay-as-you-use scheme, and COGNITUS[24], a media AI software platform for crowdsourcing and enhancing video for broadcast. Concerning free video aspect ratio transformation tools, there is the online video smart-cropping service of [25], a service that lets you submit videos and transform them to a different aspect ratio.

3.4 Applications for TV content personalisation

One of the current major developments is the use of OBM [214–216] which allows individual media assets, for example, sound clips, video clips, specific video objects to be composed and rendered into the playout viewed by the end-user. It also allows these individual clips to be exchanged between complete works. This may offer advanced services (some are described below) which are directed towards end-viewers or may allow flexible media production practices. The BBC is working on a set of tools and workflows for OBM that eventually can support scalable use in production settings [217, 218]. [219] presents a demo of how OBM could be made universally accessible across devices using a cross-platform approach, a games engine-like runtime and cloud-based rendering. As reported on the BBC's OBM webpage[26], possible scenarios for both audiences and production studios are listed yet also marked as purely illustrative. There has been one episode of BBC Click in 2019 which was interactive in nature[27]. Netflix had actually already provided an interactive episode ("Bandersnatch" as part of the series "Black Mirror") at the end of 2018. However, such programs were manually composed (in terms of the alternative paths through the story). The ReTV project also demonstrated an interactive TV program using the RBB children's show Sandmännchen which functions via a smart speaker application [220]. Children can speak out loud which characters

[20] https://retv-project.eu/content-adaptation-publication-online/
[21] http://multimedia2.iti.gr/videosummarization/service/start.html
[22] https://www.vsn-tv.com/en/products/vsn-wedit/
[23] https://www.pixop.com/blog/super-resolution-in-broadcasting
[24] http://cognitus-h2020.eu/
[25] http://multimedia2.iti.gr/videosmartcropping/service/start.html
[26] https://www.bbc.co.uk/rd/object-based-media, accessed 24 January 2022
[27] https://www.bbc.co.uk/rd/blog/2019-04-object-based-media-click-interactivity-tv

or objects they want to see in an episode played out just for them. The interactivity is enabled by the video fragmentation and concept detection analyses applied to an archive of Sandmännchen episodes. We consider further applications for TV content personalisation under different headings (application categories):

TV personalisation and interaction: One of the biggest service sectors enabled by object-based media is for personalisation of and interaction with TV [221, 222]. In these scenarios, a user may have content personalised for them on the basis of a number of factors, for example, a user profile, previous interactions with the media, etc. The use cases include personalised training and learning, documentaries, advertising, entertainment, etc. In each case objects of still-image, video segments or audio segments may be inserted into the video playout at specific positions in time and space in the content.

For example, a car may be introduced as an advertising placement. This needs to be inserted in a relevant scene, see Fig. 10. This will require the object to satisfy a number of technical matches: The car needs to fit the scale of the frame; it needs to match the lighting; it needs to match the editorial aspects of the production. For example, it needs to be a specific colour to fit the plot of the film, it may need to be a certain model and date, etc. Consequently, any placement opportunity in the production needs to be able to identify a candidate placement object, and to match this with the integrity of the source (original) video production. This should be able to occur in near-real-time. Objects may be acquired from external libraries (e.g. advertising agencies, etc) and allow the placements to occur on the fly. For this reason, a lightweight trading platform is required to support the near real-time demands of trading objects into the content.

The use of media objects facilitates interactions. For example, the author of a video object allows a user to interact with the object to cause on-screen actions and this brings the concept of gamification into broadcast media [223]. It may allow users to manipulate three-dimensional objects to allow inspection, for example, adding additional interaction to documentaries.

Story telling: The BBC have maturing developments in the area of personalised story-telling and documentaries. The content can be tailored in terms of individual constituent topics, length of program required and fitting to the viewers' preferences. BBC have developed the StoryKit [224] to allow creation of content utilising object-based media.

Convergence of the broadcast and film market with gaming and virtual reality: The gaming market has grown to surpass the film and TV markets [225–227]. The markets will become increasingly integrated and media assets (image, audio and video clips) from video content will be utilised in gaming content, and vice versa. Moreover, this will occur with a requirement for immediate transfer of assets into recipient content. This will also require the management of these assets; their ownership; usage allowances and restrictions; payment schedules for assets; service level agreements, etc. There have also been a number of initiatives to converge Virtual Reality (VR) into TV services.

Fig. 10 User Interaction with on-screen objects

[228] which would allow TV content to be played in VR environments with enhancements from additional objects authored for the purpose, or to utilise VR objects from other content.

Media re-use: An increasing requirement of broadcasters is to capitalise on their growing archive stock, or to keep program content topical. This is achievable by swapping out existing content objects and replacing them with new or updated content audio and/or video objects.

One of the key requirements in these scenarios is the likelihood of media being constructed and authored from multiple input sources. An example is a background scene with additional video objects imported to create a personalised scene, as discussed above. The background and the objects need alignment in space and time and need scene management to integrate and coordinate into the scene. The associated objects need methods to inter-relate them and to dictate to a receiver (client device) how these need to be rendered into the final video for viewing.

All of these use cases demonstrate the requirement for metadata to provide extremely tight descriptions of media content, down to the level of an object with a scene. All of these media elements are likely to be individually traded media assets requiring contractual support within the media value chain as they are sought, located, acquired and rendered into a composite scene for a specific video or audio feature for the end-user, which in turn raises the need for appropriate standards as we will see in the next chapter.

4 Metadata standards for the media value chain

Over the last two decades, a range of standards have aimed to define descriptive formats, including ontologies to support activities in the media trading and value chain. However, the standards that exist are still not fully supportive of the range of advanced features and services. The standards aim to cover the details of how media is traded in terms of i) media instantiation and format, ii) media ownership, iii) media transactions to new owners, iv) allowances and restrictions of media usage and re-use, and other features. MPEG have already defined ontologies to support semantics in traditional text readable contracts in the media value chain. The use of ontologies adds value to trading activities. In the move to the use of Smart Contracts on Distributed Ledgers Technology platforms (i.e. Blockchain) a method is required to support the facilities of current contract types and translate them into required smart contract languages. This is the focus of Part 23 of the MPEG-21 standard suite [229]. In general, the use of metadata in broadcasting (for a number of applications) has been sparse until around ten years ago. Descriptive metadata was limited mainly to the type of detail required for a brief entry in an Electronic Programming Guide (EPG). The rapid increase in services (especially those requiring semantic processing, for example for cross-platform services) has seen large growth over recent years. Unfortunately, many of these new services have been rolled out partly experimentally before maturing to full service. They were also rolled out at great haste without much planning and not always managed in an integrated way. As a result, many broadcasters have upwards of 50 different metadata applications and formats. Proprietary data formats are a major problem when evolving services further in a clean manner. There is also a problem when integrating with services in external organisations.

4.1 Interworking between media asset platforms

Applications for trading of media assets were identified in Section 3.4 above. Media assets may be full-length programs or component media objects. It is likely that there will be a range of independent media asset trading platforms emerging in the current traditional media trading market. For example, platforms that are likely to be blockchain-based and trading media assets via smart contracts [230]. They need to interwork with each other in the current market, and also with other media trading platforms. Standards need to cover the areas of:

- Description of content
  - Full programs items
  - Program elements
- Description of value chain operations
  - Parties involved (e.g. creator, vendor, purchaser, etc)
  - Content formats / instantiations
  - Contractual (Service level agreements, obligations, etc)

4.2 Ontologies in content description and media asset trading

To facilitate the depth of detail required for services described in Section 3.4, it is necessary for scene and object description to be defined to a tight level of detail, including permitted activities on any object sought, traded and utilised in a derivative production, for example for scene decomposition in Sections 2.3 and 3.2; or for the re-purposing of content as described in Section 3.3. Within MPEG standards a number of ontologies have been developed [231, 232], to allow semantic operations and processing on media for search, retrieval, reasoning for personalisation and interaction, and other services in media value chains as previously discussed. The issue arises here that ontologies representing media require conversion to smart contracts supported on blockchain platforms. This gives the platforms the additional features and benefits of the MPEG ontologies. The details of these conversions between traditional narrative contracts and smart contracts and the use of MPEG ontologies are discussed below.

4.3 Metadata Candidates

4.3.1 Introduction to descriptive metadata for broadcast

Metadata in film and broadcast media has traditionally focused on the necessary structural aspects required by receiving equipment to decode and playback the media. Structural aspects include resolution(s), frame rate, coding, audio aspects (stereo, 5.1, etc) and other parameters. For descriptive metadata the standards have traditionally been very simplistic and have been limited to the basics of the content, typically sufficient for an EPG: title, length, simple plot-line, leading performers, etc. Initial descriptive metadata formats have been built on the DublinCore standard [233], and evolved into DVB-SI [234], TV-Anytime [235], BMF [236] and, by the late 1990s, MPEG-7 [237–239].

As discussed below, MPEG-7 is a complex standard and has gained little support in the industry, largely due to the lack of immediate requirement for a solution without the initial problem. In the years since the publication of MPEG-7, there have been a number of advances in media features (some are discussed in Section 2 above), and technologies to achieve immersive media have required a revised approach to content description. This has resulted in the development of MPEG-I Part 14 [240, 241] "Scene Description for Immersive Media". This standard is one of many parts which make up the suite of standards for Immersive Media [242], including emerging media coding formats, for example, part 3 [243] is the part that describes the requirements to support Versatile Video Coding (VVC) [244].

4.3.2 EBU Core

EBU Core is introduced at [245] and the ontologies to define the Core operations is given in [246]. The standard defines concepts, relationships, and

properties that apply to broadcast media to describe program content. It is based on the Dublin Core metadata model [247]. EBU Core was first published in 2000 as a set of definitions for audio archives. At the time of its introduction XML was an emerging standard but its use has increased since, requiring a more structured approach to audio-visual content description. A range of more semantic languages has been developed which have influenced the way of modelling audio-visual objects. EBU Core has followed this development. The first representation of EBU Core comprised the 15 ontological elements of Dublin Core.

Following the development of semantic representation on the Web, Web Ontology Language (OWL) [248] is used as the semantic basis. This facilitates machine-operable queries for items based on semantic understanding.

The suite of standards includes the related EBU specifications:

- Tech 3293 - EBU Core [249]
- Tech 3293 - RDF/OWL [250]
- Tech 3332 - Music [251]
- Tech 3336 - Classification Schemes [252]
- Tech 3349 - Acquisition Metadata [253]
- Tech 3351 - CCDM [254]
- Tech 3352 - Identifiers in BWF [255]

The purpose of the scheme is to classify content at a comparatively high level, like EPG applications, etc. However, it is not considered that the scheme is sufficiently fine-grained to support the ranges of next generation services envisaged.

4.3.3 MPEG-I and content description

An implementation of personalised video is likely to be based on the compositing of audio, video, and data objects into an integrated scene as described in subsection 2.4. Tools to support these processes include support for scene description, object description and scene graphing. MPEG-7 had been the only standard offering this level of detail, but to date, such detail has not been required in traditional media, and MPEG-7 remained largely dormant as a standard. The standard was initiated in the 1990s for the range of multimedia presentations envisaged in MPEG-4. As such MPEG-7 gave extremely fine-grained content description down to objects within a frame-level of detail. However, in the context of the next generation services envisaged, MPEG-7 is now superfluous along with other scene description standards LASeR (MPEG-4 Part 20, Lightweight Application Scene Representation) and BIFS (Binary Interchange Format for Scenes, MPEG-4 Part 11). Developments in immersive media require support for 3D scenes, games, and other next generation features. Recent MPEG activity has been in an ad-hoc group developing proposals and standards to support these requirements, and these are emerging as the evolving standard MPEG-I Part 14 [256, 257] - scene description. This part defines extensions to existing metadata scene description schemes to support these new features.

The MPEG strategy has been to extend an existing technology rather than redefining new standards from scratch. The most suitable candidate identified is Graphics Language Transmission Format, glTF [258]. The glTF specification is royalty-free and intended for the transmission and loading of 3D scenes and models used by graphics engines and applications. The format minimises the size of 3D assets, and the runtime processing needed to unpack and use them. An extensible, format for publishing is defined that streamlines authoring workflows and interactive services by enabling the interoperable use of 3D content. The application and specifications are aimed across the industry. However, there are issues that needed to be addressed to extend to format to be suitable for scene description required in MPEG-I. These included support for audio; timed media, including dynamic meshes, point clouds and video textures; scene updating; decoupling of media access from rendering. MPEG has formulated extensions to address these gaps in the original standard. MPEG has also defined a Media Access Function (MAF) API which decouples rendering from media acquisition (fetching). MAF also provides the API to the Presentation Engine to request media, and the associated metadata. MPEG-I also specifies a node hierarchy as a model to relate spatial media in terms of meshing, light, material, shaders, texture, etc. An example is the MPEG_Media Extension to allow timed and non-timed media (compressed and non-compressed), and to support a range of delivery formats including DASH/CMAF, HLS/CMAF, WebRTC and also for local storage in ISO Base Media File Format (ISO BMFF). Other MPEG extensions to glTF define circular buffers for read/write media access, allow materials in scenes to use textures and allow spatial audio. The specification also details the management of media pipelines for the timed delivery of media to the Presentation Engine.

4.3.4 MPEG-21 part 23: smart contracts for media

The MPEG standards addressing the usage and management of media assets are mainly part of the MPEG-21 standards. The MPEG-21 suite of standards is defined as the "Multimedia Framework," and addresses digital asset management from two main considerations:

- The definition of a Digital Item (i.e. as a fundamental unit of distribution and transaction)
- Users interactions with Digital Items; viz. their roles, allowances, restrictions of usage, etc.

The standard is a suite of parts that were originally defined in the 1990s as a framework for digital works authored and described from the evolving MPEG-4 multimedia standard. In recent years the methods envisaged for support of digital media assets have been revised. A key technology to the automated contractual trading of media is part 23 of MPEG-21, "Smart Contracts for Media". This is described below with reference to associated parts of the MPEG-21 standard suite. The approach envisaged is that media contracts are initially constructed using XML schema and RDF from parts 19, 20 and 21 of the standard (ISO/IEC 21000-19, 21000-19/AMD1, 21000-20, and 21000-21).

These form the basis for smart contracts. Part 23 refers to a set of ontologies for the coding of media asset Intellectual Property Rights (IPR), and lists the object hierarchies which are to be utilised to construct a smart contract.

Two semantic formats have been developed in the MPEG-21 suite:

- The Contract Expression Language (CEL) [259] - a language for representing media contracts with XML
- The Media Contract Ontology (MCO) [232] - a language for representing media contracts as ontologies with RDF (OWL).

The IPR ontologies also includes:

- The Media Value Chain Ontology (MVCO) [260]. This facilitates rights tracking for transparent payment of royalties
- The Audio Value Chain (AVCO) [261] - extends MVCO functionality - IP entities in the audio domain.

The principles of these ontologies can manage the execution of rights-related processes in platforms that support the contractual activities, specifically Distributed Ledger Technologies (DLTs), i.e. blockchain platforms.

The aim of MPEG's work is to develop the protocols and APIs for converting between ontologies and contract languages to smart contracts. Translation of MPEG-21 contracts to smart contracts will ensure a clean correlation between human-readable MPEG-21 contracts and smart contracts.

Maintaining a standard for conversion of contracts (in both directions between MPEG-21 ontologies and smart contracts) will ensure MPEG-21 ontologies and languages prevail as the interlingua for transferring verified contractual data from one DLT to another.

The main parts of MPEG-21 which support the smart contract work in Part 23 are:

- ISO/IEC 21000-3, Information technology - Multimedia framework (MPEG-21) - Digital Item Identifier
- ISO/IEC 21000-19, Information technology - Multimedia framework (MPEG-21) - Media Value Chain Ontology
- ISO/IEC 21000-19/AMD1, Information Technology – Multimedia Framework (MPEG-21) – Part 19: Media Value Chain Ontology / AMD 1 Extensions on Time-Segments and Multi-Track Audio', June 2018.
- ISO/IEC 21000-20 (2nd Ed), Information technology - Multimedia framework (MPEG-21) - Contract Expression Language
- ISO/IEC 21000-21 (2nd Ed), Information technology - Multimedia framework (MPEG-21) - Media Contract Ontology

MPEG-21 Part 23 provides methods to allow conversion from MPEG-21 semantic descriptions to a smart contract, and also the conversion in the reverse direction.

4.3.5 MPAI standards

The MPAI Community [262] are addressing a number of advanced technologies for next generation media. For scene description their initiative is MPAI-OSD - Visual Object and Scene Description [263]. They see visual object and

scene description as a collection of Use Cases sharing the goal of describing visual objects and locate them in a space. Scene description includes the usual description of objects and their attributes in a scene and the semantic description of the objects.

They have derived application notes to consider the use cases. The "object and scene description" component of several use cases is used to indicate a description (language) of objects and their attributes, and the semantic description of the individual objects in a scene. Proprietary solutions can address the needs of the example use cases, as follows:

- Vision-to-sound transformation
- Integrative genomic/video experiments
- Audio Recording Preservation
- Person movement description, specifically Multiplayer online gaming, Posture analysis
- Scene description, specifically AI-assisted driving, Integrative genomic/video experiments
- Generic object description for gaming and automotive applications
- Person identification, Person matching
- Conversation with emotion
- Multi-modal Question Answering
- Movement description; Human, animal; Integrative genomic and video experiments

In conclusion, MPEG is not the only approach to standardising media assets for the next generation of personalised, interactive and immersive media. However, it is a well-known family of standards with a proven and well-understood ecosystem. The standards are maintained and continually developed by academics and practitioners across the globe and across the media industry and span the breadth and depth of all media activities. The standards addressing immersive media description and smart contract deployment and management are already well developed. The industry would be hindered by a disparate range of approaches and so it is recommended that standards such as MPEG continue to be adopted to enable efficient globally deployed media services for interaction and personalisation, especially for commercial use.

5 Open problems & future directions

5.1 TV content decomposition

Future directions in temporal video fragmentation include the support for more abstract yet semantically-coherent notions of video segments. Cognitive science has shown that humans consistently **segment videos into meaningful temporal chunks**, without strongly-defined types of segments [264]. Computer-based methods replicating this human behaviour could result in video segments that encapsulate more generic "events" - even combining temporal video segmentation with notions from the event detection domain [265] -

instead of detecting more strictly-defined domain-specific segments (e.g. movie segmentation based on cinematography editing rules [266, 267], play-break segmentation of soccer [268]). One way this could be achieved is by automatically adjusting the level of granularity, e.g. to produce coarser-than-shot segments when dealing with visually and semantically similar shots, or finer-than-shot segments when needed to account for quick changes in the visual content, e.g. a fast-moving foreground object. Given that temporal video fragmentation is usually the first step of a video analysis pipeline, it is easily understood that the subsequent stages in most video analysis pipelines could potentially benefit from such a flexible temporal video fragmentation.

Regarding object detection (including semantic segmentation), the most discussed open problem in the literature is implementing a well-performing **few-shot training scheme**, i.e. incrementally learning to detect new classes, with very few training examples. This can greatly reduce the effort required to learn new object classes, especially if achieved in a weakly supervised way [269]. Additionally, using spatial and temporal relationships between the frames for **video object recognition**, instead of processing each frame as a separate still image, is an open problem [270]. A common solution is to employ object tracking techniques in order to perform object detection in a sparse sampling of frames and then propagate the results to nearby frames, estimating the new positions of objects with object tracking techniques. Future research should focus on deep-learning models that will unify object detection and tracking. Finally, **employing models resulting from neural architecture search for object detection** is already an actively growing area (as discussed in Section 2.1.2) of great potential, yet still in its nascency.

5.2 TV content annotation

It is a commonly known fact that deep learning requires a large amount of, most often labelled, training data. Such a large amount of data may not always be accessible which can lead to a scarcity of publicly available data for training neural networks. Indeed, in practice, researchers often fine-tune existing pre-trained models, instead of training deep CNNs from scratch [271]. Therefore, future research should focus on devising **new data augmentation methods for expanding limited available training data**, or explore methods for unsupervised learning, such as GANs, where a generator network works in partnership with a discriminator network [271]. Unsupervised learning avoids the need for manual data labelling by automatically discovering patterns in the data, such that the network model can generate new outputs that could have been drawn from the original real dataset. In this regard, it is rather like generating augmented data. Impressive examples of the latter exist, where GANs have been used to create very realistic, but completely artificial, fake human faces [272].

Concerning cross-modal representation and retrieval, besides algorithmic improvements, e.g. designing effective yet efficient methods, a major open concern is the limited use of information from more than two modalities [273].

Several existing benchmark datasets, such as ActivityNet [274], Vatex [275], MSR-VTT [95], YouTube8M [276], are made of video, audio and textual (e.g. caption) information. On the other hand, social networks such as YouTube, Facebook, Instagram, and Twitter, accumulate large amounts of even richer multi-modal data (i.e. text, video, audio, and related sentiment, popularity, usage-related etc. information), data which could be exploited for training cross-modal retrieval models. The key question is how to use restricted and/or noisy data as multi-modal annotations and ultimately learn semantic relations among different media modalities [277]. Cross-media retrieval method performance is directly proportional to the nature of the dataset used for training, as argued in [278], therefore future **cross-media techniques should also investigate the use of such less-structured data sources**.

Research in the area of deep learning for classification and annotation tasks will continue to bring forward ever more (marginally) accurate systems. Both Vision Transformers [279] and CoAtNets [280] are currently the best performers, but this can change quickly. Research on video annotation still lags behind image annotation, where features such as temporality or movement in the video may also be an input to the network (as opposed to still images taken from the video). The latest research has shifted towards multi-modal networks as part of Multi-Task Learning (MTL), where the same network can classify different inputs which are of different modalities such as text, image or video equally (e.g. data2vec[28]). However, what we also want to see is networks optimised to use a combination of modalities (i.e. text, audio and video as features of the same input) which is needed for instance-level annotation. Combined with unsupervised learning and the Web as a large-scale real-time knowledge source, such networks could even come to annotate correctly previously unseen instances in a video. The holy grail would be an annotation system able to **generalise from the known classes to also classify accurately unknown classes** (where there is no training data) as, in the context of annotating of TV content with broad topics, the lack of sufficient training examples is still a bottle-neck for supervised approaches. Either synthetic data needs to be generated (e.g. [281]) to cover the missing classes or state-of-the-art approaches to unsupervised learning - few-shot, one-shot and especially zero-shot - need to be employed [282, 283].

5.3 TV content re-purposing

Starting with technologies to find the right content, i.e. text-to-video techniques, an open problem concerns the **detailed temporal alignment** of a diverse textual query to the visual semantic exploited from the given video [284, 285]. A coarse matching between two modalities (e.g. matching a short sentence to a video shot) is not most effective for real-world applications since it fails to localise the exact moment expressed by a detailed query. An envisaged approach to alleviate this is combining ad-hoc video search

[28] https://ai.facebook.com/research/data2vec-a-general-framework-for-self-supervised-learning-in-speech-vision-and-language

with dense video captioning, a direction being explored in the literature very recently ([286, 287]). A **more fine-grained multi-modal analysis**, i.e. the ability to identify more concepts in the visual content as well as in the text query, is also an important objective. This may be achieved by additionally employing object detection models' results as the features that describe visual content ([288, 289]).

Moving on to content transformation and specifically regarding video summarization, the research community seems to be lately putting effort towards the development of deep learning architectures that can be trained in a **fully-unsupervised**, or at least weakly-supervised, manner so as to completely overcome the need for large collections of human-annotated training data [62]. Additionally, some recent summarization methods aim to meet additional criteria about the content of the generated summary, e.g. its visual diversity [132] and its uniformity [290]. Such efforts towards **offering more control in the process of summarization** using easily understood human notions, are promising future research directions.

In the field of highlight detection, there is a noticeable trend towards **embracing multi-modality** (e.g. [291, 292]). Relying solely on visual features is often insufficient for capturing the highlights of videos with complicated semantics, e.g. a political debate video. Furthermore, utilising transfer learning approaches, i.e. deriving an effective highlight detector on a target video category by transferring the highlight knowledge acquired from a source video category with a large collection of training data available, seems a promising direction when devising **domain-agnostic video highlight detection systems**.

Concerning super resolution techniques, a currently discussed open problem regards the **excessive computational requirements** [293]. While many solutions have been proposed already for deployment on resource-limited devices (e.g. [294, 295]), these are usually not optimised yet for common smartphones, not to mention more constrained smart TV platforms [296]. To address this, [297] introduces the first Mobile AI challenge, where the target is to develop end-to-end deep learning-based video super-resolution solutions that can achieve real-time performance on less powerful hardware. The outcome of such efforts might also benefit TV applications of AI techniques, enabling the transfer of the computational load of super-resolution (as well other tasks, such as aspect-ratio adaptation) to client devices, like smart TVs and tablets. Finally, the short list of methods for video super resolution methods, discussed in Section 2.3.2, highlights the need for **approaches specifically designed for videos**, i.e. that take into consideration the temporal continuity of video frames to offer gains in both speed of execution as well as the quality of the final results.

5.4 In-stream personalisation

Traditional personalisation and recommendation systems are faced with two long-standing obstacles, namely, data sparsity and cold-start problems [298].

Increasing concern by consumers about data collection and tighter laws on data privacy (e.g. GDPR) mean that future systems should be able to **personalise media without needing to build individual user profiles** which touch on the area of personal data. Collaborative filtering is one working solution to this, i.e. viewers are categorised into groups that closely share viewing patterns (or other signals for determining interest in particular media items). Google has been exploring such an approach on the Web known as Federated Learning of Cohorts[29] (FLoC) for targeted advertising without personal data. In the ReTV project, the audience was similarly clustered into groups that shared similar viewing patterns. Models were evaluated for predictive ability, i.e. given the past viewing history, to predict which program they would watch next. This prediction then acted as a recommendation. A baseline model using Non-Negative Matrix Factorization (NNMF) was compared to Field-aware Factorization Machines (FFM) where additional features learned from content analysis was used (i.e. topics, genres and concepts in the program), leading to an over 100% improvement on predictive ability (i.e. the prediction of which program would be next watched matches the actual program that was next watched) [299]. This work suggests that content feature extraction (the classification of media discussed in Section 2.2.1) can be used in personalised systems to find relevant media for viewers separate from individual profiling of their personal interests, an activity which increasingly concerns viewers. Apart from avoiding the cold-start problem, it might also ensure a broader range of media items than traditional classification-based approaches which tend to develop a "filter bubble" (users watch a lot of x so they are recommended more of x, at the expense of all other content). Future work on TV personalisation must leverage information collected from other domains to alleviate both problems (i.e. move towards an approach to cross-domain recommendation) and discover the link among media content, activities and user interaction as well, e.g. formulating this as a relationship between triggered emotions and user expectations [300].

Despite advances in recommendation systems through AI and deep learning, the focus has remained to date on the whole media item as the object of recommendation. In the future, while the subject of recommendation is becoming less granular (from the individual viewer to a viewing group), the object of recommendation needs to become more granular - **the recommended content may be a dynamically composed media item from a set of different decomposed media objects** (e.g. video segments). While standalone research efforts have been performed to show some of the possibilities of in-stream personalisation (e.g. putting together a news program made up only of segments of interest to the viewer, such as [301]) truly personalised TV will only emerge when these personalisation technologies are combined with solutions for the other aforementioned technical requirements - the media decomposition, annotation, and re-purposing - and deployed in a production setting. A key prerequisite to this will be sufficient accuracy and relevance in

[29] https://blog.google/products/ads-commerce/2021-01-privacy-sandbox/

results to not only retain but to attract new viewers to what will emerge as a radically new form of media consumption. One where **no one viewer might see exactly the same content stream as another**. Where even the clothes on the program characters, the cars they drive or the dialect they use when speaking might be different according to the preference of the viewer.

5.5 Standards

This paper has considered the range of interactive and personalised media services where scenes can be composed of media objects. Media trading has also been described along with the technologies and standards required to ensure the contracted commercial acquisition of objects on a future object market. Section 3.4 described TV personalisation where objects are commercially traded in near real-time.

The procedures and standards described in Section 4 can allow for this, but to date have not specifically addressed the real-time object trading issue. An issue to be addressed here is the **dynamic derivation of scene graphing to allow for objects to be "null"** in the original production, and awaiting import of objects for the personalised resulting playout to the end viewer. In these cases, the metadata needs to describe content parameters to allow for editorial control by the original producer. For example - allowing or disallowing specific objects to be used in personalised selections, for example on the basis of ethical or moral preferences by the content author or production team.

Another issue is the potential for **tracking content versioning in a process where there may be limitless versions derived**. This may affect industry standards such as Application Specification AS-02 of Material eXchange Format, MXF [302] or the Interoperable Media Format, IMF [303].

In summary, standards activities need to address some new issues. These include standards for support of external library content traded into a video (in near real-time), tracking media component assets (e.g. audio and video objects) throughout the media content value cycle and the application of production versioning to personalised interactive media.

6 Conclusions

In this work, we provided a review of the AI and data technologies landscape that can aid the data-driven personalisation of television content. We identified classes of relevant technologies and we discussed how these technologies have evolved over the last years, what is the current state-of-the-art and what is the potential for the future. Additionally, we provided examples of tools/services/products that are already in use to support advanced personalisation functionalities. Furthermore, we analysed the existing relevant standards, again looking at open problems and future directions. We hope that this survey can encourage further efforts, both by the research community towards advancing the relevant technologies, and the content owners/media

organisation/TV broadcasters in order to adopt and integrate such technologies, putting them to test under realistic conditions and using them in real-life applications.

Acknowledgments

This work was supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-780656 ReTV and H2020-951911 AI4Media.

Declarations

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Apostolidis, E., Mezaris, V.: Fast shot segmentation combining global and local visual descriptors. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6583–6587 (2014). IEEE

[2] Tsamoura, E., Mezaris, V., Kompatsiaris, I.: Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In: 2008 15th IEEE International Conference on Image Processing, pp. 45–48 (2008). IEEE

[3] Xiao, Z.-m., Lin, K.-h., Zhou, C.-l., Lin, Q.: Shot segmentation based on HSV color model. Journal of Xiamen University (Natural Science) **5** (2008)

[4] Küçüktunç, O., Güdükbay, U., Ulusoy, Ö.: Fuzzy color histogram-based video segmentation. Computer Vision and Image Understanding **114**(1), 125–134 (2010)

[5] Baber, J., Afzulpurkar, N., Dailey, M.N., Bakhtyar, M.: Shot boundary detection from videos using entropy and local descriptor. In: 2011 17th International Conference on Digital Signal Processing (DSP), pp. 1–6 (2011). IEEE

[6] e Santos, A.C.S., Pedrini, H.: Shot boundary detection for video temporal segmentation based on the weber local descriptor. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1310–1315 (2017). IEEE

[7] Hassanien, A., Elgharib, M., Selim, A., Bae, S.-H., Hefeeda, M., Matusik, W.: Large-scale, fast and accurate shot boundary detection

through spatio-temporal convolutional neural networks. arXiv preprint arXiv:1705.03281 (2017)

[8] Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW), pp. 117–122 (2018). IEEE

[9] Gygli, M.: Ridiculously fast shot boundary detection with fully convolutional neural networks. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–4 (2018). https://doi.org/10.1109/CBMI.2018.8516556

[10] Souček, T., Lokoč, J.: Transnet v2: An effective deep network architecture for fast shot transition detection. arXiv preprint arXiv:2008.04838 (2020)

[11] Lokoč, J., Kovalčík, G., Souček, T., Moravec, J., Čech, P.: A framework for effective known-item search in video. In: In Proceedings of the 27th ACM International Conference on Multimedia (MM'19), October 21-25, 2019, Nice, France, pp. 1–9 (2019). https://doi.org/10.1145/3343031.3351046

[12] Lei, X., Pan, H., Huang, X.: A dilated CNN model for image classification. IEEE Access **7**, 124087–124095 (2019)

[13] Tang, S., Feng, L., Kuang, Z., Chen, Y., Zhang, W.: Fast video shot transition localization with deep structured models. In: Asian Conference on Computer Vision, pp. 577–592 (2018). Springer

[14] Gushchin, A., Antsiferova, A., Vatolin, D.: Shot boundary detection method based on a new extensive dataset and mixed features. arXiv preprint arXiv:2109.01057 (2021)

[15] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., Trancoso, I.: Temporal video segmentation to scenes using high-level audiovisual features. IEEE Transactions on Circuits and Systems for Video Technology **21**(8), 1163–1177 (2011)

[16] Kishi, R.M., Trojahn, T.H., Goularte, R.: Correlation based feature fusion for the temporal video scene segmentation task. Multimedia Tools and Applications **78**(11), 15623–15646 (2019)

[17] Baraldi, L., Grana, C., Cucchiara, R.: A deep siamese network for scene detection in broadcast videos. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1199–1202 (2015)

[18] Rotman, D., Porat, D., Ashour, G., Barzelay, U.: Optimally grouped

deep features using normalized cost for video scene detection. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 187–195 (2018)

[19] Apostolidis, K., Apostolidis, E., Mezaris, V.: A motion-driven approach for fine-grained temporal segmentation of user-generated videos. In: International Conference on Multimedia Modeling, pp. 29–41 (2018). Springer

[20] Peleshko, D., Soroka, K.: Research of usage of haar-like features and AdaBoost algorithm in viola-jones method of object detection. In: 2013 12th International Conference on the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), pp. 284–286 (2013). IEEE

[21] Nguyen, T., Park, E.-A., Han, J., Park, D.-C., Min, S.-Y.: Object detection using scale invariant feature transform. In: Pan, J.-S., Krömer, P., Snášel, V. (eds.) Genetic and Evolutionary Computing, pp. 65–72. Springer, Cham (2014)

[22] Bouguila, N., Ziou, D.: A dirichlet process mixture of dirichlet distributions for classification and prediction. In: 2008 IEEE Workshop on Machine Learning for Signal Processing, pp. 297–302 (2008). IEEE

[23] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

[24] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)

[25] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)

[26] Pramanik, A., Pal, S.K., Maiti, J., Mitra, P.: Granulated RCNN and multi-class deep sort for multi-object detection and tracking. IEEE Transactions on Emerging Topics in Computational Intelligence (2021)

[27] Yao, Y.: Granular computing: basic issues and possible solutions. In: Proceedings of the 5th Joint Conference on Information Sciences, vol. 1, pp. 186–189 (2000). Citeseer

[28] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788

(2016)

[29] Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)

[30] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

[31] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016, pp. 21–37. Springer, Cham (2016)

[32] Sanchez, S., Romero, H., Morales, A.: A review: Comparison of performance metrics of pretrained models for object detection using the tensorflow framework. In: IOP Conference Series: Materials Science and Engineering, vol. 844, p. 012024 (2020). IOP Publishing

[33] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

[34] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[35] Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)

[36] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014). Springer

[37] Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)

[38] Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M.: You only learn one representation: Unified network for multiple tasks. arXiv preprint arXiv:2105.04206 (2021)

[39] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

[40] Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic

segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528 (2015)

[41] Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3194–3203 (2016)

[42] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)

[43] Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: European Conference on Computer Vision (ECCV), pp. 173–190 (2020). Springer

[44] Jain, J., Singh, A., Orlov, N., Huang, Z., Li, J., Walton, S., Shi, H.: SeMask: Semantically masked transformers for semantic segmentation. arXiv preprint arXiv:2112.12782 (2021)

[45] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. arXiv preprint arXiv:2111.09883 (2021)

[46] Hao, S., Zhou, Y., Guo, Y.: A brief survey on semantic segmentation with deep learning. Neurocomputing **406**, 302–321 (2020)

[47] Lan, Z.-z., Bao, L., Yu, S.-I., Liu, W., Hauptmann, A.G.: Multimedia classification and event detection using double fusion. Multimedia tools and applications **71**(1), 333–347 (2014)

[48] Daudpota, S.M., Muhammad, A., Baber, J.: Video genre identification using clustering-based shot detection algorithm. Signal, Image and Video Processing **13**(7), 1413–1420 (2019)

[49] Gkalelis, N., Mezaris, V.: Subclass deep neural networks: re-enabling neglected classes in deep network training for multimedia classification. In: International Conference on Multimedia Modeling, pp. 227–238 (2020). Springer

[50] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[51] Pouyanfar, S., Chen, S.-C., Shyu, M.-L.: An efficient deep residual-inception network for multimedia classification. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 373–378 (2017).

IEEE

[52] Shamsolmoali, P., Jain, D.K., Zareapoor, M., Yang, J., Alam, M.A.: High-dimensional multimedia classification using deep cnn and extended residual units. Multimedia Tools and Applications **78**(17), 23867–23882 (2019)

[53] Dai, X., Yin, H., Jha, N.K.: Incremental learning using a grow-and-prune paradigm with efficient neural networks. IEEE Transactions on Emerging Topics in Computing (2020)

[54] Gkalelis, N., Mezaris, V.: Structured pruning of lstms via eigenanalysis and geometric median for mobile multimedia and deep learning applications. In: 2020 IEEE International Symposium on Multimedia (ISM), pp. 122–126 (2020). IEEE

[55] Chiodino, E., Di Luccio, D., Lieto, A., Messina, A., Pozzato, G.L., Rubinetti, D.: A knowledge-based system for the dynamic generation and classification of novel contents in multimedia broadcasting. In: ECAI 2020, pp. 680–687 (2020)

[56] Doulaty, M., Saz-Torralba, O., Ng, R.W.M., Hain, T.: Automatic genre and show identification of broadcast media. In: INTERSPEECH (2016)

[57] Yadav, A., Vishwakarma, D.K.: A unified framework of deep networks for genre classification using movie trailer. Applied Soft Computing **96**, 106624 (2020)

[58] Mills, T.J., Pye, D., Hollinghurst, N.J., Wood, K.R.: AT_TV: Broadcast television and radio retrieval. In: RIAO, pp. 1135–1144 (2000)

[59] Smeaton, A.F., Over, P., Kraaij, W.: High-level feature detection from video in TRECVid: a 5-year retrospective of achievements. In: Multimedia Content Analysis, pp. 1–24 (2009)

[60] Rossetto, L., Amiri Parian, M., Gasser, R., Giangreco, I., Heller, S., Schuldt, H.: Deep learning-based concept detection in vitrivr. In: International Conference on Multimedia Modeling, pp. 616–621 (2019). Springer

[61] Agarwal, A., Mangal, A., et al.: Visual relationship detection using scene graphs: A survey. arXiv preprint arXiv:2005.08045 (2020)

[62] Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Video summarization using deep neural networks: A survey. Proceedings of the IEEE **109**(11), 1838–1863 (2021). https://doi.org/10.1109/JPROC.2021.3117472

[63] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)

[64] Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E.: Deep learning for computer vision: A brief review. Computational intelligence and neuroscience **2018** (2018)

[65] Touvron, H., Vedaldi, A., Douze, M., Jégou, H.: Fixing the train-test resolution discrepancy: Fixefficientnet. arXiv preprint arXiv:2003.08237 (2020)

[66] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). IEEE

[67] Gkalelis, N., Goulas, A., Galanopoulos, D., Mezaris, V.: Objectgraphs: Using objects and a graph convolutional network for the bottom-up recognition and explanation of events in video. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3370–3378 (2021). https://doi.org/10.1109/CVPRW53098.2021.00376

[68] Pouyanfar, S., Chen, S.-C.: Semantic event detection using ensemble deep learning. In: 2016 IEEE International Symposium on Multimedia (ISM), pp. 203–208 (2016). IEEE

[69] Marechal, C., Mikolajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., Wegrzyn-Wolska, K.: Survey on AI-Based Multimodal Methods for Emotion Detection. (2019)

[70] Kwak, C.-U., Son, J.-W., Lee, A., Kim, S.-J.: Scene emotion detection using closed caption based on hierarchical attention network. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1206–1208 (2017). IEEE

[71] Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 467–474 (2015)

[72] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing **10**(1), 60–75 (2017)

[73] Vandersmissen, B., Sterckx, L., Demeester, T., Jalalvand, A., De Neve, W., Van de Walle, R.: An automated end-to-end pipeline for fine-grained

video annotation using deep neural networks. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 409–412 (2016)

[74] Haynes, M., Norton, A., McParland, A., Cooper, R.: Speech-to-text for broadcasters, from research to implementation. SMPTE Motion Imaging Journal **127**(2), 27–33 (2018). https://doi.org/10.5594/JMI.2018.2790658

[75] Sharma, D.P., Atkins, J.: Automatic speech recognition systems: challenges and recent implementation trends. International Journal of Signal and Imaging Systems Engineering **7**(4), 220–234 (2014)

[76] Radzikowski, K., Wang, L., Yoshie, O., Nowak, R.: Accent modification for speech recognition of non-native speakers using neural style transfer. EURASIP Journal on Audio, Speech, and Music Processing **2021**(1), 1–10 (2021)

[77] Nixon, L., Mezaris, V., Thomsen, J.: Seamlessly interlinking tv and web content to enable linked television. In: ACM Int. Conf. on Interactive Experiences for Television and Online Video (TVX 2014), Adjunct Proceedings, Newcastle Upon Tyne, UK, p. 21 (2014)

[78] Liu, A.H., Jin, S., Lai, C.-I.J., Rouditchenko, A., Oliva, A., Glass, J.: Cross-modal discrete representation learning. arXiv preprint arXiv:2106.05438 (2021)

[79] Guo, W., Wang, J., Wang, S.: Deep multimodal representation learning: A survey. IEEE Access **7**, 63373–63394 (2019). https://doi.org/10.1109/ACCESS.2019.2916887

[80] Wang, Y.: Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(1s), 1–25 (2021)

[81] Jin, W., Zhao, Z., Zhang, P., Zhu, J., He, X., Zhuang, Y.: Hierarchical cross-modal graph consistency learning for video-text retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1114–1124 (2021)

[82] Habibian, A., Mensink, T., Snoek, C.G.M.: Video2vec embeddings recognize events when examples are scarce. IEEE Trans. Pattern Anal. Mach. Intell. **39**(10), 2089–2103 (2017). https://doi.org/10.1109/TPAMI.2016.2627563

[83] Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2VV++: Fully deep learning for ad-hoc video search. Proceedings of the 27th ACM International

Conference on Multimedia (2019)

[84] Dong, J., Li, X., Snoek, C.G.: Word2visualvec: Cross-media retrieval by visual feature prediction. arXiv preprint arXiv:1604.06838 **2** (2016)

[85] Galanopoulos, D., Mezaris, V.: Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 336–340 (2020)

[86] Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9346–9355 (2019)

[87] Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7464–7473 (2019)

[88] Ruan, L., Jin, Q.: Survey: Transformer based video-language pre-training. AI Open **3**, 1–13 (2022). https://doi.org/10.1016/j.aiopen.2022.01.001

[89] Li, L., Chen, Y.-C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: HERO: Hierarchical encoder for video+ language omni-representation pre-training. In: EMNLP (2020)

[90] Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7331–7341 (2021)

[91] Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743 (2019)

[92] Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8746–8755 (2020)

[93] Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)

[94] Gao, Z., Liu, J., Chen, S., Chang, D., Zhang, H., Yuan, J.: CLIP2TV: An

empirical study on transformer-based methods for video-text retrieval. arXiv preprint arXiv:2111.05610 (2021)

[95] Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296 (2016)

[96] Kim, C., Hwang, J.-N.: Object-based video abstraction for video surveillance systems. IEEE Transactions on Circuits and Systems for Video Technology **12**(12), 1128–1138 (2002). https://doi.org/10.1109/TCSVT.2002.806813

[97] Ejaz, N., Tariq, T.B., Baik, S.W.: Adaptive key frame extraction for video summarization using an aggregation mechanism. Journal of Visual Communication and Image Representation **23**(7), 1031–1040 (2012). https://doi.org/10.1016/j.jvcir.2012.06.013

[98] Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: Stimo: STIll and MOving Video Storyboard for the Web Scenario. Multimedia Tools Appl. **46**(1), 47–69 (2010). https://doi.org/10.1007/s11042-009-0307-7

[99] de Avila, S.E.F., Lopes, A.P.B.a., da Luz, A. Jr., de Albuquerque Araújo, A.: Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn. Lett. **32**(1), 56–68 (2011). https://doi.org/10.1016/j.patrec.2010.08.004

[100] Almeida, J., Leite, N.J., Torres, R.d.S.: Vison: VIdeo Summarization for ONline Applications. Pattern Recogn. Lett. **33**(4), 397–409 (2012). https://doi.org/10.1016/j.patrec.2011.08.007

[101] Chu, W., Song, Y., Jaimes, A.: Video co-summarization: Video summarization by visual co-occurrence. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3584–3592 (2015). https://doi.org/10.1109/CVPR.2015.7298981

[102] Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: Sparse modeling for finding representative objects. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1600–1607 (2012). https://doi.org/10.1109/CVPR.2012.6247852

[103] Ma, M., Mei, S., Wan, S., Wang, Z., Feng, D.: Video summarization via nonlinear sparse dictionary selection. IEEE Access **7**, 11763–11774 (2019). https://doi.org/10.1109/ACCESS.2019.2891834

[104] Zhao, B., Xing, E.P.: Quasi real-time summarization for consumer videos. In: 2014 IEEE Conference on Computer Vision and Pattern

Recognition, pp. 2513–2520 (2014). https://doi.org/10.1109/CVPR.2014.322

[105] Lai, J.-L., Yi, Y.: Key frame extraction based on visual attention model. Journal of Visual Communication and Image Representation **23**(1), 114–125 (2012). https://doi.org/10.1016/j.jvcir.2011.08.005

[106] Ejaz, N., Mehmood, I., Baik, S.W.: Feature aggregation based visual attention model for video summarization. Computers & Electrical Engineering **40**(3), 993–1005 (2014). https://doi.org/10.1016/j.compeleceng.2013.10.005. Special Issue on Image and Video Processing

[107] Zhang, Y., Tao, R., Wang, Y.: Motion-state-adaptive video summarization via spatiotemporal analysis. IEEE Transactions on Circuits and Systems for Video Technology **27**(6), 1340–1352 (2017). https://doi.org/10.1109/TCSVT.2016.2539638

[108] Gygli, M., Grabner, H., Gool, L.V.: Video summarization by learning submodular mixtures of objectives. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3090–3098 (2015). https://doi.org/10.1109/CVPR.2015.7298928

[109] Li, X., Zhao, B., Lu, X.: A general framework for edited video and raw video summarization. IEEE Transactions on Image Processing **26**(8), 3652–3664 (2017). https://doi.org/10.1109/TIP.2017.2695887

[110] Elfeki, M., Borji, A.: Video summarization via actionness ranking. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, January 7-11, 2019, pp. 754–763 (2019). https://doi.org/10.1109/WACV.2019.00085

[111] Panda, R., Das, A., Wu, Z., Ernst, J., Roy-Chowdhury, A.K.: Weakly supervised summarization of web videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3677–3686 (2017). https://doi.org/10.1109/ICCV.2017.395

[112] Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, pp. 358–374. Springer, Cham (2018)

[113] Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., Remagnino, P.: Summarizing videos with attention. In: Carneiro, G., You, S. (eds.) Computer Vision – ACCV 2018 Workshops, pp. 39–54. Springer, Cham (2019)

[114] Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., Yokoya, N.: Video

summarization using deep semantic features. In: The 13th Asian Conference on Computer Vision (ACCV'16) (2016)

[115] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[116] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1179. https://www.aclweb.org/anthology/D14-1179

[117] Zhang, K., Chao, W.-L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016, pp. 766–782. Springer, Cham (2016)

[118] Ji, Z., Xiong, K., Pang, Y., Li, X.: Video summarization with attention-based encoder-decoder networks. IEEE Transactions on Circuits and Systems for Video Technology, 1–1 (2019). https://doi.org/10.1109/TCSVT.2019.2904996

[119] Fu, T., Tai, S., Chen, H.: Attentive and adversarial learning for video summarization. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, January 7-11, 2019, pp. 1579–1587 (2019). https://doi.org/10.1109/WACV.2019.00173. https://doi.org/10.1109/WACV.2019.00173

[120] Feng, L., Li, Z., Kuang, Z., Zhang, W.: Extractive video summarizer with memory augmented neural networks. In: Proceedings of the 26th ACM International Conference on Multimedia. MM '18, pp. 976–983. ACM, New York, NY, USA (2018). https://doi.org/10.1145/3240508.3240651. http://doi.acm.org/10.1145/3240508.3240651

[121] Zhao, B., Li, X., Lu, X.: Hierarchical recurrent neural network for video summarization. In: Proceedings of the 2017 ACM on Multimedia Conference. MM '17, pp. 863–871. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3123266.3123328. http://doi.acm.org/10.1145/3123266.3123328

[122] Zhao, B., Li, X., Lu, X.: HSA-RNN: Hierarchical structure-adaptive rnn for video summarization. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '18 (2018)

[123] Zhang, Y., Kampffmeyer, M., Liang, X., Zhang, D., Tan, M., Xing,

E.P.: Dtr-gan: Dilated temporal relational adversarial network for video summarization. CoRR **abs/1804.11228** (2018)

[124] Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. IEEE Transactions on Circuits and Systems for Video Technology (2020)

[125] Jung, Y., Cho, D., Kim, D., Woo, S., Kweon, I.S.: Discriminative feature learning for unsupervised video summarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8537–8544 (2019)

[126] Jung, Y., Cho, D., Woo, S., Kweon, I.S.: Global-and-local relative position embedding for unsupervised video summarization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, pp. 167–183 (2020). Springer

[127] Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Unsupervised video summarization via attention-driven adversarial learning. In: International Conference on Multimedia Modeling, pp. 492–504 (2020). Springer

[128] Apostolidis, E., Metsai, A.I., Adamantidou, E., Mezaris, V., Patras, I.: A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization. In: Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery, pp. 17–25 (2019)

[129] Wang, J., Wang, W., Wang, Z., Wang, L., Feng, D., Tan, T.: Stacked memory network for video summarization. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 836–844 (2019)

[130] Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., Remagnino, P.: Summarizing videos with attention. In: Asian Conference on Computer Vision, pp. 39–54 (2018). Springer

[131] Liu, Y.-T., Li, Y.-J., Yang, F.-E., Chen, S.-F., Wang, Y.-C.F.: Learning hierarchical self-attention for video summarization. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3377–3381 (2019). IEEE

[132] Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., Shao, L.: Exploring global diverse attention via pairwise temporal relation for video summarization. Pattern Recognition **111**, 107677 (2021)

[133] Ji, Z., Jiao, F., Pang, Y., Shao, L.: Deep attentive and semantic preserving video summarization. Neurocomputing **405**, 200–207 (2020)

[134] Apostolidis, E., Balaouras, G., Mezaris, V., Patras, I.: Combining global and local attention with positional encoding for video summarization. In: 2021 IEEE International Symposium on Multimedia (ISM), pp. 226–234 (2021). IEEE

[135] Xu, M., Jin, J.S., Luo, S., Duan, L.: Hierarchical movie affective content analysis based on arousal and valence features. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 677–680 (2008)

[136] Xiong, B., Kalantidis, Y., Ghadiyaram, D., Grauman, K.: Less is more: Learning highlight detection from video duration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1258–1267 (2019)

[137] Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T.S.: Highlights extraction from sports video based on an audio-visual marker detection framework. In: 2005 IEEE International Conference on Multimedia and Expo, p. 4 (2005). IEEE

[138] Tang, H., Kwatra, V., Sargin, M.E., Gargi, U.: Detecting highlights in sports videos: Cricket as a test case. In: 2011 IEEE International Conference on Multimedia and Expo, pp. 1–6 (2011). IEEE

[139] Wang, J., Xu, C., Chng, E., Tian, Q.: Sports highlight detection from keyword sequences using HMM. In: 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), vol. 1, pp. 599–602 (2004). IEEE

[140] Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for tv baseball programs. In: Proceedings of the Eighth ACM International Conference on Multimedia, pp. 105–115 (2000)

[141] Sun, M., Farhadi, A., Seitz, S.: Ranking domain-specific highlights by analyzing edited videos. In: European Conference on Computer Vision, pp. 787–802 (2014). Springer

[142] Petkovic, M., Mihajlovic, V., Jonker, W., Djordjevic-Kajan, S.: Multi-modal extraction of highlights from tv formula 1 programs. In: Proceedings of IEEE International Conference on Multimedia and Expo, vol. 1, pp. 817–820 (2002). IEEE

[143] Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 982–990 (2016)

[144] Gygli, M., Song, Y., Cao, L.: Video2gif: Automatic generation of animated gifs from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1001–1009 (2016)

[145] Jiao, Y., Li, Z., Huang, S., Yang, X., Liu, B., Zhang, T.: Three-dimensional attention-based deep ranking model for video highlight detection. IEEE Transactions on Multimedia **20**(10), 2693–2705 (2018)

[146] Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: European Conference on Computer Vision, pp. 540–555 (2014). Springer

[147] Yang, H., Wang, B., Lin, S., Wipf, D., Guo, M., Guo, B.: Unsupervised extraction of video highlights via robust recurrent auto-encoders. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4633–4641 (2015)

[148] Panda, R., Das, A., Wu, Z., Ernst, J., Roy-Chowdhury, A.K.: Weakly supervised summarization of web videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3657–3666 (2017)

[149] Hong, F.-T., Huang, X., Li, W.-H., Zheng, W.-S.: Mini-net: Multiple instance ranking network for video highlight detection. In: European Conference on Computer Vision, pp. 345–360 (2020). Springer

[150] Rochan, M., Reddy, M.K.K., Ye, L., Wang, Y.: Adaptive video highlight detection by learning from user history. In: European Conference on Computer Vision, pp. 261–278 (2020). Springer

[151] Wu, L., Yang, Y., Chen, L., Lian, D., Hong, R., Wang, M.: Learning to transfer graph embeddings for inductive graph based recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1211–1220 (2020)

[152] Xu, M., Wang, H., Ni, B., Zhu, R., Sun, Z., Wang, C.: Cross-category video highlight detection via set-based learning. arXiv preprint arXiv:2108.11770 (2021)

[153] Mundnich, K., Fenster, A., Khare, A., Sundaram, S.: Audiovisual highlight detection in videos. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4155–4159 (2021). IEEE

[154] Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE transactions on image processing **13**(10), 1327–1344 (2004)

[155] Farsiu, S., Elad, M., Milanfar, P.: Multiframe demosaicing and super-resolution from undersampled color images. In: Computational Imaging II, vol. 5299, pp. 222–233 (2004). International Society for Optics and Photonics

[156] Farsiu, S., Robinson, D.M., Elad, M., Milanfar, P.: Dynamic demosaicing and color superresolution of video sequences. In: Image Reconstruction from Incomplete Data III, vol. 5562, pp. 169–178 (2004). International Society for Optics and Photonics

[157] Yang, C.-Y., Huang, J.-B., Yang, M.-H.: Exploiting self-similarities for single frame super-resolution. In: Asian Conference on Computer Vision, pp. 497–510 (2010). Springer

[158] Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Computer graphics and Applications **22**(2), 56–65 (2002)

[159] Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)

[160] Wang, Z., Bovik, A.C.: Mean squared error: Love it or leave it? a new look at signal fidelity measures. IEEE signal processing magazine **26**(1), 98–117 (2009)

[161] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

[162] Rad, M.S., Bozorgtabar, B., Marti, U.-V., Basler, M., Ekenel, H.K., Thiran, J.-P.: Srobb: Targeted perceptual loss for single image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2710–2719 (2019)

[163] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., *et al.*: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)

[164] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0 (2018)

[165] Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with VQ-VAE-2. In: Advances in Neural Information Processing Systems, pp. 14866–14876 (2019)

[166] Gatopoulos, I., Stol, M., Tomczak, J.M.: Super-resolution variational auto-encoders. arXiv preprint arXiv:2006.05218 (2020)

[167] Atwood, J., Towsley, D.: Diffusion-convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1993–2001 (2016)

[168] Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems **34** (2021)

[169] Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. arXiv preprint arXiv:2106.15282 (2021)

[170] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636 (2021)

[171] Chadha, A., Britto, J., Roja, M.M.: iseebetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks. Computational Visual Media **6**(3), 307–317 (2020)

[172] Isobe, T., Zhu, F., Jia, X., Wang, S.: Revisiting temporal modeling for video super-resolution. In: Proceedings of the 31st British Machine Vision Conference (BMVC) (2020)

[173] Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3897–3906 (2019)

[174] Rozumnyi, D., Oswald, M.R., Ferrari, V., Matas, J., Pollefeys, M.: DeFMO: Deblurring and shape recovery of fast moving objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3456–3465 (2021)

[175] Liu, H., Ruan, Z., Zhao, P., Dong, C., Shang, F., Liu, Y., Yang, L.: Video super resolution based on deep learning: A comprehensive survey. arXiv preprint arXiv:2007.12928 (2020)

[176] Nam, H., Park, D., Jeon, K.: Jitter-robust video retargeting with kalman filter and attention saliency fusion network. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 858–862 (2020). https://do

i.org/10.1109/ICIP40778.2020.9191354

[177] Lee, H.-S., Bae, G., Cho, S.-I., Kim, Y.-H., Kang, S.: Smartgrid: Video retargeting with spatiotemporal grid optimization. IEEE Access **7**, 127564–127579 (2019)

[178] Rachavarapu, K.-K., Kumar, M., Gandhi, V., Subramanian, R.: Watch to edit: Video retargeting using gaze. In: Computer Graphics Forum, vol. 37, pp. 205–215 (2018). Wiley Online Library

[179] Jain, E., Sheikh, Y., Shamir, A., Hodgins, J.: Gaze-driven video re-editing. ACM Transactions on Graphics (TOG) **34**(2), 1–12 (2015)

[180] Deselaers, T., Dreuw, P., Ney, H.: Pan, zoom, scan – time-coherent, trained automatic video cropping. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). https://doi.org/10.1109/CVPR.2008.4587729

[181] Liu, F., Gleicher, M.: Video retargeting: automating pan and scan. In: Proceedings of the 14th ACM International Conference on Multimedia, pp. 241–250 (2006)

[182] Kaur, H., Kour, S., Sen, D.: Video retargeting through spatio-temporal seam carving using kalman filter. IET Image Processing **13**(11), 1862–1871 (2019)

[183] Wang, S., Tang, Z., Dong, W., Yao, J.: Multi-operator video retargeting method based on improved seam carving. In: 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), pp. 1609–1614 (2020). https://doi.org/10.1109/ITOEC49072.2020.9141774

[184] Wang, Y.-S., Lin, H.-C., Sorkine, O., Lee, T.-Y.: Motion-based video retargeting with optimized crop-and-warp. In: ACM SIGGRAPH 2010 Papers, pp. 1–9 (2010)

[185] Kopf, S., Haenselmann, T., Kiess, J., Guthier, B., Effelsberg, W.: Algorithms for video retargeting. Multimedia Tools Appl. **51**(2), 819–861 (2011). https://doi.org/10.1007/s11042-010-0717-6

[186] Kiess, J., Guthier, B., Kopf, S., Effelsberg, W.: SeamCrop for image retargeting. In: Multimedia on Mobile Devices 2012; and Multimedia Content Access: Algorithms and Systems VI, vol. 8304, p. 83040 (2012). International Society for Optics and Photonics

[187] Nam, S.-H., Ahn, W., Yu, I.-J., Kwon, M.-J., Son, M., Lee, H.-K.: Deep convolutional neural network for identifying seam-carving forgery. IEEE Transactions on Circuits and Systems for Video Technology (2020)

[188] Apostolidis, K., Mezaris, V.: A fast smart-cropping method and dataset for video retargeting. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 2618–2622 (2021). IEEE

[189] Chou, Y.-C., Fang, C.-Y., Su, P.-C., Chien, Y.-C.: Content-based cropping using visual saliency and blur detection. In: 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), pp. 1–6 (2017). IEEE

[190] Zhu, T., Zhang, D., Hu, Y., Wang, T., Jiang, X., Zhu, J., Li, J.: Horizontal-to-vertical video conversion. IEEE Transactions on Multimedia (2021)

[191] Smyth, B., Cotter, P.: Case-Studies on the Evolution of the Personalized Electronic Program Guide, vol. 6, pp. 53–71 (2004). https://doi.org/10.1007/1-4020-2164-X_3

[192] Kim, E., Pyo, S., Park, E., Kim, M.: An automatic recommendation scheme of tv program contents for (ip) tv personalization. IEEE Transactions on Broadcasting **57**(3), 674–684 (2011)

[193] Soares, M., Viana, P.: Tv recommendation and personalization systems: integrating broadcast and video on-demand services. Advances in Electrical and Computer Engineering **14**(1), 115–120 (2014)

[194] Hsu, S.H., Wen, M.-H., Lin, H.-C., Lee, C.-C., Lee, C.-H.: Aimed-a personalized tv recommendation system. In: European Conference on Interactive Television, pp. 166–174 (2007). Springer

[195] Aharon, M., Hillel, E., Kagian, A., Lempel, R., Makabee, H., Nissim, R.: Watch-it-next: a contextual tv recommendation system. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 180–195 (2015). Springer

[196] Aroyo, L., Nixon, L., Miller, L.: NoTube: the television experience enhanced by online social and semantic data. In: 2011 IEEE International Conference on Consumer Electronics-Berlin (ICCE-Berlin), pp. 269–273 (2011). IEEE

[197] Veloso, B., Malheiro, B., Burguillo, J., Foss, J., Gama, J.: Personalised Dynamic Viewer Profiling for Streamed Data, pp. 501–510 (2018). https://doi.org/10.1007/978-3-319-77712-2_47

[198] Gonçalves, D., Costa, M., Couto, F.M.: A flexible recommendation system for cable tv. arXiv preprint arXiv:1609.02451 (2016)

[199] Maccatrozzo, V., Terstall, M., Aroyo, L., Schreiber, G.: Sirup: Serendipity in recommendations via user perceptions. IUI '17, pp. 35–44. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3025171.3025185

[200] Armstrong, M., Brooks, M., Churnside, A., Evans, M., Melchior, F., Shotton, M.: Object-based broadcasting-curation, responsiveness and user experience (2014)

[201] Cox, J., Jones, R., Northwood, C., Tutcher, J., Robinson, B.: Object-based production: a personalised interactive cooking application. In: Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video, pp. 79–80 (2017)

[202] Ursu, M., Smith, D., Hook, J., Concannon, S., Gray, J.: Authoring interactive fictional stories in object-based media (OBM). In: ACM International Conference on Interactive Media Experiences, pp. 127–137 (2020)

[203] Silzle, A., Weitnauer, M., Warusfel, O., Bleisteiner, W., Herberger, T., Epain, N., Duval, B., Bogaards, N., Baume, C., Herzog, U., *et al.*: Orpheus audio project: piloting an end-to-end object-based audio broadcasting chain. In: IBC Conference, Amsterdam, The Netherlands, September, pp. 14–18 (2017)

[204] Chen, X., Nguyen, T.V., Shen, Z., Kankanhalli, M.: Livesense: Contextual advertising in live streaming videos. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 392–400 (2019)

[205] Akgul, T., Ozcan, S., Iplik, A.: A cloud-based end-to-end server-side dynamic ad insertion platform for live content. In: Proceedings of the 11th ACM Multimedia Systems Conference, pp. 361–364 (2020)

[206] Carvalho, P., Pereira, A., Viana, P.: Automatic tv logo identification for advertisement detection without prior data. Applied Sciences **11**(16), 7494 (2021)

[207] Park, S., Cho, K.: Framework for personalized broadcast notice based on contents metadata. In: Proceedings of the Korea Contents Association Conference, pp. 445–446 (2014). The Korea Contents Association

[208] Hunter, J.: Adding multimedia to the semantic web: Building an MPEG-7 ontology. In: Proceedings of the First International Conference on Semantic Web Working. SWWS'01, pp. 261–283. CEUR-WS.org, Aachen, DEU (2001)

[209] EBU-MIM: EBU-MIM semantic web activity report. Technical report,

EBU-MIM (Aug 2015). Accessed: 30 September 2021

[210] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust Disambiguation of Named Entities in Text. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland, pp. 782–792 (2011)

[211] Brasoveanu, A.M., Weichselbraun, A., Nixon, L.: In media res: A corpus for evaluating named entity linking with creative works. In: Proceedings of the 24th Conference on Computational Natural Language Learning, pp. 355–364 (2020)

[212] Nixon, L., Troncy, R.: Survey of semantic media annotation tools for the web: towards new media applications with linked media. In: European Semantic Web Conference, pp. 100–114 (2014). Springer

[213] Collyda, C., Apostolidis, K., Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V.: A web service for video summarization. In: ACM International Conference on Interactive Media Experiences, pp. 148–153 (2020)

[214] R&D, B.: Object-Based Media. https://www.bbc.co.uk/rd/object-based-media. Accessed: 30 September 2021

[215] Jackson, W.: Object-Based Media Transforms Audio Content Creation. https://www.radioworld.com/news-and-business/objectbased-media-transforms-audio-content-creation. Accessed: 30 September 2021 (2017)

[216] Axonista: Object-based broadcasting. https://medium.com/axonista-hq/object-based-broadcasting-e4dd91b2b2e9. Accessed: 30 September 2021 (2016)

[217] Armstrong, M.: Object-based media: A toolkit for building responsive content. In: Proceedings of the 32nd International BCS Human Computer Interaction Conference 32, pp. 1–2 (2018)

[218] Cox, J., Brooks, M., Forrester, I., Armstrong, M.: Moving object-based media production from one-off examples to scalable workflows. SMPTE Motion Imaging Journal **127**(4), 32–37 (2018)

[219] Carter, J., Ramdhany, R., Lomas, M., Pearce, T., Shephard, J., Sparks, M.: Universal access for object-based media experiences. In: Proceedings of the 11th ACM Multimedia Systems Conference, pp. 382–385 (2020)

[220] Zwicklbauer, M., Lamm, W., Gordon, M., Apostolidis, K., Philipp, B., Mezaris, V.: Video analysis for interactive story creation: The

sandmännchen showcase. In: Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery, pp. 17–24 (2020)

[221] Veloso, B., Malheiro, B., Burguillo, J.C., Foss, J., Gama, J.: Personalised dynamic viewer profiling for streamed data. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) Trends and Advances in Information Systems and Technologies, pp. 501–510. Springer, Cham (2018)

[222] Veloso, B., Malheiro, B., Burguillo, J.C., Foss, J.: Product placement platform for personalised advertising. New European Media (NEM) Summit 2016 (2016)

[223] Malheiro, B., Foss, J., Burguillo, J.: B2B platform for media content personalisation. In: B2B Platform for Media Content Personalisation (2013)

[224] R&D, B.: StoryKit. https://www.bbc.co.uk/rd/projects/object-based-media-toolkitJune2021. Accessed: 30 September 2021

[225] Stewart, S.: Video game industry silently taking over entertainment world. Verfügbar unter ejinsight. com/eji/article/id/2280405/20191022 (2019)

[226] Witkowski, W.: Videogames are a bigger industry than movies and north american sports combined, thanks to the pandemic. MarketWatch. MarketWatch, December **22** (2020)

[227] Ward, L., Paradis, M., Shirley, B., Russon, L., Moore, R., Davies, R.: Casualty accessible and enhanced (A&E) audio: Trialling object-based accessible tv audio. In: Audio Engineering Society Convention 147 (2019). Audio Engineering Society

[228] Montagud, M., Núñez, J.A., Karavellas, T., Jurado, I., Fernández, S.: Convergence between tv and vr: Enabling truly immersive and social experiences. In: Workshop on Virtual Reality, Co-located with ACM TVX 2018 (2018)

[229] Kudumakis, P., Wilmering, T., Sandler, M., Foss, J.: MPEG IPR Ontologies for Media Trading and Personalization. In: International Workshop on Data-Driven Personalization of Television (DataTV2019), ACM International Conference on Interactive Experiences for Television and Online Video (TVX2019) (2019)

[230] MAP: MAP Marketplace. https://map-marketplace.mog-technologies.com/makefilmhistory/auth/login. Accessed: 28 October 2021 (2021)

[231] ISO/IEC: Information technology - multimedia framework (MPEG-21) - part 19: Media value chain ontology / amd 1 extensions on time-segments and multi-track audio'. standard, International Organization for Standardization (June 2018). Accessed: 30 September 2021

[232] ISO/IEC: Information technology - multimedia framework (MPEG-21) - media contract ontology. standard, International Organization for Standardization (May 2017). Accessed: 30 September 2021

[233] Core, D.: Dublin Core Media Initiative. https://dublincore.org/. Accessed: 30 September 2021

[234] dvb.org: DVB-SI, (Service Information), DVB,. https://dvb.org/?standard=specification-for-service-information-si-in-dvb-systems. Accessed: 30 September 2021

[235] etsi.org: TV-Anytime, ETSI. https://www.etsi.org/deliver/etsi_ts/102800_102899/1028220301/01.07.01_60/ts_1028220301v010701p.pdf. Accessed: 30 September 2021 (2001)

[236] Keltsch, M.: BMF – Metadata Exchange Format Of The German Public Broadcasters. https://tech.ebu.ch/publications/bmf--metadata-exchange-format-of-the-german-public-broadcasters. Accessed: 30 September 2021 (2019)

[237] ISO/IEC: MPEG-7, part 1 et seq. standard, International Organization for Standardization. Accessed: 30 September 2021

[238] Chang, S.-F., Sikora, T., Purl, A.: Overview of the MPEG-7 standard. IEEE Transactions on circuits and systems for video technology **11**(6), 688–695 (2001)

[239] ISO/IEC: Introduction to MPEG-7, coding of moving pictures and audio. standard, International Organization for Standardization (March 2001). Accessed: 30 September 2021

[240] ISO/IEC: MPEG-I: Scene description for MPEG media, MPEG group, MPEG-I part 14. standard, International Organization for Standardization. Accessed: 30 September 2021

[241] ISO/IEC: Coded representation of immersive media – part 14: Scene description for mpeg media, ISO,. standard, International Organization for Standardization. Accessed: 30 September 2021

[242] Group, M.: MPEG group, coded representation of immersive media. standard, MPEG standards (September 2020). Accessed: 30 September 2021

[243] Group, M.: MPEG-I: Versatile video coding, MPEG-I part 3, MPEG group,. standard, MPEG standards. Accessed: 30 September 2021

[244] Wieckowski, A., Ma, J., Schwarz, H., Marpe, D., Wiegand, T.: Fast partitioning decision strategies for the upcoming versatile video coding (VVC) standard. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 4130–4134 (2019). IEEE

[245] EBU: EBU Core. https://tech.ebu.ch/MetadataEbuCore. Accessed: 30 September 2021

[246] EBU: EBU Ontologies. https://www.ebu.ch/metadata/ontologies/ebucore/. Accessed: 30 September 2021

[247] Core, D.: Dublin Core Media Initiative. https://dublincore.org/. Accessed: 30 September 2021 (2021)

[248] W3C: Web Ontology Language (OWL). https://www.w3.org/OWL/. Accessed: 30 September 2021

[249] EBU: EBU Tech 3293 – EBUCore. http://tech.ebu.ch/docs/tech/tech3293.pdf. Accessed: 30 September 2021 (2020)

[250] EBU: EBU Tech 3293 - RDF/OWL. http://www.ebu.ch/metadata/ontologies/ebucore/. Accessed: 30 September 2021

[251] EBU: EBU Tech 3332 - Music. http://tech.ebu.ch/docs/tech/tech3332v1_1.pdf. Accessed: 30 September 2021 (209)

[252] EBU: EBU Tech 3336 - Classification Schemes. http://tech.ebu.ch/docs/tech/tech3336.pdf. Accessed: 30 September 2021 (2011)

[253] EBU: EBU Tech 3349 - Acquisition Metadata. http://tech.ebu.ch/docs/tech/tech3349.pdf. Accessed: 30 September 2021 (2012)

[254] EBU: EBU tech 3351 - ccdm. Technical report, EBU (August 2020). Accessed: 30 September 2021

[255] EBU: EBU Tech 3352 - Identifiers in BWF. http://tech.ebu.ch/docs/tech/tech3352.pdf. Accessed: 30 September 2021 (2012)

[256] MPEG-I: MPEG-I: Scene Description for MPEG Media, MPEG Group, MPEG-I Part 14. https://www.mpegstandards.org/standards/MPEG-I/14/. Accessed: 30 September 2021

[257] ISO/IEC: Coded representation of immersive media – part 14: Scene description for mpeg media, iso,. standard, International Organization for Standardization. Accessed: 30 September 2021

[258] Khronos.org: glTF – GL Transmission Format. Khronos. https://www.khronos.org/api/index_2017/gltf. Accessed: 30 September 2021 (2017)

[259] ISO/IEC: Information technology - multimedia framework (MPEG-21) - contract expression language. standard, International Organization for Standardization (December 2016). Accessed: 30 September 2021

[260] Rodríguez-Doncel, V.e.a.: Overview of the mpeg-21 media contract ontology. In: Overview of the MPEG-21 Media Contract Ontology (2016)

[261] mpeg.chiariglione.org: Media Value Chain Ontology. https://mpeg.chiariglione.org/standards/mpeg-21/media-value-chain-ontology. Accessed: 30 September 2021 (2011)

[262] community, M.: Picture, Audio and Data Coding by Artificial Intelligence (MPAI). https://mpai.community/. Accessed: 30 September 2021

[263] org., M.: MPAI - Visual Object and Scene Description. https://mpai.community/standards/mpai-osd/. Accessed: 30 September 2021

[264] Shou, M.Z., Ghadiyaram, D., Wang, W., Feiszli, M.: Generic event boundary detection: A benchmark for event segmentation. CoRR **abs/2101.10511** (2021)

[265] Krishna, M.V., Bodesheim, P., Körner, M., Denzler, J.: Temporal video segmentation by event detection: A novelty detection approach. Pattern recognition and image analysis **24**(2), 243–255 (2014)

[266] Serrano, A., Sitzmann, V., Ruiz-Borau, J., Wetzstein, G., Gutierrez, D., Masia, B.: Movie editing and cognitive event segmentation in virtual reality video. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)

[267] Shou, M.Z., Lei, S.W., Wang, W., Ghadiyaram, D., Feiszli, M.: Generic event boundary detection: A benchmark for event segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8075–8084 (2021)

[268] Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4508–4519 (2021)

[269] Verschae, R., Ruiz-del-Solar, J.: Object detection: current and future directions. Frontiers in Robotics and AI **2**, 29 (2015)

[270] Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B., Li, L., Tang, X.: New generation deep learning for video object detection: A survey. IEEE Transactions on Neural Networks and Learning Systems, 1–21 (2021). https://doi.org/10.1109/TNNLS.2021.3053249

[271] Smith, M.L., Smith, L.N., Hansen, M.F.: The quiet revolution in machine vision - a state-of-the-art survey paper, including historical review, perspectives, and future directions. Computers in Industry **130**, 103472 (2021). https://doi.org/10.1016/j.compind.2021.103472

[272] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)

[273] Kaur, P., Pannu, H.S., Malhi, A.K.: Comparative analysis on cross-modal information retrieval: a review. Computer Science Review **39**, 100336 (2021)

[274] Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–970 (2015)

[275] Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4581–4591 (2019)

[276] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, A.P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. In: arXiv:1609.08675 (2016). https://arxiv.org/pdf/1609.08675v1.pdf

[277] Rehman, S.U., Waqas, M., Tu, S., Koubaa, A., ur Rehman, O., Ahmad, J., Hanif, M., Han, Z.: Deep learning techniques for future intelligent cross-media retrieval. Technical report, CISTER-Research Centre in Realtime and Embedded Computing Systems (2020)

[278] Tu, S., ur Rehman, S., Waqas, M., Rehman, O.u., Yang, Z., Ahmad, B., Halim, Z., Zhao, W.: Optimisation-based training of evolutionary convolution neural network for visual classification applications. IET Computer Vision **14**(5), 259–267 (2020)

[279] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers

for image recognition at scale. CoRR **abs/2010.11929** (2020)

[280] Dai, Z., Liu, H., Le, Q., Tan, M.: CoAtNet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems **34** (2021)

[281] Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.-C., Kamalzadeh, M., Li, B., Leal, S., Parisi, P., et al.: Unity perception: Generate synthetic data for computer vision. arXiv preprint arXiv:2107.04259 (2021)

[282] Tan, C., Xu, X., Shen, F.: A survey of zero shot detection: Methods and applications. Cognitive Robotics **1**, 159–167 (2021)

[283] Wang, W., Zheng, V.W., Yu, H., Miao, C.: A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(2), 1–37 (2019)

[284] Hu, Y., Nie, L., Liu, M., Wang, K., Wang, Y., Hua, X.-S.: Coarse-to-fine semantic alignment for cross-modal moment localization. IEEE Transactions on Image Processing **30**, 5933–5943 (2021)

[285] Hu, Y., Nie, L., Liu, M., Wang, K., Wang, Y., Hua, X.-S.: Coarse-to-fine semantic alignment for cross-modal moment localization. IEEE Transactions on Image Processing **30**, 5933–5943 (2021). https://doi.org/10.1109/TIP.2021.3090521

[286] Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Jointly localizing and describing events for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7492–7500 (2018)

[287] Chen, S., Jiang, Y.-G.: Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8425–8435 (2021)

[288] Dong, C., Chen, X., Chen, A., Hu, F., Wang, Z., Li, X.: Multi-level visual representation with semantic-reinforced learning for video captioning. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4750–4754 (2021)

[289] Francis, D., Anh Nguyen, P., Huet, B., Ngo, C.-W.: Fusion of multimodal embeddings for ad-hoc video search. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2019)

[290] Yaliniz, G., Ikizler-Cinbis, N.: Using independently recurrent networks for reinforcement learning based unsupervised video summarization. Multimedia Tools and Applications **80**(12), 17827–17847 (2021)

[291] Mundnich, K., Fenster, A., Khare, A., Sundaram, S.: Audiovisual highlight detection in videos. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4155–4159 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413394

[292] Hu, L., He, W., Zhang, L., Xu, T., Xiong, H., Chen, E.: Detecting highlighted video clips through emotion-enhanced audio-visual cues. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2021). https://doi.org/10.1109/ICME51207.2021.9428252

[293] Lee, R., Venieris, S.I., Lane, N.D.: Deep neural network–based enhancement for image and video streaming systems: A survey and future directions. ACM Comput. Surv. **54**(8) (2021). https://doi.org/10.1145/3469094

[294] Xiao, Z., Fu, X., Huang, J., Cheng, Z., Xiong, Z.: Space-time distillation for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2113–2122 (2021)

[295] Chu, X., Zhang, B., Ma, H., Xu, R., Li, Q.: Fast, accurate and lightweight super-resolution with neural architecture search. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 59–64 (2021). https://doi.org/10.1109/ICPR48806.2021.9413080

[296] Ignatov, A., Timofte, R., Denna, M., Younes, A.: Real-time quantized image super-resolution on mobile NPUs, mobile AI 2021 challenge: Report. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2525–2534 (2021)

[297] Ignatov, A., Romero, A., Kim, H., Timofte, R.: Real-time video super-resolution on smartphones with deep learning, mobile ai 2021 challenge: Report. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2535–2544 (2021)

[298] Zang, T., Zhu, Y., Liu, H., Zhang, R., Yu, J.: A survey on cross-domain recommendation: Taxonomies, methods, and future directions. CoRR **abs/2108.03357** (2021)

[299] Nixon, L., Ciesielski, K., Philipp, B.: AI for Audience Prediction and Profiling to Power Innovative TV Content Recommendation Services,

pp. 42–48 (2019)

[300] TALUĞ, D.Y.: User expectations on smart TV; an empiric study on user emotions towards smart TV. The Turkish Online Journal of Design Art and Communication **11**(2), 424–442 (2021)

[301] Borgotallo, R., Pero, R.D., Messina, A., Negro, F., Vignaroli, L., Aroyo, L., Aart, C.v., Conconi, A.: Personalized semantic news: Combining semantics and television. In: International Conference on User Centric Media, pp. 137–140 (2009). Springer

[302] AMWA: AMWA Application Specification - AS-02 MXF Versioning. https://static.amwa.tv/as-02-mxf-versioning-spec.pdf. Online; accessed 3 February 2022 (2011)

[303] Telestream, Inc.: A Guide To The Interoperable Master Format (IMF). http://www.telestream.net/pdfs/datasheets/App-brief-Vantage-IMF.pdf. Online; accessed 3 February 2022 (2019)