https://doi.org/10.2352/EI.2023.35.10.HVEI-254

Biosensors For Landing Creative Intent

Scott Daly, Evan Gitterman, Dan Darcy, and Shane Ruggieri
Dolby Labs, Inc., San Francisco, USA

Abstract

The motivation for use of biosensors in audiovisual media is made by highlighting the problem of signal loss due to wide variability in playback devices. A metadata system is proposed that allows creatives to steer signal modifications as a function of audience emotion and cognition as determined by biosensor analysis. This is needed because today's audiovisual ecosystem includes such a wide variety of playback devices (i.e., display plus sound) that the audience's experience can change substantially for the same source content. In many cases, such changes may cause a distortion of the creative intent. In this proposed system, the rendering is affected by assessment of the viewer's internal state (i.e., emotion and cognition) which is obtained by various types of physiological monitoring. This assessment is combined with metadata for narrative and emotional expectation that is inserted during the content production stages. Computational models are used to consolidate multiple sources of physiological signals as well as the interactions with the metadata and the playback device used for final transduction of the source signals to the viewer. As a result, the system allows for creative intent to be scalable as best as possible across many types of playback systems.

Author Keywords

Creative Intent, Playback Technology Variability, Biosensors, Physiological Assessment, Displays, Audio Systems,

Introduction

Bioinstrumentation technology is steadily entering the consumer market, such as smart rings using coarsely sampled spectrophotometers to assess blood oxygen level and pulse rate. Smart watches have even more biosensor capability including photoplethysmography, skin temperature measurement, clinically accurate ECG, ego movement, optical or oscillometric blood pressure, respiration rate, EDA or GSR, and BIA. Smartphones routinely have facial cameras using visible light, structured light, and LIDAR. Some TV models have room-facing cameras to allow them to behave more like smartphones, thus seeing the viewers, and a newer trend is to use radar. While facial cameras and depth sensors are not typically categorized as biosensors, when they are used to determine facial expression, pupil size, and gaze, we consider them to be acting as biosensors. Prototypes showing advanced biosensing include headphones and smart earbuds that can measure heartrate, temperature, GSR, and EOG for eye tracking. Smart AR glasses typically have eye tracking. Gaze position, pupil tracking, and facial expression recognition are already present in recent consumer HMDs for AR, VR, and XR [1]. There is even a VR HMD on the market with built-in EEG sensing [2,3] for localization and timing of brain activity. Currently it is only considered laboratory equipment as opposed to a consumer product due to initial price. Large-scale group responses to cinema have also been explored in the laboratory, including facial expression recognition by Disney [4] while CO_2 gas sensors have been used by Dolby to determine overall audience arousal levels [5]. Using GSR and heart rate tracking, Warner Media has studied how audiences engage with cinema narrative [6]. In combination, these biosensors and cameras can be used to assess an individual's or group's mental state, including stress, arousal/valence, emotion, cognitive load, ROI, and attention locus. While the accuracy, precision, and cost of some of these biosensors are open to criticism, we can expect improvement along all three of those dimensions.

(Acronyms: ECG = electrocardiogram, EDA = electrodermal activity, GSR = galvanic skin response, BIA = bioelectric impedance analysis, LIDAR = light detection and ranging, SLAM = simultaneous localization and mapping, AR = augmented reality, TV = television, EOG = electro-oculogram, HMD = head-mounted display, VR = virtual reality, XR = extended reality {includes VR, AR, and Mixed Reality}, EEG = electroencephalogram, CO_2 = carbon dioxide, ROI = region of interest)

Another key trend, occurring in consumer entertainment technology, is the continual improvement in capabilities of playback devices (displays, TVs, speakers, and headphones) as well as signal formats. This leads to improved quality of experience for consumers and allows for a wider gamut of expression by the creative production community. Examples include HDR, WCG, and UHDTV, providing more realism and expression over SDR-SDTV, while for audio, both 3D object-oriented sound and binaural HRTF rendering allows for better realism and immersion over stereo or 5.1. However, there are some negative consequences in the overall media ecosystem due to these improvements. The significant variability of the capabilities of playback devices for video and audio can cause a loss of information intended to be perceived by consumer, for those whose playback devices or viewing environments are of lower capability than those used in production. In many cases this can affect the creative intent. A famous example of creative intent failure due to technology which had financial, legal, and even political implications will be described, as well as more recent examples of widespread creative intent failure due to playback technology variability.

(Acronyms: HDR = high dynamic range, WCG = wide color gamut, UHDTV = ultra-high-definition TV {>= 3840 x 2160}, SDR = standard dynamic range, SDTV = standard definition TV {<= 480 lines}, 3D = three-dimensional, HRTF = head related transfer functions, 5.1 = planar surround sound)

We are currently exploring how biosensors can be leveraged to help maintain creative intent over the wide capability of perceived signal ranges in the media ecosystem. These include bioinstrumentation that is currently available in consumer products, as well as those that are still limited to the lab, but anticipated to emerge in products in several years. One facet is that we have developed a metadata system for conveying creative intent that can be used by production teams. Directors, editors, colorists, sound designers and other creatives will be able to embed instructions in the content for manipulating or adaptively rendering the signal if the expected or hoped-for mental state of the audience does not occur, such as due to loss of perceived information in the signal. These mismatched states can include cognitive load, misplaced attention, lack of arousal and incorrect valence direction. A simple example of signal modification is adjusting the ratio of dialogue to foley and music

SPLs when a viewer's cognitive load exceeds expectations. Cognitive load can indicate difficulty understanding dialogue, which is commonly due to a lower quality playback sound system and/ or higher ambient noise level. Another example is zooming into an actor's face when the subtle expressions are not visible due to a smaller playback display FOV than the cinema's wide FOV that the creative intent was designed for. Other examples will be described as well as the details of the metadata system and its integration into the content timeline, and types of signal modifications allowed by the system. Lastly, theories of creative intent will be discussed, including quotes from artists that span the spectrum from those having a specific message and hoping to change minds to those who could care less about the audience.
(Acronyms: SPL = sound pressure level, FOV = field of view)

By rendering, we mean image and audio processing, where the image processing includes spatiotemporal, color, depth, cropping, and steering the image signal across multiple playback devices if needed. The audio processing includes positional (i.e., directional or spatial), equalization, reverberation, timbre, phase, loudspeaker selection, and volume. Both the image and audio processing can be nonlinear and adaptive.

The overall topic can be considered to be an application of *affective computing*, also known as artificial emotional intelligence (or emotion AI), which aims to develop systems and devices that can recognize, interpret, and simulate human emotion, understanding and behavior.

Creative Intent

To understand creative intent in an actionable structural framework, we can look to academia for theories of creative intent from the philosophy of art. At the most basic level, there is a spectrum which historically begins at Authorial Intent, which has become known a *Intentionalist***.** In this thinking, the aartist has a specific message, aesthetic, or emotional goal that the audience should get. This was espoused by Aristotle for Greek tragedies in 330 BCE, and generalized to literature by Tolstoy in 1897. Newer advocates of this philosophy include Farrell 2017 [7. The other end of the spectrum was referred to as Manifest Intent, now known as *Non-intentionalist*. In this philosophy, the aartist may not fully be aware of the intent. Reasons include that the intent comes from the subconscious, or the divine, and it is believed that the message and meaning can change over time as society changes. Advocates of this viewpoint include T.S. Eliot 1919, Beardsley 1946, Barthes 1967 and others. In this philosophy, one can't look to clues from the artist to ascertain the meaning, such as diaries, interviews, or portions of other works. For nearly a hundred years, this was the key academic divide and papers advocated one side or the other in more detail. Today's most common viewpoints are more of a middle ground, and most philosphers on this topic work on painting a more complete picture of the full spectrum. Examples of the newer or more advanced thought on creative intent include these theories: Contextualist Hermeneutic [8], Reader Response [9], Extreme Intentionalism [10], etc.

Information vs. Subjective Continuum

Rather than drawing solely from academia, we thought it important to ask practicing artists directly. We did this in the form of extracting ideas from interviews with artists, across all the fields, ranging from literature, visual art, theater, music, and cinema. We focused on sections of the interviews that discussed creative intent. We found about 25 quotes directed at this topic (a topic most artists don't prefer to discuss). While this was not enough for quantitative analysis, it became apparent that there is key *continuum of intents ranging from informational to subjective.* We use *informational* to indicate intent that aims for a specific response in the audience, whether it is an intellectual or emotional response. Some example quotes of artists at this end of the continuum include:

Art is a human activity consisting in this, that the artist consciously, by means of certain external signs, hands on to others feelings he has lived through, and that other people are infected by these feelings and also experience them.
-Tolstoy

I need to talk about Chinese culture. We have a deep strong philosophy and culture. I want to share some information, tell the worldwide audience
- Jet Li

I want to take the audience through a sequence of emotions.
-Anthony Doerr

The *subjective* end of the continuum is where the intent has no message or expected response of the audience, and it is fine with that. Some believe it is impossible to get a specific response out of the audience. Here are a few examples of quotes lying at the subjective end of the continuum:

Interviewer: Do people try to ascribe meaning to your movies?
Oh my God, I hope not.
-Robert Downey, Sr.

Sometimes songs are not what they are meant to mean, but rather what they need to mean to someone
– Bono (of U2)

The extreme end of this continuum is exemplified by the following quote:

The dream is to keep surprising yourself , never mind the audience
- Tom Hiddleston

While there are many artists who have worked across the entire continuum, whether for specific projects or a shift over the course of their career, those lying firmly at one end often fail to understand those at the other end. Common criticisms lobbed across the continuum include those at the subjective end using the deriding term *Didactic* to describe those who work at the informational end, while those informationally motivated artists often use the *Decadent* to describe work made without a message or expected response by those favoring subjectivism.

Variations on Creative Intent

Creative intent is one of several similar terms that have not been defined rigorously, also including *artistic intent, authorial intent, director's intent, producers' intent* and *approvers' intent*. Creative Intent (or creator's intent) is the most general term, and the most

vague. The term *artistic intent* arose from the world of painting, sculpture, and art philosophy and was originally used for a solo artist. Artistic Intent (or artist's intent) is mainly used in the visual arts sand explored by Beardsley, Clement Greenberg, Rosenburg, etc. Authorial Intent (or author's intent) is mainly used for literature with key thinking by Tolstoy, Beardsley, Hirsch, Carroll , etc., especially including the most thought put in on the Intentionalist vs. Nonintentionalist debates, including the key term Intentional Fallacy. Composer's Intent is used for Music and includes thinking from Huffman, Hegel, Adorno, D. Byrne, etc. One of the key aspects is Program vs. Absolute Music, as well as Period instruments' authenticity. The term that is particularly relevant to cinema, stage shows, and other content that requires team efforts is Director's Intent. For higher budget team projects involving companies, the term Approver's Intent is important (SMPTE glossary 2014, incomplete) where the approver could be the producer or a trusted expert. This term acknowledges the wide variability across businesses of who makes the final decision for determining the version that will be distributed. This includes editing, overall look, and sound of the media content, as well as variations intended for specific markets. For cinema and broadcasting business ecosystems, the term Rendering Intent is used for algorithms, display mapping and CG (computer graphics).

Cinema Creative Intent

Since we are mostly interested in creative content for cinema, our concern is with the details of the production stages, such as requiring interactions with the skilled technical and artistic staff. We will use the term 'creative intent' to describe the various goals of the media work from its creative staff. For narrative cinema, creative intent can comprise various elements, but most fall in one of the following categories:

a. **Information as needed by the story -** Is the viewer able to perceive the various information elements that make up the narrative ? Is it visible, is it audible and in the right location, & how does the viewing/hearing playback device and ambient surroundings affect these?
b. **Emotions of the story -** Is the viewer experiencing the emotions that are intended? How does color/ contrast/timbre/dynamics affect these? Both in terms of range and accuracy? We classify the lower-level responses/reflexes, such as due to jarring moments, as a type of emotion.
c. **Aesthetics -** Artists may choose a certain color on purpose, whether by whim, feeling, personal color harmony understanding, symbolism, etc. In most cases, we aim to match these. Are there other acceptable renderings that still meet intentions of (a.) above? For example, changing the musical key to accommodate a singer's range is often acceptable. Are there visual analogies? Asymmetry of aesthetics is an important issue. That is, it more accepted to reduce the color saturation on playback if the device is limited, but it is generally not accepted to boost the color saturation if the playback device had a larger range than that used to set the creative intent
d. **Message -** The combination of the storyline, evoked emotions, and aesthetics may be for the overall purpose of conveying a message. Of course, there are certainly cases where the content only intends to focus on any of these three aspects above, and is free of an overall message.
e. **Resumé -** Content can contain aspects of technical quality that may only be appreciated by experts . Such technical quality indicates level of budget of project. Portfolios of prior work enable creatives to work their way up to larger budgets, so technical quality is very important for content creator buy-in.

Of these five aspects of creative intent, this paper will be directed to the first two: *narrative* information and *emotional* effect. The third, aesthetics, will be covered by either the narrative info when deemed important to the story (e.g., symbolic colors) or as an emotion effect when the aesthetics are deemed critical to that.

Creative Intent Failures

Failures are not uncommon in art. In fact they are more common than not. Some of the causes are that the Artist is not skilled enough (yet; we give them the benefit of doubt to improve), the Artist misreads the audience (e.g., out of step with current society; see manifest intent), and Technology. Examples of failures by technology include work created on one technology, played back on another. For example, the capability is much better on the mastering technology than on the audiences' playback. Another reason is that the ambient conditions are much better during mastering than during playback, such as darker ambient allowing the display to have better performance, or quieter and with a better acoustic space with less reflections and frequency response coloration [ref- Storr] . In addition, there are now ttransmission step losses. In all these technological causes, significant 'signal' portions of the intent are lost due to the above.

To those who think that creative intent is a minor and negligible aspect of life, we give a classic example of intent failure involving substantial financial, legal, and political aspects. A musician from Asbury Park is renowned for rejuvenating folk concepts in the rock genre. One of their hits essentially was criticizing the lack of opportunity in the United States. Verses contained descriptions of having to work in chemical refineries near superfund sites with the threat of jail shadowing their efforts to escape that fate. Others verses criticized the Vietnam War (e.g., no GI Bill for vets). However, the chorus shifted the mood facetiously with a simple uplifting four-word phrase. Within several years, politicians began using this song in their rallies, and the crowd would sing this chorus pumping their fists in the air enthusiastically. However, the politicians using this song were of a strongly different political bent than the musician, as the song was being used in a jingoistic fashion. This upset the musician, and over time he learned he needed to have a lawyer issue a cease and desist letter, and follow up with court action if needed. To keep it simple, he applied these orders to any politician using his song. Over time he accumulated a team of lawyers to carry out this action as it was occurring across the entire country. In the year 2023, the musician is still active and it is known his tour has the most expensive tickets of the season, no doubt in part for paying the team of lawyers. The cause of this expensive intent failure lies not with the musician, nor the engineer who mixed the recording, but rather the technology differences between the studio and the rally arena. The PA systems combined with the poor acoustics of these venues (many multiple reflections with various time delays and reverb) caused the lyrics in the verses (critical of the country) to be unintelligible, while the simpler uplifting lyrics in the chorus were understood, and without the underpinnings of the lyrics, their ironic intent was lost.

That source recording was several decades ago, so one wonders if we still have such problems. It is not hard to find criticism on the web today about not understanding dialogue in movies, which has its source as a similar technology difference between the mastering technology and the large differences in playback technology across households. Likewise for imaging, the most common problem is visibility of dark detail. Current examples of creative intent failure for imagery include the dark scenes of Game of Thrones and its sequel House of the Dragon. To be more illustrative, I'll describe a particular scene from the 2018 movie Welcome to Marwen. In the movie, the main character played by Steve Carell had a brain injury and falls in love with a woman in his neighbourhood. Over time as she spends a a lot of time with him, he comes to believe the feeling is mutual. In a pivotal scene, he learns through their conversation that she was rather viewing him as a rescue dog/ Florence Nightingale relationship. As he realizes this, his loving optimistic face struggles to keep composure in the midst of this massive disappointment and ego landslide. The cinematographer placed Carell against an outdoor window so that he is backlit, and his face is washed amidst shadows, possibly so that the darkness symbolizes the state he is reduced to. I first saw this on an HDR TV in a dark room, and the facial expression transitions were subtle and sublime. I later rewatched this scene on an SDR TV in a bright ambient room, and his facial expression were clipped to perceptual black, entirely missing. While this could have been a valid different intent, the effort in getting the nuanced acting would have been a waste. .

Playback Technology Capability Differences

Spatial resolution

SDTV resolution at 640x480 up to UHDA 8k at 7860 x 4320. Resolution has a strong effect on visibility of higher frequencies, perception of sharpness and physiological effects is triggers.

Temporal resolution

Frame rates now range from traditional cinema at 24 FPS (frames per second) to gaming displays, currently at 360 FPS, with many options between. In addition, there is a wide range of temporal response across different display technologies, from the relatively slow IPS once having times of 20-50ms, to the very fast response times of OLED and uLED, which are less than 0.2 ms. The temporal responses affects sharpness for moving objects, and could be a key factor in ability to portray micro expressions.

Display dynamic range

SDR has a dynamic range of ~100:1 (2log10), with 100 nits max, consumer HDR ranges from 3log10 to 4log10 with maximum luminance of 600 -15000 nits nowadays, and professional HDR displays exceeds 4log10 dynamic range, with max luminances from 1200-5000 nits. The BT. 2100 spec [12] which is intended to be more future-proof extends up to 10,000 nits and minimum luminance level of 0. Different display technologies excel at the dark end, giving high contrast, while others excel at high brightness. The dark end capabilities allow for strong depth effects, as well as image frame border disappearance, which aids immersion. The high brightness technologies aid in realism of specular reflections and emissive sources, as well as in creating high APL scenes for proprioceptive realism (pupil sizes matching real scene behaviour).

Color gamut and volume

Three key color gamuts achievable today are 709, as codified in the sRGB computer display spec, P3 which is larger and nearly encompasses all the color gamut of cinema film, and the 2020 color gamut in the BT 2100 spec, which exceeds almost all current display capabilities, and generally requires laser primaries.

Depth

There is now a very wide range of depth options that can be transmitted and displayed, ranging from traditional flat screen 2D, but extending in depth capability to curved screen 2D, glasses-based 3D (shutter, polarization, and interference filters technologies). Glasses-free AS3D (autostereoscopic) using lenticular screen layers reached a commercial peak in 2010, but is on a comeback due to availability of 8K display panels allows for more views and smooth viewpoint transitions (at the slight expense of spatial resolution). In addition, there is now VR HMDs, which offer a range of depth-environment interactions going from seated 3DOF (Degrees of freedom), 3DOF+, Windowed 6DOF, Restricted Omni 6DOF, and full DOF.

FOV

Field of View (FOV) can have substantial effect on creative intent goals and opportunities, as is well known to the action-field expressionist painters who use very large cases to affect the viewers' light adaptation as well as overall general impact. For digital television, the FOV ranges from 15 degrees for SDTV when viewed at 6 picture heights (6H), 35 degrees for HDTV when viewed at 3H, 65 degrees for 4k UHDTV when viewed at 1.6H, and the very immersive 100 deg when viewing 8k UHDTV at 0.75H. In all these cases listed, the optimal viewing distance is set by mapping the display Nyquist (max spatial frequency) to approximately 30 cy/deg, the presumed visual system cut-off frequency for adaptation luminance levels in the SDR range. In addition, we have the traditional wide FOV format of cinema (>100 degrees for front row seating, typically at 0.5H), including its variations such as Cinerama which exceeded 180 degrees. The newer VR technology using HMDs allows for FOV of 100 to approaching 180 degrees. CAVE technologies for VR can allow 360 vertically and horizontally.

Audio DR and accuracy

In audio, the playback device capabilities have long been described by frequency rage, frequency response, and dynamic range (in dB; not to be confused with image DR). Speaker cost and size limitations are key factors to achieve dynamic range. Dynamic range has long been known to have an effect on creative intent, as pop music is mastered to have a small dynamic range to match the limited dynamic range possible in an automobile (of that era), while classical music, opera, progressive rock, and some jazz and musicals are mastered to have a large dynamic range, allowing for quiet sections to contrast against loud passages in the composition. This type of mastering is expecting the playback system to have

substantial dynamic range and quietness (e.g., headphones). Details involve phase accuracy, coding losses such as overall sampling rate, baseband bit-depth, and codec. In addition the listening environment ambient conditions play a strong role in the perception of audio signals, such as white or pink noise, dense transient noise or sparse noise.

Sound field

For the last several decades, the audio options ranged from mono to stereo to surround, with mono almost disappearing from the consumer ecosystem. But now with audio interface portals (Amazon Echo), the proportion of mono listening has increased. In addition, there is a common behaviour of listening to earbuds, but with placement in only one ear. This results is what is referred to as half-stereo, as the signal is still typically stereo, and one of the two channels is simply lost. These are examples of reduced capability than stereo. Examples of increased capability include surround sound which keeps the sound field in a single plane, typically horizontal, of which 5.1 and 7.1 are the most common. Newer formats add vertical speakers and expand into a volumetric space. These advancements in 3D sound include Ambisonics, Atmos, and Binaural (which is the version of 3D sound that works for headphones or earbuds).

Technology Capability and Creative Intent

The two main factors here are expressive gamut and audience reach. Nowadays a creative must decide how to render the content, in consideration of the various formats and display capabilities mentioned in this section. On the one hand, the improved capabilities allow for more expression, and more dramatic effect on the audience which can trigger immersion, emotion, as well as cognitive effect. It is known that the highly immersive VR capability causes emotional responses and memories that are stronger than can be elicited on a SDTV 2D screen. On the other hand, the creative needs to consider the reach of the technology, whether that reach is motivated by message or financial reasons. There is an optimum technology for reach, but that shifts in time. For example, designing content for SDR SDTV has the widest reach currently, but that is quickly shifting to higher DR and resolutions. Drawing from the past for examples, there was a point in time where analog VHS was the technology that offered the widest reach, but it would be hard to find an audience now that would watch such a low quality. Going further back, achromatic movies (i.e., black and white, B&W) had a wider reach than color movies at one point in time, but it is typically hard to find significant audiences who will watch B&W movies today, with the exception of cineastes and historians. We are at a stage where the technology capability is not expected to condense to a single format again, as the modes of viewing can be so different (viewing on phone, on a laptop, on a large TV, in a cinema, in VR, or entoptically [23]) and not expected to change.

Emotions

There are several taxonomies on emotion, ranging from a small number such as the six from Elkman theory [13], to others containing more nuances and including almost thirty different emotions. Some of these emotions have corresponding facial expressions, while others involve deeper internal feelings without visible signs. It is important to consider the emotions having facial expressions as a distinct set, since those can be the most easily assessed. For example, it may only require a camera pointed at the audience (or solo viewer) and expression software to obtain estimates of those emotions. Table 1 below shows four key taxonomies as well as the subset that can be determined from facial expressions [14, 15, 16, 17].

Other familiar emotions not cited by these specific theories include: vigilance, grief, rage, loathing, ecstasy. These may be approximately mapped to corresponding synonyms on list: vigilance is a stronger form of interest, grief is a stronger form of sadness, rage is a stronger form of anger, loathing is an extrapolated contempt, and ecstasy is the extrapolation of romance or sexual desire or amazement.

Immersiveness is not on any of these lists, but is very important to media ecosystems and playback devices. It is usually assumed to mean that the viewer feels placed in the world of the story, and techniques to achieve this require more realistic imagery and sound capabilities, such as wider field of views (FOV), wider color gamut, increased dynamic range, higher bit-precision, higher fidelity positionalized sound. Another viewpoint is that these improvements are mainly to avoid the distractions of lower capabilities, for example, constantly seeing the image border from a narrow FOV presentation on a small screen display. Technologically-achieved immersiveness and viewer engagement generally go hand-in-hand, but there are exceptions (being immersed in a compelling book, or bored in a derivative VR game). In this work, we look at immersiveness as a magnifier of all the possible emotions, so we don't try to measure it directly.

Another approach at modelling emotion is to have just two Cartesian axes (arousal and valence), where arousal is essentially a magnitude of intensity, and valence is whether the feeling is positive or negative. This type of model is important because some of the physiological measurements can only identify arousal and valence levels. Galvanic Skin Response (GSR) is an example where only these two can be assessed. For the arousal/valence approach to emotions, there is a standardized model called IAPS (interactive affective picture system) [18]. Both types of emotional representation are used in this system.

It is known that the effects of image processing, including the passive effects on images cause by resolution loss or low amplitude detail loss, can alter the perception of facial expressions [19,20, 21]. Viewing distance and the extended FOV of the face also can affect expressions, as well.

Biosensors (Physiological Assessment Technology)

There is now a host of technologies that can be used to assess the emotion and cognitive state of the viewer. Some can be incorporated directly on the playback display, while some may require auxiliary devices, such as smart earbuds or a smartwatch. In applications with large audiences, such as a cinema, assessment technologies range from a camera facing the audience [22], to sensors placed in the seat, to measurements of the overall theater such as gas content or temperature [23]. These will be discussed in

more detail, later in Figure 4 & 5, for individual and group physiological assessment.

Some of the current technologies for assessing a viewer's internal state include eye gaze via electro-oculogram (EOG) [24,25], cognitive state via electroencephalogram EEG [26,27], and auditory attention via EEG [28] . Some of these will be discussed in more detail in the description of the system.

It is useful to break the physiological assessments into two categories, cognition and emotion, as these map to the two distinct aspects of creative intent in the list above (info for story, emotion of story). Cognition includes cognitive load, which indicates the viewer is struggling to comprehend something that may be important to the storyline. The internal state of attention is now becoming relatively straightforward to measure through eye trackers, and mapping the gaze position onto the content. Eye trackers may be built into the display (TV, mobile display, and computer monitors) and they are now considered mandatory for state of the art VR and AR. Engagement is an important internal state that has both emotional and cognitive aspects. It can be assessed with EEG, such as with the P300 evoked potential response (where reduction indicates more engagement). Rather than considering it as emotion, we consider it here in the framework of metadata to be related to attention, and thus a subset of cognition.

Emotions can be read from imagery of the viewers face, including both visible-light and thermal cameras. As of this writing there are several systems and algorithms for assessing the visible-light facial expression [29]. Readings of emotion from thermal imagers allows for deeper understanding of the internal state than possible with visible imaging, which may be masked by a 'poker face'. We are unaware of products offering this, but research has been published on this topic [30,31].

The primary way to assess non-visible emotions requires EEG (electroencephalography), and this technology is advancing from the era of requiring a skullcap of dozens of electrodes, to just a handful of electrodes touching the head at a few places. These include a headband, over-the-ear headphones (cans) or part of a hat. For VR applications, there are already systems that build a multi-sensor EEG system into the HMD [32, 33]. More innocuous ways to access the EEG are being developed that use electrodes placed in smart earbuds.

As mentioned, some of the technologies only allow for readings of arousal & valence, such as GSR, which is also known as ectodermal activity (EDA), skin conductance, electrodermal response (EDR), psychogalvanic reflex (PGR), skin conductance response (SCR), sympathetic skin response (SSR) and skin conductance level (SCL). Heart-rate and respiration monitoring are other examples that can only assess arousal levels.

Image and Video rendering under metadata control

The concept of processing video corresponding to metadata inserted into the content are becoming established, with examples being Dolby Vision, Samsung's HDR10+, and Technicolor Advanced HDR [34]. In all these cases, the metadata from the content is used in conjunction with metadata from the playback display to alter dynamic range, color saturation, hue angle, and spatial filtering.

Metadata System for Creative Intent Scalability via Physiological Monitoring

This system is intended for an audience ranging from a single viewer to a large group in a theater. Some applications may involve only audio, but we won't call out distinctions between 'viewers' and 'listeners' in this description, and refer to all cases as 'viewers'. Figure 0 shows a viewer with possible physiological monitoring sensors embedded in portions of the playback system, as well as on an accompanying smartwatch. We won't illustrate the monitoring devices for the large group applications, but they will be described later.

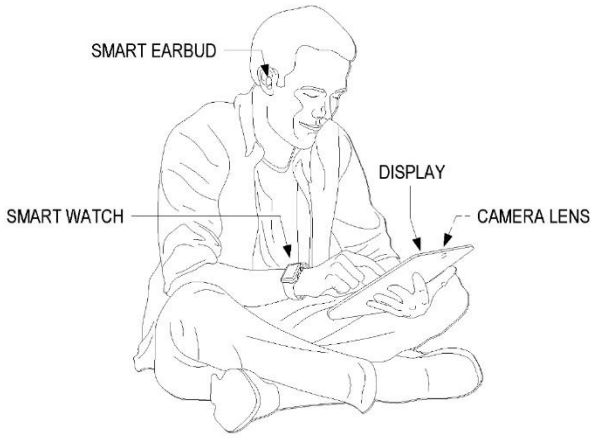

Figure 0 : solo audience watching tablet with key physiological monitoring locations

Metadata inserted in the production stages of content creation is essential to this system, and there are two key types: emotional and narrative. For the metadata describing the intended emotions, we include all 27 emotions as listed by Cowan's theory (from Table 1), but it is important to note that technologies that extract facial expressions are limited to the nine in the rightmost column of table 1. The others can only be estimated from EEG and combinations of others such as thermal and GSR. Not all emotions are required to be used by the creatives, and not all emotions will be able to be estimated by some of the playback devices.

In the technologies associated with this system, there tend to be terms that used by creatives, and terms that are used by experts in the cognitive or neurosciences. Often these terms can be synonymous or have substantial overlap in their meaning. The specific terms for each field have advantages in that they have more specificity than more colloquial usage. We try to use terms that are most appropriate to those interacting with each particular portion of the system. That is, for steps involving the insertion of metadata, as would be done by creatives in the production stages, we use terms that would be more familiar to them, while for terms that involve processing of the physiological signals, we use terms that are more appropriate to neuroscience. An example is cognitive state and confusion estimate. 'Confusion' is a more appropriate term to use with creatives, while neuroscientists would use the term 'cognitive load' for describing the level of confusion. Cognitive load has additional specificity as it includes gradations from very stressed confusion to mental states simply requiring attention.

Figure 1 shows the overall system, with two key stages: the steps in the production process, and the steps that would occur in the consumption process, that is, when the audience is watching the content. The upper half of the diagram are the steps that occur during production. Since storyboards contain useful storyline info and often emotional expectations, they can be used in the system to help populate the metadata when available. Digital storyboards can be processed to extract key info like chief characters, regions of interest, storyline connectivity, etc. These are optional, since they are not always available. The audiovisual content timeline is shown as an arrow, and while this is being produced, metadata describing the emotions and narrative key points can be inserted by those in the production staff, i.e., the creatives. The small arrows indicate these steps are repeated throughout the content's timeline. The metadata input by humans are then consolidated and formatted, and finally bound to the content.

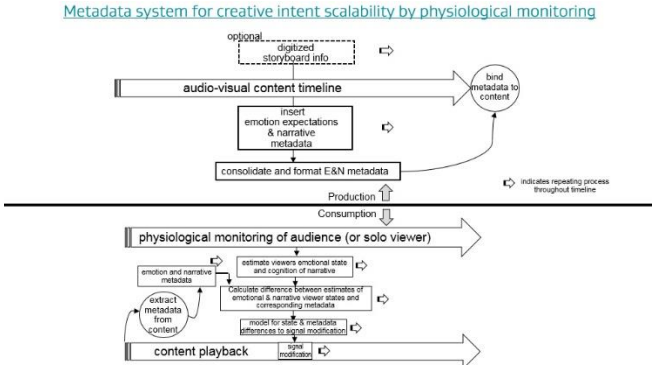

Figure 1 : simplified summary of the overall system

The bottom half of the figure describes the steps occurring during playback. There are two key time-dependent processes. One of these is the physiological monitoring of the audience, which is ideally continuous, but may be either finely or coarsely sampled, depending on the components. The other is the playback and modification of the content. The modification is accomplished by using the estimates of the viewer's emotional state, as well as cognitive state. The cognitive state is used to assess effectiveness of the narrative info. Both arise from the physiological monitoring and processing of those signals. At each given point in the content timeline, the corresponding emotional and narrative (abbreviated as E & N in subsequent diagrams, for brevity) metadata are extracted and compared to the estimates from the physiological monitoring. Next, the difference between the viewers' actual emotional state and the intended state as extracted from the metadata are calculated. Similar differences are made for the narrative metadata. We go into more detail on what kind of info are in the narrative metadata later. The calculated differences are then input to a model whose output describes what signal modifications should be applied to the content to reduce the ineffectiveness of the content with respect to its creative intent as described in the metadata. Lastly, these signal modifications are applied to the content, as these various steps are repeated during the timeline.

Attention can be considered a subset of cognition, so in some of the diagrams we collapse attention-based processes into the cognition processing blocks, and refer these to the creative intent aspect (narrative state). In others diagrams we specifically call out attention as a separate process with its own block processes.

Figure 2 goes into more detail, and for production stages, it breaks out individual storyboard pages, content timelines, edits, and various key moment and arcs in the story. For example, there may be more edits tan storyboard pages, soothes e do not necessarily need to align . In addition, metadata may only be inserted during key scenes between the edits.

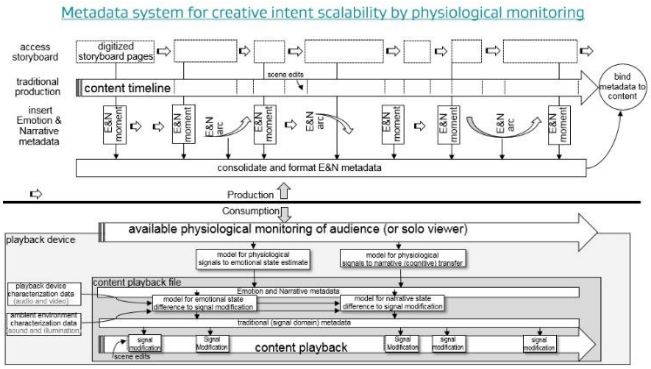

Figure 2 : more detailed version of the overall system

The consumption side of the diagram in Fig 2 shows that the physiological monitoring may be a part of the playback device (indicated as a light gray background box), or it may be standalone, and just transmitting the info to the playback device. Blocks show the models used to convert the raw physiological signals to models of the emotional state, as well as the cognitive state of the viewer which relates to the narrative metadata. The content playback file is shown as a darker gray, consisting of the content, the narrative metadata, and the emotion metadata. Processes that use the data from this file, and then in turn the modify the content are shown as blocks with drop shadows, to distinguish that they are acting on, but not part of the content file. These steps include the models for how to change the content based on the differences from the intentions as decoded in the metadata, as well as the actual signal modification. On the left side of the lower diagram, there are blocks that describe the static metadata of the playback device, such as its dynamic range, color gamut, number of speakers or positional rendering capability. In the bottom content playback timeline, the signal modifications are shown, which are generally held constant between edits, but not necessarily.

Figure 3 shows the basic format of the metadata, and how it is inserted during content production, and utilized at playback (content consumption). It is broken down into fields for emotion and fields for narrative info. Each of these are further subdivide into the expected states and intended modifications. The narrative metadata is not necessarily fully semantic, but rather aims for key low-level aspects, such as image regions of interest (per frame, or tracked across the scene), audio objects of interest, and a confusion index. The confusion index is expected to be sparsely used, but inserted when critical storyline info must be understood. It is needed because sometimes confusion intended, such as a chaotic action scene. As is common for metadata used in media content, it is only present when needed, and objects associated with metadata need not persist if they are not used. The metadata is envisioned to be inserted at edit junctions, and persist across the frames until the next edit junction.

Flags are available for continuation of metadata to avoid the overhead bits of repeating metadata per frame. However, the metadata edit junctions can be at the frame resolution if needed.

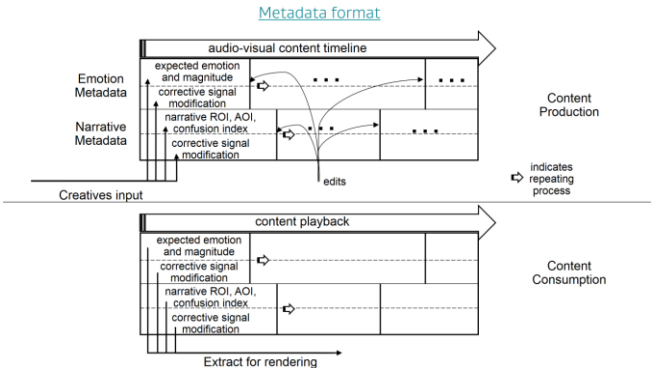

Figure 3 : Format of metadata used to describe the expected emotional state, and key narrative info

Figure 4 shows the physiological monitoring and processing for a solo audience. This includes the kinds of sensors that can be placed on handheld displays, earbuds, and smartwatches. TVs may also be used, and the simplest case is when there is just a single viewer. Each of these locations afford a certain array of sensing, which is illustrated in the figure. The playback device will typically consist of display and sound source, which may be wirelessly connected. Smartwatches are currently not considered part of the playback device, so they can be considered as auxiliary components. In Figure 1, there is a block shown for 'estimate viewers emotional state and cognition of narrative' and in Figure 2 there are blocks for 'model for physiological signals to emotional state estimate' and 'model for physiological signals to cognitive transfer'. These steps are shown in the pentagon in the figure, and described as sensor fusion and segregation. The output of this fusion and segregation state is a consolidated model of the emotional state, and narrative transfer state. The narrative transfer state includes cognitive load and also where the viewer is paying attention.

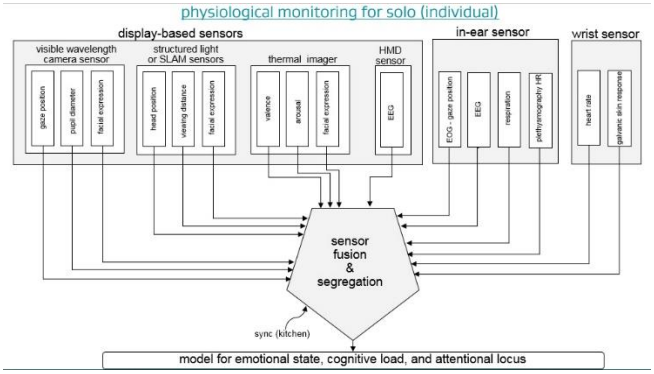

Figure 4 : various physiological sensors and their consolidation into a representation of the viewer's state, for applications with a solo viewer

Figure 5 shows the physiological monitoring for group audiences. These could be large audiences in a theater, as well as smaller groups in a home. These sensors are located in the overall room, and in the seats. The sensor fusion and segregation consolidation step is also shown on the pentagon block.

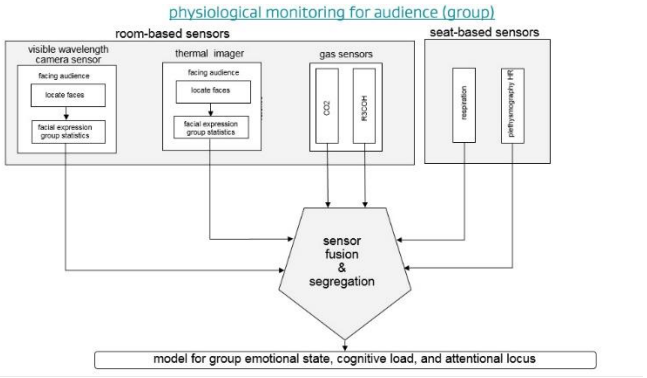

Figure 5 : various physiological sensors and their consolidation into a representation of the viewer's state, for applications with a group of viewers.

Figure 6 goes into more detail of the sensor fusion and segregation for the solo audience application. There would be a similar diagram for the group viewing application, which would only differ in details. The various physiological signals from the different sensors shown in figure 4 convey both info the viewers emotional state and ongoing success of narrative transfer. In this diagram, we have separated the two key elements of narrative transfer assessment, which is the cognitive load and the attentional locus (to what the viewer is paying attention to). Sensors coming from a particular component of the playback device can contribute to both the emotional state, cognitive load, and attentional locus. In addition, there may be duplication from differing sensors on a given state estimate, such as eye gaze position via a display-based camera as well as from the EOG signal from an earbud. These multiple signals must be consolidated, as shown.

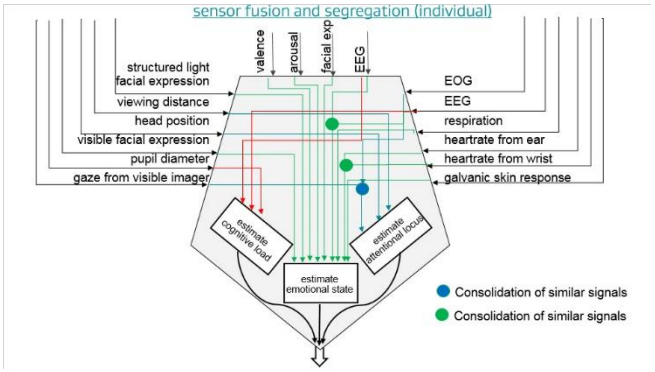

Figure 6 : Details of the use of physiological signals to determine an estimate of the cognitive load, the emotional state, and the attentional locus

Table 2 is another way to list the various physiological signals in terms of their physical location, the type of sensor, and what the signals can be used to estimate. The table conveys the same info as Figure 6, but may be easier for some readers to digest than tracing the signal lines in the block diagrams. It is for the single viewer case,

and Table 3 is similar, but describes the sensors and usage for the group viewing applications.

Physiological Signal	Source Location	Source Sensor	Estimator using Signal
Gaze position	Display (e.g., phone, TV, tablet)	Visible wavelength camera	Attentional locus
Pupil diameter	Display	Visible wavelength camera	Attentional locus & Cognitive load
Facial expression	Display	Visible wavelength camera	Emotional state
Head position	Display	Structured light or SLAM	Cognitive load (& vision thresholds)
Viewing distance	Display	Structured light or SLAM	Cognitive load (& vision thresholds)
Facial expression	Display	Structured light or SLAM	Emotional state
Valence	Display	Thermal camera	Emotional state
Arousal	Display	Thermal camera	Emotional state
EEG	Display	HMD sensors	Emotional state & Cognitive load
EOG gaze position	In-ear (e.g., smart earbud)	Earbud ERG dipole electrode	Attentional locus
EEG	In-ear	Earbud dipole electrode	Emotional state & Cognitive load
Respiration	In-ear	Earbud microphone or accelerometer	Emotional state
Heart rate (Plethysmography)	In-ear	Earbud microphone, accelerometer, passive infra-red (PIR)	Emotional state
Heart-rate	Wrist (e.g., smartwatch)	PPG (photo sensor)	Emotional state
Galvanic skin response	Wrist	Skin conductance sensor	Emotional state- Arousal

Table 2 : Physiological signals associated with individual viewers and for which internal state they can estimate

Audience (group) signals and usage

Physiological signal	Source location	Source sensor	Estimator using signal
Facial expression group stats	Room	Visible camera	Emotional state
Facial expression group stats	Room	Thermal camera	Emotional state
CO2	Room	Gas sensor	Emotional state
R3COH	Room	Gas sensor	Attentional locus
Respiration	Seat	??	Emotional state
Heart rate	Seat	??	Emotional state

Table 3 : Physiological signals associated with group audiences and for which internal state they can estimate

Example cases

The remaining figures go through examples of emotional, cognitive, and attentional data and corresponding signal modification. Further development would be best achieved by collaboration with creatives on the production staffs of studios and other media production facilities.

Figure 7 shows the general processing for emotional state and corrsponding metadata. The emotion metadata consisting of an expected state and signal modifciation options is shown in the upper left block. Physiological signals of a facial emotion estimate and an EEG emotion estimate are consolidated into a single emotional state estimate. This is compared with expected state (part of the creative intent) from the metadata, and a difference is formed . This difference is fed into the model for signal modifciation based on the state difference, along with the possible options from the metadata. The model determines the magnitude of specific paramters of the signal modifciation changes, based on the magnitude of the state difference. Other inputs to this block may include narrative metadata, such as the image region of interest (ROI) and the audio object of interest (AOI). From these inputs, the parameters of the signal modifications are determined, and it is used to modify the actual content being played back (either through image or audio processing, or both).

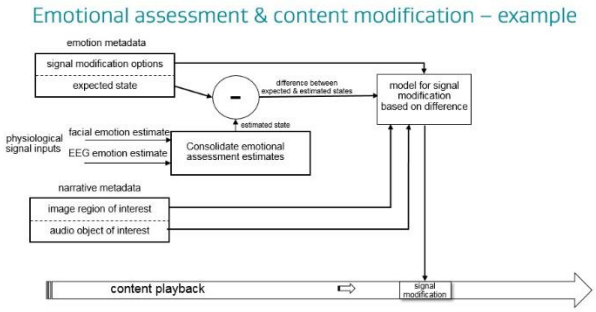

Figure 7: Determining signal modification based on the emotional metadata and viewers' physological data

Figure 8 describes a specific example of the processes in Figure 7. The content example is a critical scene where the central character may be saying one thing, but their facial expression belies a different emotion. The viewer is watching on a mobile display held at such a distance that the field of view (FOV) is small. As a result, the subtle facial expressions cannot be seen due to perceptual resolution limits (i.e., the pixel Nyquist frequency exceeds the visual cutoff frequency). The signal modification metadata indicates the that zooming into or out of a specific region-of-interest (ROI) is the suggested option. The two physiological input signals give somewhat conflicting info. The camera-based facial expression estimates the viewer is calm, while the EEG emotion estimate is that there is the emotion of interest. These two signals are consolidated to output a signal gradation along the neutral-to -interest emotional vector that is smaller than intended, and that difference is input to the signal modification model. Meanwhile, the narrative metadata has info on the image ROI The metadata of signal modification for a specific emotional state difference includes the image ROI, which is the pixel locations of the character's face. It also has info on the relative ranking of the audio-object-of-interest (AOI)s, but in this case that data is unused. The model for signal modification takes the magnitude of the emotion difference, the narrative ROI data, and the signal mod data of zooming into the ROI, to calculate the emotion state of the viewer can be increased (according to intent) by zooming into the character's face. This information is used for the specific signal modification, which would be to zoom into the pixel position centered at I(x,y***).***

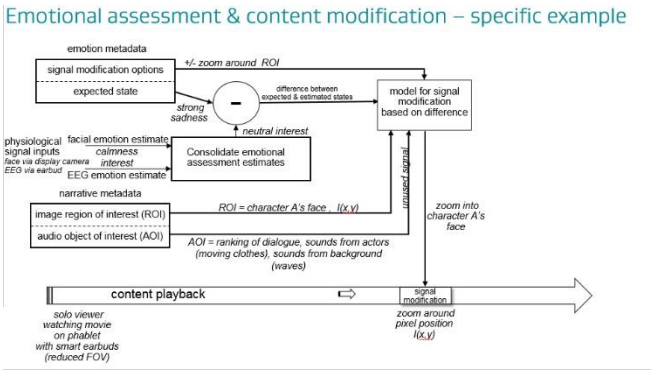

Figure 8: Specific example of determining signal modifciation based on the emotion metadata and viewers' physological data.

The next example, shown in Figure 9, is for a narrative metadata change, specifically, the cognitive portion. In this case, the narrative metadata consists of the signal modification options and a confusion index. In general, it may also include narrative ROIs and AOIs but they are not used in this example. In this case, the relevant physiological signals include a pupil diameter estimate coming from the eye tracker in the display-sited camera, as well as an EEG signal coming from a smart earbud. These are both used to estimate the cognitive state, which is compared against the expected cognitive state from the narrative metadata (i.e. confusion index). Like the case for emotional metadata, the differences in expected viewer state is then input to the model for determining the signal modification specifics. That model also takes input from the emotional metadata, but those are secondary and minor contributors. The computed overall signal modification is applied to the content for rendering modification, as shown in the content at the very bottom of the figure.

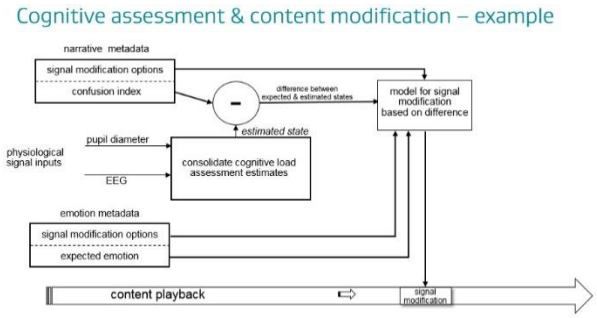

Figure 9: Determining and performing signal modification based on narrative metadata and viewer's physiological data

Figure 10 shows a specific example for the process described in Figure 9. The scene contains a character explaining something critical to the story in a scene with a lot of auxiliary sounds. The viewer is watching on a tablet using smart earbuds, but the environment is noisy enough that the earbuds do not sufficiently block the external sounds. Consequently, they are missing parts of the dialogue that are critical to the storyline. The specific signal modification metadata indicates that increasing the volume of the speaking voices should be done if the confusion index is high. The confusion index is set to zero since it is an important dialogue scene, the creatives desire the viewer have complete understanding. Note, that in most cases the confusion index would default to zero, but there would be certain scenes where it may be set for a higher value, such as in scenes that are meant to be overwhelming in complexity (e.g., action scenes, political drama of many arguing voices, etc.) The physiological signal of pupil diameter indicates stress and the EEG signal indicates cognitive load, so the output of the consolidation of cognitive load indicates viewer confusion. Since the viewer's confusion is higher than the confusion index, the difference signal input to the signal modification model modulates the increase in dialogue volume relative to the other audio objects of the soundtrack. In this specific example, there are emotional, metadata with flags set for the emotion of anticipation, and compensation steps of increasing the image contrast. However, they are not used in this instance due to the available physiological signals, and thus do not affect the resultant signal modification. The output from the signal modification model is used to change the ratio of the dialogue volume objects over those audio objects corresponding to Foley sounds and background music.

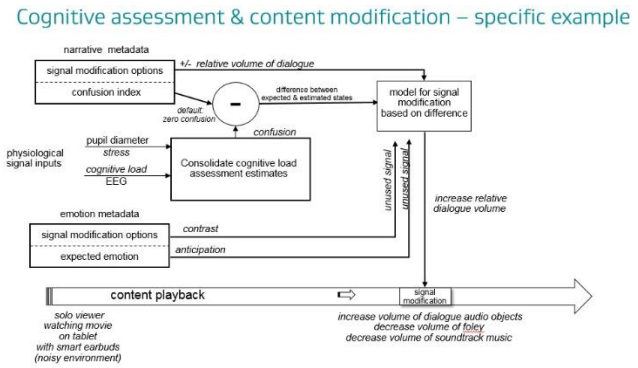

Figure 10: Specific example of determining signal modification based on the narrative metadata and viewers' physological data

In the last general example, Figure 11 shows another case for the use of narrative metadata, but where the specific process is viewer attention. The narrative metadata fields for assessed viewer state indicate specific image ROIs and AOIs. There are two physiological signals describing gaze position, as mapped to the content image. In this case, there is no emotional metadata inserted by the creatives. The two gaze positions are consolidated into a single gaze position, and that is compared with the intended image ROI from the narrative metadata. For some reason, the viewer is fixating a non-essential portion of the image, and thus the consolidated gaze position results in a difference when compared to the ROI metadata. The difference is used to control the signal modification which is intended to shift the viewers gaze back toward the ROI.

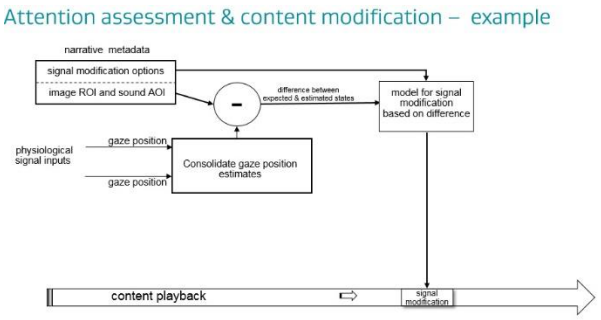

Figure 11: Determining and performing modifications for the attention aspect of narrative metadata

Figure 12 is one specific example for Fig 11. The narrative metadata for compensation of the gaze position and ROI mismatch is to apply a localized sharpening filter centered at the ROI. The physiological signals providing gaze estimations include the eye tracker in the display-based camera, and an output from the EOG which is located on the smart ear bud. The magnitude of the difference is used to control the strength, the spread, and the feathering (gradation) of the localized sharpening filter.

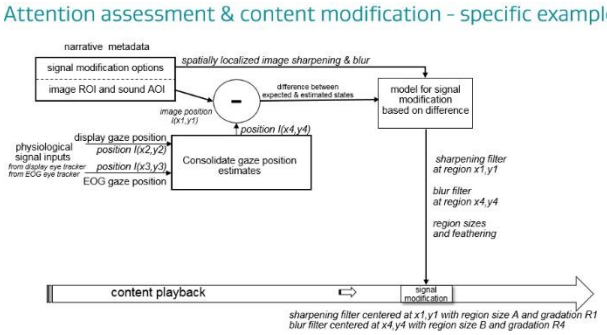

Figure 12: specific example for the attention aspect of narrative metadata

Figure 13 is another specific example for usage of attentional metadata, as part of narrative metadata. The viewer is watching in a home theater with a full immersive sound system (e.g., Atmos), plus a large modern display (105") that uses standing glass vibration to have sound emanate directly from the screen, with a 3x3 positional grid resolution (e.g., Crystal Sound technology).

In a scene from a movie on Newton's experimentations with alchemy, he is exploring the vegetation of metal. The candlelit scene in a cathedral late at night shows sprawled across the marble floor a complex crystalline silver texture, all in motion, with accompanying metallic crinkling sounds. One portion is changing shape from crystalline to biomorphic dendritic shapes, and the corresponding sounds from that activity changes to more of fluidic pitch-bending having subtle human voice undertones (implying the 'vital spirit' he was seeking). These are localized to the image region of the dendritic growth. Before the camera slowly zooms into the anomalous region to eventually show a convex reflection of Newton's entranced face, it only occupies a small part of the screen image, and may easily be overlooked. Since the display is so large, the viewer is looking in the general neighborhood of the dendritic region, but is slightly off so the region falls just outside the viewer's perifovea, where the distinction between the crystalline and more biomorphic textures cannot be distinguished. So, the viewer's attention needs to be guided to the mysterious growing dendritic region that looks alive.

The same physiological signals are used as in Figure 12, as well as the ROI and AOI metadata in the 'state' portion of the metadata field. However, in this case, the creative has decided that the signal modification used to redirect attention would be audio processing. Specifically, they have chosen to have a low tolerance snap option in immersive audio processing, which is to favor the use of single speaker. This is known to better preserve the timbre aspects. It is part of SMPTE ST 2098, where there is a metadata field for describing whether the timbre or position is more important. The terms 'snap' means to snap a sound position to nearest positioned speaker, since that is known to best preserve timbre. In this case, the use of a snap setting would place the sound into one of the nine positions on the screen (from the 3x3 sound grid of the glass panel speaker). That discretization would cause the sound position to be mismatched from its actual place on the screen. Because the snap (or timbre distortion) *tolerance* has been set high by the creatives, the audio processing favors using speaker interpolation to more exactly place the sound on exact screen position, at the expense of timbre distortion. In addition, since reverb also causes sound position diffusion, it is decreased from its default setting which was high due to the cathedral setting.

This would be contrasted to someone viewing the same scene on a small display, where most of the image is falling in the perifovea anyway, and thus the dendritic shapes would be noticed without having to resort to the advanced audio compensation processing.

Note: The audio space representation, A(x,y,z), needs to be converted to the image space representation of I(x,y,z) , where I(x,y,z) = A(x,z,y). That is depth from screen is z in the image space representation, while it is y in the audio space.

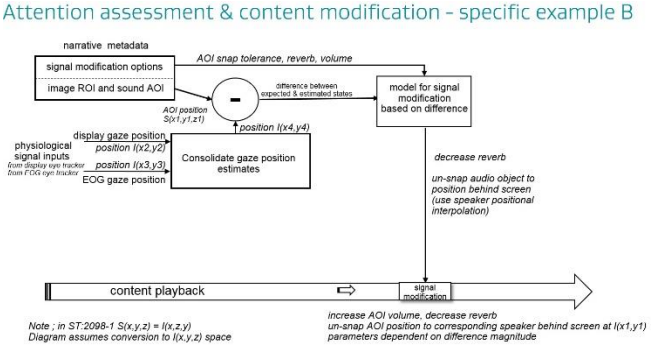

Figure 13: another specific example for the attention aspect of narrative metadata

As mentioned, many other examples can be devised, with similar processing, but differing in the specific emotions described by the metadata, as well as signal modification options.

Example case – Timeline and Adjustment Levels

By way of example of the timeline aspects, a content creator, i.e., a cinema director (aka Director), establishes their creative intent by mapping their desired emotional and/or narrative viewer states, like arousal, over the duration of the entire movie which for our purposes here is represented as a thin solid line Fig 14. The horizontal axis is the movie timeline, and the vertical axis indicates the viewer's emotional and narrative state (E&N), in this case, the emotional state of arousal. It also indicates the signal level which an assumed trigger factor toward the specific E&N state, The viewer's actual assessed emotional and/or narrative responses i.e., arousal in this case, over the duration of the entire movie is represented in a thick solid line Fig 14. As the viewer deviates from the expected state(s), while taking several factors into account including the rendering environment, audience state metadata is generated and used to generate corresponding adjustments to re-align the viewer's arousal with the intended emotional and or narrative viewer state metadata , i.e., arousal state in this case. To simplify this on Fig 14 we have limited the signal adjustments to Luminance, however, the possible adjustments span both audio and video characteristics.

At some time points or time intervals (or some entire scenes), the viewer may be expected to be more excited, whereas at some other time points or time intervals (or some other scenes), the viewer may be expected to be less excited, even subdued or quiet. An example would be to warm up or prepare for a massive shock or an elevation of interest or emotional arousal. Similarly, at some time points or time intervals (or some scenes), the viewer may be expected to be more engaged, whereas at some other time points or time intervals (or some other scenes), the viewer may be expected to be less engaged, even relaxed. But first, let look at some options dealing with an over-stimulated viewer.

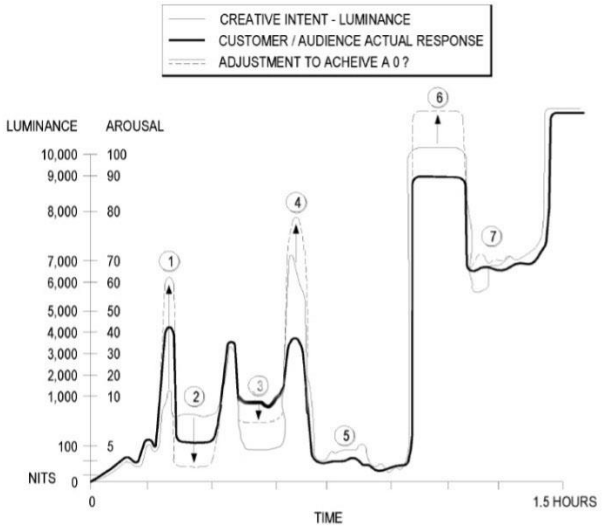

Figure 14: Timeline example of expectation for local luminance-triggered arousal and adjustment for a 90-minute movie

Over-stimulated audience. Looking at the time point corresponding to the circle 3 in Figure 14, in response to determining that the viewer's assessed state(s) such as assessed arousal as estimated by physiological monitoring (as indicated in the thin solid line of Figure 14) is over-responsive as compared with the viewer's expected state. Based on predetermined input embedded or passed to the signal real-time from the Director, the playback device can apply a signal modification (dashed line) to cause the viewer's assessed state or arousal to move toward the viewer's expected state or arousal (toward achieving a zero difference or *0 L1),* or to become less aroused. It is key to note that the Director has discretion to adjust many variables and setup contingencies and second order contingencies. These would include adjustments to amounts, modes, assessments' priority, among others. Other tools create a threshold that if breached halts further adjustment(s) regardless of system, in this example, luminance capabilities. These tools address factors which come into play when addressing an ecosystem with various tiers of playback devices with various capabilities and technologies and other factors such as a limited set of processing resources or even a system-determination that for this viewer, additional adjustments would not be fruitful. While the Director oversees this activity, the actual work would likely be carried out by an editor, colorists, or possibly new technician/craftsperson specialists. In the time point at circle 3 the Director has chosen to set a threshold in this range of time that is less than a zero difference or 0L1 because she deemed this section to only require a 50% or "half-way there" response, or colloquially, "the scene is more informational rather than emotional, only becoming non-effective in a viewer when they are outside the threshold.

Under-stimulated audience. At the time point corresponding to the circle 4 in Figure 14 it is determined that the viewer's assessed state is under-responsive as compared with the viewer's expected state. The playback device can apply a luminance adjustment/modification (or other signal modification) to change or raise an original or even a pre-adjusted APL (Average Picture Luminance), peak brightness, black point, contrast, or several other pre-assigned signal ranges of the scene or within regions of interest within an image. The first adjusted signal as raised from the original or pre-adjusted APL may be used to cause the viewer's assessed state or arousal to move toward the viewer's expected state or arousal (toward achieving a zero difference or *0 L1),* or to become more aroused. Additionally or optionally, the media content adjustments/modifications may be generated based at least in part on models similar to those used in translating, converting and/or implementing the viewer's specific expected emotional and/or narrative states in the production stage.

Expected Audience Response- No Adjustments Needed. Shift your attention at time points 5 and then 7 in Figure 14. In these two cases no adjustment is made by the playback device in response to determining that the difference between the viewer's expected and assessed state is smaller than an E&N state difference threshold.

Summary

This paper explored several aspects and proposed a system that will be socialized in the creative production community. It covered:

1. A system for optimizing creative intent playback by physiological monitoring
2. Metadata for creative intent that includes emotion and narrative goals
3. Metadata for the emotional and narrative goals that include instructions for signal modification
4. A system that modifies signal rendering bases on metadata for creative intent

A key question is whether creatives will be interested in working with biofeedback, generation of metadata, and understanding emotion and narrative effects on other the gut level instincts or traditional craft understanding. One viewpoint is that this requires the creatives to micromanage intent. At this point it is hard to predict the level of interest or acceptance of these ideas in the creative community. We expect of the information-subjective continuum, it will be those who desire to get a across a specific message who have the most interest. But we no there are often snowball effects , so it could be that a few influential creatives motivate others to work in this space. There are also management and business motivations, where if adding such creative intent metadata allows the content to be experienced on a wider range of devices increases the market, and such forces can motivate the creatives, or instigate a new type of craftsperson. Lastly, we wrap up with a centennial example of expansive creative intent metadata, which is the famous poem by T.S. Elliot , *The Wasteland*. This long poem was first published in 1922 by the author in the form of a small book. One surprise in the publication, was that the author included metadata intended to both inform the audience (the reader) of historical details and referenced poems that would be helpful to understand the poem. In addition, he even explained his symbolism, his goals, and his intent behind juxtapositions that were novel at that time. Of course, the term metadata was not existent in 1922, so the term footnotes were used, and of course they were not digitally interactive as described in our proposal, but they were emotionally and cognitively interactive if the reader chose to read the footnotes. As an example of one of these footnotes, here is one for line 218 that addresses the key theme and take-home message;

Tiresias, although a mere spectator and not indeed a "character," is yet the most important personage in the poem, uniting all the rest. Just as the one-eyed merchant, seller of currants, melts into the Phoenician Sailor, and the latter is not wholly distinct from Ferdinand Prince of Naples, so all the women are one woman, and the two sexes meet in Tiresias. What Tiresias sees, in fact, is the substance of the poem. The whole passage from Ovid is of great anthropological interest: [read Ovid for the rest]

Acknowledgements

We particularly thank Dr. Poppy Crum for realizing it is time to move these ideas of real-time biofeedback beyond the world of experimental art, as well as her steering us to Tolstoy's work on creative intent which helps provide actionable structure to an often fuzzy topic.

References

1. https://www.roadtovr.com/htc-facial-eye-trackers-vive-focus-3/
2. https://galea.co/#home
3. https://arpost.co/2022/06/01/varjo-opencbi-vr-headset-with-neurotech/
4. Z. Deng *et al*., "Factorized Variational Autoencoders for Modeling Audience Reactions to Movies," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6014-6023, doi: 10.1109/CVPR.2017.637.
5. https://www.ted.com/talks/poppy_crum_technology_that_knows_what_you_re_feeling
6. C. Mosher and B. Wellner (2022) Biometric signals reveal how audiences engage with stories. SMPTE Motion Imaging Journal. V131 #2
7. J. Farrell (2017) Th varieties of authorial intention: literary theory beyond the intentional fallacy. Macmillan
8. Livingston, Paisley (2009) Poincaré's ''Delicate Sieve'': On Creativity and Constraints in the Arts.' The Idea of Creativity. Eds. M. Krausz, et al.. 129–46.
9. Huddleston, Andrew (2012) "The Conversation Argument for Actual Intentionalism", British Journal of Aesthetics **52**(3):241–256.
10. Stock, Kathleen (2017). Only Imagine: Fiction, Interpretation and Imagination (1st ed.). Oxford University Press.
11. J. Storyk (2020) recording studio and architectural room acoustic design , tutorial. Walters- Storyk Design Group
12. Image parameter values for high dynamic range television for use in production and international programme exchange, Recommendation ITU-R B. 2100 , 7/2018
13. *Paul Ekman". American Psychologist. 47 (4): 470–471. April 1992.* doi*:*10.1037/0003-066x.47.4.470.
14. Plutchik- *Plutchik, Robert (1980), Emotion: Theory, research, and experience: Vol. 1. Theories of emotion, 1, New York: Academic*
15. Alan S. Cowen and Dacher Keltner (2017) Self-report captures 27 distinct categories of emotion bridged by continuous gradients. PNAS September 19, 2017. 114 (38) E7900-E7909; published ahead of print September 5, 2017. https://doi.org/10.1073/pnas.1702247114
16. https://www.humintell.com/2010/06/the-seven-basic-emotions-do-you-know-them/
17. D. Rufenacht and A. Shaji (2022) Customized facial expression analysis in video. SMPTE Motion Imaging Journal , April
18. P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report A-6," Univ. Florida, Gainesville, FL, Tech. Rep. 6, 2005
19. Livingstone, MS (2001) Is it warm? Is it real? Or just low spatial frequency? Science 290:1299.
20. L. Kontsevich and C.W. Tyler (2004) what makes Mona Lisa Smile?, Vis. Res.
21. Daly and Feng S. Daly and X. Feng (2004) "Decontouring: Prevention and removal of false contour artifacts" , SPIE Proc. 5292, pg. 130-149.
22. https://qz.com/1039102/disney-can-now-use-infrared-cameras-to-track-your-reactions-to-films-in-darkened-cinemas/
23. A. Salah and A. Salah (2008) Technoscience Art: A bridge between neuroesthetics and art history. Review of General Psychology. V12 #2, See pg. 7 for entoptics example .
24. A. Favre-Félix ; C. Graversen ; T. Dau ; T. Lunner (2017) real time estimation of eye gaze by in-ear electrodes. IEEE EMBC annual conf.
25. Hladek, Porr, Brimijoin (2018) Real-time estimation of horizontal gaze angle by saccade integration using in-ear electrooculography, PLOS
26. Antonenko, Paas, Grabner, van Gog, (2010) Using Electroencephalography to Measure Cognitive Load, Educational psychology review V22 , #4, pp 425-438.

27. Muhl, Jeunet, and Lotte (2014) EEG-based workload estimation across affective contexts. Front. Neuroscience V 12.
28. Wong, Daniel E., Hjortkjær, Jens; Ceolini, Enea; Nielsen, Søren Vørnle; Griful, Sergi Rotger; Fuglsang, Søren; Chait, Maria; Lunner, Thomas; Dau, Torsten; Liu, Shih-Chii; Cheveigné, Alain de (2018). A closed-loop platform for real-time attention control of simultaneous sound streams., ARO Midwinter meeting (abstract).
29. https://nordicapis.com/20-emotion-recognition-apis-that-will-leave-you-impressed-and-concerned/
30. Nhan and Chau (2010) Classifying Affective States Using Thermal Infrared Imaging of the Human Face, IEEE Tran Biomedical engineering, V57.
31. P. Shem, S. Wang, and Z. Liu (2013) facial expression recognition from infrared thermal videos, In: Lee S., Cho H., Yoon KJ., Lee J. (eds) Intelligent Autonomous Systems 12. Advances in Intelligent Systems and Computing, vol 194. Springer, Berlin, Heidelberg.
32. https://medium.com/inborn-experience/behold-the-next-generation-vr-technology-part-6-brain-interface-89b1d31a0a96
33. http://neurable.com/
34. https://www.cnet.com/news/dolby-vision-hdr10-advanced-hdr-and-hlg-hdr-formats-explained/

Author Bios

Scott Daly is an applied vision scientist at Dolby Laboratories, with specialties in spatio-chromatic-temporal vision and auditory-visual interactions. He has a BS in electrical engineering from NCSU and an MS in bioengineering from the University of Utah. Past accomplishments led to the Otto Schade award from SID in 2011, a team technical Emmy in 1990, and he recently completed the 100-patent dash in just under 30 years.

Evan Gitterman is a researcher at Dolby Laboratories, where his work includes EEG- and physiology-based assessment of multimedia technology experiences. He received his BS from Stanford University, where his research was primarily in neuromusic.

Dan Darcy received his PhD in neuroscience from the University of California, San Diego and did his postdoctoral fellowship at the University of California, San Francisco. His research interests include synaptic physiology, sensory perception and integration, and neuroplasticity. He is currently a Senior Staff Scientist at Dolby Laboratories in San Francisco, CA.

Shane Ruggieri is the Advanced Imaging Systems Creative Lead in Dolby's office of the CTO. He has contributed to several imaging tools and technology patents, and his color work is cited by MPEG to examine new video standards. He is a published author with "Breaking Out of the 100nit Box: A Colorist's View of HDR Grading" and "A Perceptual EOTF for Extended Dynamic Range Imagery" – both presented at SMPTE Annual Technical Conferences. He produced and directed the short film, One-way Ticket, which highlights concepts of interscene high dynamic range. His career spans 23 years including credits on television commercials, music videos and feature films. He has contributed his creative talents to projects for Apple, ARRI, California Music Awards, Dockers, Fox Interactive, Universal Studios, LG, NETFLIX, Sony, Square, Twitter, VISA, and featuring artists such as Carrie Underwood, Avicii, Enrique Iglesias, Elton John, Green Day and Ellie Goulding. His feature film color work includes the multiple-award-winning Nowhere Girl (2014), What Are The Chances (2016) NETFLIX's Nocturne (2018), and The Weight of Success (2018). Shane is a SMPTE San Francisco Section Manager and has degree in Philosophy/Law & Society. He holds 1 issued and 2 pending patents in the field of creative technology.

Elkman theory (6)	Plutchik theory (8)	Core -anonymous (7)	Cowan theory (27)	Facial expressions (9)
Joy	joy	happiness (joy)	Joy	happiness (joy)
Surprise	surprise	surprise	Surprise	surprise
Sadness	sadness	sadness	Sadness	sadness
Anger	anger	anger	Anger	anger
Disgust	disgust	disgust	Disgust	disgust
Fear	fear	fear	Fear	fear
	anticipation		excitement (anticipation?)	
	trust			
		contempt	Contempt	contempt
			Calmness	neutral
			Boredom	boredom
			Awkwardness	
			Anxiety	
			Horror	
			Romance	
			sexual desire	
			Nostalgia	
			Confusion	
			entrancement (amazement?)	
			Amusement	
			Adoration	
			Admiration	
			Awe	
			aesthetic appreciation	
			Craving	
			interest (anticipation?)	
			Satisfaction	
			Relief	

Table 1: Theories of Human Emotions