AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "MEDIA SUMMARIZATION – THE CASE FOR VIDEO"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Evlampios Apostolidis** (Centre for Research and Technology Hellas Hellas – Information Technologies Institute) **Vasileios Mezaris** (Centre for Research and Technology Hellas Hellas – Information Technologies Institute)

This report is part of the deliverable D2.3 - "*AI technologies and applications in media: State of Play, Foresight, and Research Directions*" of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

Copyright

Media summarization – The case for video

Current status

Video summarisation technologies aim to create a short synopsis that conveys the important parts of the full-length video. In terms of presentation format, the produced summary can be either static, composed of a set of representative video frames (a.k.a. ***video storyboard***), or dynamic, created by stitching video's most important and informative fragments in chronological order to form a shorter video (a.k.a***. video skim***). One advantage of video skims over static sets of frames is the ability to include audio and motion elements that offer a more natural story narration and potentially enhance the expressiveness and the amount of information conveyed by the video summary. Furthermore, it is often more entertaining and interesting for the viewer to watch a skim rather than a slide show of frames. On the other hand, storyboards are not restricted by timing or synchronisation issues and, therefore, they offer more flexibility in terms of data organisation for browsing and navigation purposes.

During the last couple of decades, several attempts were made by the relevant research community to automate video summarisation. Currently, the focus is mainly put on methods that try to learn how to perform video summarisation by exploiting the learning capacity of deep network architectures. Most of these methods rely on datasets with ground-truth human-generated summaries (such as, SumMe[1] and TVSum[2]), based on which they try to discover the underlying criterion for video summarisation. However, the amount of currently-available data is relatively small, and the generation of ground-truth data (usually in the form of video summaries or annotations indicating the importance of video frames) is a time-consuming and tedious task.

These limitations resulted in a constantly-increasing interest of the relevant research community, on the development of deep-learning-based approaches that can be trained without the use of extensively-annotated ground-truth data. A recent survey on deep-learning-based video summarisation methods[3], showed that ***unsupervised video summarisation*** methods can be highly-competitive compared to the best-performing supervised approaches. Moreover, the use of less-expensive weak labels - with the understanding that they are imperfect compared to a full set of human annotations - can be another option to build good summarisation models. Figure 1 presents the typical analysis pipeline of deep-learning-based video summarisation methods.

[1] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating Summaries from User Videos," in European Conf. on Computer Vision (ECCV) 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 505–520.
[2] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 5179–5187.
[3] E. Apostolidis, E. Adamantidou, A. Metsai, V. Mezaris, I. Patras, "Video Summarization Using Deep Neural Networks: A Survey", Proceedings of the IEEE, vol. 109, no. 11, pp. 1838-1863, Nov. 2021. DOI:10.1109/JPROC.2021.3117472.

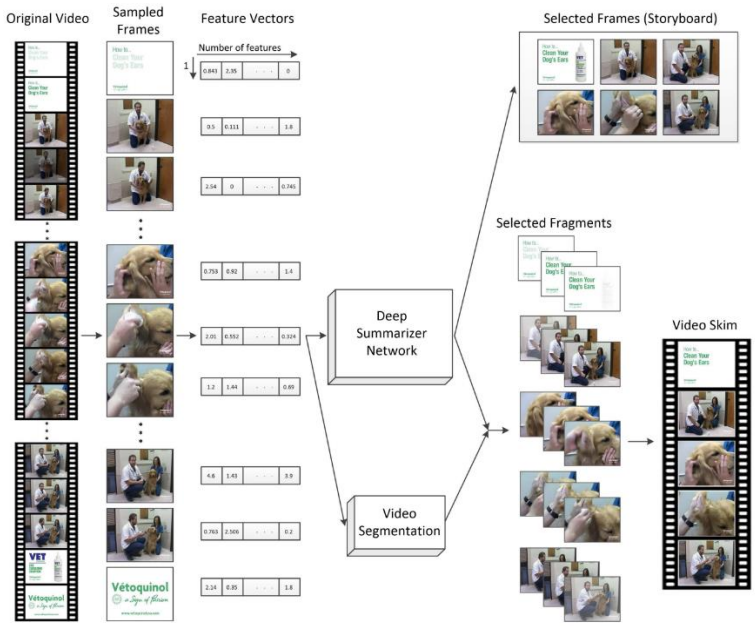

Figure 1: High-level representation of the analysis pipeline of deep-learning-based video summarisation methods for generating a video storyboard and a video skim.[4]

Despite the fact that some video summarisation methods already exhibit good performance according to the established evaluation protocols and datasets, nowadays the practical use of integrated tools and applications for video summarisation is at a very early stage and quite far from being a common procedure of the media management workflow. In most cases, the preparation of a video summary requires the observation of the full-length video and the manual selection of the most suitable pieces of video content by a human expert; and depending on the length of the video this procedure can be a really time-demanding one. Taking into account the number of video material created or obtained on a daily basis by the stakeholders of the media sector, the production of video summaries could require considerable amounts of human and time resources, and possible shortages in these resources could significantly harm the revenue generation potential from the use of these video materials.

This problem is further emphasised by the growing use of different communication channels with varying requirements (such as social networks and video sharing platforms) by media organisations that necessitates the production of ***different summaries for the same piece of video content***. These requirements are usually related to the needs of the targeted audience; for example Twitter users are used to get very short videos, the users of Facebook are familiar with a bit longer videos, while the users of the YouTube platform can spend even more time when watching a video. All the above point out a serious lack of mature video summarisation technologies, that could significantly assist professionals by facilitating and, most importantly, accelerating video summary production.

[4] Figure provided by the authors. Source: DOI:10.1109/JPROC.2021.3117472.

A paradigm of an integrated tool which can assist the production of video summaries that are tailored to the specifications of different communication channels, is the "On-line Video Summarisation Service"[5] (Figure 2). This tool harnesses the power of artificial intelligence[6] to automatically generate video summaries. It takes as input a video and produces different versions of a video summary with adapted length and format for publication on different social media platforms. Based on its functionality, the user can accelerate the production of engaging video summaries for multiple on-line audiences.

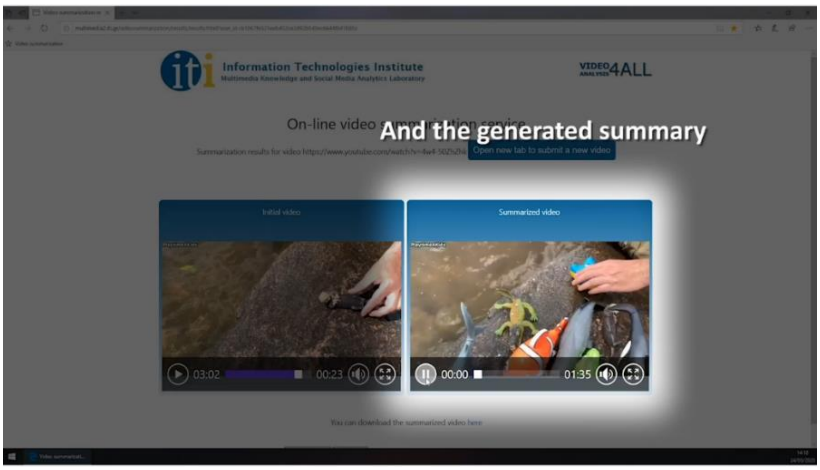

Figure 2: An online video summarisation service [7].

Research challenges

To identify the research challenges in the field of video summarisation, one should consider the current state of the art in terms of summarisation methods, and the current (and possibly future) requirements of the media sector for video content adaptation and re-purposing. With respect to the former, the major research direction is towards the development of supervised algorithms. However, there is an ongoing and increasing interest in the design and development of ***unsupervised video summarisation methods***, mainly propelled by the ***limited amount of training data***. With regards to the latter, the traditional communication channels of media organisations (e.g., TV streams, webpages and online archives) have been framed by the adoption and wide use of modern communication instruments that include social networks and video sharing platforms. Nevertheless, each one of these new instruments is associated with different (mainly audience-driven) requirements about the ***format of the distributed content***. Hence, experts working in the media sector should deal with the preparation of different

[5] C. Collyda, K. Apostolidis, E. Apostolidis, E. Adamantidou, A. Metsai, V. Mezaris, "A Web Service for Video Summarization", Proc. ACM Int. Conf. on Interactive Media Experiences, Barcelona, Spain, June 2020 DOI:10.1145/3391614.3399391. Online demo available at http://multimedia2.iti.gr/videosummarization/service/start.html

[6] E. Apostolidis, E. Adamantidou, A. Metsai, V. Mezaris, I. Patras, "AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization", IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 8, pp. 3278-3292, Aug. 2021. DOI:10.1109/TCSVT.2020.3037883

[7] Screenshot taken from: YouTube, CERTH ITI Video Summarization Service v1.0 (March 2020) https://www.youtube.com/watch?v=LbjPLJzeNII

adapted versions of a given piece of video, in order to reach different audiences and increase the potential for effective media re-use.

Given the above described landscape, we believe that a core research challenge is the development of effective video summarisation methods that can be trained without or using limited supervision. In this way, the research community will be able to tackle issues associated with the restricted amount of annotated data, and to significantly diminish (or even completely eliminate) the need for laborious and time-demanding data annotation tasks. Moreover, such developments can be a game changer considering the need to train different models for each different type of video, as they can provide a solution for building powerful summarisation models that can be easily adapted to the requirements of different application domains (e.g. targeting the production of movie trailers, or the generation of videos with the highlights of a sports event).

With respect to the development of ***fully-unsupervised video summarisation methods***, most of the existing approaches learn summarisation by trying to increase the representativeness of the generated summary with the help of summary-to-video reconstruction mechanisms[8]. So, the main criterion for building the summary is the coverage of the full-length video. The challenge for the relevant research community is to identify additional ***criteria for selecting the key-parts of the video***, and to formulate these criteria in ways that enable their integration into the unsupervised learning process. Such criteria can be associated, for example, with the diversity of the visual content of the summary (to avoid the existence of similar and possibly redundant information), its temporal coherence (to provide a meaningful and appealing presentation of the video story), or its alignment with the core semantics of the video (to focus on parts of the video that are associated with the video's topic or subject).

With regards to the development of ***weakly-supervised video summarisation methods***, the challenge is to discover effective ways that allow editors to intervene in the summary production process, so that the produced video summary is aligned with user-specified rules and requirements. By taking into account user-profile data (e.g., indicating a user's preferences), the research community could provide solutions that facilitate the provision of personalised video summaries. Another, more aspiring scenario would involve the use of an on-line interaction channel between the user/editor and the trainable summariser. So, the challenge here is to build solutions that combine video summarisation and active learning algorithms, in order to incorporate the ***user's/editor's feedback*** with respect to the generated summary[9]. Such developments will be extremely useful for the practical application of summarisation technologies in the media sector, where complete automation that diminishes editorial control over the generated summaries is not always preferred.

[8] E. Apostolidis, E. Adamantidou, A. Metsai, V. Mezaris, I. Patras, "Unsupervised Video Summarization via Attention-Driven Adversarial Learning", Proc. 26th Int. Conf. on Multimedia Modeling (MMM2020), Daejeon, Korea, Springer LNCS vol. 11961, pp. 492-504, Jan. 2020. https://doi.org/10.1007/978-3-030-37731-1_40

[9] A. G. del Molino, X. Boix, J. Lim, and A. Tan, "Active Video Summarization: Customized Summaries via On-line Interaction," in Proc. of the 2017 AAAI Conf. on Artificial Intelligence. AAAI Press, 2017.

Given the plethora of online-available examples of video summarisation, another research challenge could be the design of an effective methodology for ***learning from unpaired data*** (i.e., using raw videos and video summaries with no correspondence between them)[10]. In this way, weak supervision is associated to the collection of the appropriate sets of data based on the targeted application domain (e.g., unpaired groups of movies and movie trailers). Such a data-driven weak-supervision approach would eliminate the need for defining hand-crafted functions that model the domain rules (which in most cases are really hard to obtain), and would allow a deep learning architecture to automatically learn a mapping function between the raw videos and the summaries in the targeted domain.

Last but not least, researchers working in the field of video summarisation should target the development of mechanisms that ***remove bias and provide human-interpretable explanations*** about the decisions made by an AI-based video summarisation method. In this way, the provided solutions for automated video summarisation will offer the needed transparency to the end-users, increasing in this way the level of trust between machines and humans, and improving user experience.

Overall, the modern deep learning architectures have already shown their great potential to learning the main principles of generic video summarisation. From now on, the goal for the research community is to effectively tackle the aforementioned challenges, and push the barriers for making these architectures easily adaptive to the video summarisation needs of several domains and application scenarios. In this way, building on existing technologies with demonstrated content adaptation and re-purposing capabilities[11], it will provide mature solutions that meet the requirements of the media industry, and highly accelerate the video content adaptation and distribution tasks of media professionals.

Societal and media industry drivers

Vignette: Creating teasers, trailers and video summaries for the promotion of TV shows in TV and social media

Jane is a media and social media professional with an expertise on the preparation of highly-engaging video content. Over the last five years, she has been working in a large TV network with an active presence in social media networks (Facebook and Twitter) and video sharing platforms (YouTube). As part of her daily job, Jane deals with the production of different types of summaries for the episodes of five popular TV series. More specifically, a complete video summary is created in order to be added in the beginning of each episode and provide a synopsis of the previous one; a shorter video trailer is produced to advertise each new episode via the TV program, the YouTube channel and the Facebook account; and a very shot teaser video is generated to promote the release of new episodes on Twitter. A couple of years ago, this procedure used to be quite laborious and time-consuming. For every single video, Jane had to

[10] M. Rochan and Y. Wang, "Video Summarization by Learning From Unpaired Data," in 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2019, pp. 7894–7903.
[11] L. Nixon, K. Apostolidis, E. Apostolidis, D. Galanopoulos, V. Mezaris, B. Philipp, R. Bocyte, "Content Wizard: demo of a trans-vector digital video publication tool", Proc. ACM Int. Conf. on Interactive Media Experiences (IMX), June 2021. DOI:10.1145/3452918.3468083.

watch the entire content, spot the most appropriate parts of it, and then produce summaries of different length by deciding which of these parts will be included in each different type of summary. Usually, she was starting by producing the longest and more complete one that was used during the production of the next episode. Then, she was preparing the shorter trailer video for the TV program, the YouTube channel and the Facebook account. Finally, she was working on the creation of the short teaser video that would be posted on Twitter. In total, for a 45-minute episode, Jane spent approximately three hours to prepare all the different summaries.

However, the last two years, Jane's work has been significantly accelerated, as Elena, the director of the Media Management Dept., decided to buy a tool for video summarisation that is based on AI technologies. Using this tool, Jane is now able to analyse an episode and quickly get recommendations about the most informative and story-telling parts of it. Based on the automatically provided explanations about the recommended pieces of information, she can easily decide whether she needs to check other parts of the video as well. If there is such a need, then using the generated storyboard of the episode (i.e., a static visual summary composed of a set of representative frames of the video) Jane is able to quickly inspect other parts of the video and replace some of the recommended ones, by other, manually selected by her. Having this first version of the video summary (i.e., the most complete one) available for the production of the next episode, Jane can then create shorter versions of this summary, by simply adjusting a parameter in the interface of this tool, that is associated with the summary duration. So, after a few clicks and editing checks, she can create the trailer and the teaser of the episode. Moreover, for each type of generated summary, she can immediately create different versions of the video file (with different size and bit-rate) that are compatible with various devices (laptop, table, smartphone) and networks (5G, Wi-Fi, networks of limited bandwidth). Based on the process described above, the production of all these different types of summaries for a 45-minute episode now takes approx. 30 minutes. So, using the AI-based video summarisation tool, Jane needs significantly less time to produce effectively-customised and highly-engaging video summaries; and after spending some of the saved time on the design of the social media campaigns for these five TV series, she achieved much higher audience engagement!

Future trends for the media industry

The development of mature AI-based technologies for video summarisation will be a game changer in the media industry. In the following, we discuss some cases that highlight how these technologies will be used as part of the data analysis workflows of media organisations to accelerate video content adaptation, re-purposing and re-use:

- Process different types of already existing proprietary video data (including both full-length videos and their summarised versions) with powerful AI-based video summarisation tools that can automatically identify the main summarisation patterns for each different type of video content.
- Utilise the trained version of these tools to analyze a new video and quickly get recommendations about the key parts of it that should be taken into account when producing the video summary, and check the automatically provided explanations about the tools' choices with respect to the key parts of the video.

- Constantly improve the performance of the AI-based video summarisation workflow by providing feedback about the provided recommendations about the key parts of the video, based on the ability of the video summarisation tools to actively learn from and adapt to the user's preferences.
- Produce different summarised versions of the same video according to different criteria about the content and duration of the summary, thus accelerating the production of e.g., teasers, trailers and summaries for an episode of a TV series.
- Create different versions of the produced summaries for distribution and consumption via different communication channels (e.g., TV, web-TV, account in social media networks, channels in video sharing platforms) and different devices (e.g., smart-TVs, laptops, tablets, smartphones).
- Enhance the AI-based video summarisation workflow by integrating the available profile data about the viewers/subscribers of the different on-line channels of the media organisation, to offer highly-personalised summaries that match each viewer's interests; e.g., given a 2-hour video of a TV show about traveling to different cities, generate a summary for food lovers, that focuses on the visits at the city restaurants, and another summary for people interested in nature, that shows scenes from rivers and mountains near the city

Goals for next 10 or 20 years

With respect to the goals for the next one or two decades, we foresee that: i) ***repositories of well-trained models*** for the most common video summarisation scenarios (e.g., targeting the production of summaries for news shows, documentaries, sitcoms and the generation of movie trailers and teasers) will be available; ii) AI video summarisation systems will be able to ***discover more complex, domain-specific rules***, and facilitate, for example, the production of short videos with the highlights of sports, music or other events; iii) AI video summarisation systems will be quickly ***adapted to new types of data*** (e.g., medical videos, educational videos, videos captured from surveillance cameras) based on the transfer learning mechanisms of the integrated tools; iv) AI video summarisation systems will be capable of ***summarising more complex types of video content, such as 360-degree video, AR video and XR video***; and v) ***highly compact versions of powerful AI video summarisation methods*** will be available for use in mobile devices, allowing the instant generation of summaries for any recorded video; thus enabling new forms of user-generated content and empowering citizen journalism.

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu