The Virtual Director Concept: Data-Driven Adaptation and Personalization for Live Video Streams

Rene Kaiser
rkaiser@know-center.at
Know-Center – Research Center for Data-Driven Business & Big Data Analytics
Graz, Austria

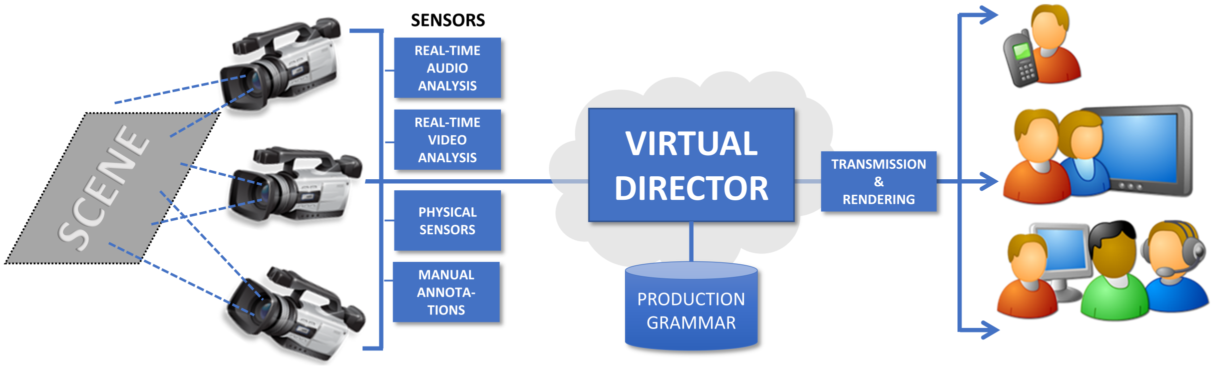

Figure 1: A simplifying illustration of the Virtual Director metaphor. In a broadcast production process workflow from left to right, a Virtual Director software replaces the decision making tasks of a human TV broadcast director. Multiple live cameras capture a scene. The Virtual Director mixing these available views for each individual playout device.

ABSTRACT

This paper gives a comprehensive overview of the Virtual Director concept. A Virtual Director is a software component automating the key decision making tasks of a TV broadcast director. It decides how to mix and present the available content streams on a particular playout device, most essentially deciding which camera view to show and when to switch to another. A Virtual Director allows to take decisions respecting individual user preferences and playout device characteristics. In order to take meaningful decisions, a Virtual Director must be continuously informed by real-time sensors which emit information about what is happening in the scene. From such (low-level) 'cues', the Virtual Director infers higher-level events, actions, facts and states which in turn trigger the real-time processes deciding on the presentation of the content. The behaviour of a Virtual Director, the 'production grammar', defines how decisions are taken, generally encompassing two main aspects: selecting what is most relevant, and deciding how to show it, applying cinematographic principles.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Information systems** → **Multimedia streaming**; *Multimedia content creation.*

KEYWORDS

Virtual Director, interactive media, immersive media, video personalization, AI director, broadcast, telepresence, cinematography, HCI, real-time decision making

DataTV'19 at ACM TVX, June 05, 2019, Manchester, UK

1 INTRODUCTION

A *Virtual Director* is a software component that automatically takes real-time decisions on the mixing and presentation of remote video streams in setups where a scene is covered by multiple cameras. As the metaphorical name of the Virtual Director concept suggests, it addresses the multifaceted challenge of automating the tasks that typically a human broadcast director conducts, in collaboration with further members of the production crew. A scene may consist of a single confined space like a sports arena, or consist of several related spaces as for example in videoconferencing (see Figure 2).

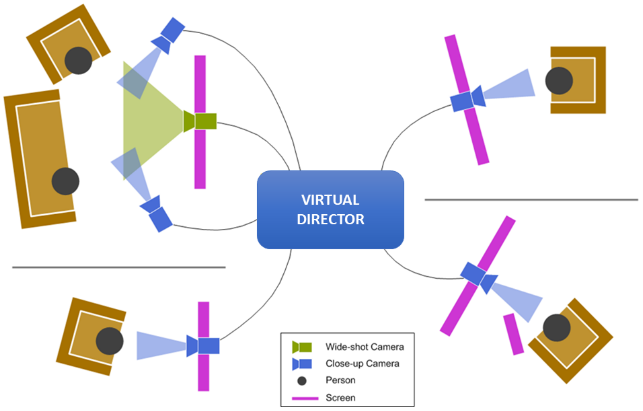

Figure 2: Bird-eye view of a schematic videoconferencing setup involving 4 locations, with conference participants sitting on chairs and sofas. ***Diagram adapted from [45].***

Users are watching actions in the scene via some multimedia system or service, with one or multiple devices like screens for playout. While the Virtual Director concept focuses on traditional 2D video, it can be extended to process as well other forms of video content, or further content modalities such as audio.

The multiple available camera views are combined over time by mixing them, i.e., there are switches from one view to another, using cinematic elements such as cuts and transitions. At each point in time, one or multiple (arranged in some layout) camera views are shown. Cameras may be fixed or moving (robotic, or moved by camera operators). Further components which are part of the overall multimedia system take care of transmitting the content and executing the Virtual Director's decisions such that what it decides is indeed rendered on the users' screens.

With respect to certain aspects, the Virtual Director approach is related to concepts such as *shape-shifted TV* [43] [40], *non-linear interactive narratives* [20] [24], *object-based broadcasting* [38] [2] [22] [33] and *perceptive media* [10].

The Virtual Director concept can be regarded as a concept that combines aspects and capabilities from the above. In a nutshell, the following four aspect define what it enables:

- *Automation* of director's decision making
- *Adaptation* e.g. to playout device capabilities
- *Personalization* to user needs and interests
- *Interaction* during content consumption

A key difference to most related approaches is that it is targeting *live* application scenarios. Like shape-shifting TV, it is suited to take automatic decisions on a very granular level – but in live setups, it reasons with *streams* instead of recorded video *clips*, and with *mixing* instead of *editing* grammar. Real-time constraints provide for especially ambitious research challenges in this realm. In this respect, the Virtual Director concept fills a research gap.

A Virtual Director capable of automating the decisions of a human director may not only create a single output, but take personalized decisions for each individual user or playout device.

This granular personalization capability can be the basis for intriguing features, as intelligent personalization tailored to user needs and preferences can provide added value for media services. The Virtual Director approach allows to serve each user individually. It scales very well and its output is not constrained to a manageable set of pre-authored branches. In order to enable intelligent personalization, two aspects need to be covered: first, the Virtual Director's behaviour – referred to as the *production grammar* – needs to be defined and engineered. Second, the Virtual Director needs to reason about what is happening in the scene, hence it needs access to sensors which inform it about current actions taking place, emitting real-time (meta)data. A Virtual Director is hence a data-driven service.

'Virtual Director' in this paper and underlying research stream is referred to as a *concept*, *approach*, *paradigm*, or a *technology* or *software component*, to indicate a less technical or more technical viewpoint when referring to it.

A Virtual Director is part of an overall multimedia system which is enabling some kind of multimedia service to users. The multimedia system consists of multiple components which may operate in a distributed manner over a (local or Internet) network. The Virtual Director can as a metaphor be referred to as the *brain* of the system taking decisions the results of which are directly visible to its users.

The Virtual Director concept can be applied as well to domains which typically do not involve a human director for vision mixing, such as group video communication [42] [41] [9] where Virtual Director features are hypothesized to be especially valuable for larger groups and groups where participants embody specific communication roles (e.g. a teacher presenting to students).

This paper builds on previous publications (most notably [17]) in a long-term research stream, provides an updated, comprehensive definition for the general Virtual Director concept, and elaborates on many aspects which characterize it. It does not present new experimental results.

The description in this paper is the synthesis of extensive research. Individual aspects of the concept, research prototypes and formal experiments have been documented in a set of publications [42] [41] [9]. See [15] for an overview and further references.

The Virtual Director concept has been developed iteratively. It has matured over the extended timespan of a decade, incorporating inputs and feedback from potential users as well as experts at all stages. Experts included researchers from both technical and non-technical (e.g. video production) backgrounds, as well as representatives of industry stakeholders.

The main idea behind the Virtual Director approach is not entirely new. Research addressing this space so far, however, resulted in dedicated individual solutions for particular application domains and setups. The Virtual Director concept, in contrast, is a generic concept. It fills a research gap and contributes to the research and practitioner community and is the basis for a flexible, re-usable technical approach and software framework.

Apart from this theoretical concept, a technical approach has been proposed and validated as well. It is based on complex event processing (CEP) and rule-based computing. See Figure 3 which illustrates the two main sub-processes of a Virtual Director: the *Semantic Lifting* process which is responsible with deriving a higher-level understanding of what is happening in the scene, and the *Director* decision making process which ultimately decides what to show on a particular screen. See [17] and [15] for further details on this end.

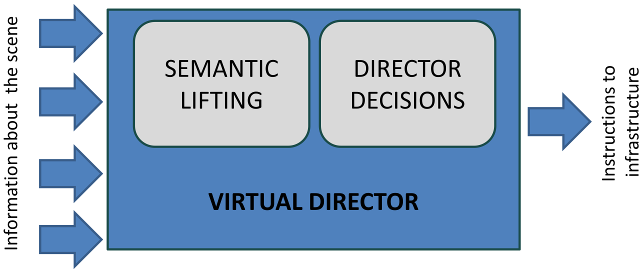

Figure 3: Pattern of decoupling the Semantic Lifting from the director decision making sub-processes.

A near future vision for Virtual Director technology is to utilize it for simple setups with limited user expectations. Such setups may have only few cameras at their disposal, and users may be provided with limited personalization options only. Nevertheless, it could enable a more or less automatic service requiring little effort to set up and operate.

For live event broadcast applications, automatic mixing of multiple camera views should have considerably more visual appeal compared to a continuous live stream of a single (wide field of view) camera. Ad-hoc events or smaller events of limited supra-regional interest may be made accessible for live viewing. Personalization capabilities transform a broadcast into what could be called a *narrowast.*

As a blue sky long term research vision, the intelligent adaptation and personalization features could become standard features in everyday media consumption. Should a further development of this concept be established as a standard technology one day, the Virtual Director could also become the basis for future forms of media content, novel forms of consumption formats engaging audiences in ways which do not exist yet.

2 DIRECTING LIVE TV BROADCAST

The term *Virtual Director* is a metaphor (see Figure 1) for replacing certain tasks of a human broadcast production crew, most notably those of the director. Naturally, every broadcast team may consist of a different number of people with a different distribution of roles. To simplify, the main prototypical roles which a Virtual Director sets out to replace are the director and in some cases and to some extent camera operators (i.e., with respect to the animation of virtual cameras). Beyond, there may be further production crew members with additional specific roles such as the preparation of instant replays or the insertion of pre-recorded content – such roles may be taken up by a Virtual Director as well but are considered additional features beyond the basic concept.

How are the tasks and responsibilities of a human television director defined in literature? Referring to material from a media course [29], the director is responsible for the realisation of the *script* which corresponds to a Virtual Director's production grammar definition. The director also needs to instruct further members of the production crew, and as appropriate also the cast (actors, but not athletes in a sports broadcast):

"*The director is responsible for making the script come alive. They will be in charge of directions for the cast and crew and will give them instructions on how to move around the set with the right attitude towards the scene and script.*" [29]

The description further acknowledges that the concrete responsibilities of a director depend on the type of production, and there is a difference between live (broadcast) production, which is the focus of this paper, and recording video for offline editing:

"*The duties of a television director vary depending on whether the production is live, as in television news, a televised sporting event or recorded to tape, digital video or video server, as in a dramatic or interview production.*" [29]

What is naturally beyond the scope of a digital director service is the placement and setup of the physical equipment:

"*In both types of productions, the director is responsible for supervising the placement of professional video cameras (camera blocking), lighting equipment, microphones, and props.*" [29]

A further section from the same task description explains why the director's job is a very intense one requiring a great amount of concentration. This person must take decisions very quickly with low delay, maintain the overview of a complex set of production equipment, think ahead and steer further members of the production crew:

"*Other than quickly calling out commands, the television director is also expected to maintain order among the staff in the control room, on the set, and elsewhere. A news studio might have multiple cameras and few camera movements. In a sports broadcast, the director might have 20 or 30 cameras and must continuously tell each of the camera operators what to focus on.*" [29]

Directors are using dedicated hardware and software tools to conduct their tasks. For directing live broadcast, the arguably most prominent tool is the vision mixer, a physical switching desk which allows to prepare shots, watch several live streams in parallel on adjacent screens, and execute cuts from one shot to another.

To obtain a basic understanding of how a vision mixer is handled, see the video in [27], part of a larger collection of videos [6] documenting professional production processes which was published by the ADAPT[1] research project. Researchers have previously targeted the further development of such tools as well as their adaptation to certain usage setups, see [7] for an example of a vision mixing system that can be used on mobile devices.

3 SENSORS

To be able to metaphorically 'understand' what is happening in the scene, and to take high-quality decisions, a Virtual Director needs access to information which contain relevant facts either directly, or allow to refer them. The Virtual Director hence needs to be informed by sources of such information, which we refer to as *sensors* in this research. Without any information from the outside, a Virtual Director could not take meaningful but only random decisions. Sensors may be physical devices or 'soft' (software) sensors.

Information is required about what is dynamically going on in the scene, hence also sensor information needs to be provided continuously, e.g. as an event stream or sequence of notifications. Sensor information needs to be generated and transmitted with very low delay due to the real-time requirements of the Virtual Director and the multimedia services it enables. The term *cues* is used in this paper to refer to such sensor information.

Any sensor might be useful, as long as the information provided can be exploited in the production grammar of the Virtual Director. Even redundant sensors and cues might help to filter noise and raise confidence in the correctness of information. Cues may explicitly contain uncertain information which may even be equipped with confidence estimation values. A Virtual Director may have interfaces to several kinds of sensors which may cover multiple modalities. Sensor streams may subject to multimodal fusion when being processed by a Virtual Director.

When cues are processed as *events* with an event processing paradigm, there are some implicit advantages. A key one is that with most event processing frameworks the information received automatically gets associated with a timestamp. Efficient event management and processing algorithms of such frameworks may automatically dismiss cues which are no longer relevant.

Note that the processing of timestamped information can be more complicated than one may intuitively assume. First of all, besides the timestamp stating when the event was received, there may be different timestamps stating when the real-life event it represents has taken place, or when the cue was originally created. Further, the transmission of cues from the sender to the Virtual Director may vary from sensor to sensor, and even from transmission to transmission. This circumstance may be very relevant when executing real-time pattern matching algorithms to analyse the cue streams. Specific out-of-order event processing strategies may have to be applied. The detection of a pattern that consists of a strict sequence of certain events may require a certain time period to be waited to make sure no delayed cue is received which breaks the pattern. However, the use of deliberate waiting periods entails a difficult trade-off with respect to the real-time requirement of the system.

For the data representation of cues, metadata standards such as the the MPEG-7 [26] format or the Media Value Chain Ontology [32] may be useful. Alternatively, self-designing a metadata schema using RDF/OWL [13] is a candidate format with advantages concerning component integration and interoperability. Due to the requirement of very fast processing, leaner formats such as XML or JSON snippets with a self-defined structure and required fields may also be chosen.

The following gives a brief overview on potential sources.

Content analysis

A Virtual Director is part of an overall multimedia system which processes audio and video content streams. As these content streams are anyhow available, they can be accessed with reasonable effort. What is hence an obvious candidate to be integrated and act as a sensor are real-time content analysis modules, i.e. computer vision and audio analysis algorithms that extract information from the content streams on the fly.

From a system architecture point of view, the content analysis modules are regarded as not being part of the Virtual

[1] http://www.adapttvhistory.org.uk/

Director. Instead, they are exchangeable external component with an interface to the Virtual Director. A sensor emits information about one particular scene location. When a Virtual Director's application involves multiple scenes at the same time, such as in videoconferencing, it may fuse cues from multiple locations for processing.

What a typical content analysis sensor informing a Virtual Director would emit are features of rather low abstraction level, simple bits of information describing what is happening in the scene – i.e. not very low-level information on an isolated pixel level and also not complex information derived from multiple aggregated or fused modalities.

Even though real-time content analysis algorithms are typically used, usually there is some, if minimal, processing delay involved. These delays can be critical and need to be considered when designing a Virtual Director service and the workflow and processing chains of the overall multimedia system. Delays may occur especially when analysing high-resolution video, while audio processing is often computationally cheaper and hence causes less delay.

Another limiting aspect of content analysis is the expected quality of its results. Usually, 100% correctness cannot be achieved. Algorithms may be configured to balance the trade-off between false positive and missed detections. There are specific algorithms for specific purposes, e.g. retrieval from large datasets. What can be expected for real-time applications, however, may be considerably less, since

- the algorithm does not have any means to take information about the future into consideration, except for a few frames at most, and
- certain algorithms cannot be used since they simply take too long to provide a result such as a detection[2] of a certain object.

What in recent years triggered considerable advancements in the computer vision research field are deep learning approaches [46] – in some research communities these vast improvements are even referred to as the 'age of deep learning'. A good overall understanding of the state-of-the-art can further be obtained by looking at the results of evaluation competitions and benchmarking events such as the *MediaEval Benchmarking Initiative for Multimedia Evaluation* [21] or the *TREC Video Retrieval Evaluation (TRECVID)* [28].

Examples what content analysis sensors may extract are:

- The current number and exact location of persons or faces in the video.
- The location and movement directions of objects such as a ball in a sports match.
- The gaze direction of a certain person.
- Events when a person starts or stops talking.
- Specific sound events such as a local music concert crowd cheering loudly.
- etc.

Besides these examples, rather different properties which content analysis may extract are quality aspects – see [44] for examples. A Virtual Director may exploit such information by filtering out content streams that are below certain standards.

How is the information contained in these data streams exploited? Since this (meta)data typically contains low-level information as mentioned above, it needs to be translated to a higher abstraction level. In a nutshell, firstly this information is fused over multiple sensors and analysed. This process may be referred to as *Semantic Lifting* (see [17] for details). The result of it are higher-level cues, facts and states. The abstraction level of this higher-level information describes what is happening in the scene in more abstract terms and fits with the second processing step. Hence it can be used there as a decision trigger or decision parameter: the Virtual Director is taking decisions on how to present the available content on the individual screens, as defined by its production grammar.

As an example, an audio analysis sensor may detect a specific sound such as the sound of a whistle in a basketball match. Based on fusing this cue with further available sensor data, it can be inferred that a free throw is going to be awarded on the left side of the court.

Viewer A is in the midst of a training program to become a referee for minor leagues. She is eager to watch the movements and gestures of professional colleagues to learn from them. Watching a game via a Virtual Director service, she sets a preference for this game to closely watch their gestures whenever the game is interrupted by e.g. a foul, a timeout or the ball falling out of bounds.

Viewer B, however, may use different personalization preferences and at the same time see a close-up shot of the player who was fouled and is preparing to attempt the free throw. Would another player of whose team viewer B is not a fan of do the same, the likelihood would increase that the Virtual Director is showing the coaches' reaction instead of focusing on the player.

Manual annotation

Because of the aforementioned deficiencies of automatic content analysis algorithms, such sensors may be replaced or combined with manual annotation interfaces. This is obviously only an option for professional applications where the integration of a manual human annotation task can be afforded to begin with.

[2] Note that while the temporal delay of detections is critical with respect to the overall multimedia system's real-time constraints, the time required for training machine learning algorithms is not relevant.

Dedicated annotation tools with well-designed work processes and user interfaces are required which allow to efficiently enter what is observed. Operating such tools may require extensive training.

The key difference between (a) an operator taking decisions when to cut from one camera view to another using a vision mixing tool, and (b) an operator conducting manual annotation, is that the latter input is not final and can be used for scalable personalization and adaptation.

An obvious annotation strategy is to manually annotate high-level actions that are not directly observable by automatic sensors. Another example task would be identity assignment for persons that are automatically tracked in a scene, e.g. the most interesting people in a concert or sports match. To enable users to follow their favourite characters with close-up shots, a system needs to be aware of their current location.

A manual annotation task could also serve as a tool for validating and correcting the results of content analysis. An interface may also serve for selection of, or re-prioritisation between shot options determined automatically by the Virtual Director system.

Physical sensors

Apart from the above, any further physical sensor may also be used to inform a Virtual Director. It may extract and emit any relevant information from the real-life scene, the physical world. Examples are the location of people or objects (e.g. sports game ball location within a certain area), the state of a certain object (e.g. door open or closed), the distance and speed of certain objects, or real-time 'quantified self' sensors (e.g. heart rate of a person playing a game in a videoconference).

Application context

Further, cues may be extracted from the application context of the multimedia service which the Virtual Director is part of. As an example, let's assume a Virtual Director is part of a videoconferencing tool which is integrated with a social network platform. A sensor in this case could extract characteristics of the relationships of the videoconference participants which the Virtual Director can exploit to bias camera selection towards close ties. This personalization may improve the shared experience of the participants (see the approach in [39] and [37]). Such a sensor would only have to emit periodic updates since this kind of information does not change very rapidly.

4 VISION AND FEATURES

The following discusses key features a Virtual Director may enable in detail. Examples are discussed which implicitly sketch a vision of how Virtual Director technology enhance viewing experiences.

Automatic camera/viewpoint selection

A key question which a Virtual Director needs to continuously evaluate is which part of the scene is currently most relevant to a particular user, which camera views cover it (possibly multiple), and if a switch should be issued to show this camera view instead of the current one.

A Virtual Director is automatically mixing the available content streams and shall do that in a visually engaging manner, providing a (possibly even personalized) viewpoint on the scene and conveying what is happening in the scene by appropriate cinematographic means.

Any switch to another camera view can be initiated by the Virtual Director (i) reacting to what is happening in the scene, (ii) reacting to the current camera no longer being available or, (iii) following cinematographic principles (simple example: maximum shot duration exceeded).

Beyond this basic behaviour, there are many more capabilities and decision factors, depending on the concrete implementation, features enabled, and application context in general. See Figure 4 for an overview of common factors which may influence the decisions a Virtual Director is taking.

Automatic execution of cinematographic/cinematic principles

A basic idea behind the Virtual Director concept is to automate the execution of cinematographic principles – in a nutshell, how the content is framed and how camera views are put into a sequence over time.

Since different terms are used by different communities with inconsistent definitions, to simplify, in this paper the term *cinematographic principles* is mostly used in a generalizing manner to describe cinematographic and cinematic rules, principles, techniques and common conventions, i.e. elements of film language and visual storytelling – see for example [1] [5] [4] [34] [25].

A basic element of cinematography is frame composition [23]. "*Composition refers to the placement of objects and subjects within the frame (...)*" [23] which can be used as a storytelling tool to convey or underline the meaning of a situation, or suggest deeper meaning. Adherence to well-known principles such as symmetry, good headroom, the golden section or the rule of thirds (see [4, pp. 42–44]) may not only be perceived and understood by viewers as a sign of cinematographic quality, but in specific situations also mediate a harmonious atmosphere or favourable sentiment. The breaking of such standard rules the meaning of which is understood and expected by most viewers can be an intentional instrument as well, however. Unconventional composition may

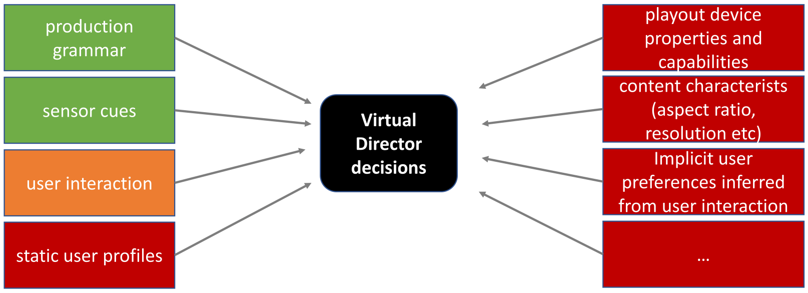
production grammar
sensor cues
user interaction
static user profiles
Virtual Director decisions
playout device properties and capabilities
content characterists (aspect ratio, resolution etc)
Implicit user preferences inferred from user interaction
...

Figure 4: Common factors influencing the decisions a Virtual Director component takes.

enhance user experience when watching dramatic moments or following a surprising twist to the course of events.

A specific instrument which may be used to intensify the meaning of certain situations is to vary the ratio of person's size in contrast to the shot. For example, a narrow close-up of a person may help convey the sense of notable individual achievement, closely showing facial expressions in moments of joy. Showing a single person much smaller, in contrast, may help transport the sense of that person experiencing failure, feeling small, of feeling isolated.

Any principles may be specific to content genres and application domains, yet in general, the utilization of such common conventions is useful since they are proven to be effective and, as mentioned above, commonly understood by most users. Most users are familiar with basic principles and their intended meaning by watching edited content in their everyday lives (television, cinema, online, etc.).

Just like a human broadcast director understands which principles shall be applied in a certain situation, the Virtual Director shall be capable of automatically deciding when to apply certain principles. In contrast to human professionals, a Virtual Director's behavioural repertoire may of course be limited, yet sufficient. A cinematographic principle in that sense could be described as the combination of a cinematographic technique and an understanding in which specific situations and how exactly to apply it.

There are two main facets where a Virtual Director is executing cinematographic principles:

- When selecting camera views to show, considering which type of shot currently fits, considering the temporal sequence.
- For the definition and animation of virtual cameras as digital croppings.

Naturally, a third facts would be for camera steering, but this is not considered in scope.

Cinematographic techniques may help to communicate meaning, entertain, or evoke a particular emotional or psychological response by the audience. This includes e.g. lighting, using depth of field, focus, camera position, camera movement, framing, special effects, cutting effects, etc.

However, a concrete Virtual Director implementation may have means to steer and influence only a subset of these aspects. It may for example not have any mandate and interface to steer the physical cameras and move them around. It may, however, define virtual cameras as croppings of the original video streams and animate these virtual cameras by panning and zooming. While the Virtual Director may not be able to instruct a certain camera to capture a certain type of shot, it may be able to choose from multiple available streams and select one with a shot type desired.

As an introduction with very limited depth, the following, adapted from [17], lists and explains commonly used shot types.

Shot Types by Size. A shot (see [4, pp. 8–24]) can be of any reasonable temporal extent. Its type is defined by the distance between the camera and the subject (in relation to the subject size). It defines the ratio of the size of the visible part of the object to the total area in the shot. Shot types by size may be interpreted differently in different contexts and content genres. For example, a camera operator would use a different focal length for a wide shot when showing an open landscape than when showing a wide shot of an indoor theatre stage.

The following suggests a taxonomy that is based on and a subset of the shots described in [4, pp. 8–24] by Bowen and Thompson.

See Figure 6.10 in [17] for further example shots for six of these shot types, taken from a soccer match. This content was used for a Virtual Director research prototype [18]. The subjects in these examples are a single or multiple athletes.

While most literature like [4] is primarily intended for film production, i.e. recording content of scenes that can be planned ahead, many of the same principles and rationales apply as well to capturing live video, or even photography.

Extreme long shot: An *extreme long shot* shows a large amount of the environment and is taken from a distance where the subject(s) are not clearly visible in detail. It is often used as an 'establishing shot' to show the audience where the action is taking place and shows e.g. the landscape or the surroundings of the scene.

Wide shot: In a *wide shot* the subject(s) are usually shown in full height. While the subject shall be the focus of attention, the surroundings are still visible. For sports content, for example, a group of persons instead of a single one may be of interest and hence framed by a wide shot. That shot may be somewhat wider than for other domains and show several athletes and their position and movement relative to each other.

Medium long shot: A *medium long shot* is something between a wide shot and a medium shot and shows "*more of who than where and can still show when*" [4], i.e. the focus typically is on human subjects while a sense of the surroundings is still conveyed. A medium shot does not show an entire person but cut off e.g. below the knees. However, the exact framing depends on several factors, such as the amount of subjects in focus or their movement.

Medium shot: A *medium shot* typically shows a single subject very prominently, cut off below the waist. A medium shot may well be used for dialogue scenes since it shows the subject at a natural distance for such situations. How narrow the framing can be depends among other factors on the amount of movement and gesturing.

Over the shoulder shot: A shot (not listed in [4]) often used for dialogue scenes, typically showing two people talking to each other, both visible, yet suggesting the perspective of one of them.

Close-up: The *close-up* shot is a full face shot that typically shows the person's face above the upper shoulders. It may cut off the top of a person's hair or head. This shot allows the viewer to focus on a subject's face, subtle emotions, eye gaze and facial expressions.

Extreme close-up: Another shot type is the *extreme close-up* shot which is a very detailed shot of an object or parts of a person, such as the person's eyes, mouth or hand. The viewer is lacking context since no surroundings are identifiable. It can be challenging to use such extreme close-up shots in Virtual Director implementations, and to maintain proper framing when the person is moving or turning.

Shot Types by Angle. Another aspect to consider when positioning a camera are the angles of the shot in relation to the main subject and its own natural direction [4, pp. 45–61].

On one hand, the *horizontal angle* refers to the horizontal angle of the camera to the subject, for example a person standing up straight and firmly looking ahead. Depending on this angle the person may be shown directly from the front, slightly to the side, in a profile view from 90 degrees to the side, or even directly from the back.

The *vertical angle*, on the other hand, refers to the height and vertical angle of the camera. This angle may be horizontal, i.e. parallel to the floor, or looking upwards or downwards. A very common view and natural shot is the *eye level* shot where the camera shows the subject as humans of average height would see it. A slightly lower position and upwards angle may be used to make the subject appear taller and more significant. Vice versa, a higher camera position and downwards angle may enhance the opposite impression. The *bird's eye view* is a very high camera angle which shows a subject or scene from above. While for many content domains this is a very unnatural angle, it can be useful especially in sports and documentaries to give an overview.

Camera Movements. Apart from objects and subjects moving within a static shot, also the physical camera may be moved to create a shot [4, pp. 165–181]. The camera may be mounted on some kind of *tripod* device which allows it to be moved in certain directions. Cameras can also be mounted on a *dolly* vehicle with wheels or rails to smoothly move it along a main direction. The camera may also be moved around feely by a human camera operator who may just carry it or mount it on a *steadicam* device to stabilize it and make sure any movement is perfectly smooth. Cranes may be used as well. Some cranes are robotic and allow to remotely steer and pre-program motion sequences.

The following terms describe the arguably most important camera movements. A *pan* is a horizontal camera movement where there the camera remains in a fixed location. Vertical camera movement is referred to as a *tilt* where the camera points upwards or downwards from a stationary spot. A *truck* is related to a pan but instead of turning the camera, the camera is physically moving sidewards. The *pedestral* analogously is related to the tilt but instead of tilting the camera to look upwards or downwards, the camera is physically moved up or down while the angle remains constant. A *zoom* is technically not a camera movement but produces an effect similar to moving the camera closer or further away from an object or subject by changing the focal length.

With respect to the Virtual Director concept, first of all, steering of physical cameras is considered not in scope for this concept. What is however relevant, is the definition of static and animated *virtual cameras* as croppings of high-resolution source video streams. For these, a Virtual Director needs to automatically take decisions on how to frame and how to animate. Instead of physically moving a camera, it replicates the camera movements mentioned above mainly by a combination of digital (i.e., not optical) zoom, truck and pedestral movements. See [17] for a research prototype with such features.

Shot angle selection

A very specific decision aspect when selecting one from multiple available cameras is when multiple options are available showing e.g. the same person, but from different angles. A video analysis sensor for example could provide metadata about from which angle each person is currently filmed, e.g. stated in degrees, with 0 degrees for a fully frontal shot.

In that case, the Virtual Director can consider the horizontal angle discussed above, and infer some shot types such as a frontal, sideways or profile shots. Information about which shot types are available can be a key decision factor for shot selection. See the Geometric Reasoner in [8] as an example, a reasoning sub-process of a Virtual Director prototype capable of exactly that.

As an example, this feature could be used by a Virtual Director in a videoconferencing system. Let's assume a session involving three people where from one of the two people located remotely the two shots depicted in Figure 5 are available in parallel – a frontal shot, and a sideways shot.

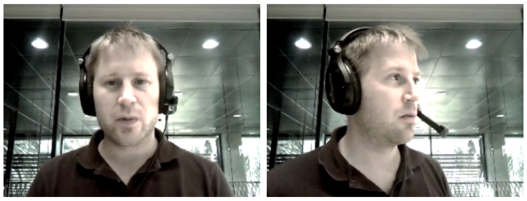

Figure 5: Two shot candidates: a frontal view of a person, and a sideways view. ***Screenshots taken by the author of this paper, and re-used from [16].***

From basic audio analysis cues, the Virtual Director can detect interaction patterns over time and infer turn taking behaviour, specifically who is currently talking directly to whom. The Virtual Director's production grammar defines the following logic: when the local participant is talking to this remote person directly, the frontal shot is shown, since this is the most natural view when directly talking to somebody. However, when the local participant is passively following a conversation between the two remote participants, the sideways shot is preferred since this in turn is a more natural view when passively following a conversation among other people.

Executing that logic, the Virtual Director essentially implements a personalization feature. At each point in time, the participant may get to see different views, those which are currently most suitable.

Enabling natural communication can be an important success factor for mediated communication and telepresence systems (see [36]), something that can be measured via user experience evaluation. To this end, a body of research is available, including the media naturalness theory and naturalness scale proposed by Ned Kock [19].

Definition and animation of virtual cameras

Multimedia services featuring a Virtual Director component likely use multiple physical cameras to capture a scene from different viewpoints. Dynamically switching from one viewpoint to another allows to provide suitable shots of the current actions taking place in the scene. To an appropriate degree and in line with cinematographic principles and proven practices, visual variety adds to the appeal of the resulting content consumption experience.

However, it is also possible to use a single source camera only. One possibility would be to just display the remote video stream like a typical remote webcam view, but this would not be appealing enough for many application scenarios and would not require intelligent content selection, adaptation and personalization.

What a Virtual Director can support to nevertheless add some variety in shot types is to use different croppings of the single camera source – *virtual cameras* – which show different parts of the scene, possibly in different zoom levels and shot types. The camera source could for example be a panoramic camera or one with a large field of view such that a whole music concert stage or sports court fits into the shot. A schematic example of concurrent virtual camera cropping candidates from a high-resolution panoramic video stream is visualised in Figure 6.

Naturally, the resolution of the source camera and the crop ratio need to be high and low enough such that the resolution of the resulting virtual camera is still useful. Virtual cameras can be static or dynamic, i.e. the Virtual Director animates them moving across the image.

Altogether, while the use of virtual cameras can be an inexpensive option, there are inherent limitations. Apart from specific camera types like light field (plenoptic) cameras, the focus can not be adjusted after the recording. Viewers will notice the difference between digital croppings and real zooming by physical cameras.

Layout composition of multiple streams

A Virtual Director may show only one single camera view at a time, like it is the case in traditional TV most of the time. The image may fill the whole screen or parts of it, leaving

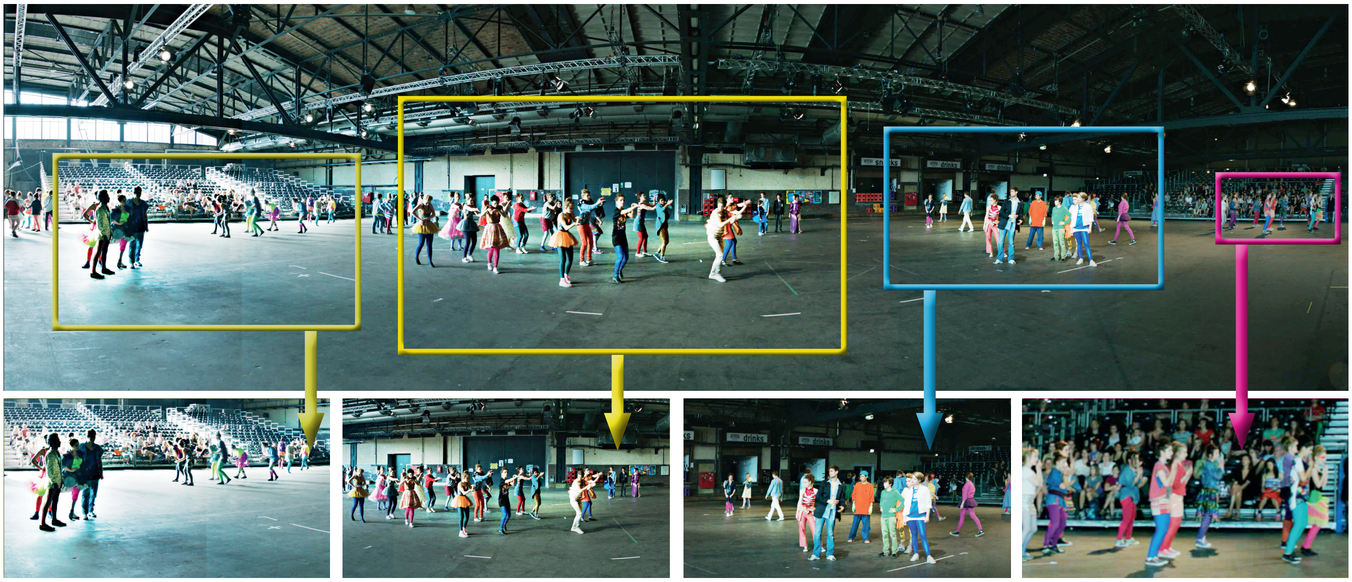

Figure 6: Different virtual cameras can be positioned within a high-quality panoramic video stream, covering different aspects of the scene and the actions within, with shots that differ in type and aspect ratio. ***Image re-used from [17].***

space for e.g. menu and control elements which are part of a multimedia service. It may, however, also show multiple camera views at the same time, arranging them in some kind of static or dynamically changing layout.

For the definition of layouts, templates may be used. See Figure 7 for four examples: one example shows a single view while the other combine multiple.

When combining multiple streams, there are several options and parameters: the camera views can be of equal or varying size, possibly cropped. They can be aligned in an adjacent manner or overlap. Layouts may cover the whole available space or leave parts of the area blank.

See Section 3.2.2 of [45] for more details regarding the practical realization in a research prototypes.

Towards a practical solution of layout adaptation several challenges need to be solved, including how to fit the source streams into the predefined areas within the layout template. Towards that end, see also the Bachelor thesis of Thomas Popp [31].

Content adaptation

Involving a Virtual Director enables a multimedia system to handle the adaptation of content. The perhaps most obvious use case for that is to adapt to the concrete properties of playout devices.

The Virtual Director's role within this process is not to alter the content directly, but to take decisions on how exactly it needs to be altered or selected. Decisions are instructing components (e.g. content transmission, content processing, viewpoint framing, rendering/playout) to execute the adaptation required. The decision can obviously differ for users or groups of users – which effectively translates to playout devices and groups thereof. Depending on where in the processing chain the adaptation is put into effect, there are different advantages and disadvantages – close to the capture site, in the network, or only at playout time [11]. Naturally, the Virtual Director needs to be able to obtain the playout device properties via some interface.

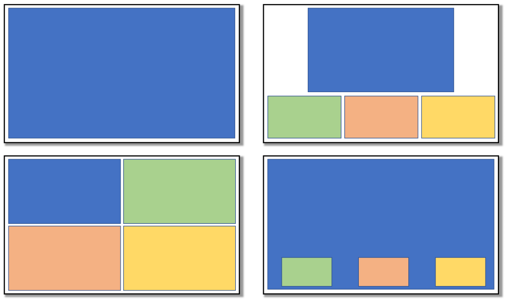

Figure 7: Four example layouts, showing how a single or multiple video streams can be displayed.

With respect to video, the Virtual Director's decision process may take the screen resolution into account when selecting camera sources and their current shot types. It can take the playout screen's aspect ratio into account when cropping the original camera feed. The current screen orientation (portrait or landscape orientation, or even square) is another related aspect, one which is gaining more and more importance as more and more content is recorded, streamed and consumed on mobile devices in portrait orientation.

It can also bias or restrict the shot types used. The detail level of extreme close-up shots, the speed of camera pans, or certain patterns of shot sequences may be appropriate for some playout screen and not for others. Fast camera movement like a pan for example may look appropriate on a small screen like on a mobile phone while the same may appear overwhelming on a large projection.

As a side note, there are also (broadcast production) systems which dedicatedly handle content adaptation to playout device properties and capabilities as a special feature. Examples of such systems are in some cases referred to as *format-agnostic production systems*. One of the research projects in which research related to the Virtual Director concept was conducted has addressed such systems. For more details on format-agnostic approaches, see the book *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media* which a Virtual Director chapter [17] is part of.

Content personalization

Another feature which may be very beneficial for users' viewing experience is content personalization: a software component's ability to automatically decide which camera viewpoint to show and when to cut is the basis for the ability to take different decisions for different users in parallel, in line with their individual preferences. Personalisation may concern individual users or groups of users. Users may express preferences both before and during the consumption of content, or the use of a multimedia service.

Theoretically, personalization may concern any aspect of a Virtual Director's decision making processes, by filtering, introducing a bias, etc. The following lists some examples for the application domain of watching a live event such as a sports competition or music concert from remote:

- Bias camera selection to focus on a certain person, i.e. increase the likelihood that this person is visible. Such a personalization feature could be used by fans of an athlete, or relatives of a particular musician.
- Bias camera selection towards a group of persons, such as athletes of a particular team, or musicians playing a certain instrument.
- Focus on a certain object which a user intends to observe closely, unless something very relevant is happening elsewhere.
- Bias towards certain types of actions, e.g. when multiple competitions may take place in parallel, as during a track-and-fields meeting or race competitions with a large number of participants. The Virtual Director's production grammar can ensure that actions of particularly high priority nevertheless may overrule specific personalization preferences. For example, for a viewer who chose to follow all athletes of a certain country across all disciplines at a track and field competition, whatever she is currently watching would be left in favour of the start of the 100m sprint final, regardless of the competitor's nationality, to the benefit of the viewing experience.
- Preference towards certain types of cinematographic shots such as close-ups.
- etc.

Further examples for the application domain of participating in a group video communication session:

- Every participant may get to see something different at each point in time to begin with, as a Virtual Director can distinguish between a self-view and views of remote participants. The latter group is obviously different for every participant.
- A participant in a *teacher role* giving a talk or explaining something may see several remote participants in parallel, arranged in a mosaic layout, to be able to observe their non-verbal reactions, and assess if the contents of what has been told was understood by the participants.
- In the previous example, all other participants in a *student role* may see a close-up shot of the teacher most of the time, in order to concentrate on what is being said, and being able to follow also the non-verbal communication such as gestures.
- In sessions involving larger groups of participants, camera selection bias towards persons the local user has particularly strong social ties with, based on social network data.
- etc.

While intelligent personalization may create added value for users, and in turn for content/service providers, it shall be noted that to personalisation there may be disadvantageous aspects as well. For certain content genres at least, potential users expressed concerns towards this Virtual Director feature since they feared it might be detrimental to enabling *shared experiences*. Some mentioned it may be difficult to talk about strongly personalized content with friends, for example when meeting in the office the next day, since they might not have seen the same content, at least on a detailed level.

Enabling personalization of a media service that is perceived by users as valuable and worthwhile is not a straightforward task. "*For personalization, you have to understand* *people*", Lora Aroyo mentioned in her inaugural speech[3] '*Data Science with Humans in the Loop*' at Vrije Universiteit Amsterdam in 2017.

Another challenge towards the design of an overall multimedia system which includes a Virtual Director is how to make the overall workflow of reacting to users' preference inputs work. For this, a whole chain of components needs to work together smoothly, including the user interaction per se, the interpretation of the user input, and the decision making processes of the Virtual Director. Any preference changes by users shall make an immediate effect, as far as applicable. Noticing the changes take effect should make users confident that the system work correctly.

Real-time processing

Processing data and reacting with decisions with minimal, real-time delay, is one of the most essential features of a Virtual Director. One of the main challenges with a fully automatic Virtual Director approach are implications of the delays that occur and accumulate over the whole processing chain within the overall, distributed multimedia system.

A Virtual Director is only one component embedded in the system architecture of a distributed multimedia system. It depends on further components and the sum of all components needs to function together. The interplay of a Virtual Director with related components such as sensors or renderers needs to be designed thoroughly, in order to be able to meet the users' expectations with respect to service delays.

Comparing with the current state-of-the-art of broadcast solutions (TV/Internet), there are significant delays which consumers notice and are bothered with especially when co-located people watch for example the same live event via broadcast TV and there are several seconds or even minutes time difference. Instant updates via social media contribute to this issue, as people already may find out what is going to happen before they actually see it. Such delays are detrimental to their viewing experience.

Media industry players seek to reduce the overall delay of processing and transmitting content, see e.g. the approaches discussed in [30] and [12]. The latter White Paper by Harmonic Inc discusses how latency adds up over the broadcast chain. It also suggests a latency sensitivity scale where viewer sensitivity depends both on the type of content and the type of service. Not surprisingly, worldwide premium events and sports in general are considered most crucial. The White Paper does not seem to consider symmetric setups beyond broadcast like for example videoconferencing where content is both sent and received. For the Virtual Director approach, however, the real-time requirement is especially crucial in such setups.

Supporting active and passive consumption

A Virtual Director capable of fully automatic decision making may enable fully passive content consumption where the user just watches the default coverage and does not interact at all with the system to change preferences. Conversely, it is also a feature of the Virtual Director approach to enable very active consumption, i.e. to enable users to interact with the system to take decisions themselves or express preferences which influence future decisions taken by the Virtual Director. We refer to these two opposing viewing modes as *lean forward* and *lean backward* consumption.

It is assumed that over a considerable amount of time, most users prefer not to stick to either extreme but change their mode of consumption. Apart from the two extremes, the Virtual Director may also support a combination or compromise thereof, e.g. by influencing tendencies rather than hard rules, or by stating abstract preferences which the Virtual Director itself translates to particular viewing situations.

Support for multiple degrees of automation

There are two distinct aspects where a Virtual Director can support varying degrees of automation.

First of all, a Virtual Director may take decisions fully automatically on one hand, or support and even require user input on the other hand. From a different point of view, this relates to shades between active and passive consumption as described in the previous section above.

Second, from the multimedia service provider's point of view, a Virtual Director component not necessarily implements a fully automated process. Members of a production team could steer or inform it to some degree, e.g. via manual real-time annotation.

A simple illustrating example would be to have an operator annotate fouls during the broadcast of a basketball match in a setup where these actions cannot be detected automatically. Another example would be to have an operator continuously adjust the framing of shots, or filter out inadequate shot from a list of shot candidates.

It is essential to note that the production professionals are not taking final decisions in these examples. They are not taking decisions what is eventually shown on a particular playout device, but are injecting metadata that can be used for the personalisation process which in parallel creates potentially many different versions for the many parallel viewers. To be able to do that, operators require dedicated user interfaces – see [3] for an example and general considerations.

For the users, the degree of automation on this end is likely transparent, i.e. not known and not relevant. Either degree

[3] See: https://av-media.vu.nl/mediasite/Play/5745f2482d3f4fe7a547458393af322a1d

of automation may, however, enable different features which users experience and benefit from.

User preference expression and user interaction

A fundamental capability of the Virtual Director vision is that users may express preferences before and during content consumption such that the Virtual Director can take them into account for decision making. Preference expression and user interaction can be designed in manifold ways, not constrained by the concept presented in this paper.

How can these inputs eventually be taken into account in technical terms? This may be realized as a parameterization of production grammar, as a switching to different, distinct production grammar definitions, or by triggering a certain behaviour within a production grammar definition.

Viewers may influence any aspect of the content selection and content presentation. Preferences may be expressed in any form and on any abstraction level, for example:

- Meta-decisions, such as the selection of a general cinematic style.
- Individual detailed 'branching' decisions – *do you want A or B?* – that can be taken any time or only when prompted for input.
- Adjusting bias settings, e.g. setting a level for a certain kind of bias on a normed scale such as 0 − 100.

Apart from direct interaction, a Virtual Director may infer user preferences from external sources, such as social media profiles. Preferences may further be inferred from previous interactions analogous to the concept of *relevance feedback*. For a recent survey on interaction methods for interactive media access, see [35]. Independent of the concrete design, privacy concerns have to be considered and respected.

Even though user interaction to dynamically express and change preferences while using a Virtual Director service is a key part of the concept, this aspect was not a focus of the research prototypes developed alongside this work.

Predictive behaviour

The general concept of a Virtual Director is that it derives an 'understanding' of what is happening in the scene from its sensor cues and *reacts* to it by taking appropriate decisions. Due to the real-time setup, it cannot look into the future, in contrast to applications based on recorded content. A Virtual Director may create added value for its users based on reactive behaviour only.

However, in certain application contexts such purely reactive behaviour might not be sufficient and may not satisfy user expectations. Implementing *predictive* behaviour is typically very challenging, though. Predictive decision making requires reasoning with assumptions and uncertainty, in a sense rather different than reasoning with uncertain cues. A predicted event may or may not happen eventually. Odd behaviour can be the result of the latter case.

In technical terms, predictive production behaviour may be realised for example by partial recognition of patterns on the sensor cue streams.

5 ECONOMIC PERSPECTIVE

Research on the Virtual Director concept is encouraged by the author's belief that there is a lot of potential for this kind of technology to make a business impact. Numerous research and industry experts who got in touch with this string of research have acknowledged the appeal of its application opportunities. Yet, this technology is in its infancy and only time will tell how successful and wide-spread it may become.

While the original idea behind the Virtual Director concept was to automate the decision making task of a human director, its impact is not limited to the idea of automating human labour, automating this complex task. The Virtual Director concept can also be applied in numerous application domains where currently no human director is involved. Further, it could be a basis for the development of entirely new media consumption formats, enabling more interactive and more personalized content consumption.

A more short term vision is that it could first be utilized in rather simple, low-profile setups with less demanding requirements with respect to the cinematic and aesthetic quality expected by users. For broadcasters, the availability of Virtual Director technology may make the coverage of smaller regional events more attractive and economically appealing. Instead of deploying a large broadcast crew, a Virtual Director working with a set of fixed (not moved around by operators) cameras could be the basis for a low-cost solution that would allow to cover such events. Production cost could be minimized by putting a set of cameras in place, configuring a Virtual Director's production camera to the specifics of the physical setup, and relying on re-usable production grammar to conduct the task automatically. Providers of media content and media services may embrace the Virtual Director technology since

- intelligent features can help differentiate their product from those of competitors,
- or users are willing to pay extra for advanced features.

Already simple forms of personalization may have considerable added value for users, for example in application scenarios where there are many actions going on in parallel and users want to influence what is prioritized, selected and shown. One example are track and field competitions where important actions may happen in parallel and users may favour certain disciplines, individual athletes, representatives of certain nations, etc. Users may bias content presentation and their preferences may change during content consumption, by means of interacting with the system.

Last but not least, there are many application scenarios where typically there is no human director involved currently, but adding a virtual one may significantly improve the experience. Video communication is one such potential domain, see [8].

A Virtual Director allows to parallelise the decision making process. The approach scales while a human production crew could for economic reasons could not produce a large number of different outputs in parallel.

One way to monetize Virtual Director services and the additional effort required to set them up compared to traditional broadcast is to include personalized advertising as well. "*The power of individual targeting – the technology will be so good it will be very hard for people to watch or consume something that has not in some sense been tailored for them*", Google's Eric Schmidt told the Wall Street Journal in 2010 to that end [14].

6 DISCUSSION AND CONCLUSION

This paper presents and dissects the Virtual Director concept, a concept for technology capable of automating the decision making tasks of a human TV broadcast director, and taking personalized live video stream mixing decisions for individual users. A Virtual Director is a data-driven content personalization and adaptation service, as it relies on sensors which inform it about what is happening in the scene. Bits of low-level information are possibly fused, analysed and interpreted, such that higher-level cues, facts and states can be inferred. This in turn is triggering or parametrizing the decision making processes of a Virtual Director which eventually decide what is shown on a particular screen.

The Virtual Director approach can also be applied to further, non-broadcast application domains, such as videoconferencing. For this theoretical concept, an implementation approach has been proposed and validated as well [17].

From a research point of view, there is scope to extend this basic and generic concept to other forms of media and to add further intelligent features. One of the future challenges towards enabling high-end applications is to support the production grammar engineering process. Especially, tools and design patterns for authoring, structuring, adapting, testing and re-using complex production grammar definitions need to be developed.

ACKNOWLEDGMENTS

The Know-Center is funded within the Austrian COMET Program – Competence Centers for Excellent Technologies – under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

REFERENCES

[1] Daniel Arijon. 1976. *Grammar of the Film Language.* Focal Press.

[2] Mike Armstrong, Matthew Brooks, Anthony Churnside, Michael Evans, Frank Melchior, and Matthew Shotton. 2014. Object-based Broadcasting – Curation, Responsiveness and User Experience. *BBC R&D White Paper WHP 285* (2014).

[3] Werner Bailer, Marco Masetti, Goranka Zoric, Marcus Thaler, and Georg Thallinger. 2014. *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media.* Wiley, Chapter Semi-Automatic Content Annotation, 166–208.

[4] Christopher J. Bowen and Roy Thompson. 2013. *Grammar of the Shot* (3rd ed.). Taylor & Francis.

[5] Blain Brown. 2002. *Cinematography: Theory and Practice: Imagemaking for Cinematographers, Directors and Videographers.* Focal Press.

[6] John Ellis, Amanda Murphy, and Nick Hall. 2018. Discover research from ADAPT, on figshare. Retrieved 23 April, 2019, from https://royalholloway.figshare.com/Adapt.

[7] Arvid Engström, Goranka Zoric, Oskar Juhlin, and Ramin Toussi. 2012. The Mobile Vision Mixer: A Mobile Network Based Live Video Broadcasting System in Your Mobile Phone. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (MUM'12).* ACM, New York, NY, USA, Article 18, 4 pages. https://doi.org/10.1145/2406367.2406390

[8] Manolis Falelakis, Rene Kaiser, Wolfgang Weiss, and Marian Ursu. 2011. Reasoning for Video-Mediated Group Communication. In *Proceedings IEEE International Conference on Multimedia & Expo (ICME'11).* 1–4.

[9] Manolis Falelakis, Marian F. Ursu, Erik Geelhoed, Rene Kaiser, and Michael Frantzis. 2016. Connecting Living Rooms: An Experiment In Orchestrated Video Communication. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX'16).* ACM, New York, NY, USA, 49–58. https://doi.org/10.1145/2932206.2932215

[10] Adrian Gradinar, Daniel Burnett, Paul Coulton, Ian Forrester, Matt Watkins, Tom Scutt, and Emma Murphy. 2015. Perceptive Media – Adaptive Storytelling for Digital Broadcast. In *IFIP Conference on Human-Computer Interaction (INTERACT'15).* Springer International Publishing, 586–589. https://doi.org/10.1007/978-3-319-22723-8_67

[11] Simon N. B. Gunkel, Jack Jansen, Ian Kegel, Dick C. A. Bulterman, and Pablo Cesar. 2013. The Optimiser: Monitoring and Improving Switching Delays in Video Conferencing. In *Proceedings of Workshop on Mobile Video Delivery (MoViD'14).* ACM, New York, NY, USA, Article 1, 6 pages. https://doi.org/10.1145/2579465.2579472

[12] Harmonic. 2018. DASH CMAF LLC to Play Pivotal Role in Enabling Low Latency Video Streaming. *A white paper from Harmonic, with contributions from Akamai, THEOplayer, Viaccess-Orca and NexStreaming* (2018), 1–12.

[13] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. 2012. *OWL 2 Web Ontology Language: Primer (Second Edition).* W3C Recommendation. http://www.w3.org/TR/owl2-primer/ Available at http://www.w3.org/TR/owl2-primer/.

[14] Holman W. Jenkins Jr. 2010. Google and the Search for the Future. The Wall Street Journal interview with Eric Schmidt of Google, retrieved 19 May, 2019, from https://www.wsj.com/articles/SB10001424052748704901104575423294099527212.

[15] Rene Kaiser. 2016. Virtual Director: Towards Automatic Real-Time Viewpoint Selection. In *Adjunct Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX'16), Doctoral Consortium.* https://doi.org/10.6084/m9.figshare.3457466.v1

[16] Rene Kaiser and Wolfgang Weiss. 2013. Using Cinematic Techniques to Improve Video Communication. In *6. Forum Medientechnik*. http://fmt.fhstp.ac.at/

[17] Rene Kaiser and Wolfgang Weiss. 2014. *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media*. Wiley, Chapter Virtual Director, 209–259.

[18] Rene Kaiser, Wolfgang Weiss, and Gert Kienast. 2012. The FascinatE Production Scripting Engine. In *Proceedings of the 18th International Conference on Advances in Multimedia Modeling (MMM'12)*. Springer-Verlag, 682–692. https://doi.org/10.1007/978-3-642-27355-1_73

[19] Ned Kock. 2004. The Psychobiological Model: Towards a New Theory of Computer-Mediated Communication Based on Darwinian Evolution. *Organization Science* 15, 3 (2004), 327–348. https://doi.org/10.1287/orsc.1040.0071

[20] Hartmut Könitz and Noam Knoller. 2017. *Interactive Digital Narratives for iTV and Online Video*. Springer Singapore, Singapore, 1097–1126. https://doi.org/10.1007/978-981-4560-50-4_44

[21] Martha Larson. 2019. MediaEval Benchmarking Initiative for Multimedia Evaluation. Retrieved 23 May, 2019, from http://www.multimediaeval.org/.

[22] Jie Li, Zhiyuan Zheng, Britta Meixner, Thomas Röggla, Maxine Glancy, and Pablo Cesar. 2018. Designing an Object-based Preproduction Tool for Multiscreen TV Viewing. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA'18)*. ACM, New York, NY, USA, Article LBW600, 6 pages. https://doi.org/10.1145/3170427.3188658

[23] Daniel Maddock. 2018. Uncomposed: Unconventional cinematographic composition in cinema and television. *Australian Art Education* 39, 2 (2018), 268–287.

[24] Britta Meixner. 2014. *Annotated Interactive Non-linear Video – Software Suite, Download and Cache Management*. Ph.D. Dissertation. University of Passau.

[25] Soja-Marie C Morgens and Arnav Jhala. 2014. EduCam: Cinematic Vocabulary for Educational Videos. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[26] MPEG. 2001. Information Technology – Multimedia Content Description Interface. ISO/IEC 15938.

[27] Amanda Murphy, John Ellis, and Nick Hall. 2018. outsidebroadcast-rigging-bitesize18-cameraorder.mp4. Retrieved 23 April, 2019, from figshare, https://royalholloway.figshare.com/articles/outsidebroadcast-rigging-bitesize18-cameraorder_mp4/6159563/1. https://doi.org/10.17637/rh.6159563.v1

[28] NIST. 2019. TREC Video Retrieval Evaluation: TRECVID. Retrieved 19 May, 2019, from https://trecvid.nist.gov/.

[29] Pearson Education Ltd. 2014. The Development and Principles of Editing. Retrieved 23 April, 2019, from http://bcsmedia2014-15.weebly.com/stephen-barrett1.html.

[30] Chris Poole. 2019. Reducing Latency – Video Streaming Without the Delay. BBC R&D blog post, retrieved 30 January, 2019, from https://www.bbc.co.uk/rd/blog/2018-09-latency-video-streaming.

[31] Thomas Popp. 2012. Visualisierung von Videokonferenzen mit mehreren Teilnehmern. Bachelor thesis at Georg-Simon-Ohm Hochschule Nürnberg.

[32] Victor Rodriguez-Doncel and Jaime Delgado. 2009. A Media Value Chain Ontology for MPEG-21. *IEEE MultiMedia* 16, 4 (2009), 44–51. https://doi.org/10.1109/MMUL.2009.78

[33] Thomas Röggla, Jie Li, Stefan Fjellsten, Jack Jansen, Ian Kegel, Luke Pilgrim, Martin Trimby, Doug Williams, and Pablo Cesar. 2019. From the Lab to the OB Truck: Object-Based Broadcasting at the FA Cup in Wembley Stadium. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA'19)*. ACM, New York, NY, USA, Article CS16, 8 pages. https://doi.org/10.1145/3290607.3299038

[34] Rémi Ronfard. 2012. A Review of Film Editing Techniques for Digital Games. In *Workshop on Intelligent Cinematography and Editing*, R. Michael Young Arnav Jhala (Ed.). ACM, Raleigh, USA. http://hal.inria.fr/hal-00694444

[35] Klaus Schöffmann, Marco A. Hudelist, and Jochen Huber. 2015. Video Interaction Tools: A Survey of Recent Work. *ACM Comput. Surv.* 48, 1, Article 14 (2015), 34 pages. https://doi.org/10.1145/2808796

[36] Abigail J. Sellen. 1995. Remote Conversations: The Effects of Mediating Talk with Technology. *Hum.-Comput. Interact.* 10, 4 (1995), 401–444. https://doi.org/10.1207/s15327051hci1004_2

[37] Tim Stevens, Ian Kegel, Douglas Williams, Pablo Cesar, Rene Kaiser, Nikolaus Färber, Pedro Torres, Phil Stenton, Marian F. Ursu, and Manolis Falelakis. 2012. Video Communication for Networked Communities: Challenges and Opportunities. In *Proceedings of 16th International Conference on Intelligence in Next Generation Networks*. 148–155.

[38] Philip N. Tudor, Peter J. Brightwell, and Robert N. J. Wadge. 2016. Future Models for Live Event Broadcasting. *SMPTE Motion Imaging Journal* 125, 3 (2016), 40–45. https://doi.org/10.5594/JMI.2016.2534318

[39] Marian Ursu, Pedro Torres, Vilmos Zsombori, Michael Franztis, and Rene Kaiser. 2011. Socialising Through Orchestrated Video Communication. In *Proceedings of the 19th ACM International Conference on Multimedia (MM'11)*. ACM, 981–984. https://doi.org/10.1145/2072298.2071918

[40] Marian F. Ursu, Jonathan J. Cook, Vilmos Zsombori, Robert Zimmer, Ian Kegel, Doug Williams, Maureen Thomas, John Wyver, and Harald Mayer. 2007. Conceiving ShapeShifting TV: A Computational Language for Truly-Interactive TV. In *Interactive TV: a Shared Experience*, Pablo Cesar, Konstantinos Chorianopoulos, and Jens F. Jensen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 96–106.

[41] Marian F. Ursu, Manolis Falelakis, Martin Groen, Rene Kaiser, and Michael Frantzis. 2015. Experimental Enquiry into Automatically Orchestrated Live Video Communication in Social Settings. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX'15)*. ACM, 63–72. https://doi.org/10.1145/2745197.2745211

[42] Marian F. Ursu, Martin Groen, Manolis Falelakis, Michael Frantzis, Vilmos Zsombori, and Rene Kaiser. 2013. Orchestration: TV-like Mixing Grammars Applied to Video-communication for Social Groups. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*. ACM, 333–342. https://doi.org/10.1145/2502081.2502118

[43] Marian F. Ursu, Ian C. Kegel, Doug Williams, Maureen Thomas, Harald Mayer, Vilmos Zsombori, Mika L. Tuomola, Henrik Larsson, and John Wyver. 2008. ShapeShifting TV: interactive screen media narratives. *Multimedia Systems* 14, 2 (2008), 115–132. https://doi.org/10.1007/s00530-008-0119-z

[44] Stefanie Wechtitsch, Hannes Fassold, Marcus Thaler, Krzysztof Kozłowski, and Werner Bailer. 2016. Quality Analysis on Mobile Devices for Real-Time Feedback. In *International Conference on Multimedia Modeling MultiMedia Modeling (MMM'16)*. Springer International Publishing, Cham, 359–369.

[45] Wolfgang Weiss, Manolis Falelakis, Rene Kaiser, and Marian F. Ursu. 2014. Models for Decision Making in Video Mediated Communication. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions (UM3I'14), in conjunction with the ACM International Conference on Multimodal Interaction (ICMI'14)*. ACM, 45–50. https://doi.org/10.1145/2666242.2666250

[46] Wei Zhang, Ting Yao, Shiai Zhu, and Abdulmotaleb El Saddik. 2019. Deep Learning-Based Multimedia Analytics: A Review. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1s, Article 2 (2019), 26 pages. https://doi.org/10.1145/3279952