Shaping the Video Conferences of Tomorrow With AI

Paulo Renato C. Mendes
paulo.mendes@telemidia.puc-rio.br
PUC-Rio, Rio de Janeiro, Brazil

Eduardo S. Vieira
eduardo@telemidia.puc-rio.br
PUC-Rio, Rio de Janeiro, Brazil

Pedro Vinicius A. de Freitas
pedropva@telemidia.puc-rio.br
PUC-Rio, Rio de Janeiro, Brazil

Antonio José G. Busson
busson@telemidia.puc-rio.br
PUC-Rio, Rio de Janeiro, Brazil

Álan Lívio V. Guedes
alan@telemidia.puc-rio.br
PUC-Rio, Rio de Janeiro, Brazil

Carlos de Salles Soares Neto
salles@ufma.br
UFMA, São Luís, Brazil

Sérgio Colcher
colcher@inf.puc-rio.br
PUC-Rio, Rio de Janeiro, Brazil

ABSTRACT

Before the COVID-19 pandemic, video was already one of the main media used on the internet. During the pandemic, video conferencing services became even more important, coming to be one of the main instruments to enable most social and professional human activities. Given the social distancing policies, people are spending more time using these online services for working, learning, and also for leisure activities. Videoconferencing software became the standard communication for home-office and remote learning. Nevertheless, there are still a lot of issues to be addressed on these platforms, and many different aspects to be reexamined or investigated, such as ethical and user-experience issues, just to name a few. We argue that many of the current state-of-the-art techniques of Artificial Intelligence (AI) may help on enhancing video collaboration services, particularly the methods based on *Deep Learning* such as face and sentiment analyses, and video classification. In this paper, we present a future vision about how AI techniques may contribute to this upcoming videoconferencing-age.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Web conferencing**.

KEYWORDS

Video conference, Artificial Intelligence, Deep Learning, Covid-19

1 INTRODUCTION

Even before the COVID-19 pandemic, video already was one of the main media used on the internet. Youtube, for example, has accounted that more than 500 hours of video were uploaded to their platform every minute during 2019.[1] During the pandemic, video services became even more important, coming to be one of the main instruments to enable most social and professional human activities. According to J.P. Morgan Consulting,[2] given the social distancing at home, consumers are spending more time online to virtually work, connect with others, and stream entertainment. In work and education, video conferencing software tools became the communication standard for home-office and remote learning. For entertainment purposes, the social distancing has led to a higher engagement, better audience and subscriber growth for streaming services.[3,4] In particular, user-generated content was broadly distributed in live streaming (*e.g.* music, life-style and gaming) and enabled social interaction.[5]

This new video-conferencing-based age, however, still needs improvement. After some months in home-office or remote classes during the pandemic, there are a lot of facets regarding communication, ethical and user-experience issues that can be already pointed out. In education, there is an engagement problem in keeping children's attention.[6] Moreover, there are cases of unappropriated content (nudity or intercourse) being accidentally (or maliciously) inserted during video conferences.[7]

Artificial intelligence (AI) methods, specially those based on *Machine Learning* (ML) became the state-of-the-art in various segments related to automatic media analysis. We argue that these techniques can strongly contribute to the aforementioned videoconferencing context. In this work, we present a future vision on how AI may contribute to the future of video conferences services and collaboration.

To present our vision, the reminder of this article is organized as follows. Section 2 describes state-of-the-art in the tasks of image and video classification, face analysis, video summarization, and natural language processing. Then, section 3 presents our future vision of AI in video conferences. Finally, section 4 contains our final considerations.

[1] https://kinsta.com/blog/youtube-stats/

[2] https://www.jpmorgan.com/global/research/media-consumption

[3] https://itbrief.com.au/story/video-streaming-services-to-see-a-boom-as-people-stay-at-home

[4] https://www.economist.com/graphic-detail/2020/03/27/covid-19-is-a-short-term-boon-to-streaming-services

[5] https://www.theverge.com/2020/5/13/21257227/coronavirus-streamelements-arsenalgg-twitch-youtube-livestream-numbers

[6] https://indianexpress.com/article/trending/trending-globally/kindergartner-falls-asleep-online-class-6549986/

[7] https://www.dailymail.co.uk/news/article-8636161/Get-Zoom-Couple-caught-having-SEX-Rio-Janeiro-council-video-conference.html

In: VII Workshop "O Futuro da Videocolaboração" (WCT-Video 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.

ISSN 2596-1683

2 STATE OF THE ART

To support creating our future vision for video video conferences, we select some AI topics applied to video, and briefly summarize their main state-of-the-art techniques in the following subsections.

2.1 Video Classification

Unlike images, videos have not only visual, but also audible information. Current methods for classifying video are generally divided into two stages: (1) Convolutional Neural Networks (CNNs) stages, called *backbones*, used to extract the audio-visual *features* from the video content; (2) After extraction stages, that apply sophisticated methods for aggregating *features* such as NetVLAD [1] and NetFV [18] to undermine audio-visual *features* and perform classification. These methods achieve state-of-the-art performance on the YouTube-8M video classification task [26].

To extract the visual *features* from a video, CNNs (*e.g.* Inception [24], ResNet [9]) pre-trained in the ImageNet dataset can be used. As for extracting *features* from audio, CNNs adapted for the audio domain, such as AudioVGG [10] or WaveNet [19] pre-trained in AudioSet dataset.

Currently, the video analysis community is devoted to classifying video at the segment level (*i.e.*, temporally locate and classify the segments in the videos). Deep recurrent models such as LSTMs [11] and GRUs are commonly used for video segment classification as they are well suited to extract temporal features across time. ActivityNet [3] is a large-scale video benchmark for human activity classification, introduced to boost segment proposal and temporal localization algorithms; BSN [14] is the state-of-art model for temporal action proposal generation; the Video Action Transformer Network [8] is a proposal to localize and classify actions in space and time.

2.2 Face Analysis

Face-Analysis-based systems can be used in a wide variety of fields [17, 22], from security to entertainment. Two tasks regarding face analysis have been gaining the attention of researchers for many years: *Face Detection* [22] and *Face Recognition* [16].

Face Detection is the process of detecting faces in a given scene and distinguishing them from other objects. Many methods have been proposed to solve this problem [22]. Viola and Jones [27], for instance, proposed a framework for real-time face detection in early 2003 using an image representation called "Integral Image" and a classifier using AdaBoost [7]. Their work was capable of performing face detection at a rate of 15 frames per second. More recently, with the advance of deep-learning-based methods, works have been conducted aiming at performing *Face Detection* in more difficult scenarios. As an example, we cite Zhang *et al.* [28], whom has achieved a high *Average Precision* in the *Wider Face* dataset, which is a benchmark composed of 32,203 images and 393,703 faces with a high degree of variability in scale, pose, occlusion and illumination.

Face Recognition is the task of making a positive identification of a face in a photo or video image against a pre-existing database of faces. It is used for video surveillance and security systems, video analytics systems, smart shopping, automatic face tagging in photo collections, investigative tools that search for identities in social networks based on face images, and in thousands of other applications in our daily lives. Traditional deep learning models for face recognition such as DeepFace [25] and DeepID [23] use a CNN with fully-connected layer output to produce a representation of high-level features (face embeddings) from an input image, followed by a softmax layer to indicate the identity of classes. Other approaches, such as FaceNet [21], can directly measure the similarity among faces using euclidean space. Inspired by DeepID, this model uses the *triplet loss* as the loss function to estimate similarity to one character's face to a collection of other faces.

2.3 Video Summarization

Several authors have addressed this topic with supervised and unsupervised solutions [6, 13]. Initially, the researchers used unsupervised methods, developing low-level space-time resources and reducing dimensionality with clustering techniques [13]. For this, the models only need to define distance/cost functions between the frames in relation to the original video. However, this method tends to choose frames that are too homogeneous due to the type of resources used, such as semantic and pixel intensities [6]. On the other hand, current methods implement bi-directional recurring networks, such as BiLSTM, however, these networks are complex to implement and demand high computational cost [6].

In this sense, since the launch of the attention mechanism, it has become an important topic in research on neural networks and have proved useful in a wide variety of tasks [15]. It is simpler to implement and unlike recurring networks, it allows parallelization, which reduces the computational cost. Using only attention in the task of video summarization, Fajt's *et al.* [6] work became the state of the art and obtained benchmarks in *TVSum* and *Summe* datasets.

2.4 Natural Language Processing

It is a research area that focuses on processing, labeling, and generating speech or textual content. Some Natural Language Processing (NLP) models, such as BERT [4] and GPT-3 [2] can achieve impressive feats in question answering and perform arithmetic operations just from being trained on text. From speech-to-text to smart assistants, NLP has a wide range of applications. Tasks in NLP, such as *named entity recognition* and *sentiment analysis*, can become a vital part of interactions between the video conference system and the participants.

3 FUTURE VISION

To present our future vision, we discuss a set of envisaged use cases (UC) as follows.

UC-01 Sensitive/Inappropriate Video Classification. There are many examples of uncomfortable situations during video conferences. For instance, when a participant forgets to disable his/her microphone or camera and goes to the bathroom or change clothes. Sensitive/inappropriate video detection could improve trust and comfort during the video conference, since the participants could be confident that no inappropriate content will be shared.

UC-02 Audience Counting. More than one person may be using the same camera. Counting the number of attendants present in a video conference might be useful to many different purposes.

Real-time face detection systems could be used to address this requirement.

UC-03 Identity Verification. In the context of meetings where only authorized people can participate, it is important to guarantee that only those with the right permissions are allowed to enter and stay. Face recognition can be used to verify the identity of participants and compare them with a previously authorized list. Another example is a class where only the students that have signed to a term of confidentiality are authorized to participate. In this case, a dataset of authorized faces can be created. Then, each face detected in the video of the participants is compared against the dataset to verify whether it matches one of the authorized participants.

UC-04 Automatic Layout. The most common visual layout used in video conferencing systems is to show participants within a multiple video grid, which is far from being a kind of immersive visualization. There are some current initiatives to improve the users' immersion experience, such as the one proposed in the Microsft Teams' virtual auditorium, for example.[8] It removes users' background and displays them together in auditorium chairs, giving the sensation of a classroom. However, considering that users' streams are produced within different angles and distances form the camera, their bodies and other surrounding objects or even working tables are usually not aligned. All this "misalignment" is often responsible for a poor immersion experience. Automatic layout techniques may improve this perception, by using object detection together with automatic zoom, or even replacement of objects, in a way that might give a better sense of continuity among different physical spaces.

UC-05 Automatic Accessibility. The World Health Organization (WHO) indicates that there are more than 2bi visually impaired people[9] and 466mi with some kind of hearing loss[10] in the world. During a video conference, instructors often present videos and images, which cannot be seen by visually impaired people. In this context, audio description becomes an important resource. It consists of a narrator who speaks during the presentation, describing what is happening on the screen. For the ones with hearing loss, subtitles and sign languages are critical resources. During a video conference, speeches must be subtitled and translated into sign language, which is performed by an interpreter who expresses himself/herself through a set of gestures, facial and body expressions along the presentation. Thus, through the text extracted from the audio of the video, we can generate, in addition to subtitles, the necessary input for the recognition and translation of the spoken text into an appropriate sign language.

UC-06 Participant Anonymization. In many videoconferencing scenarios, like education, for example, lives sessions are recorded and shared to a broader audience after finished, making it available for people who were not able to participate in real time, for example. In these sessions, many of the participants that opened their cameras are not comfortable with their image being shared in the final video file. Therefore, it is important to have a mechanism that makes those who are not willing to appear in the final video file anonymous. Similar to the *Identity Verification* use case, face recognition, and detection mechanisms could be used to detect and recognize those participants. Once recognized, an image filter could be applied to the students' faces to make them anonymous. Moreover, this kind of anonymization process can also be done live if each participant had previously informed this desire.

UC-07 Video Summarization. In some situations, when a participant needs to quickly find on a set of recorded video conferences which are the most adequate for his/her interests, video summarization can be useful. It allows creating a preview by extracting the main frames of the video and thus creating a visual history with the most important moments of the video, facilitating the search and reducing the time spent on it. However, summarizing a video can be more than just a simple preview with main frames. Participants may also want to perform other types of filtering, such as excerpts from the most important videos, a summary of the content taught by the instructor, and the questions asked during the video conference. In this context, extracting highlights from videos is an extremely complex task. In the context of sports videos, Khan *et al.* [12] proposed a method that generates highlights by extracting audio and visual resources from sports video. Thus, in video conferences, a similar technique could be applied to identify and recognize the most important moments and generate a summary video.

Other situations require a textual summary of the content taught during the video conference as well as the main questions asked by the listeners. In this sense, Diao *et al.* [5] perform text summarization with a neural network model using the attention mechanism to capture context information and relationships between sentences improving the performance of phrase regression for text summarization.

UC-08 User Attention/Engagement Detection. Besides detecting users in their streams, video conferencing services could also classify participants' current engagement along the meeting. Sentiment Analysis using users' faces [20] can be done to detect emotions such as sadness and happiness. Thus, such emotions may help classifying the attention and engagement of users during the video conference. This information is particularly useful in the educational context, during, and after video class. The teacher may adapt his/her presentation style and classroom teaching according to this kind of feedback.

UC-09 Live Comments Sentiment Analysis. Sentiment analysis may also be done over users' information shared in video conference chats, like the textual information, links, and images, in order to classify the interest in the main speaker. For instance, this kind of analysis might be used indicate which topics were the ones that caused more doubts during a lecture.

UC-10 Personalized Content. Given the users identification (face recognition) or context information (sentimental analyses), some pre-configured actions could be performed. Thus, the video conference may deliver personalized content to those who have different preferences and requirements. For instance, teachers could program questions that pop up for less engaged students, different content may be presented to different learning profiles, or complementary slides may be presented to students with lower grades.

[8] https://www.microsoft.com/en-ww/microsoft-365/microsoft-teams/group-chat-software

[9] https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[10] https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

UC-11 Virtual Conference Assistant. A Bot is a conversational software that interacts with users. It uses natural language processing in users' textual input, or intents, to perform actions or responses with information. Today, they are used in different industry fields and especially in services regarding consumer support. A Bot may act as a Virtual Conference Assistant in a video conference. It may support meeting actions such as stop/start recording, perform/accept users' intentions, or even configure background music. Moreover, and more interestingly, the aforementioned use cases can be also configured as actions. For instance, the bot may turn specific users anonymous, create video summaries or inform about users engagement.

We must highlight that all aforementioned use cases are not limited to be performed just during the video conference. Most of them could be done during and after the meeting, such as the ones regarding classification, detection, sentiment analyses, and anonymization. On one hand, when performed during the video conference, it enables the main lecturer to react and make decisions. On the other hand, when performed later, after the conference, they enable the main lecturer (or the organizer) to create analyses and reports that may improve the group activities or the video conference itself. Particularly in the case of *Personalized Content* and *Virtual Conference Assistant*, action triggers and customization actions should be prepared before the meeting.

4 FINAL REMARKS

Video conferences have become the main communication and interaction method after the emergence of the COVID-19 pandemic. This paper presented the state of the art of Artificial Intelligence techniques for video that are specially important when considering and designing the much needed new generation of video conferencing systems. Based on the literature review, we could draw a picture for the future regarding the usage of techniques, and list a set of envisaged use cases. These use cases, which are strongly based on AI techniques – such as face and sentiment analyses, video classification, object detection, bot agents etc – might be used to shed some light to the process of improving current and future video conferencing systems regarding the user's experience, privacy and ethical concerns.

This work is part of an ongoing research which aims to improve video conferences using Artificial Intelligence. We mention as future work promoting participatory design sessions with developers and administrators of other videoconferencing services in order to identify and validate requirements for our use cases.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5297–5307.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Yufeng Diao, Hongfei Lin, Liang Yang, Xiaochao Fan, Yonghe Chu, Di Wu, Dongyu Zhang, and Kan Xu. 2020. CRHASum: extractive text summarization with contextualized-representation hierarchical-attention summarization network. *Neural Computing and Applications* (2020), 1–13.

[6] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Summarizing videos with attention. In *Asian Conference on Computer Vision*. Springer, 39–54.

[7] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[8] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 244–253.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://arxiv.org/abs/1609.09430

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997).

[12] Abdullah Aman Khan, Jie Shao, Waqar Ali, and Saifullah Tumrani. 2020. Content-Aware Summarization of Broadcast Sports Videos: An Audio–Visual Feature Extraction Approach. *Neural Processing Letters* (2020), 1–24.

[13] Krishan Kumar, Deepti D Shrimankar, and Navjot Singh. 2016. Equal partition based clustering approach for event summarization in videos. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 119–126.

[14] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.

[15] Yunfei Long, Rong Xiang, Qin Lu, Chu-Ren Huang, and Minglei Li. 2019. Improving attention model based on cognition grounded data for sentiment analysis. *IEEE transactions on affective computing* (2019).

[16] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. 2018. Deep Face Recognition: A Survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 471–478.

[17] Paulo RC Mendes, Eduardo S Vieira, Álan LV Guedes, Antonio JG Busson, and Sérgio Colcher. 2020. A Clustering-Based Method for Automatic Educational Video Recommendation Using Deep Face-Features of Lecturers. *arXiv preprint arXiv:2010.04676* (2020).

[18] Antoine Miech, Ivan Laptev, and Josef Sivic. 2017. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905* (2017).

[19] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[20] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca. 2017. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Transactions on Cybernetics* (2017), 1–11.

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[22] A. Srivastava, S. Mane, A. Shah, N. Shrivastava, and B. Thakare. 2017. A survey of face detection algorithms. In *2017 International Conference on Inventive Systems and Control (ICISC)*. 1–4.

[23] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1891–1898.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[25] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[26] Yongyi Tang, Xing Zhang, Lin Ma, Jingwen Wang, Shaoxiang Chen, and Yu-Gang Jiang. 2018. Non-local netvlad encoding for video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.

[27] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

[28] Faen Zhang, Xinyu Fan, Guo Ai, Jianfei Song, Yongqiang Qin, and Jiahong Wu. 2019. Accurate face detection for high performance. *arXiv preprint arXiv:1905.01585* (2019).