PROCEEDINGS OF THE IEEE

Computational Media Intelligence: Human-centered Machine Analysis of Media

Journal:	*Proceedings of the IEEE*
Manuscript ID	Draft
Manuscript Categories:	Special Issue Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Somandepalli, Krishna; University of Southern California, Electrical Engineering Guha, Tanaya; University of Warwick, Kumar, Naveen; Disney Research Adam, Hartwig; Google Inc Narayanan, Shrikanth; University of Southern California, Electrical Engineering
Keyword:	multimodal machine learning, cross-modal machine learning, media content analysis, media narrative understanding, media intelligence

JOURNAL OF LaTeX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015

Computational Media Intelligence: Human-centered Machine Analysis of Media

Krishna Somandepalli, Tanaya Guha, Naveen Kumar, Hartwig Adam, Shrikanth Narayanan

Abstract—Media is created by humans for humans to tell stories. There exists a natural and imminent need for creating human-centered media analytics to illuminate the stories being told, and to understand their impact on individuals and the society at large. Objective understanding of media content has numerous applications for different stakeholders, from creators and decision/policy makers to consumers. Advances in multimodal signal processing and machine learning enable detailed and nuanced characterization of media content (of who, what, how, where and why) at scale, and help understand its impact ranging from individual experiences to behavioral, cultural and societal trends to commercial outcomes. Modern deep learning algorithms combined with audiovisual signal processing can analyze entertainment media (movies, TV) and quantify gender, age and race representations to create awareness in objective ways that was hitherto impossible. Text mining and natural language processing allow nuanced understanding of language use and spoken interactions in media to track patterns and trends across different context. Moreover, advances in human sensing have enabled us to directly measure the influence of media on an individual's physiology (and brain), while social media analysis enable tracking societal impact of media content on different cross-sections of the society. This paper reviews representative methodologies and algorithms, tools and systems advancing the area of human-centered media understanding through machine intelligence.

Index Terms—Media intelligence, cross-modal and multimodal modeling, media content analysis, media narrative understanding.

1 Introduction

Technology has a rich and longstanding history in the creation, production, manipulation, distribution, sharing, archival, synthesis and display of multimedia content. They occur across different modalities (sound/audio, print/text and visuals), formats, platforms and content types (short, long, live action, graphics and animated content) in conventional venues such as newspapers, radio, film and television to contemporary streaming and social media platforms. We use digital media to create, capture and experience *stories*. These stories permeate our daily routines and impact what we know, and how we think, form and communicate ideas and opinions. They cover an amazing range of domains: arts and entertainment (movies, television, games, user-generated stories e.g., on YouTube, Instagram), education and research (lectures, scholarly archives), information sharing (news) and commerce (advertisements). There is a rich variety and wide variability in the purpose, type and quality of the media content in terms of what stories are being told, and why (e.g., movies entertain and try to be commercially successful, documentaries try to educate and create awareness, ads try to be compelling and help market/sell products, media archives offer a platform for scholarly research and education), as well as how and in what form and through what platform they are communicated. Finally and critically there is the important "human" dimension: the story creators, the story audience and the story subjects–who is telling the story and how, and for whom and about whom. The theme of this paper is centered on supporting the understanding of the stories we tell through media, and quantifying their impact on individuals and society through human-centered machine intelligence methods, tools and systems. This entails not only enabling rich multimodal analysis of media content–of the people, places and their interactions in stories–but in connecting it to the related human experience and behavior e.g., felt emotions, and broader impact such as commercial outcomes e.g., predicting the success of an ad, and societal trends e.g., delineating the impact of violent media content on youth.

A range of research questions and applications motivate the development of human-centered media intelligence techniques and tools. The most well-established, but still active, domain of these relates to retrieving and interacting with media content which attempts to answer basic questions related to the *who, what, where, how, when* using audio, speech, text, image and video signal processing and machine learning. But numerous other domains inspire continuing technical advances and their applications such as in

- *Understanding nuanced representations and portrayals of people* along dimensions of character traits such as age, gender, appearance and race, the interaction between characters and their environment, including identifying any biases along these human dimensions, and creating objective measures of diversity and inclusion of individuals in media
- *Modeling and predicting human media experiences* both proximal effects e.g., emotion, engagement, attention, and boredom; decisions and behavior (more distal) e.g., willingness to buy a product based on an ad; and long term societal trends/outcomes (distal) e.g., new trends: behavior change, culture shifts
- *Methods and tools for "closing the loop" with human stakeholders* including for personalizing media experiences e.g., age, culture, relevant/appropriate content, and designing novel ways of connecting intended (creative output) and actual human experiences e.g., tools for mediating story telling: modifying scripts

and seeing narrative structure/flow changes editing tools combined with analytics.

Sample media creation and consumption statistics

• 786 movies made in Hollywood in 2019 with box office revenue more than $10 billion [1]

• 560 billion USD/year spent globally on advertisements: an individual is exposed to 4,000-10,000 ads/day [2]

• Social Media Users: 2.6 billion Facebook, 2 billion YouTube, 1 billion Instagram, 800 million TikTok [3]

• 500 hours of video uploaded every minute on YouTube [4]

• $134,000$ hours of sports media content created (2017) from mass media to social media and direct connection to fans.

The scale of global media content production is staggering, and so is its consumption. Some illustrative statistics can be seen in the inset above. Despite the variety and scale of media content creation and consumption and its impact on society, it has not been studied systematically from a human-centered computational perspective. For example, efforts to study diversity in the representations of people in media have been largely qualitative and require immense manual work with human annotations and/or surveys, which cannot match the scale of media content production or consumption. Hence, such methods have been unable to produce systematic data for both science and media scholarship at scale, as well as for actionable intelligence. Traditional computational media content analysis has been largely focused on addressing the needs of organizing, indexing and navigating through large multimedia data corpora. However, it is critical that computing effort is driven not only toward personalizing interaction experience and generating insights and human-centered analytics, but in quantifying the very *stories* that the media tell, and their impact of media on individuals and society. *Computational media intelligence* (CMI) aspires to achieve these goals. In particular, CMI aims to answer the following research questions:

- How do media stories represent and reflect society along human dimensions, such as gender, race, ethnicity, age, ability, profession and socioeconomic status? How do these representations evolve over time?
- How are media portrayals and representations perceived and experienced by individuals, and in the light of the inherent diversity and variability across humans?
- How does media impact and influence individuals, society and culture, both short term and long term? How can we computationally measure such impact and influences?

Creating such machine intelligence requires capabilities to process, model and analyse media content across multiple modalities (audio, video, language), both individually and jointly. These modalities are heterogeneous, noisy, and have dynamic, complex relations among them. Often, they offer only partial information about the story being told. On the other hand, the information from these channels need to be connected to seemingly abstract attributes such as human representations, perception, behavior and impact. The primary objective of CMI are twofold: (i) developing algorithmic capabilities to analyse multimodal media content for deriving human-centered analytics, and (ii) creating methodologies to quantify and measure content's influence and impact on individuals, groups and society.

With these goals in mind, the rest of the paper is organized as follows. Section 2 outlines the three main components of CMI: individual *identity and representation* in media portrayals, the story *context* notably the dynamic scenes, and *interaction* between individuals and their story world, and provides the necessary social science and media studies background for these components. Section 3 describes the available video datasets that can be used for developing CMI and current methodologies and algorithms in representation, context and interaction components of CMI. Section 4 presents an illustrative case study on gender bias analysis in media within the CMI framework. Section 5 discusses the open challenges and future opportunities in this emerging area.

2 Components of Media Intelligence

In order to create comprehensive media intelligence, we first want to understand the overall process of media creation and consumption. Figure 1 illustrates the key stages in this process that we call the *life-cycle* of media development and consumption.The life-cycle of systematically-crafted media forms such as movies, television/streaming media shows and advertisements typically begins with script development. At every stage of this media life-cycle, humans are involved. The focus of CMI is to understand, measure and quantify the various human-centric aspects across the different stages of the media life-cycle - from creation to consumption. Toward this end, we identify three aspects of media content that serve as the primary *components* of media intelligence:

Representation and identity: How do we describe characters in media? Can we obtain a dimensional representation of the characters for studying their portrayal objectively and at scale?

Context: What roles do the characters play in the content? In what environment do they appear, how are they portrayed? What are their actions and what actions do they receive?

Interaction: How do characters interact with each other within a narrative? How do they interact with their environment? How do their relationships evolve through the length of the content?

First, we provide a social science/humanities or a media studies perspective toward developing engineering methodologies and solutions to model, quantify and measure the three aforementioned components of media intelligence. To illuminate the scope of understanding different components of human-centered media intelligence, let us consider an excerpt from the script of an Academy award winning movie, `Thelma and Louise (1991)`.

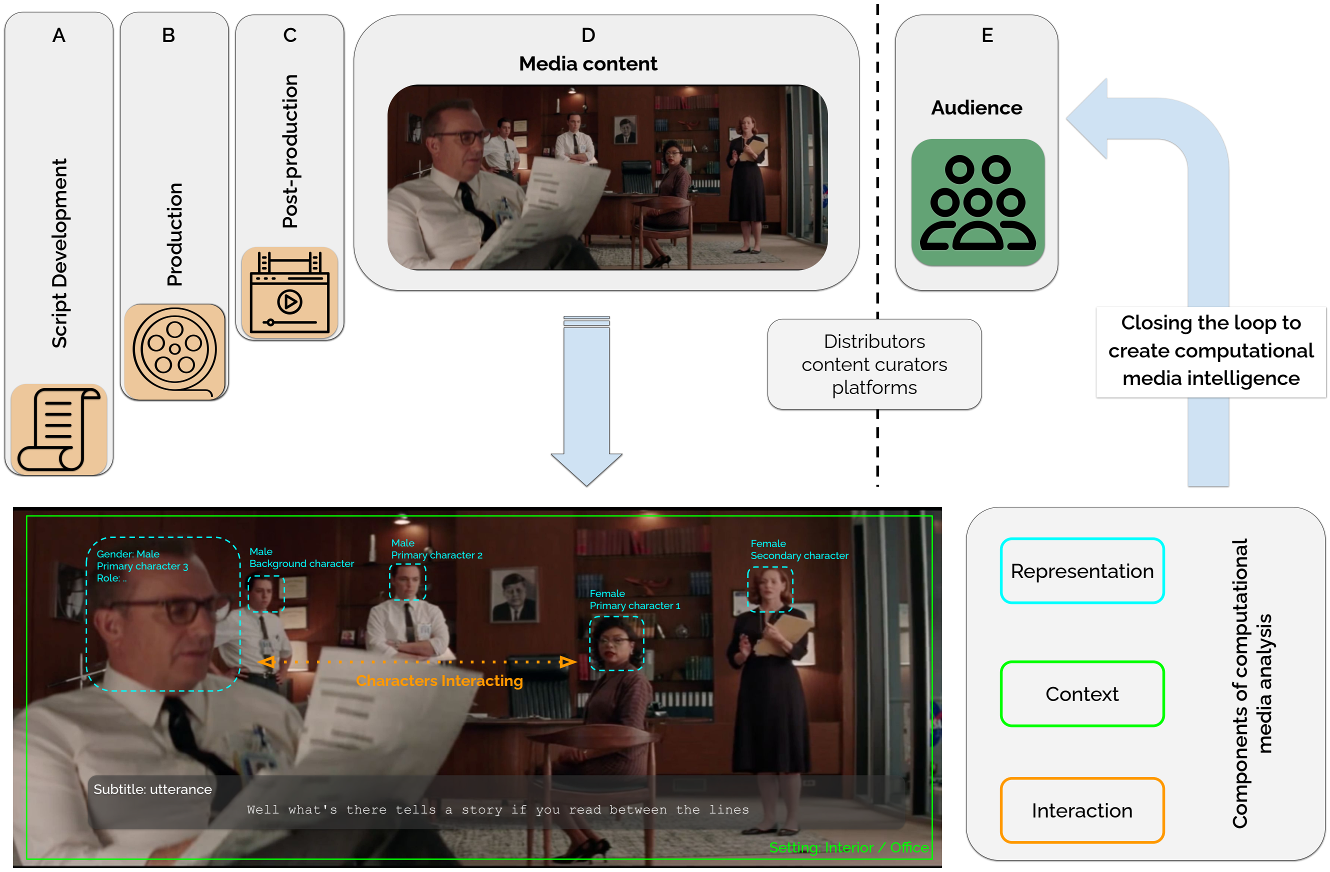

Fig. 1. Five key stages in the media life-cycle: (A) Script development: creating the blueprint for the content (e.g., screenplay, casting, location scouting, budgeting) (B) Production: Creating raw materials for the finished content (e.g., shooting scenes) (C) Post-production (e.g., film, sound and music editing and mixing, visual effects) (D) Produced media content: audio-visual content with dialogue and (E): Release, distribution across various platforms and audience outreach (e.g., trailers, ads) and consumption.

```
LOUISE is a waitress in a coffee
shop. She is in her early thirties,
but too old to be doing this. She
is very pretty and meticulously
groomed, even at the end of her shift
[...] THELMA is a housewife. It's
morning and she is slamming coffee
cups from the breakfast table into
the kitchen sink, which is full
of dirty breakfast dishes and some
stuff left from last night's dinner
which had to "soak". She is still
in her nightgown. The TV is ON in
the background. From the kitchen, we
can see an incomplete wallpapering
project going on in the dining room,
an obvious "do-it-yourself" attempt
by Thelma.
```

We will use this example excerpt throughout this section to illustrate the different media components of interest.

2.1 Representation and Identity

Let us understand how people are described or represented in the above excerpt. It is apparent to a human reader that it describes two characters (named Louise and Thelma), and serves as an exposition for their roles in the rest of the story. The author (screenwriter) describes different facets of the characters by *identifying* personal attributes that may be physical (e.g., appearance) or functional (e.g., profession) or more abstract (e.g., outlook, personality, temperament). The author may use attributes, such as gender, race/ethnicity, socio-economic descriptors, age and body type. These facets used to describe a character form the core of quantifying the *representation* of a character [8]. The intersectionality of these identity dimensions along with the story plot associated with the character also form the basis of character tropes and stereotypes in media content [9].

A first step toward describing media representations of individuals is to understand how we identify ourselves and others individually and socially. We thus ground the definition of representation within the social-scientific concepts of identity and self-identity [10], [11], [12]. Media creates characters with different identities to connect to their target audience [13]. In social science, identity is conceptualized based on a set of identity attributes referred to as the *dimensions of identity*. The dimensions are largely based on how we define ourselves internally and externally [14]. This gives rise to self-identity (how we understand ourselves and experience others) and social identity (how we express ourselves to others). A third set of identity

Fig. 2. Word clouds for location settings from (A) Paces365 dataset [5], (B) Holistic video understanding dataset [6], and (C) Settings/location tags mined from scene headings of the scripts used in Ramakrishna et al. [7]. The size of the words is proportional to the frequency of the label in the corresponding dataset.

attributes include factors influencing the formation of our identity (that we may be unaware of) [15]. A few prominent individual identity dimensions are age, race, gender, socio-economic status, sexual orientation, (dis)ability and physical appearance. Some of these dimensions of identity, e.g., gender [16], [17], race (see survey in [18]) and age [19], [20] have been widely studied in the speech processing and computer vision community, but as isolated topics, notably to ensure robust performance of algorithms e.g., automatic speech recognition, due to variability across some of these dimensions e.g., gender, age. It is only recently that these methods are being employed in the context of media understanding [21], [22]. None of the existing works have approached representation analysis in media from the perspective of self-identity.

For computational understanding of identity and representations, we identify three key challenges: (i) *Dimensions*: the aspects of identity that are relevant in a given context (e.g., gender, race/ethnicity), (ii) *Classification* taxonomies: the categories or classes for a given identity dimension (e.g., different categories of race), and (iii) *Identifiability*: whether or not we can identify a dimension of identity reliably and without systematic biases. To illustrate these challenges, let us return to the movie script excerpt: Louise's gender can be inferred through the use of the pronoun 'she' (not considering for the normative use of the name Louise). There are several other widely used dimensions of identity such as race/ethnicity and body type that are not specified although other aspects of the person's appearance (e.g., 'meticulously groomed') are described.

While we may attempt to computationally model some of these dimensions such as gender using a classification system (e.g., two classes: male/female), these systems often rely on social normative based on broad, often incomplete taxonomies for each dimension. This presents an important and open challenge of having to determine dimensions of identity and also to have a classification system, as we are interested in developing computational means to quantify representation. Even if we are given dimensions and classification taxonomies, there still remains the challenge of *identifiability*, that is, to understand the limitations of the computational means in identifying those dimensions reliably from observable data from either the media content or associated metadata.

2.2 Context

Quantitative representation analysis often results in counts or frequencies of appearance of the identity dimensions in a media story. No matter how sophisticated, without an understanding of the context, environment and backdrop in which the story is situated, representation analysis is unlikely to be meaningful by itself. Consider a movie, where women appear frequently on screen, but they are only shown as caregivers, maids and waitresses, thereby reinforcing a negative stereotype about normative professions held by women. Additional information about the scene environment (e.g., kitchen) is thus crucial to contextualize the counts and frequency statistics. Media context or simply context refers to this *when* and *where* aspects of the story.

Context in media stories is primarily conveyed through the visual modality. In media studies, this is summarized by the term 'Mise en scéne' [23] i.e., 'placing on stage'. For films, this term refers to the composition, sets, props, actors, costumes and lighting [24], essentially all elements that appear on camera. This is often augmented by music and sound effects in the audio modality, but not necessarily related to the events explicitly shown on screen. In scripts, some context information is available about scene locations from the 'scene headings' (e.g., `INT. CAR` as a scene heading refers to the scene set being the interior of a car).

Two key challenges in computational modeling of context are: (i) Building taxonomies for different media domains, and (2) multimodal dynamic scene understanding. To illustrate the challenges, let us consider two recent datasets created for scene understanding: Places365 [5] and holistic video understanding (HVU) dataset [6]. We visualize the distribution of location tags for these two datasets in Fig. 2A and 2B, respectively. In a similar vein, we mined the location tags from a corpus of about 1000 movie scripts [7] and visualize them in Fig 2C. The juxtaposition of the different classification taxonomies in Fig 2 reveals that existing datasets and taxonomies from Places365 and HVU may not generalize well to the media domain which represents a richer variety and range of contexts in which stories unfold. This highlights the need for building data-driven and domain-specific taxonomies for application to different media content. While there have been significant advances in automatic video understanding [25], works that augment information from other modalities (audio, subtitles) is somewhat limited. Recent efforts in archiving large scale audio datasets, such as VGGsound [26] can

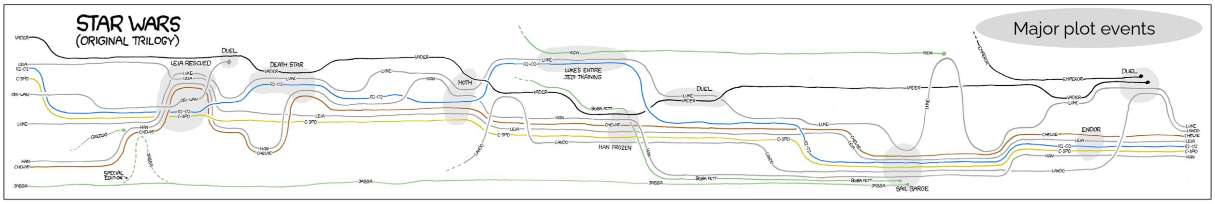

Fig. 3. Movie narrative charts from the comic xkcd.com/657. Each line corresponds to the 'journey' of a character. Notice that the major plot events are centered around a confluence of appearances from multiple characters. Understanding how characters interact with each other can shed light into the prominence of character roles in media content such as Film and TV.

help build multimodal models for visual scene/ambience understanding. In this respect, multimodal modeling is key to a holistic understanding of media context.

2.3 Interaction

Understanding how characters *interact* with each other and with their environment is another crucial feature of the CMI framework. We classify interactions into two categories: character-to-character (C2C) interactions and character-to-environment (C2E) interactions.

In the media domain, C2E interactions are somewhat less studied although there is a wealth of related work in the broader video understanding and video summarizing literature [27]. To illustrate C2C interactions, let us take an example from the XKCD comic[1] presented in Fig. 3. It visualizes the appearance of characters in the original *Star Wars* movie trilogy [28]. Simply examining the character tracks together can help us localize the major events (shaded gray regions in Fig. 3) in the story as a confluence of the prominent characters in the movies. C2C interaction modeling can answer summative questions about a narrative, such as which character is central to the story. From a computational perspective, modeling C2C interactions heavily relies on constructing narrative structures of the stories told. A theoretical framework to represent narrative structures [29] can help C2C interaction modeling, and even can be extended to model C2E interactions as it includes mechanisms to quantify which characters interact with each other, in which manner they express themselves and how their expressions are received. Challenging open questions in the realm of C2C interactions modeling requiring further research include (1) how to automatically extract salient events in a plot from a dynamic graph constructed from character interactions? Recent efforts show early promise in this direction [30]. (2) how to quantify the interactions between characters along dimensions of agency, power or emotional expression? (3) how to identify causal relationships in narratives to determine most 'influential' characters? A computational modeling of C2C and C2E interactions helps us to understand the relation between characters and their environment. Such modeling can help shed light on systematic portrayals or narratives in media.

2.4 Impact and Experience

Media is ultimately targeted towards humans, hence, quantifying how media affects humans in a systematic manner is a key objective of CMI. This is also an essential element for closing the loop with the users in providing personalized experiences e.g., recommender systems with increased awareness of not just user needs and preferences but their experiences [31]. This includes quantifying felt experience such as emotional response, elicited experience such as likability, consumption patterns and even, longer term impact media has on human behavior and societal trends. We broadly categorize media impact into *immediate*, *proximal* and *distal* impact. Immediate impact includes direct human sensory and affective influences e.g., audience emotions while proximal impact involves both direct effects such as propensity to purchase a product after watching an ad and related effects including financial outcomes (e.g., box Office returns) and popularity/viral spread of specific content. Distal impact refers to how broader and more enduring societal perceptions are shaped by media narratives; for instance, how behavioral changes in youth are influenced by violent media content or how specific news media shapes opinions towards historical events or toward specific communities such as persons of color and other minorities. It is important to note that these three categories are not mutually exclusive and can all contribute to the overall influence of media on the society in the long term.

Immediate sensory-affective impact: We can categorize three different "types" of affect influences of media content [32]: *intended*, *expected* and *experienced*. Intended affect is what the content creators attempt to evoke in their audience, experienced affect describes the emotion an individual *actually* feels when consuming the content, while expected affect is the expected value of experienced affect in a population. Although some prior work has considered the intended and expected emotions to be the same [33], this is not generally true for media content. It is possible that a movie is unsuccessful in conveying the intended affect to its audience. In fact, this *mismatch* is often used to assess movie success and quality [32]. Affect is also understood and conveyed differently for shorter media content, such as advertisements, a topic that has received limited attention so far [34], [35].

One common way to quantify the affective impact is by mapping media content to affective dimensions such as intended arousal (or strength) and polarity (positive/negative) or categories (happiness, sadness etc). There are also specific affective constructs of relevance to specific media domains such as violence [36], [37], [38] and humor [35] in entertainment media like movies and TV shows, and attention grabbing elements or likability of ads [39]. These tasks are typically multimodal [40], [41] and require

1. https://xkcd.com

an understanding of what modality evokes what affective dimension and how they interact with each other.

Proximal Impact refers to the shorter term effect of media, outside its direct consumption. An objective measurement and categorization of proximal impact is an open problem. Nevertheless, we can examine proxies,for example, box office ratings can be used as a proxy for movie's popularity [42], and views and likes can do the same for user-created YouTube content [43].

Long-term distal impact refers to a media impact at relatively larger spatio-temporal scales that can lead to broad and enduring societal effects. In addition to measuring the temporal variation of impact over a long period of time, the aim is to understand its relationship, if any, with various societal behavior and cultural shifts, and even to specific events. The challenges associated with assessing and predicting long-term media impact are complex and enormous. It is important to be mindful of sampling bias, and several other factors, such as style, form, genre and topic, and the general prevalent social context that is implicit to an audience. A few studies have examined the longitudinal effect of media content, such as relation between violent content to aggressive behavior in children [44], and exposure to certain media and alcohol use by adolescents [45]. These studies are largely inconclusive; revisions of earlier claims about violent games and agression in a recent meta-analysis reveal "..negligible relationships between violent games and aggressive or prosocial behavior, small relationships with aggressive affect and cognitions, and stronger relationships with desensitization" [46]. Such longitudinal analysis, is also of interest in a commercial sense such as in understanding advertisements, because this reflects the evolution of advertising and marketing strategies over the years [47]. Objective computational approaches offer new possibilities to pursue some of these questions in increasingly data-driven ways in the future.

3 Methods and Algorithms

In this section, we first describe various databases that can facilitate the design and evaluation of CMI approaches. Next we discuss representative algorithmic and methodological work for each component of CMI identified in Section 2.

3.1 Data resources

The availability of large, curated and labeled corpora is essential for enabling the multimodal machine learning and data analytics tasks of computational media research. We discuss recent advances in creating relevant large scale databases that can facilitate the goals of CMI.

Analysis of media content such as film, TV, ads or news require labeled audio/video resources corresponding to the domain-of-interest for ensuring robust model learning and performance. Significant advancesin person (face) identification have been made possible through the rich datasets such as VGGFace2 [48], MS-Celeb-1M [49], IMFDB [50] and CelebA [51]. Domain-matched video datasets with visual character tracking are only recently being compiled as shown in Table 1. For audio, there are large scale speaker verification and recognition datasets, such as VoxCeleb [52] and VoxCeleb2 [53]. However, no large scale datasets for other tasks such as speech activity detection are available, leading to novel ways of creating self-supervised data resources such as the Subtitle-aligned Movie (SAM) Corpus [54]. This dataset will be discussed further in this section.

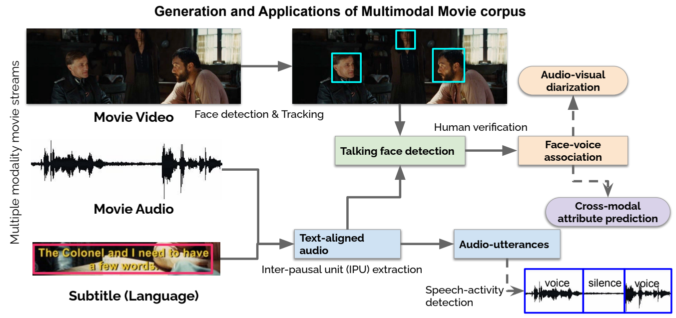

Fig. 4. SAIL Multimodal Movie Corpus: Illustration of a processing pipeline to extract domain-matched training data in an semi-automatic fashion. Different shades indicate different operations for generating labeled data, and the broken lines indicate different end goals. The lower half of the figure illustrates the process for generating Subtitled-Aligned Movie corpus (SAM) which was used to develop state-of-the-art movie speech activity detector in [58].

There is also a need for data with richly diverse people, context, interaction related attributes including representing diverse individual attributes (age, gender, appearance, etc), cultures and languages; for example, it is well demonstrated that models fail to generalize on data originating from different race or culture not included in the design [55], [56]. Recent efforts such as FairFace [55] have compiled datasets that are balanced across attributes such as race, gender and age from face images. Another effort [57], listed in Table 1 created a benchmark dataset for movie character identification in movies with a more racially diverse cast. However, there is a general lack of such well curated resources for the media domain.

Toward filling this gap, several open-source *video* databases are being released recently for research in the area of media understanding. Table 1 and Table 2 provide a list of currently available media databases. The tables also provide other details about the databases, such as database size, original tasks, attributes and labeling schemes. It should be noted that the majority of the large scale video datasets in movie and TV domains have only been released since 2015, underscoring a growing interest in computational media research. We also observe that the labeling efforts can be scaled up for long-form content (movie and TV shows) with semi-automatic methods and human-verification. More recently, self-supervision approaches are also being used to obtain cross-modal labels automatically. Such approaches are going to be increasingly needed to afford not only scale but process complex media data spanning multiple modalities.

Subtitle-aligned Movie Corpus (SAM)

The recently developed subtitle-aligned movie (SAM) corpus [54] is an example of data curation using self-supervision between different modalities with some human-in-the-loop verification. The approach for generating labels

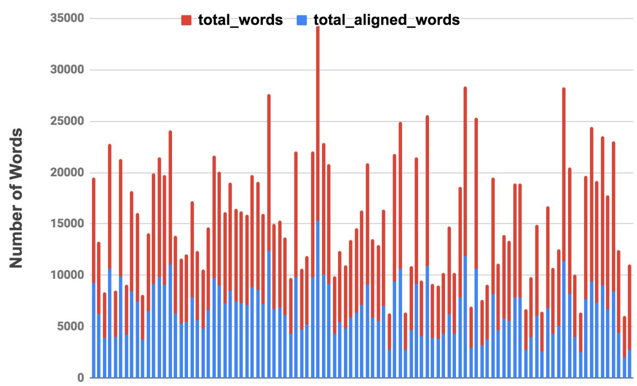

Fig. 5. Distribution of number of total words aligned from the subtitiles to audio. Overall (N=120), $76.4\% \pm 8.6\%$ of words across the 95 movie subtitles were successfully aligned using the system. This resulted in about 52 hours of speech data and 156 hours of non-speech data in the SAM corpus.

from movie data is illustrated in the Fig. 4. It employs a combination of automatic tools and human verification tasks to generate these labels at scale. For the example shown in Fig. 4, labels are generated for three tasks: speech activity labels from audio, active speaker localizationfrom video, and cross-modal gender identification. To illustrate the label generation process, just the speech activity detection task is further elaborated below.

To generate precise labels for speech activity in the movie audio, automatic text-to-audio alignment is used. We use the subtitles generated automatically as transcripts for alignment[2]. These subtitles provide approximate starting and ending time-stamps corresponding to each single dialog. The audio segments between two successive timestamps are then used to form the non-speech segments. An open-source speech-to-text alignment tool [59] was used to align speech segments at word-level. Note that both subtitle generation and alignment are completely automated. See Fig. 5 for illustrative details on the per-movie distributions of percentage words successfully aligned. We used this measure as a proxy to understand the effectiveness of our self-supervised approach in mining speech labels from subtitles.

After speech-to-text alignment, speech and non-speech *segments* are obtained as follows: Speech regions corresponding to consecutive gentle-aligned words were accumulated to form segments of length t_{seg}. First, a heuristic threshold of t_{break} seconds (duration of pause) was used to chunk consecutive aligned words into inter-pausal units (IPU). Hence, two consecutive aligned words were considered to belong to the same IPU if they were no more than t_{break} seconds apart. Finally, these IPUs were split into non-overlapping segments of t_{seg} seconds each. Both t_{seg} and t_{break} can be tuned to obtain speech segments at different scales for different tasks. We then trained models from data generated using this approach for speech activity detection in movies. Our detailed performance evaluation showed that models trained with this data were able to achieve state-of-the-art performance in speech activity detection for

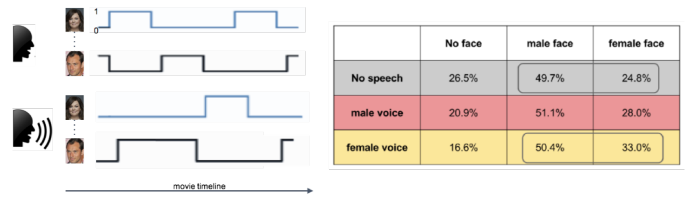

Fig. 6. Audio-visual co-occurrence analysis (on 17 Hollywood movies) shows disparities in portrayals of women on screen even while speaking.

movie audio [54]. Such curated data can readily be used to derive a variety of CMI constructs such as talk time of characters in the story, vocal arousal patterns and speech interaction dynamics including interpersonal vocal synchrony, and other dimensions of identity from speech.

3.2 Representation and identity

As outlined in Section 2.1, there are several targets for CMI in the context of illuminating media portrayals of people through identification for the purposes of quantifying representation in media. We consider two examples below.

Gender representation in media

One of our early efforts toward studying identity and representation in media was motivated by the need of enabling an objective understanding of gender portrayals[3] in media [22]. Social scientists and media experts have repeatedly noted that women are highly underrepresented in popular films and media [60]. Motivated by the aspects that media researchers and practitioners consider important, we proposed to automatically estimate the on-screen time (from video) and speaking time (from audio) of male and female characters in movies. The video processing pipeline follows a simple approach, where on-screen time is computed as the percentage of time female faces are detected over all faces using a standard face detector and face-based gender classifier. The audio processing pipeline performs a speech activity detection followed by an utterance-level gender classification. This simple framework could reveal interesting aspects of gender gaps in Hollywood: Women are seen and heard significantly less amount of time as compared to their male counterparts [61]. An audiovisual co-occurrence analysis (Fig. 6) revealed that women are seen less even when they are speaking [22].

Our initial effort to quantify female and male representation in movies exposed a number of weaknesses of the standard algorithms as they were applied to complex media data. Due to the huge variability and complex nature of the data itself, high accuracy was not attainable even for 'routine' tasks, such as gender classification. This motivated us to revisit some of the fundamental problems in audio and video analysis.

Improving gender recognition with cross-domain data

While the complexity of media data poses additional challenges, they also offer cross-domain information (e.g., subtitles accompanying audio) that can be leveraged for better

2. github.com/ruediger/VobSub2SRT

3. DEMO: shorturl.at/mnoCE

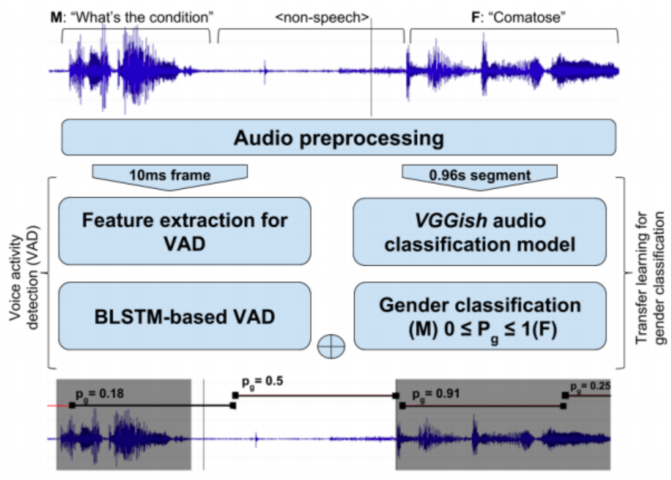

Fig. 7. Leveraging cross-domain and cross-corpus information to improve audio analysis tasks (reproduced with permission from [21])

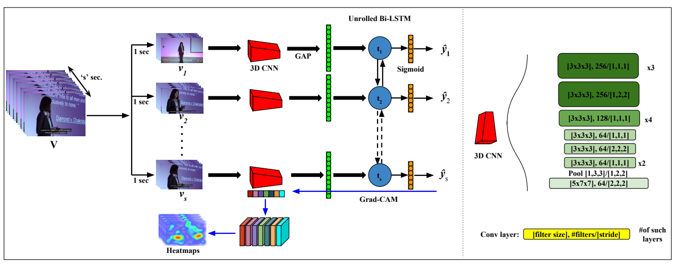

Fig. 8. Hierarchical Context-Aware deep neural network for cross-modal speech activity detection.

performance on fundamental tasks such as gender identification [21], [54], [62] integral to many media intelligence frameworks and models.

To be able to use powerful supervised learning methods, labels are required for large-scale data. However, manually labeling large volumes of data is tedious and expensive. Hence, we proposed to use cross-modal supervision, and also to leverage labeled data from other corpus, where available; for example speech gender labels from AudioSet [63]. While one of our recent works suggest the use of visual information for reliable speech activity detection [64], another recent work aligned the movie subtitles to the movie audio to obtain coarse VAD label (segment-level as opposed to frame-level) [54]. A bidirectional long short term memory network (BLSTM)-based architecture was then developed to analyze the log-spectrograms of an audio segment to detect speech activity within the given segment. A transfer learning technique was used to improve gender detection in movie audio using the VGGish architecture, where a large-scale audio event detection database was used for pretraining [21] (see Fig. 7 for an overview of the method).

In a similar vein, multichannel information present in different language channels (e.g., English, Spanish, French) for a movie can be used to improve the robustness of gender classification [62]. We exploit the fact that the speaker labels of interest in this case co-occur in each language channel. This work fused the predictions from different channels using a method called recognition output voting error reduction (ROVER), which can handle labels even when they are not exactly temporally aligned (as we would expect to happen in different language channels).

The strategy of leveraging cross-corpus and cross-domain information described above helps to achieve state-of-the-art performance for both the tasks, and subsequently improved the accuracy of gender-specific speaking-time estimations in movies.

3.3 Context

Computational techniques offer ways for characterizing the ambient context and backdrop of the unfolding media stories. Some examples are highlighted.

Active speaker localization

The multimedia content in movies is *unconstrained* yet *structured* across multiple modalities, enriching the possibilities for analytical insights. As such, we can leverage both the audio and visual modalities to model tasks such as speech activity detection, speaker classification, speaker gender classification and other video-assisted audio tasks. Recently, there has been a growing number of studies looking at predicting one modality from another [65]. Although some cross-modal methods have been applied to constrained settings such as news media [66], they have not been studied in widely varying settings such as movies and TV shows. In a recent work by Sharma et al. [67], cross-modal supervision was applied between video stream and audio speech labels obtained from the SAM corpus (see Section 3.1) to explore the following questions: 1) Can we predict audio speech-activity-detection from video only? 2) Would such a video-speech-activity model learn to localize the talking faces?

We proposed an end-to-end trainable hierarchical context-aware (HiCA) deep neural network to predict coarse speech activity labels using just the visual information. In order to enable the network to learn from a longer context, which is a necessity in case of videos, we decentralize the temporal context in form of local 3D convolutions and a global LSTM. We do not explicitly detect the face of a speaker or extract facial features, neither for training nor for inference. We evaluate the proposed architecture with videos from Hollywood movies, which is a challenging domain due to its relatively uncontrolled settings in form of frequent shot changes and varying camera dynamics, and the variety and variability in the depiction of speaking characters. The proposed HiCA architecture is illustrated in Fig. 8. For further details about analysis of the HiCA architecture, we refer the reader to the original work [67].

In order to understand the visual constructs captured by the 3D convolutional layers in the HiCA architecture, we modified the Grad-CAMs [68] to accommodate 3D convolutions. The results of the visualization for a few samples are shown in Fig 9. Although the proposed model learns to attend to faces, the speech activity detection performance itself from video frames is at 66.1% accuracy. More recently, we showed that by modeling audio features with the cross-modal HiCA representations in a late-fusion fashion, the overall speech activity detection as well as active speaker classification performance is comparable to the state-of-the-art in the respective domains [69].

Fig. 9. Qualitative localization performance of the proposed HiCA network for various videos. Notice that the network learns to primarily attend to faces and movement in some cases (for example: **c**)

Understanding movie narrative structures

One way to understand the broad context in media is by investigating its narrative structure or storytelling paradigm. Popular films and screenplays follow a well-defined storytelling paradigm that comprises three essential segments or acts: exposition (act I), conflict (act II) and resolution (act III). Act I introduces the main characters in a movie, and presents an incident (plot point 1) that drives the story; this leads to a series of events in Act II including a key event (plot point 2) that prepares audience for the climax. Act III features the climax and the resolution of the story. The 3-act structure provides an important basis for comparing different movies and evaluating relative importance of the characters. We developed an automated system that is able to provide an estimate of the act boundaries using features from visual, music and text modalities [70]. Hand-crafted features like shot length, motion activity, presence of music and speaking rates were extracted from the different features and were linearly combined to obtain a continuous measure of story intensity. This plot was used to detect the act boundaries, plot points and climax in mainstream Hollywood movies to assist in further critical analysis of the narrative structure and form [70].

3.4 Interaction

Character graphs for interaction modeling

Automated analysis of media content, such as movies has traditionally focused on learning and using low level features from audio/video scenes and key events. For humans, however, it is the characters that usually play the most important role in storytelling. To understand and model how characters interact within a story, an efficient approach is to build a character graph or network. A character network usually has the major characters as its nodes where the edges summarize the relationship between character pairs. The relationship between characters though can be defined in various ways, one prevalent approach is to measure co-occurrence [7], [30], [71].

One of our works builds a character interaction network using scripts from movies[4], where an edge between two characters (nodes) is added if the characters have consecutive dialogs (at least once) in a movie [7]. This network uses different graph-theoretic metrics (such as betweenness

4. DEMO: shorturl.at/szST8

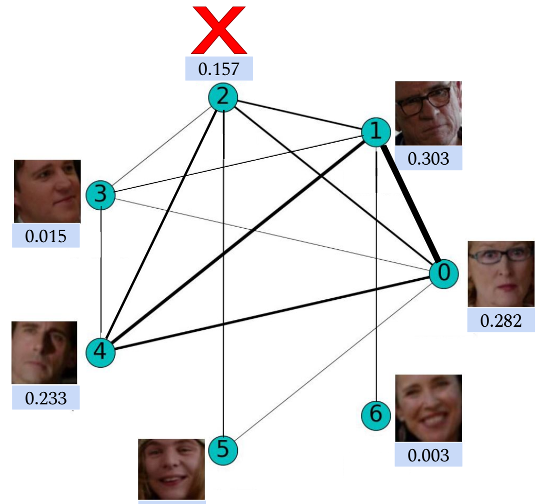

Fig. 10. Character interaction graph for the movie *Hope Springs* constructed via face clustering. The numbers below the faces are the character importance score, and 'X' denotes a noisy cluster.

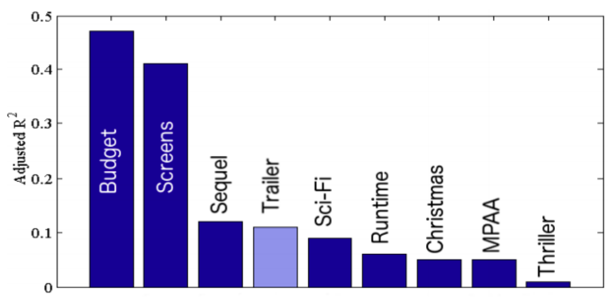

Fig. 11. Comparison of the predictive performance of various metadata and trailer media content on box-office success.

and centrality) to measure the importance of each character, which are then used to examine the character analytics across more than 1000 movies based on gender, race and age. These analytics from the interaction graph showed that women characters have prominent presence only in Horror movies. Latino and native American characters, though present in movies, do not have much interaction with other characters.

Another approach to building such character network is through using the visual stream [30]. In another recent work, we construct a dynamic character network for a given movie through a novel online face clustering algorithm [72]. The relationship between two characters is modeled as their temporal co-occurrence, i.e., if they appear in the same or consecutive shots. The dynamic aspect of the network offers an effective way to capture the variations in character interactions over time. As this work relies on face clustering, it could discover only the major characters. Similar to the script-based approach [7], this work too computes character importance scores, and can easily output the screen-time for each character.

3.5 Impact and experience

Modeling affective media experiences

Understanding the multimodal affective experiences of humans evoked by different components of media is a com-

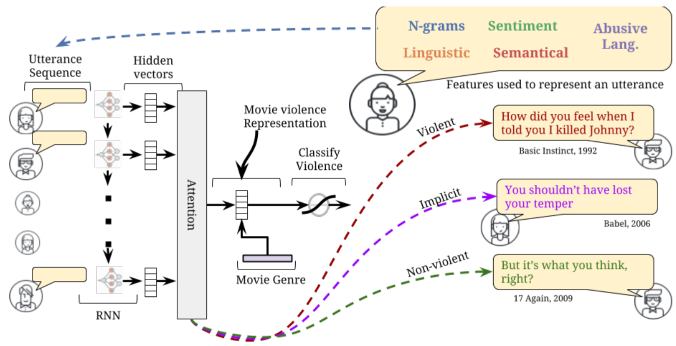

Fig. 12. Violence rating classification model. A sequence of characters' utterances (each represented by a concatenation of language features) is constructed. This sequence is then taken as input to a Recurrent Neural Network with Attention to learn a movie representation. By its attention weights, this model is capable of discerning when a character's utterance is explicitly, implicitly, or non-violent. Reproduced with permission from [38].

Top utterances in movies predicted to be HIGH violent
How did you feel when I told you Johnny Boz had died (Basic Instinct, 1992)
Do you want me to wring that creature's neck? (Batman Returns, 1992)
Tina —You motherf***er!! (A nightmare on Elm Street, 1984)
I knew it was going to end this way. (The Bourne Ultimatum, 2007)
You shouldn't have lost your temper. (Babel, 2006)
We win with hitting, running and fielding, nothing else (42, 2013)

Lowest utterances in movies predicted to be LOW violent
For God's sakes, Alvy, even Freud speaks of a latency period. (Annie Hall, 1977)
No, but it's what you think, right? (17 Again, 2009)
She wadn't home. (The Blind Side, 2009)

Fig. 13. Examples of utterances with highest and lowest attention weights for a few movies. **green** - correctly identified, **blue** - depends on context (implicit), **red** - miss identified. Reproduced with permission from [38]

plex task. For example, music is known to evoke powerful emotional experiences in humans affecting brain activity, physiological response and behavior. In order to analyse this complex interplay, our recent work [73] explored the possibility to predicting brain activity and physiological response from music features. We developed computational methodology that uses auditory features (related to its dynamics, timbre, harmony and rhythm) to predict brain activity (phase synchronizations in bilateral Heschl's gyri and superior temporal gyri), physiological response (galvanic skin response, heart activity), and human emotion in the form of continuous, subjective descriptions reported by music listeners. Multivariate time series models with attention mechanisms are developed for effective prediction of emotional ratings, and vector-autoregressive models are proposed to predict the brain activity and physiological response [73].

In the context of illuminating general multimedia content experience, continuous emotional dimension ratings of activation and valence provided by the viewers can be predicted using audio, visual and even, language features. To combine heterogeneous mulitmodal information, we developed a mixture of experts (MoE) model, where two experts (audio, video) contribute towards the prediction of viewers' emotion ratings while watching movie clips. Our MoE model uses a time-varying attention mechanism for information fusion [41], where the attention component controls the contribution of each expert based on their features at a given time instant. This component is computed through a hard expectation-maximization (EM) algorithm. In another work, we developed a deep autoencoder approach to learn an audiovisual representation of video advertisements (ads) or TV commercials [35]. This work focused on classifying ads in terms of categories such as funny or exciting from viewers' perspective.

Violence prediction from movie scripts

Another key affective construct of interest is the depiction of violence; computational methods can offer ways of understanding this aspect even before the media content is fully produced. -As such, identifying attributes such as violence in the earlier stages of the movie development can have an immense effect on the subsequent steps of movie production, and the ultimate audience experience. Toward this end, we proposed a model to predict violent language from movie scripts in [38].

The proposed architecture is shown in Figure 12. This model was designed to capture two forms of context: *conversational context* and *movie genre*. The former refers to what is being said in relation to what has been previously said. This follows from the fact that most utterances are not independent from one another, but rather follow a thread of conversation. The latter takes into account that utterances in a movie follow a particular theme set by the movie's genre (e.g., action, sci-fi). Conversational context is captured by the recurrant neural network (RNN) layer (left part of Fig. 12). It takes all past utterances as input to update the representation for the utterance-of-interest. This allows our model to learn that some utterances that are violent for a particular genre may not be considered violent in other genres. One additional benefit of using this architecture comes from analyzing the attention weights after training the complete system. The system assigns a higher attention weight to those utterances it considers to be violent (see right side of Fig. 12). By exploring the utterances with the highest and lowest attention weights, we can get an idea of utterance-level violence contained in scripts. A few examples are illustrated in Fig. 13). Our approach appears to pick up on more subtle indications of aggression such as "losing one's temper".

Subsequently, in a recent work [74], we explored the directionality of violence among the character utterances, i.e., identifying the victim and perpetrator by analyzing the subject-verb-object relations in the language use in movie scripts. On an open-source dataset of nearly 1000 movie scripts, our analysis revealed two significant differences in the frequency of portrayals and the character demographics in the interactions between victims and perpetrators : (1) female characters appear more often as victims, and (2) perpetrators are more likely to be White if the victim is Black or Latino. Besides violence, such large-scale studies can

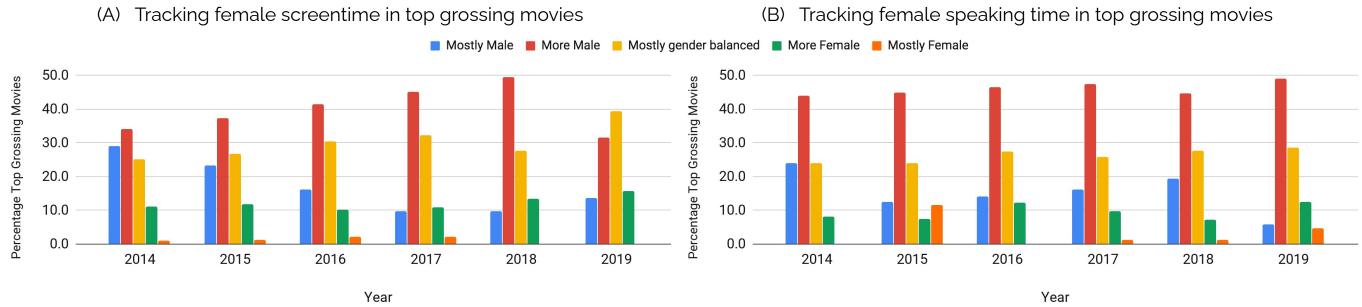

Fig. 14. Distribution of female screentime and speaking time measuring the representation of female characters in top Box Office grossing Hollywood Films. Notice that in general, in general there are significantly more movies with more male participation than that of female. Over the past two years, there is a small uptick in the number of movies with more female participation. However, within the years examined, we do not see a significant change in the overall trend.

also reveal systematic patterns in movie character portrayals along dimensions such as power and agency [75].

Predicting financial outcomes

Connecting media content to financial success is another important aspect of measuring media's impact. To draw insights from what makes a movie financially successful, in a case study we computationally analyzed a database of 474 movie trailers along with their meta-data available from IMDB [42]. Using simple regression models, we investigated the most important predictors of movie's commercial success. We observed that trailer content (handcrafted audiovisual features) play a significant role in determining a movie's success (first week's box office income), even more than its genre and cast (see Fig. 11).

4 Illustrative Real World Case Studies

In this section, we highlight case studies of applying computational media intelligence in the real world context of movies and ads with a focus on diversity and inclusion in media representations.

4.1 SeeJane: Tracking female participation in top Box Office grossing Hollywood films

SeeJane is a collaborative effort led by the Geena Davis Institute for Gender in Media (GDIGM[5]) to understand female representation in top Box Office grossing Hollywood movies with the support of computational media intelligence methods [22].

The analysis focuses on the top 100 grossing Hollywood movies sample for each year to get a representative sample of the most viewed movies in that year. To date, we have analyzed about 600 movies till date (2013 - 2019). As described in Section 2, we automatically extracted two measures to quantify female presence: (1) **Female Screen time:** Proportion of all the faces shown on screen that belong to female persons, and (2) **Female Speaking time:** Proportion of all voices heard in a Film that are classified as belonging to female persons. Female screen time was estimated per [22] and female speaking time per [21]. In both cases, detailed performance analysis of the systems on

5. https://seejane.org/research-informs-empowers

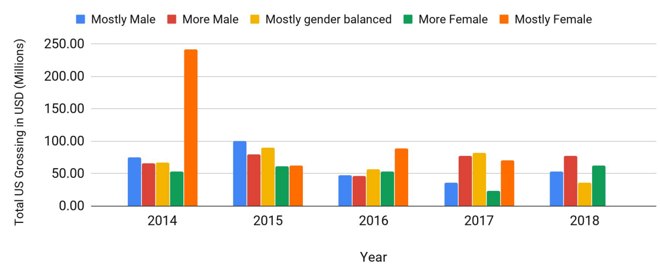

Fig. 15. Overall US Box Office grossing for movies in categories of varying female participation. Here, we show the female screen time measure; a similar trend was observed for female speaking time as well. Notice that there is a significant economic benefit from movies of higher female presence, underscoring the broad importance of diversity and inclusion with respect to gender including in economic terms.

a set of benchmark movies showed that the estimates are within 5% of the manual counts for these measures. These measures are of interest since they not only show patterns in media representations compared to the population (female persons in the US are 5̃0% [6]) in terms of diversity and inclusion, but have a direct impact on how much the actors in movies get paid. Actors tend to get paid more if they have a larger participation in a movie. To summarize the female screen time and speaking time representation, we group the measures into 20 percentage bins with *Mostly Female* when the corresponding measure is greater than 80%, more female for the bin 60%–80%, mostly gender balanced for the bin 40%–60% and more male and mostly male for the bins 20%–40% and less than 20% respectively. We use these bins consistent with another study examining diversity and inclusion in advertising [76]. Some of the key findings from the SeeJane project include:

Men are shown more often than women in top grossing Hollywood films. Fig. 14A summarizes the distribution of percentage of top grossing films per year in each of the female screentime categories. While we notice a small uptick in the number of movies with a higher female participation over the past two years, this is not statistically significant[7]

6. https://www.census.gov/quickfacts/fact/table/US/LFE046218
7. Proportion test was used to assess statistical differences

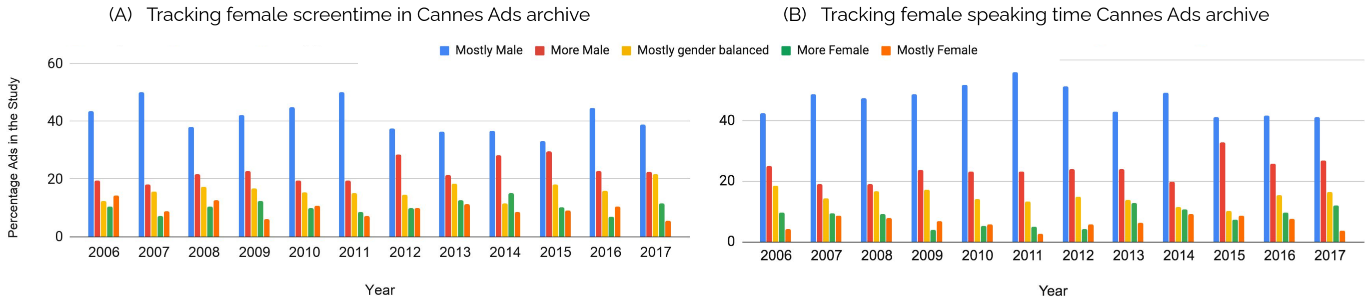

Fig. 16. Tracking female screen time and speaking time for the years 2006–2017 in the Cannes Lions archive of advertisements. Notice that men are portrayed predominantly in ads and that this trend has not changed in over a decade examined. Large scale analysis such as this can shed light on systemic representation disparities in important media such as ads.

compared to the years 2014–2016.

Men speak more often than women in top grossing Hollywood films. Similar to female screentime, movies over the past six years have a significantly higher speaking time for men compared to women. See Fig. 14B for detailed distributions.

There is a significant economic benefit for movies with larger female presence. One of the important goals of CMI is also to assess the economic impact of media. In the SeeJane project, we compared the female presence in movies with the US Box Office grossing of each film. The distribution of total grossing in USD for the different female screen time bins in shown in Fig. 15. Despite having fewer movies in the *more female* and *mostly female* categories, notice that movies with higher female presence tend to show better Box Office returns. Overall, we found that movies with diverse character representation attract large and diverse audience. While these automated tools serve an important purpose in offering useful insights about diversity in media, note that automated gender classification from audio-visual information e.g., from face or voice, is immensely complex both in terms of its conceptual framing (gender is a non-binary construct), and its assessment from measurable cues, by both human and machine observers (these privately-held social constructs are not fully observable). Hence it is critical to contextualize the design and development of such automated tools, including their limitations. The aforementioned case study for instance is restricted to the analysis of female/male representation, and for the purposes of understanding representation in movies. An understanding of portrayals of non-binary gendered characters as well as LGBTQ+ characters in the entertainment media (films, TV, streaming shows) is crucial to examining representation of broader minority demographics in media. Identifying such identities cannot be done using automated learning methods, hence these are done by manual expert annotation with access to other meta information. As an example, we refer the reader to the latest study report on this project at SeeJane2020[8]. CMI tools however can assist human experts in a collaborative fashion by providing character level computation of screen presence, talk times and interaction patterns.

8. https://seejane.org/research-informs-empowers/2020-film-historic-gender-parity-in-family-films

4.2 Longitudinal analysis of Cannes Lions Archive of advertisements

In another research study called *Unpacking gender bias in advertising*, we collaborated with GDIGM and J Walter Thompson to analyze over a decade of advertisements nominated to Cannes Lions Film Festival. We analyzed over 2000 advertisements submitted to the festival between 2006–2017 by automatically estimating measures of female screen time and speaking time, and also manually labeled perception measures of characters' appearance and portrayal of leadership roles in advertisements [77]. We highlight some of the key findings from this study here:

1) Over a decade of advertisements submitted to Cannes Lions Film Festival, the trend of female presence and portrayal has not changed. See Fig. 16.
2) Overall, about 70% of all ads have a female screentime of 50% or less. About 25% of the ads featured only men where as only 5% of the ads featured only women. These statistics were similar with respect to female speaking time as well.
3) Female characters are three times more likely to be verbally objectified than male characters in the sample studied.
4) Men are twice as likely as women to be shown as managers.
5) Analyzing the visual ambience context, we found that men are three times more likely than women to be shown in an office setting and women are twice as likely as men to be shown in a household setting.

Some of the perception measures studied here were labeled by expert annotators trained to look for the related traits of leadership and identify professions. Fully automating some of these measures, where well defined, offers new research opportunities for CMI alongside designing and building scalable tools to enable large humans or expert in the loop media analyses. Our analysis on the Cannes Lions archive also served as the framework for another recent study of over 2.7 Million ads [76]. This study also showed that commercials with *almost gender balanced* screentime/speaking time received 30% more views than commericals with either mostly male/female.

5 Summary: Challenges and Opportunities

Computational media intelligence (CMI) promises efficient, scalable and robust engineering analytic systems to enable detailed and nuanced characterization of media content. The crux of CMI is understanding the *what, who, how, where* and *why* from the multiple modalities in media, across its various forms and measure its impact on individuals and society. It has numerous applications to different stakeholders: from content creators and decision/policy makers to content curators, businesses and consumers. In particular, CMI offers supporting tools for creating awareness and change about diversity and inclusion in terms of fair representation and portrayals of people, places, organizations and other entities in the media. There are a number of technical challenges that need to be addressed in order to achieve such media intelligence capabilities.

Tremendous variety and variability. Media content is often an output of creative human processes at many stages of content creation. This makes it difficult to generalize modeling methodologies and make assumptions about data, even within a particular form of media or content genre. Additionally, many standard machine learning methods (e.g., speaker recognition for audio diarization i.e., *who* speaks *when*, video face detection and tracking) typically perform poorly on media data. Such errors may get inflated especially when used to derive higher level constructs (e.g., semantics, affective states or identity dimensions). For instance, it is challenging to generalize across cultures and time periods (e.g., media collected over time). Such analyses require complex social and cultural context to be incorporated within the models. Novel machine intelligence capabilities need to be developed to handle these requirements.

Lack of appropriate, labeled data. There is a general lack of appropriate labeled data for (supervised) learning from media content given the diversity in the descriptions that are desired (example: demographic information of movie characters including for the minor non-speaking characters). Furthermore, the now increasingly-acknowledged challenges of data bias in machine learning algorithms are especially critical in the media domain, and the inherent disparity in data can propagate into models. Another challenge is data provenance. For example copyrighted media content makes much of produced media content difficult to be annotated via scalable and less expensive efforts such as crowd sourcing methods due to distribution limitations. In addition, given the human-centric nature of media, we often contend with diverse, noisy and incomplete annotation as a proxy for human experience (e.g., movie reviews / surveys). A further fundamental challenge is the inherent subjectivity in deriving these constructs due to human variability and heterogeneity in modeling perception (experience) and action (behavior). New computing formalisms that can adequately address these challenges need to be developed.

Closing the loop with humans. Yet another hitherto problem that has not been completely solved for computational machine intelligence research is creating, measuring and influencing human experiences in predictable (causal) ways. This includes quantifying media impact and influence both at an individual level and at socio-cultural scales. Developing methodologies for modeling representation/context/interaction at scale with humans in the loop is another CMI research area which needs more development.

Machine learning fairness. A key building block of *scalable* media intelligence is the ability to automatically learn some of the identity, context or interaction attributes from media. Owing to the immense heterogeneity and variability in the media forms, we need the machine learning tools developed to be *fair* in terms of robustness of performance. For example, face detection must work regardless of the illumination of a movie shot or the cultural backgrounds of the people portrayed in it. Studying the intersection of the impact of robustness of the learning algorithms on the representations obtained is part of our ongoing work and needs much grounding in the context of developing a robust media intelligence.

TABLE 1
Long form media content datasets for computational media intelligence tasks. First part of the table shows the datasets released by our research group.

Dataset Name (year)	Description/Task	No. videos (hours)	Labels/Attributes	Label resolution	Labelling
Multi-face (2020) [57]	Must-link/cannot-link face tracks	240 movies (450)	Must-link and cannot-link face pairs	face-track level	Self-supervised
TV/Film Benchmark (2019,20) [57], [78]	Video character diarization	6 movies (11)	character ID, visual distractors	face-track level	Manual, 3 raters
Subtitle aligned movie audio (2019) [54]	Speech activity detection		Speech/non-speech	frame-level 0.64s	Self-supervised
Movie Audio Gender ID (2018) [21]	Speech gender (M/F) classification	4 (8)	M/F speech, non-speech	frame-level, 50ms	Manual, 1 expert
SAIL animation corpus (2017) [79]	Animated character identification		Character ID, non-character	track-level	Semi-automatic
AVA-Kinetics (2018) [80]	Person-centric action	230,000 (128)	Pose (1), interaction with object (upto 3), interaction with person (upto 3)	Clip-level	Manual, 3 raters
Condensed Movies (2020) [81]	Story understanding	3605 (1270)	Semantic description of clips	Clip-level	Semi-automatic
MovieNet (2020) [82]	Movie Understanding	1100	Subtitle, Genre, Cinematic style, Character bbox and ID, Action, Place, Scene boundary, synopsis alignment, Trailer	Clip-level	Semi-automatic
MovieScenes (2020) [83]	Scene Segmentation	150 (21K scenes)	Scene boundaries	Clip-level	Semi-automatic
MovieFace (2020) [84]	-	-	-	-	-
MovieShots (2020) [85]	Shot-type classification	7K trailers	5 (SS), 4 (SM)	Manual	Manual
Movie Synopses (2019) [86]	Movie Synopsis	327 (516)	Synopses paragraphs (IMDb)	Clip-level	Manual
MovieScope (2019) [87]	Multimodal Movie Trailer Analysis	5027 (195)	Trailers, Plots (Wikipedia/CMU MSC)	-	Automatic
AVA-Actions (2018) [80]	Person-centric action	430 (108)	Pose (1), interaction with object (upto 3), interaction with person (upto 3)	Frame-level, 1Hz	Manual, 3 raters
AVA-Speech (2018) [88]	Speech activity labels	185 (46)	Speech, Music, Noise	Frame-level	Manual, 3 raters
MovieGraphs (2018) [89]	Graph-based annotations of social situations in movies	7637	-	Clip-level	
LSMTD (2018) [90]	Movie Trailer Analysis	34K trailers (2200)	Genre, Plot keywords	Movie-level	Automatic
Cast in Movies (2018) [91]	Person recognition	192 (73K images)	Person bbox, Character ID	Images	Manual
Movie Cast Search (2018) [92]	Person Search in Videos	192 (127K tracklets)	Person tracklets, Character ID	Clip-level	Manual
MovieFIB (2018) [93]	Question Answering	180 (135)	Fill-in-the-blank Q&A	Clip-level	Manual
MovieQA (2018) [94], [95], [96], [97]	Question Answering	408	14944 questions	Manual	
Hollywood-2-Tubes (2016) [98]	Action Localization	32	Point annotations for action	Clip-level	Semi-automatic
MGCD (2010) [99]	Genre Classification	1239	Genre	Movie-level	Automatic
LMTD (2016) [100], [101]	Multilabel Genre Classification in Trailers	3500	Genre	Movie-level	Automatic
MPII Dataset (2015) [102]	Movie description	72 (56)	Sentence description	Movie-level	Semi-automatic
Accio (2015) [103]	Aging from video	8 (38.5K tracks)	Age (10-88yrs) for 121 characters	Clip-level	Manual
Casablanca (2013) [104]	Actor and Action Identification	1273 facetracks (1.5)	Character ID, Action	Clip-level	Semi-automatic
Hollywood Movie VAD (2013) [105]	Voice activity detection	4 (8)	speech/non-speech	frame-level	Manual
Hollywood-2 (2008) [106]	Action Recognition	32	Human action	Clip-level	Automatic

TABLE 2
Short form and TV media content datasets for CMI tasks

Dataset Name	Description	No. videos (hours)	Labels/Attributes	Label resolution	Labeling
TVSeries	Online Action Recognition	27 (16)	Action	Frame-level	Manual, 2 raters
TVQA	Question Answering	925 (450)	Question-Answers in single correct choice format	Clip-level	Manual
PororoQA	Question Answering	171 (20.5)	Question-Answers	Clip-level	Manual
DramaQA	Question Answering	18 (20.5)	Question-Answers	Clip-level	Manual
REPERE Corpus	Person Recognition	7 (6)	Head segmentation, Head description, People identification, Speaker turn segmentation, Speaker naming, Rich speech transcription	Clip-level	
Person Id	Person Recognition	6	Face Tracks, Person Tracks	Clip-level	Manual
BBT/Buffy Facetracks (6)	Face Tracking	12 (6)	Face Tracks, Speaking Tracks	Clip-level	Manual
MELD	Emotion Recognition in multiparty conversations		Multi-class emotion classification	Utterance level	Manual, 5 raters
CAER	Context based emotion recognition	13201 clips	Multi-class emotion classification	Clip level	Manual , 6 raters
Sherlock-BBC	Character recognition in unconstrained videos	2 (2)	Face detections, tracks, shots Character labels	Clip level	

REFERENCES

[1] A. Watson, "Topic: Movie industry." [Online]. Available: https://www.statista.com/topics/964/film/

[2] J. Simpson, "Council post: Finding brand success in the digital world," Aug 2017. [Online]. Available: https://www.forbes.com/sites/forbesagencycouncil/2017/08/25/finding-brand-success-in-the-digital-world/#10f71a4e626e

[3] J. Clement, "Most used social media platform," Aug 2020. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[4] J. Hale, "More than 500 hours of content are now being uploaded to youtube every minute," May 2019. [Online]. Available: https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/

[5] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[6] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. Van Gool, "Large scale holistic video understanding," *arXiv preprint arXiv:1904.11451*, 2019.

[7] A. Ramakrishna, V. R. Martínez, N. Malandrakis, K. Singla, and S. Narayanan, "Linguistic analysis of differences in portrayal of movie characters," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1669–1678.

[8] J. Cohen, "Defining identification: A theoretical look at the identification of audiences with media characters," *Mass communication & society*, vol. 4, no. 3, pp. 245–264, 2001.

[9] S. Sierra, "Linguistic and ethnic media stereotypes in everyday talk: Humor and identity construction among friends," *Journal of Pragmatics*, vol. 152, pp. 186–199, 2019.

[10] P. L. Hammack, "Theoretical foundations of identity." 2015.

[11] D. Papadopoulos, "In the ruins of representation: Identity, individuality, subjectification," *British Journal of Social Psychology*, vol. 47, no. 1, pp. 139–165, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1348/014466607X187037

[12] J. Côté, "Identity studies: How close are we to developing a social science of identity?—an appraisal of the field," *Identity*, vol. 6, no. 1, pp. 3–25, 2006.

[13] D. Popa and D. Gavriliu, "Gender representations and digital media," *Procedia-Social and Behavioral Sciences*, vol. 180, pp. 1199–1206, 2015.

[14] K. Deaux, "Reconstructing social identity," *Personality and social psychology bulletin*, vol. 19, no. 1, pp. 4–12, 1993.

[15] P. Arredondo and T. Glauner, "Personal dimensions of identity model." Boston: Empowerment Workshops, 1992.

[16] B. Golomb, D. Lawrence, and T. Sejnowski, "Sexnet: A neural network identifies sex from human faces," in *Advances in Neural Information Processing Systems 3*, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Eds. Morgan-Kaufmann, 1991, pp. 572–577. [Online]. Available: http://papers.nips.cc/paper/405-sexnet-a-neural-network-identifies-sex-from-human-faces.pdf

[17] E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541–547, 2008.

[18] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2483–2509, 2014.

[19] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images," *Computer vision and image understanding*, vol. 74, no. 1, pp. 1–21, 1999.

[20] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.

[21] R. Hebbar, K. Somandepalli, and S. S. Narayanan, "Improving gender identification in movie audio using cross-domain data." in *INTERSPEECH*, 2018, pp. 282–286.

[22] T. Guha, C.-W. Huang, N. Kumar, Y. Zhu, and S. S. Narayanan, "Gender representation in cinematic content: A multimodal approach," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 31–34.

[23] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[24] D. Bordwell and K. Thompson, *Film Art: An Introduction*. Boston: McGraw-Hill Print, 2004.

[25] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–37, 2019.

[26] A. Vedaldi, A. Zisserman, H. Chen, and W. Xie, "Vggsound: a large-scale audio-visual dataset," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2020.

[27] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 11, pp. 1993–2008, 2013.

[28] Wikipedia contributors, "Star wars — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Star_Wars&oldid=975355258, 2020, [Online; accessed 29-August-2020].

[29] E. Akleman, S. Franchi, D. Kaleci, L. Mandell, T. Yamauchi, D. Akleman *et al.*, "A theoretical framework to represent narrative structures for visual storytelling," *proceedings of bridges 2015: mathematics, Music, art, architecture, culture*, 2015.

[30] P. Kulshreshtha and T. Guha, "Dynamic character graph via online face clustering for movie analysis," p. (to appear), 2020.

[31] T. Giannakopoulos, S. Dimopoulos, G. Pantazopoulos, A. Chatziagapi, D. Sgouropoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Using oliver api for emotion-aware movie content characterization," in *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, 2019, pp. 1–4.

[32] N. Malandrakis, "Affect extraction using aural, visual and linguistic features from multimedia documents," 2012.

[33] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[34] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka, "Automatic understanding of image and video advertisements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1705–1715.

[35] K. Somandepalli, V. Martinez, N. Kumar, and S. Narayanan, "Multimodal representation of advertisements using segment-level autoencoders," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 418–422.

[36] M. Schedi, M. Sjöberg, I. Mironică, B. Ionescu, V. L. Quang, Y.-G. Jiang, and C.-H. Demarty, "Vsd2014: a dataset for violent scenes detection in hollywood movies and web videos," in *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2015, pp. 1–6.

[37] Y. Baveye, "Automatic prediction of emotions induced by movies," Ph.D. dissertation, Ecole Centrale de Lyon, 2015.

[38] V. R. Martinez, K. Somandepalli, K. Singla, A. Ramakrishna, Y. T. Uhls, and S. Narayanan, "Violence rating prediction from movie scripts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 671–678.

[39] L. Bergkvist and J. R. Rossiter, "The role of ad likability in predicting an ad's campaign performance," *Journal of advertising*, vol. 37, no. 2, pp. 85–98, 2008.

[40] S. Baruah, R. Gupta, and S. Narayanan, "A knowledge transfer and boosting approach to the prediction of affect in movies," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2876–2880.

[41] A. Goyal, N. Kumar, T. Guha, and S. S. Narayanan, "A multimodal mixture-of-experts model for dynamic emotion prediction in movies," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2822–2826.

[42] A. Tadimari, N. Kumar, T. Guha, and S. S. Narayanan, "Opening big in box office? trailer content can help," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2777–2781.

[43] R. Sharma, T. Guha, and G. Sharma, "Multichannel attention network for analyzing visual behavior in public speaking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 476–484.

[44] B. J. Bushman and L. R. Huesmann, "Short-term and long-term effects of violent media on aggression in children and adults," *Archives of Pediatrics & Adolescent Medicine*, vol. 160, no. 4, pp. 348–352, 2006.

[45] R. Hanewinkel and J. D. Sargent, "Longitudinal study of exposure to entertainment media and alcohol use among german adolescents," *Pediatrics*, vol. 123, no. 3, pp. 989–995, 2009.

[46] C. J. Ferguson, A. Copenhaver, and P. Markey, "Reexamining the findings of the american psychological association's 2015 task force on violent media: A meta-analysis," *Perspectives on Psychological Science*, vol. 0, no. 0, p. 1745691620927666, 0, pMID: 32777188. [Online]. Available: https://doi.org/10.1177/1745691620927666

[47] K. Kim, J. L. Hayes, J. A. Avant, and L. N. Reid, "Trends in advertising research: A longitudinal analysis of leading advertising, marketing, and communication journals, 1980 to 2010," *Journal of advertising*, vol. 43, no. 3, pp. 296–316, 2014.

[48] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[49] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[50] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy, R. Vaddi, V. Hemadri, J. Karure, R. Raju, B. Rajan *et al.*, "Indian movie face database: a benchmark for face recognition under wide variations," in *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)*. IEEE, 2013, pp. 1–5.

[51] Z. Li, "The "celeb" series: A close analysis of audio-visual elements in 2008 us presidential campaign ads," 2017.

[52] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[53] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[54] R. Hebbar, K. Somandepalli, and S. Narayanan, "Robust speech activity detection in movie audio: Data resources and experimental evaluation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4105–4109.

[55] K. Kärkkäinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age," *arXiv preprint arXiv:1908.04913*, 2019.

[56] H. K. Wong, I. D. Stephen, and D. R. Keeble, "The own-race bias for face recognition in a multiracial society," *Frontiers in Psychology*, vol. 11, p. 208, 2020.

[57] K. Somandepalli, R. Hebbar, and S. Narayanan, "Multi-face: Self-supervised multiview adaptation for robust face clustering in videos," 2020.

[58] R. Hebbar, K. Somandepalli, and S. Narayanan, "Robust speech activity detection in movie audio: Data resources and experimental evaluation," in *Proceedings of ICASSP*, May 2019.

[59] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[60] C. Staff and Marketing, "Improvement toward inclusion in film, but more work to be done," Sep 2019. [Online]. Available: https://annenberg.usc.edu/news/research-and-impact/improvement-toward-inclusion-film-more-work-be-done

[61] M. Ryzik, "How long is an actress onscreen? a new tool finds the answer faster." Sep 2016. [Online]. Available: https://www.nytimes.com/2016/09/15/movies/geena-davis-inclusion-quotient-research.html?_r=1

[62] N. Kumar, M. Nasir, P. G. Georgiou, and S. S. Narayanan, "Robust multichannel gender classification from speech in movie audio." in *Interspeech*, 2016, pp. 2233–2237.

[63] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[64] R. Sharma, K. Somandepalli, and S. Narayanan, "Toward visual voice activity detection for unconstrained videos," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2991–2995.

[65] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[66] B. Joosten, E. Postma, and E. Krahmer, "Visual voice activity detection at different speeds," in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[67] R. Sharma, K. Somandepalli, and S. Narayanan, "Toward visual voice activity detection for unconstrained videos," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2991–2995.

[68] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[69] R. Sharma, K. Somandepalli, and S. Narayanan, "Crossmodal learning for audio-visual speech event localization," *arXiv preprint arXiv:2003.04358*, 2020.

[70] T. Guha, N. Kumar, S. S. Narayanan, and S. L. Smith, "Computationally deconstructing movie narratives: an informatics approach," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 2264–2268.

[71] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "Rolenet: Movie analysis from the perspective of social networks," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.

[72] P. Kulshreshtha and T. Guha, "Dynamic character network via online face clustering for movie analysis," in *Multimedia Tools and Applications*, 2020, p. to appear.

[73] T. Greer, B. Ma, M. Sachs, A. Habibi, and S. Narayanan, "A multimodal view into music's effect on human neural, physiological, and emotional experience," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 167–175.

[74] V. R. Martinez, K. Somendapalli, K. Singla, A. Ramanakrishna, Y. T. Uhls, and S. Narayanan, "Victim or perpetrator? analysis of violent characters portrayals from movie scripts," *arXiv preprint arXiv:2008.08225*, 2020.

[75] M. Sap, M. C. Prasettio, A. Holtzman, H. Rashkin, and Y. Choi, "Connotation frames of power and agency in modern films," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2329–2334.

[76] "Diversity and inclusivity report: Gender in youtube advertising - think with google." [Online]. Available: https://www.thinkwithgoogle.com/feature/diversity-inclusion/

[77] "The geena davis institute on gender in media and j. walter thompson present revealing findings about women's representation in advertising at cannes lions," Sep 2018. [Online]. Available: shorturl.at/aePQX

[78] K. Somandepalli and S. Narayanan, "Reinforcing self-expressive representation with constraint propagation for face clustering in movies," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4065–4069.

[79] K. Somandepalli, N. Kumar, T. Guha, and S. S. Narayanan, "Unsupervised discovery of character dictionaries in animation movies," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 539–551, 2017.

[80] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[81] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," 2020.

[82] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "Movienet: A holistic dataset for movie understanding," in *The European Conference on Computer Vision (ECCV)*, 2020.

[83] A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin, "A local-to-global approach to multi-modal movie scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 146–10 155.

[84] Q. Huang, L. Yang, H. Huang, T. Wu, and D. Lin, "Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation," in *The European Conference on Computer Vision (ECCV)*, 2020.

[85] A. Rao, J. Wang, L. Xu, X. Jiang, Q. Huang, B. Zhou, and D. Lin, "A unified framework for shot type classification based on subject centric lens," in *The European Conference on Computer Vision (ECCV)*, 2020.

[86] Y. Xiong, Q. Huang, L. Guo, H. Zhou, B. Zhou, and D. Lin, "A graph-based framework to bridge movies and synopses," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[87] P. Cascante-Bonilla, K. Sitaraman, M. Luo, and V. Ordonez,

"Moviescope: Large-scale analysis of movies using multiple modalities," *ArXiv*, vol. abs/1908.03180, 2019.
[88] S. Chaudhuri, J. Roth, D. P. Ellis, A. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. G. Reid, K. Wilson *et al.*, "Ava-speech: A densely labeled dataset of speech activity in movies," *arXiv preprint arXiv:1808.00606*, 2018.
[89] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler, "Moviegraphs: Towards understanding human-centric situations from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8581–8590.
[90] Q. Huang, Y. Xiong, Y. Xiong, Y. Zhang, and D. Lin, "From trailers to storylines: An efficient way to learn from movies," *arXiv preprint arXiv:1806.05341*, 2018.
[91] Q. Huang, Y. Xiong, and D. Lin, "Unifying identification and context learning for person recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2217–2225.
[92] Q. Huang, W. Liu, and D. Lin, "Person search in videos with one portrait through visual and temporal links," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 425–441.
[93] T. Maharaj, N. Ballas, A. Rohrbach, A. Courville, and C. Pal, "A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6884–6893.
[94] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[95] A. Rohrbach, M. Rohrbach, and B. Schiele, "The long-short story of movie description," in *German Conference on Pattern Recognition (GCPR)*, 2015.
[96] A. Rohrbach, M. Rohrbach, S. Tang, S. J. Oh, and B. Schiele, "Generating descriptions with grounded and co-referenced people," in *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017.
[97] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, 2017. [Online]. Available: http://resources.mpi-inf.mpg.de/publications/D1/2016/2310198.pdf
[98] P. Mettes, J. C. Van Gemert, and C. G. Snoek, "Spot on: Action localization from pointly-supervised proposals," in *European conference on computer vision*. Springer, 2016, pp. 437–453.
[99] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie genre classification via scene categorization," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 747–750.
[100] G. S. Sim
oes, J. Wehrmann, R. C. Barros, and D. D. Ruiz, "Movie genre classification with convolutional neural networks," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 259–266.
[101] J. Wehrmann and R. C. Barros, "Movie genre classification: A multi-label approach based on convolutions through time," *Applied Soft Computing*, 2017.
[102] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.
[103] E. Ghaleb, M. Tapaswi, Z. Al-Halah, H. K. Ekenel, and R. Stiefelhagen, "Accio: A data set for face track retrieval in movies across age," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 455–458.
[104] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Finding actors and actions in movies," in *Proc. ICCV*, 2013.
[105] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 483–487.
[106] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

Krishna Somandepalli Krishna Somandepalli received his Masters degree from University of California at Santa Barbara, CA, USA in Electrical and Computer Engineering. Following his Masters degree, he worked as an assistant research scientist at NYU Langone medical Center, New York, NY, USA. His research interests are in multimodal analysis with image and signal data. Currently, he is a PhD candidate in the Signal Analysis and Interpretation Laboratory (SAIL) group at the department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA.

Tanaya Guha Tanaya Guha is an Asst. Professor in the Dept. of Computer Science at the University of Warwick, UK. Prior to Warwick, she was an Asst. Professor at IIT Kanpur and a postdoc at the Signal Analysis and Interpretation Lab (SAIL), University of Southern California (USC). Her research is focused on building computational models to understand, recognise and predict human behavior from multimodal data with applications to media, health and security. She received her PhD from the University of British Columbia (UBC), Vancouver. She was a recipient of Mensa Canada Woodhams memorial scholarship, Google Canada Anita Borg memorial scholarship, and Amazon Grace Hopper celebration scholarship among other awards and grants. She has served in the program committee of several conferences including ACM MM, ACM ICMI, ACII and Interspeech.

Naveen Kumar Naveen Kumar is a Research Scientist with Disney Research, Los Angeles. He received his Ph.D. in Electrical Engineering from the USC Viterbi School of Engineering, where he was a member of the Media Informatics and Content Analysis (MICA) group at the Signal Analysis and Interpretation Lab (SAIL). He received his B.Tech. degree in Instrumentation Engineering from the Indian Institute of Technology, Kharagpur in 2009. His broad research interests include machine learning and signal processing for speech, affective computing, multimedia and multimodal applications.

Hartwig Adam Hartwig Adam is an engineering director at Google Research in Los Angeles, where he is leading a group focused on developing computer vision and machine perception systems for mobile devices and Google's internal and external developer platforms. Previous projects include Lens, Photo Search, Visual Search Infrastructure, Glass and Google Goggles. Prior to Google, Hartwig was a co-founder and VP of Platform Development at Neven Vision, an early pioneer in mobile computer vision.

Shrikanth (Shri) Narayanan (StM'88–M'95–SM'02–F'09) is University Professor and Niki & C. L. Max Nikias Chair in Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical and Computer Engineering, Computer Science, Linguistics, Psychology, Neuroscience, Otolaryngology and Pediatrics, Research Director of the Information Science Institute, and director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC, he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biosignal processing and machine intelligence, and their applications with direct societal relevance. [http://sail.usc.edu]

Prof. Narayanan is a Fellow of the National Academy of Inventors, the Acoustical Society of America, IEEE, the International Speech Communication Association (ISCA), the Association for Psychological Science, the American Institute for Medical and Biological Engineering (AIMBE), and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is VP-Education, IEEE Signal Processing Society (2020-), an Editor for the Computer Speech and Language Journal and an Associate Editor for the APSIPA TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING. He was also previously Editor in Chief for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and an Associate Editor of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING (2000–2004), IEEE SIGNAL PROCESSING MAGAZINE (2005–2008), IEEE TRANSACTIONS ON MULTIMEDIA (2008-2011), the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS (2014-2015), IEEE TRANSACTIONS ON AFFECTIVE COMPUTING (2010-2016), and the Journal of the Acoustical Society of America (2009-2017). He is a recipient of several honors including the 2015 Engineers Council's Distinguished Educator Award, a Mellon award for mentoring excellence, the 2005 and 2009 Best Journal Paper awards from the IEEE Signal Processing Society and serving as its Distinguished Lecturer for 2010-11, a 2018 ISCA Best Journal Paper award, and serving as an ISCA Distinguished Lecturer for 2015-16 and the Willard R. Zemlin Memorial Lecturer for ASHA in 2017. Papers co-authored with his students have won awards including the 2020 Sustained Accomplishment Award and the 2014 Ten-year Technical Impact Award from ACM ICMI and at several conferences. He has published over 900 papers and has been granted seventeen U.S. patents.