AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

Roadmap on AI technologies & applications for the Media Industry

Section: “AI for entertainment and film production”

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	**Lorenzo Seidenari** (University of Florence) **Federico Becattini** (University of Florence) **Marco Bertini** (University of Florence)

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).

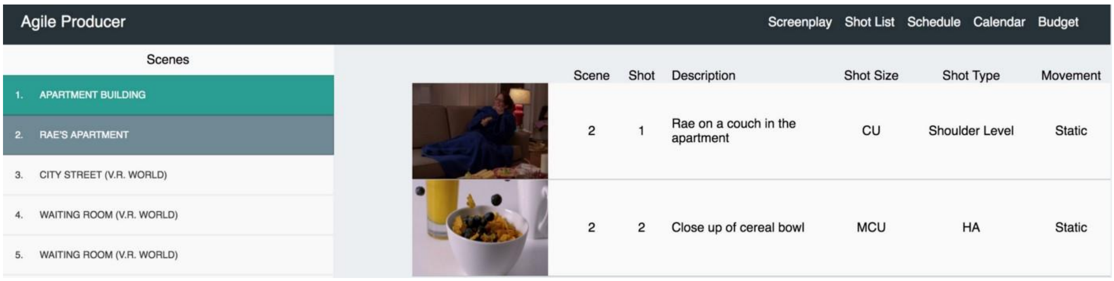
The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

AI for entertainment and film production

Current status

The production of feature movies and other entertainment content like TV series is extremely complex and composed of very diverse steps that require different types of optimisations to reduce their costs and increase the chance of success. During the last years, several techniques based on AI have been proposed for the film and TV industry, addressing the needs of pre-production (e.g. to reduce the cost of animatics storyboarding), production (e.g. via virtual cinematography), post-production (e.g. via the first deepfakes), screening and distribution (e.g. via recommendation algorithms and audience analysis). Below, we report a few notable applications that are emerging and are currently starting to be exploited in cinematography.

Currently, companies offering a complete solution[1] to support film production are also able to provide automatic storyboarding. Given a script and a dataset of already shot footage, these systems are trained to provide suggestions to directors, comprehensive of shot size, type and description (Figure 1).



Agile Producer

Scenes

1. APARTMENT BUILDING
2. RAE'S APARTMENT
3. CITY STREET (V.R. WORLD)
4. WAITING ROOM (V.R. WORLD)
5. WAITING ROOM (V.R. WORLD)

Scene	Shot	Description	Shot Size	Shot Type	Movement
2	1	Rae on a couch in the apartment	CU	Shoulder Level	Static
2	2	Close up of cereal bowl	MCU	HA	Static

Figure 1: Automatic Storyboarding tool based on AI (RivetAI).[1,2]

Moreover, current large scale language models[3] can create synthetic text either from scratch or by completion[4]. This kind of capability can provide a strong support in creation of dialogs and scripts. This kind of large-scale architectures are also able to perform machine translation tasks, given paired sets of sentences[5]. Automatic translation via Deep Learning will empower the movie industry, allowing to target the global market easily.

During the last decade thanks to diffusion and low cost equipment such as drones and 360° cameras, AI methods have been deployed to allow a certain degree of automation in the capture of video content. At the same time, deep learning techniques have matured to the point of being able to generate realistic synthetic content in different media, e.g., video and audio.

[1] RivetAI: https://www.rivetai.com/

[2] Image source: D. Ray, Data Science and AI in Film Production (2018), https://medium.com/rivetai/data-science-and-ai-in-film-production-8918ea654670

[3] OpenAI, GPT-3 Powers the Next Generation of Apps (2021): https://openai.com/blog/gpt-3-apps/

[4] Jasper - The Future Of Writing: https://www.jasper.ai/

[5] S. Edunov, M. Ott, M. Auli, and D. Grangier. 2018. Understanding Back-Translation at Scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 489–500, Brussels, Belgium. Association for Computational Linguistics.

Immersive media is also gaining popularity, particularly after the revelation of the so-called Metaverse. However, bandwidth and latency limits hinder the online experience of immersive media such as Virtual Reality (VR). 360° videos, for example, that are meant to be viewed on a Head-Mounted Display (HMD), require data rates that are around two orders of magnitude higher than standard videos to provide the same quality impression. Distributing improved quality levels in the user's Field of View (FoV) is a simple way to reduce the required data rates. When streaming, this demands anticipating where the user will look, which can be done as far ahead of time as a few seconds if needed.

AI is addressing movie production challenges in all modalities. Current state-of-the art voice generators can provide industry grade voiceovers with hyper realistic synthetic voices[6]. This allows for a great reduction in film budget by not requiring to hire multiple actors to create a multilingual product. Moreover, if this technology is combined with automated translation, the market reach of movie producers is practically global.

Recently, Neural Radiance Fields (NeRF)[7] have revolutionised the field of novel view synthesis from single images. Video generation methods allow to create new video content from some original content, changing camera views, content style and animating differently the persons framed in a scene, as shown in Figure 2.

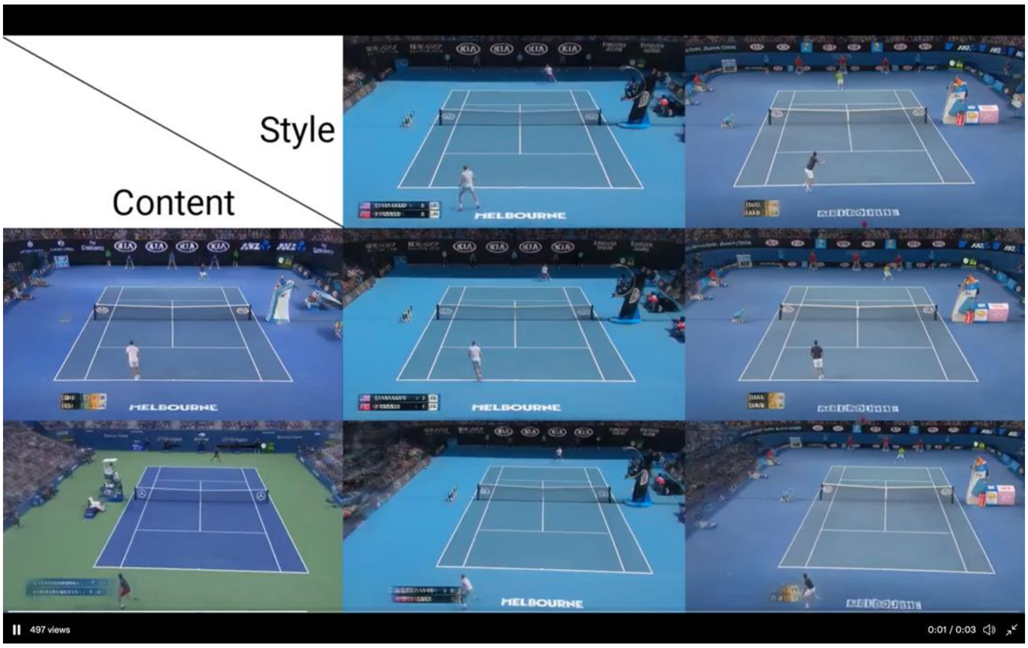


Figure 2: Example of synthetic video generation[8].

[6] Murf AI voice generator: https://murf.ai/

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, Ren Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis”, Communications of the ACM, January 2022, Vol. 65 No. 1, pp. 99-106

[8] Image source: W. Menapace, A. Siarohin, C. Theobalt, V. Golyanik, S. Tulyakov, S. Lathuilière, E. Ricci, “Playable Environments: Video Manipulation in Space and Time”, 2022, https://arxiv.org/abs/2203.01914

Also recently, thanks to AI enabled robotic pipelines automatic shooting of complex videos has become a reality (Figure 3).[9] Multiple techniques are combined allowing one or more drones to execute valid camera movements and film scenes in novel ways. Autonomous aerial cinematography has the potential to enable automatic capture of aesthetically pleasing videos without requiring human intervention.

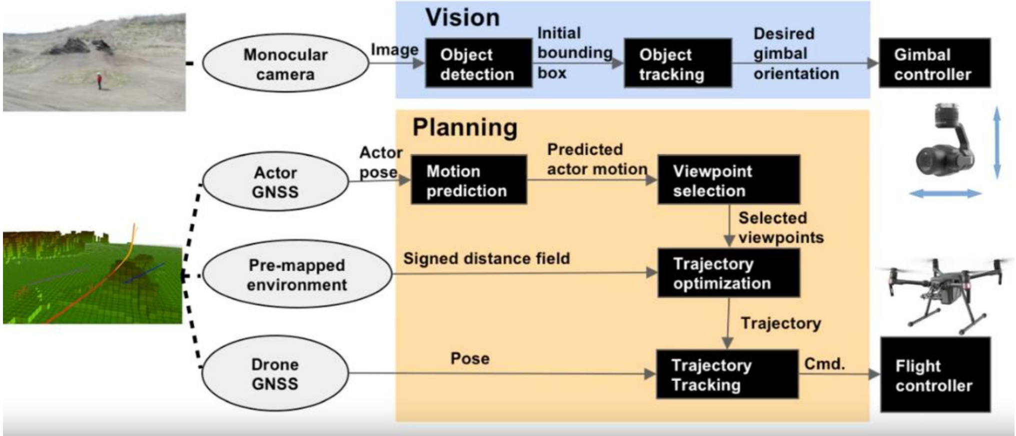


Figure 3: Example of an AI drone video filming system[10].

Deepfake technology seems to be one of the most disruptive technologies in Hollywood and it will probably drastically change the whole worldwide movie industry in the next few years. One of the first possible applications could be in film production based on the use of de-aging technology. In fact, it will be possible both to bring back performers who are no longer alive or even achieve realistic depictions of their younger selves. This opens new and exciting opportunities for film making, advertisements, historical documentaries and so on. As seen in Figure 4, deepfake technology is mature and allows face replacement, thus offering a great opportunity to, for example, reshoot scenes or work with stunt doubles in dangerous or acrobatic film sequences.

Finally, regarding content distribution, acquisition and marketing, more and more companies rely on machine learning to perform personalisation and recommendation of content, helping users find content that satisfies their needs or current mood. Thanks to the large number of customers that streaming industry leaders like Netflix and Amazon have, such recommender systems are always fed with a large set of high-quality annotations. As an example, it is estimated that 80% of the shows watched in Netflix are found through the company’s recommender system[11]. This system is fed with various data, including, interactions with the website (browsing history and ratings), relationships and similarities between the preferences of individual user groups, and information about the content itself (genre, language, actors,

[9] R. Bonatti, Y. Zhang, S. Choudhury, W. Wang, S. Scherer, Autonomous drone cinematographer: Using artistic principles to create smooth, safe, occlusion-free trajectories for aerial filming (2018), https://arxiv.org/abs/1808.09563

[10] Image source: YouTube (R. Bonatti, Autonomous drone cinematographer, ISER 2018) - https://www.youtube.com/watch?v=QX73nBBwd28

[11] A. Krysik, Netflix Algorithm: Everything You Need to Know About the Recommendation System of the Most Popular Streaming Portal (2021): https://recostream.com/blog/recommendation-system-netflix

etc.).[11] To offer a more personalised experience, additional training data include the time of the day when individual users use the service, type of device used to view content, and average viewing length[11].

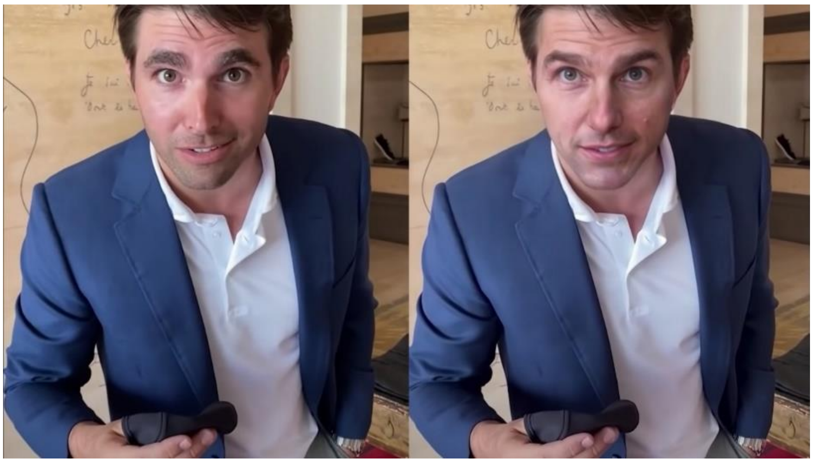


Figure 4: Example of deepfakes posted on social media[12].

Research challenges

One of the main challenges we can identify across most of the aforementioned approaches is linked with **data scarcity**. This is limiting approaches such as machine translation or speech generation to a few languages for which such aligned datasets exist. Unfortunately, this hinders the actual advantage brought by these approaches, which is the capability to post-produce movies in a large number of languages.

Video generation methods, on the other hand, are currently limited in their capabilities of content creation given that they can animate persons only in **relatively static environments**, without handling interactions with objects in the background.

Automated cinematography methodologies are still in their infancy and are currently reliable only when dealing with a **few manoeuvres**. Moreover, such methods are only applicable to camera motions derived from UAVs flying at an altitude. A lot of shooting happens in a much diversified manner, using for example dolly cameras, or handheld cameras. More complex systems and agents should be developed in order to cope with these situations.

Compressing content either in VR setups or standard streams in a way that the viewer finds pleasant without wasting bandwidth is still an unsolved challenge. Another issue that is not yet solved is the capability of codecs to be personalised for a single viewer. While this is somehow the main idea behind the delivery of 360° videos, it is extremely challenging for standard video streams. All in all, we are not yet able to characterise correctly viewer behaviour individually.

High resolution content generation from textual descriptions still produces images with a quality and resolution that do not allow to use them in production. Generation of backgrounds and settings for virtual set filming is currently still mostly performed using traditional CGI methods.

[12] Image source: TikTok (@deeptomcruise): https://www.tiktok.com/@deeptomcruise?lang=en

Developing generic image generators that have the same quality of those developed for limited domains like faces is still a challenge for the scientific community.

Regarding personalisation of content, it is still to be debated if personal preference can be derived directly from multimodal cues. We are still not able to build a model of personal likings directly from content.

Societal and media industry drivers

Vignette: Facilitating film shooting using drones and enlarging reach of content via automatic translation and dubbing

Jane is a video producer that works with a low budget. She employs innovative methodologies for filming such as drones and 360° cameras so that she can work with a limited filming crew. At the same time, she would like to extend her business and distribute content on more countries.

Recently, Jane has started to invest on AI based products that allow to speed up her filming process. Instead of requiring multiple drone pilots, Jane acquired a software to automatically shot video sequences from drones with minimal supervision. The software allows to instruct drones to follow specific subjects or to perform smart manoeuvres in order to shoot cinematographically sensible sequences.

To improve the reach of her content, she also acquired an AI based automatic dubbing and translation tool. Thanks to speech synthesis, subjects can be dubbed at low cost. Voice pitch is differentiated thanks to computer vision that automatically associates the same synthetic voice actor to the same person. Jane can still control and decide the voice pitch in order to obtain a product that suits her needs. Together with automatic voice generation, the AI tool also includes a speech recognition system and an automatic translator, allowing to extract speech from original sequences, create subtitles automatically as well as translations in multiple languages.

During the editing phase of her latest video, Jane notices that unluckily one of the scenes has not been filmed from the best point of view. Filming it again, however, is too expensive for her. Using a new tool for scene generation, Jane picks some frames from the filmed scene and recreates an animated virtual view that she can use as if it was filmed during the production.

Future trends for the media sector

In the following, we briefly summarise some the ways in which AI is expected to enhance and facilitate film production, from the pre-production stage to content distribution:

- Comprehensive **tools to organise and direct drone swarms**, aiming to improve the shooting process by filming aesthetically pleasing landscapes, action-packed film scenes or military action in war zones from novel points of view and in settings that would be difficult or dangerous for human camera operators to be.
- Comprehensive **tools to perform voice recognition, synchronisation, translation and dubbing**, aiming to address the needs of different international markets and provide localised contents.

- **Automatic content creation** (e.g. trailers for different audiences or personalised trailers, movie/TV scenarios, CGI, music to match scenes/emotion/dialogue, etc.). These techniques will significantly facilitate production by reducing costs of virtual sets, post-production through virtual cinematography and automated VFX, and marketing and promotion through targeting of demographics of clients using automated and personalised trailers.
- **Deep fakes for movies**. The current initial use of this technology aims to replace actors, e.g. to depict a younger appearance of an aged actor in a film sequel. Extension to full scenes and settings can be expected in the next few years.
- AI for **audience analysis** to provide insights on what scenarios have the potential to be popular movies, which actors to cast, etc. Automated and multimodal AI test screening allows to find the best market and demographics for each title and also to make decisions about casting or the selection of scripts, by analysing the emotions and reactions of audiences without requiring to organise a costly in-person event.
- **Automation of film directing/editing/shooting processes**. Through automatic selection of the best filmed scenes, virtual cinematography allows to create new scenes from the filmed ones while automated camera movement and tracking improve the shooting process.
- **Film aesthetics and style transfer**. Creating different types of visualisations from a common source, e.g. transforming movies into cartoons, automatic colour grading aimed at inducing different emotions, etc., may open new avenues in film production.
- **Interactive and personalised movies**. Using AI to extend certain parts and subplots of a movie with materials that are more interesting for certain audiences, or creating personalised experiences based on the reactions of the audiences (as measured e.g. by wearable sensors or cameras) can lead to innovative ways of experiencing cinema and TV in the near future.

Goals for next 10 or 20 years

AI applications are expected to become pervasive in cinematography and media content production in the next few years. A goal is to exploit the diffusion of Unmanned Aerial Vehicles (UAVs, or drones) to automatise shooting, aiming not only to facilitate the filming process but also to optimise viewer experience. In fact, we expect drone manoeuvres to be programmable based on the movie style or the emotion that the filmmaker wants to communicate. We also expect the AI piloting the drone to automatically infer such emotion from the observed scene and adapt its trajectory to convey an adequate shooting style, e.g., fast close-ups for dynamic action-packed scenes or slow movements from a distance for sentimental scenes.

Another goal is to make UAVs smarter. Onboard AI could entail a high-level understanding of the shot that could be used to enhance the quality of the footage by optimising framing position or dynamically tracking different targets depending on their estimated importance.

Another interesting application would be to shoot different versions of the same content with multiple drones that adhere to a set of shooting styles. Such diversity could be used in a post-processing step by a human or AI operator to build a rich footage or could be directly provided to users, offering different experiences depending on their personal tastes and preferences. Furthermore, drones could also be equipped with the capability of predicting trajectories of relevant subjects or objects. By being able to forecast motion patterns in the observed scene, we expect drones to anticipate movements and prepare their shooting position in advance to obtain more spectacular and memorable scenes.

Similarly to trajectory prediction, forecasting human gaze could also play an interesting role both in content generation and fruition. In fact, drones are often piloted by human operators using 360° headsets. Predicting their gaze could allow the drone to anticipate the actual command given by the operator and execute smoother manoeuvres. We expect the same kind of technology to be useful for watching 360° video content, especially when streamed online. Predicting user gaze can allow to adaptively compress the video stream, considerably reducing bandwidth. 360° content will become increasingly pervasive, also affecting how content will be produced. Videos customised for headsets in fact will open a whole new range of opportunities for content creators. For instance, videos could be shot to contain multiple narrations depending on which portion of the video is observed by the user throughout the viewing experience.

Now, it is not possible to imagine all the infinite opportunities where deepfake technology could be used besides obtaining amazing and unfeasible results such as shooting a new movie with a not-living actor. Deepfakes could be adopted to strongly reduce production costs. In fact, such a technology would also be able to replace and/or adapt the original performance for a different goal without the need of getting an entire set and crew assembled to change a part of a dialogue or to repeat a movie scene again. For instance, some dialogues and scenes could be recorded with a reduced part of the cast and with some proper stunt doubles (see Figure 4 above) and be successively completed by resorting to deepfake technologies to modify faces, facial expressions, lips and eye movements accordingly. It is easy to understand how much deepfake technologies which basically deal with faces could become crucial in post-processing operations in order to change and modify what has been recorded without reshooting. For example, it could be used by filmmakers to manipulate and alter specific dialogues without needing to do reshoots.

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