AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AUTOMATIC MULTIMEDIA CONTENT CREATION"

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) **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.

Automatic multimedia content creation

Current status

Content creation is the pillar of the multimedia industry. Media companies, broadcasters, artists, and media professionals in general all make a living out of producing a large stream of multimedia data. While most of the professional content is still manually edited and originated, several computational methodologies have risen in support of content creators (e.g. see section on "*Evolutionary learning*"). Apart from AI based methods to support editing, which have been covered in section "*Media summarization – The case for video*" of the Roadmap, a lot of effort has been made by researchers to enable machines to produce novel unseen content by directly learning from data (Figure 1).

Figure 1: Examples of faces generated using StyleGAN algorithm[1].

Generative art lays its foundation on methods that optimise large deep network models to create outputs that are reasonable with respect to a set of training data. Amazing results have been shown by methods like DALLE[2] and VQGAN+CLIP[3]. Both approaches leverage large image and language datasets. While the former usually yields precise depiction of the provided sentence, the latter has mostly been used to obtain oneiric imagery. DALLE's main strength is also its main drawback: a large 12 billion parameter model that requires 250 million image-text pairs mined from the internet. Considering current video standards (4K), one major showstopper for these methods is the high computational demand and their limited output resolution. Direct generation of video is also sought[4], which of course is affected even more by the above challenge.

A large amount of content is continuously produced and streamed online. On top of this, a bulk portion of existing digital content has largely been acquired and produced over the last decades with a plethora of not always well-performing formats. Modern video consumers are used to high quality, high resolution devices. Current broadcasters, especially on pay-per-view, aim at delivering 4K content. Media editors and producers often struggle to reuse content and keep

[1] Images generated randomly by the Random Face Generator (This Person Does Not Exist) at https://this-person-does-not-exist.com/en

[2] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." International Conference on Machine Learning. PMLR, 2021

[3] VQGAN+CLIP Notebook: https://colab.research.google.com/drive/1ZAus_gn2RhTZWzOWUpPERNC0Q8OhZRTZ

[4] VIDEO-GPT: https://wilson1yan.github.io/videogpt/index.html

the overall product quality high. Several phenomena intervene in degrading clips. A video may be acquired digitally at a low resolution with some older formats (e.g. MPEG2), some content may not be available in colour or at the desired final resolution.

Content enhancement is the task of improving quality of media from a possibly low quality or corrupted source. Recent TVs implement AI based enhancement algorithms directly in hardware. On the research end, we witness a rush towards deep learning based methods capable of blindly reconstructing missing information either by correcting defects or artefacts or by filling in lacking pixels when upsampling. First attempts at improving images and video came from classical convolutional neural networks trained to translate a low quality input image into a higher quality one[5]. More recent work leveraged techniques based on generative adversarial networks trained in a conditional fashion[6] (Figure 2). Since many distortions may be present in images, even at the same time, some effort has also been made to make a universal model[7], capable of restoring images blindly from unknown distortions.

(a) (b) (c) (d)

Figure 2: Effect of detail enhancement on a compressed picture (a). In (b) a clear degradation is shown. Standard non AI based methods are not able to improve the result (c). Finally, a GAN based enhancer shows a pleasant detail (d)[6].

Sometimes mixing black&white video with colour sources may be a style choice also guided by the will of the narrator to make the viewer focus on the change of a historic period. In some scenarios, again guided by a different artistic need, colour shall be recovered or in general improved so that the final product is consistent. A preliminary solution to this challenging problem is a technique called colourisation. In colourisation, deep networks are trained to translate images from black&white to full colour. In this task, networks shall preserve shape but add colour information pixelwise. Issues usually regard the temporal consistency of such solutions and the need of preserving semantic information that could be lost when colour is transferred wrongly.

[5] Svoboda, Pavel, et al. "Compression artifacts removal using convolutional neural networks." arXiv preprint arXiv:1605.00366 (2016).
[6] Galteri, Leonardo, et al. "Deep universal generative adversarial compression artifact removal." IEEE Transactions on Multimedia 21.8 (2019): 2131-2145.
[7] Wang, Xintao, et al. "Real-esrgan: Training real-world blind super-resolution with pure synthetic data." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021

Research challenges

The main challenge for current enhancement methods is **dealing with multiple sources of disturbance** in the input media to be improved. Since the process of degradation is non additive, meaning that a processing pipeline may involve a sequence of compression, rescaling and other kind of noises that are hard to remove individually or even detect. As an example, a video digitised from a partially compromised physical media with low quality digital coding may be hard to restore with current methods.

In respect to approaches for novel content generation, the current main limitation lies in the capability to **scale in terms of temporal and spatial resolution**. Current face generators cannot go above 1MP resolution and video generators are far behind. Methods able to create high-quality high-resolution complex scenes are not available yet.

Another challenge is that of **bias in training data**. As every data driven approach, especially when dealing with large deep learning models the issue with dataset bias can lead to extreme failures. As a very well-known example has recently shown, the lack of diversity in face datasets may significantly alter the output of enhanced faces to the end that most of the visual attributes are corrupted (see Figure 3). The ethical ramifications of such failure modes are severe and must be addressed.

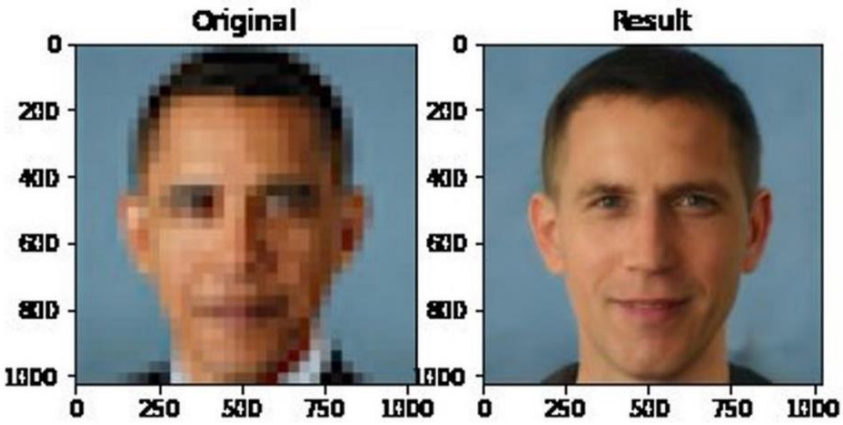

Figure 3: Example of the PULSE[8] algorithm applied to a pixelated face of Barack Obama. The algorithm reconstructs the face as that of a white man[9].

Societal and media industry drivers

Vignette: Automating video production via reuse of content and original content creation

Francesca is a video editor that works in a major broadcasting company. The company broadcasts content of all sorts from documentaries to live sport events. Francesca starts her productions designing a digital storyboard to show how she wants to visually portray the programmes before they are produced. The production and selection of the visual materials of

[8] Menon, Damian, et al. "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models", Proc. of CVPR 2020.

[9] Image source: Twitter (@Chicken3gg) - https://twitter.com/Chicken3gg/status/1274314622447820801

the storyboard sometimes is slow, since she can't find archive material that is similar enough to the description of the planned scripts and thus, she must draw or produce this storyboard content with the help of a 3D artist, resulting in a slower production or, sometimes, even in misunderstandings with the production team.

Moreover, Francesca is often tasked with the editing of clips from a diverse set of sources, especially when dealing with news reports and documentaries. To keep the final product quality high, since the editing process often has to deal with intertwining different qualities and resolutions of media, Francesca often resorts to hand-tuning the appearance of clips using video editing software. As a video editing technician, Francesca is often consulted regarding the technical process involved during live streaming events. In this situation, a bulk of streams from a very diverse set of cameras with different resolutions, framerate, format and colour calibration parameters are pooled together to be mixed in real time. Event producers are worried that the continuous shift in appearance hurts the end user experience and enrol Francesca in the process of adjusting, in a best effort manner, the appearance of all streams in a way that the final product has the highest possible quality. Unfortunately, due to many factors such as weather conditions that may affect lighting, equipment variation and also bandwidth limitation this effort is continuous and sometimes the sought result is far from the actual final product.

Recently Francesca's company established a tight collaboration with a couple of start-ups: one is a company that provides a set of tools to create stock footage materials and the other provides software for image and video enhancement. Together they started a proof-of-concept that led to a set of tools and systems that allow to automatise all processes related to video editing. Thanks to this collaboration, the companies grew and made the PoC results into actual industrial products that revolutionised the editing business. Now Francesca can rely on such products to improve the final quality of archival content prior to video production and has sped up the production of new programs, reducing the planning and design phase. Using the tools for stock footage generation, she can produce thumbnails that match the descriptions in the scripts; the same tools can produce new animated backgrounds and 3D graphics footage, thus reducing production costs. Using the tools for video enhancement, live streams improved in quality and the broadcaster widened its 4K offer increasing their subscriptions. Thanks to AI based automatic content enhancement, Francesca can spend more time following the artistic and technical process of video editing while content is processed offline. The lack of need of manual tuning makes it possible to focus on content selection. Regarding live-streamed events, producers are happier since they have less uncertainty on video quality when switching from one feed to another.

Future trends for the media sector

The automatic creation of content has a major drive in the continuous request for high quality media products. We highlight some trends that we believe will lead the development of new AI based media product:

- Developing **robust and efficient video enhancement services** allowing to edit content in real time and increase the throughput of video professionals.

- Deployment of **content enhancement AI on TV hardware** allowing end users to benefit from the maximum quality possible, independently from their connectivity and the broadcasted signal quality.
- **Repurposing and revamping of archive materials** allows to amortise the production costs by increasing the reuse; this requires adapting the low level visual features of recorded material (e.g. colour correction, colourisation, super resolution) for a new production. Automatising the process reduces costs and facilitates reuse.
- Instead of restoration, archive materials could be used as "seeds" for the **generation of new content adapting visual styles**, e.g., from movies to cartoons or 3D graphics and vice-versa.
- Integration of **automatic original content creation** into existing content production tools. Currently some softwares are already exploiting AI to provide smart layouts for presentations or to fix image appearance without user intervention. Current image generator AIs could replace the role of an **image search engine**. Instead of looking up for the best stock photo or the best clipart to integrate into a digital product, such tool could just generate a set of original images based on the query of a user.
- **Creation of deepfakes**, either as persons (full bodies or talking heads) or as settings and scenes, could lower production costs allowing also smaller companies to produce a larger variety of products, e.g., those that require expensive film sets.
- The videogame industry could benefit from **automated creation of graphical content and assets**, like backgrounds, textures, cut scenes, **or audio content**.

Goals for next 10 or 20 years

Rapid progress is expected in this field in the next 10 years. Below we summarise some goals.

One goal is the ability to **disentangle internal representations** in order to not just create realistic images and video but also to fully control the appearance via semantic and geometric attributes. This means that generated art or content should not just be pleasant and realistic but a user should be able to edit specific properties such as the arrangement of objects, their colour or texture without corrupting the overall appearance of the generated example.

Regarding content enhancement tools, we expect models to be able to **deal with multiple sources of degradation at the same** time blindly. Like current image classifiers being resilient to changes in image resolution, object scales and appearances, the same should be sought for content enhancement solutions.

Currently, the correspondence between low/high quality media is 1-to-1, while in the future, we expect to have methods able to **capture the signal from a few examples** and exploit this source to improve a larger amount of data. Currently, this is possible for video call like videos such where once every few seconds a high-quality face is transmitted and the receiver can enhance the current low quality stream incorporating information from high quality frames. This approach will likely be generalised to more complex types of videos, allowing for example to improve the quality and resolution of archival videos with few key high quality digital shots.

Finally, while current learning paradigms expect to work with paired media from different qualities, in the future **unpaired learning** will be likely possible allowing to leverage a larger amount of training data. In this scenario, enhancement tools will not just learn to invert some simulated degradation but given two large media sets with a source quality and a desired target quality will learn to restore content even without a direct pairing.

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