

## Taking a Cue From the Human: Linguistic and Visual Prompts for the Automatic Sequencing of Multimodal Narrative

 Kim Starr 

University of Surrey

 Sabine Braun 

University of Surrey

 Jaleh Delfani 

University of Surrey

---

### Abstract

Human beings find the process of narrative sequencing in written texts and moving imagery a relatively simple task. Key to the success of this activity is establishing coherence by using critical cues to identify key characters, objects, actions and locations as they contribute to plot development.

In the drive to make audiovisual media more widely accessible (through audio description), and media archives more searchable (through content description), computer vision experts strive to automate video captioning in order to supplement human description activities. Existing models for automating video descriptions employ deep convolutional neural networks for encoding visual material and feature extraction (Krizhevsky, Sutskever, & Hinton, 2012; Szegedy et al., 2015; He, Zhang, Ren, & Sun, 2016). Recurrent neural networks decode the visual encodings and supply a sentence that describes the moving images in a manner mimicking human performance. However, these descriptions are currently “blind” to narrative coherence.

Our study examines the human approach to narrative sequencing and coherence creation using the MeMAD [Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human

Citation: Starr, K. & Braun, S. & Delfani, J. (2020). Taking a Cue From the Human: Linguistic and Visual Prompts for the Automatic Sequencing of Multimodal Narrative. *Journal of Audiovisual Translation*, 3(2), 140–169.

**Editor(s):** A. Matamala & J. Pedersen

**Received:** March 02, 2020

**Accepted:** July 10, 2020

**Published:** December 18, 2020


**Funding:** This publication is part of the EU funded project MeMAD, grant agreement number 780069.

**Copyright:** ©2020 Starr & Braun & Delfani. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/). This allows for unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

---

 k.starr@surrey.ac.uk, <https://orcid.org/0000-0001-5236-1535>

 s.braun@surrey.ac.uk, <https://orcid.org/0000-0002-6187-3812>

 j.delfani@surrey.ac.uk, <https://orcid.org/0000-0003-2075-3539>

Accuracy] film corpus involving five-hundred extracts chosen as stand-alone narrative arcs. We examine character recognition, object detection and temporal continuity as indicators of coherence, using linguistic analysis and qualitative assessments to inform the development of more narratively sophisticated computer models in the future.

**Key words:** computer vision, machine learning, accessibility, audiovisual content, audio description, content description, content retrieval, video description, audiovisual translation, MeMAD, automatic captioning.

## 1. Introduction: Computer Vision and Storytelling

The quest to reduce the workload of human audio describers while broadening the range of video description services across a burgeoning range of multimedia content platforms, as well as archival material, has led to an increased interest in improving methods for automating moving image description (Bai & An, 2018). A number of competing issues compound the challenges of this task, most of which fall into the dual realms of computer vision and language processing. As Bai & An (2018) note, “[...] determination of presences, attributes and relationships of objects in an image is not an easy task itself. Organising a sentence to describe such information makes this task even more difficult” (p. 291). The reliability and accuracy of automatically generated image descriptions remains problematic even in the particular case of standard photo stills captioning (Huang et al., 2016; Krishna et al., 2017; Smilevski, Lalkovski, & Madjarov, 2018), where obscure angles and non-iconic representations of standard objects continue to confound the machine (Husain & Bober, 2016). Automating descriptions of video material such as film or television footage creates a far greater technical challenge (Xu, Mei, Yao, & Rui, 2016; Rohrbach, Rohrbach, Tang, Oh, & Schiele, 2017). There is a dearth of large training datasets to progress this task, most notably those comprising video materials with complex narrative story grammars (Mandler, 1978; Mandler & Johnson, 1980) played out through the integration of a range of audio and video inputs (image, dialogue, sound effects, music, cinematography etc.).

Current machine learning models are applied to the automation of moving image descriptions based on deep convolutional neural networks for the purposes of visual input encoding and feature extraction (Krizhevsky, Sutskever, & Hinton, 2012; Szegedy et al., 2015; He, Zhang, Ren & Sun, 2016). Recurrent neural networks, such as Long Short-Term Memory (LSTM), are used to decode these visual encodings and return a sentence that describes the multimedia content in an approximation of human captioning behaviours (Hochreiter & Schmidhuber, 1997). Although reinforcement learning (Ren, Wang, Zhang, Lv, & Li, 2017), adversarial learning and adversarial inference (Park, Rohrbach, Darrell, & Rohrbach, 2019) have been used to enhance the current captioning performance, the results are still broadly unreliable. Delivering moving image descriptions at a level of narrative sophistication that exceeds simple object-action labelling is therefore a major challenge. Wider availability of large-scale open access training data and improvements to the quality of human captioning in relation to these images are likely to be rewarded with more accurate results. However, sequencing descriptions into a cohesive, linear plot requires an understanding and interpretation of cues and prompts, which is currently only within the scope of human beings. Consequently, one avenue which has recently begun to be explored is consideration of human approaches to moving image description, and how these might be analysed and subsequently harnessed to inform the development of future machine learning models (Braun & Starr, 2019; Braun, Starr, & Laaksonen, 2020).

This paper reports on one aspect of that approach, undertaken as part of the EU-funded MeMAD (*Methods for Managing Audiovisual Data*) project, involving investigations into the patterns that human beings use in detecting and assimilating audiovisual storylines by means of narrative sequencing. It considers coherence-seeking and inferencing skills from a human perspective, exploring the way visual and linguistic prompts are conjoined to create stories from moving images. Examples are drawn from the MeMAD corpus of 500 film extracts, which includes different types of human description (e.g. audio description) along with associated machine-generated video descriptions (“captions”), and was created specifically to compare human and machine approaches to interpreting short-form narrative. We illustrate deficiencies in the computer vision model and explore avenues for enhancing the machine’s performance by replicating sequencing behaviours that are typical of sapient beings.

## 2. Background: Sequencing Film Narrative

One of the main difficulties with producing machine-generated captions is that the models and algorithms used in their production draw upon large-scale training data in order to learn the required behaviours, primarily, object recognition and feature extraction. While the datasets openly accessible to computer vision researchers—e.g. Microsoft Common Objects in Context (MS COCO) (Lin et al., 2015), TGIF [Tumblr GIF] (Li et al., 2016), Visual Genome (Krishna et al., 2017), MPII-MD [Max Planck Institut Informatik II Movie Description Dataset], Hollywood II (Rohrbach et al., 2015; Rohrbach et al., 2017)—are scaled appropriately, the crowdsourced nature of the captions and the banks from which images are drawn often result in topic bias and inaccuracies of description (Braun & Starr, 2019). Further confounding matters, many of these datasets describe still images and groups of related still images, or alternatively very limited moving image sequences containing simple actions. None of them offers data of a level of sophistication approximating the moving images found in film and television presentations.

While many of the simpler tasks involved in moving image description remain either unreliable or elusive, e.g. object recognition, action detection and character naming, the ultimate goal in the drive to automate multimodal captioning remains to achieve narratively coherent, contextually sensitive and fully sequenced computer-generated storytelling (Venugopalan et al., 2015; Smilevski et al., 2018). In order to reach this point, continuity of character identification and naming would need to be established between shot- and scene-changes, taking into account variations of camera angle, cinematographic staging and mise-en-scène, and overcoming confounds such as changing appearances (e.g. differences in costume, hair styles, body profile etc.). Object tracking, and an understanding of the relationship between multiple objects, or objects and characters, would be essential to the provision of continuity and the development of sequenced storytelling (Krishna et al., 2017). Moreover, *temporal sequencing* (Yao et al., 2015) takes on particular relevance both in terms of denoting the chronology of plot as “fabula”, and in defining the general shape of the story, or how it is told, the “syuzhet” (Propp, 1958; Shklovsky, 1965/1990). While temporal words occur in

the vernacular of computer-generated captions, they have a spurious relationship with the types of temporal words human beings draw upon when conveying narrative coherence.

In order to represent plot as occurring within the context of a temporal continuum, a variety of referential expressions including pronominalisation cues are used in both written and spoken narrative to avoid unnecessary repetition. Computer-generated descriptions, by contrast, fail to apply *personal pronouns* in a meaningful way; indeed, pronouns are only used rarely, and then within the context of a single video frame caption. Across-frame captioning is outside the computer's capacity, hence nominal and pronominal cohesion do not occur at the scene-wide level, with the result that many of the simpler clues to continuity of action are absent. Nor is the machine currently trained to seek *relevance* in meaning-making (Sperber & Wilson, 1995), with the result that the saliency of audio and visual narrative cues is not measurable, leaving literal descriptions of "object-subject-action" as the only available pathway.

In all the above aspects of multimodal storytelling the human being, as a consumer of audiovisual material, will be at a considerable advantage in the drive to comprehend the breadth and depth of a storyline from the perspective of narrative *coherence*. Although historically, studies in narrative coherence focused on semantic units connected by cohesive ties that surpassed grammatical structure in written texts (Halliday & Hasan, 1976), it has been proposed that similar ties exist across multimodal materials (Forceville, 2014; Yus, 2008). For example in the case of audio description, source-text understanding on the part of the audio describer occurs at the intersection between moving images, dialogue and non-verbal diegetic sound, assimilating information and cues, and generating material which might be considered the 'glue' that creates coherence across modalities (Braun, 2011, p. 650). Where AD is not present, the sighted viewer still performs a similar multimodal comprehension task.

For example, considering a scene from *Little Miss Sunshine* (Dayton & Faris, 2006) where a teenage boy and his young sister resolve a dispute by sitting together silently in the middle of a field, human audiences would generally choose to focus their attention on the two principal characters at the centre of the scene (the youth, Dwayne, and the young girl, Olive). Accordingly, in the "content descriptions" that were created as ground truth for the experimental audiovisual corpus in the MeMAD project, this was rendered: "Dwayne is sitting on the grass in a field, hugging his knees. He is sitting with his back to us", followed by "Olive walks towards Dwayne, who is sitting on the ground, staring at the grass". Our describers chose to mention the location, since two young people sitting in the middle of a field represents an unusual situation for a dispute to take place. They also appreciated that the silence between the two protagonists signified a moment of emotional intensity and poignancy that was narratively significant, and likely to move the story forwards. By contrast, the computer descriptions of this scene ("a man is sitting in a field" and "a man and a woman are talking to each other"), although they mention the location, fail to depict the relationship between the two individuals and the essence of their interaction at this point in the film adequately.

This brief summary provides a window on the fundamental differences between human and machine understanding of complex film narrative, suggesting that there is *value in comparing human and machine approaches* and drawing on human knowledge when looking to train deep learning models for the purposes of improving automated computer vision. Sensitivity to nuanced social situations and the ability to discern satire, sarcasm, and other non-literal interpretations is, at present, a uniquely human capability. A first step towards improving the current computer offering is to improve object and action recognition, but beyond that, the challenge is to move towards an understanding of the “choices” the computer makes and, within the limits imposed by the “black box” nature of machine learning models and convolutional neural networks, begin to address the levels of interpretation the machine presently lacks. Analysis of human behaviours has the potential to open up computer vision to new research pathways, and it is upon this basis that the study reported in this paper was conducted.

### 3. Methodology: A Corpus-Based Approach

Our comparison of human and machine-generated descriptions of audiovisual material was conducted with the aim of a) understanding the structure and limitations of current machine-generated descriptions, and b) identifying aspects and strategies of human comprehension of multimodal material which can inform automated approaches.

To facilitate the comparative analysis, we developed a *multimodal parallel corpus* consisting of audiovisual source materials and different types of human and machine-generated textual description. The audiovisual component comprised a set of 500 video extracts (“*micro-narratives*”) taken from 45 feature films. The audiovisual material was aligned with three types of textual description: (i) content descriptions (CD) were created by annotators to represent our “ground truth”, i.e. a brief summary of the narrative action as it occurred in each extract (“say what you see”); (ii) audio descriptions (AD) and dialogue were transcribed from the source materials (DVD); and (iii) machine descriptions (MD), a form of “video captions”, were generated by applying the University of Aalto’s *DeepCaption* model (Sjöberg, Tavakoli, Xu, Mantecón, & Laaksonen, 2018) and using two large-scale open access datasets for visual object recognition as training data, i.e. MS COCO (Lin et al., 2015) and TGIF (Li et al., 2016)—a combination referenced as the “dc-a3” model. In addition to the training data, Aalto’s *DeepCaption* software exploited the combined aspects of recurrent neural networks for object identification and convolutional neural networks for caption generation.

The resulting corpus is a specific type of parallel corpus, namely a unidirectional “target variant” corpus (Merkel, 1999), here with the different target text versions (AD, CD and MD) created from the same audiovisual source material through intersemiotic (image-to-text) translation. Basic information about the resulting sub-corpora is shown in Table 1. The paucity of lexical variation

within the MD corpus is striking (type-token ratio .008), with AD registering 6.84 times as lexically rich, and CD 5.27 times more diverse.

Table 1.

*Basic Corpus Information*

	AD	CD	MD
Word tokens	25,039	43,829	70,315
Types	3,969	3,061	580
<i>Type-Token Ratio</i>	0.158	0.067	0.008
Lemmas	3,108	2,356	518
Sentences	2,524	4,892	7,067

Two types of human description, i.e. AD and CD, were included, because they were regarded as having complementary benefits. The AD transcripts provide authenticity as they illustrate the strategies of intersemiotic translation employed by professional describers although, since they must fit into the audio gaps of the original dialogue, they are an incomplete rendition of mainly visual markers. The CD corpus was created specifically for our project, having more comprehensive (written) descriptions without the restriction of needing to fit into these hiatuses. They might therefore be considered a more reliable “ground truth” against which the validity of the machine descriptions can be measured. The differences between the two corpora (AD and CD) are reflected in their respective sizes (Table 1). The content descriptions were created by three independent annotators and passed to the main researcher for review to ensure consistency of descriptive/narrative style and in levels of granularity. All human descriptive texts were normalised for consistency in rendering aspects such as non-verbal utterances, abbreviated text, numeration, sound effects and other non-verbal audio elements.

The MD subcorpus contains multiple captions for each of the 500 clips, with one caption being generated by the machine at each computer-detected shot change. This means that the computer model is not applied to moving images per se, but operates on the basis of describing a single frame at a time (in our corpus, the middle frame of a shot), each of which is described in isolation from the remaining imagery and any associated context. As was pointed out earlier, the quality of the resulting captions is largely dependent on the quality of the image descriptions contained in the training data from which they are sourced, and on model feature extraction. Equally as important, by comparison with human multimodal comprehension, which integrates all sources of input to create a coherent



model of the plot, the computer model algorithms applied in this study are entirely focused on images, i.e. they do not “listen” to the audio tracks.

The decision to generate one description per shot resulted in the MD corpus being a larger-sized corpus than the two human description corpora. The reason is likely to be that human describers create summary descriptions based on units of meaning which transcend shot changes. By contrast, the MD corpus contains a relatively large amount of repetition. However, we decided not to reduce the frequency of the machine descriptions because of an observation in the pilot phase, namely that small differences between two video frames can trigger rather dramatic changes in the video captions, a phenomenon which we intend to explore further. An example is given below, in which a small change in facial angle has resulted in a different interpretation of both the character’s gender and the activity being undertaken.

Figure 1.

#### Differences in Machine Description Between Similar Video Frames



*A man is driving a car and looking at his watch*



*A woman is driving a car and smiling*

During the multi-layered approach to corpus creation, a number of software packages for aligning and analysing multimodal corpora were tested. The aim was to find a package that would enable us to align the multiple target variants simultaneously with the video clips, to allow for direct comparisons to be drawn. However, the identification of suitable software tools turned out to be one of the key challenges. None of the multimodal software packages tested to date (MaxQDA, GATE, ELAN amongst others) matched our exacting requirements for multimodal analysis. In particular, they did not support alignment of three target text variants with the same audiovisual source material, and/or their respective interfaces were too inflexible for viewing a video clip and its three descriptions simultaneously. Nevertheless, in relation to individual film extracts, we were able to run the isolated machine description (.ass) file as a set of subtitles on top of the mp4 video within the VLC



media player platform. In addition, the textual data were ingested into an established corpus analysis tool, the *SketchEngine* (Kilgarriff et al., 2014), which supports alignment of multiple subcorpora and exports them in XML format. Video clips are matched with the relevant corpus segments via encoded clip identities (IDs).

With regard to the *data analysis*, our strategy was to commence with a quantitative analysis of human and machine-generated descriptions using corpus linguistics tools and techniques, followed by a qualitative exploration to corroborate, enrich and—equally important—challenge the quantitative findings. Our overall aim was to paint a more comprehensive picture of the descriptions than a quantitative analysis alone might have revealed. The combination of quantitative and qualitative methods is also commensurate with the small size of our corpus. By combining quantitative, corpus-based inspection and qualitative, discourse-oriented analyses of our data, we were able to hone the benefits of small corpora, i.e. the possibility to combine “vertical” corpus reading techniques with whole-text reading of the larger chunks of description. These benefits of small corpora were originally highlighted in the context of corpus-based language learning (Aston 2002; Ghadessy, Henry, & Roseberry, 2001; Braun 2005) but they were deemed to be equally suitable for the exploratory stage of analysing machine-generated video captions. As it became obvious that the linguistic structure of the video captions does not follow the rules and patterns of human-generated discourse, Aston’s (2002) suggestion that whole-text reading is a useful way of “getting one’s teeth into a corpus” was adopted, allowing us to search for qualitative explanations for the sometimes curious linguistic patterns observed in the MD corpus. For example, our initial quantitative analysis of the video captions (Braun & Starr, 2019) revealed a low level of granularity in the video captions, an over-use of generic descriptions, a high level of redundancy, lexical poverty and a lack of lexical and grammatical diversity within the captions (see also Park et al., 2019). However, it was through a subsequent qualitative analysis of the data that the lack of accuracy in the captions and the large number of semantic inconsistencies became obvious, and we were able to trace some of the curious lexical choices and syntactic structures back to the training data that formed the basis for the caption generation.

In the present paper we focus on the linguistic devices available in human descriptions and video captions to support narrative sequencing for audiovisual storytelling, i.e. our focus shifts from considering lexico-grammatical devices in isolation, to analysing such devices with a view to their role in creating coherence in short-form narrative. In light of this, we explore cues for character and object identification but, equally important, we also look at the cues that support the audience in tracking characters and objects as the narrative unfolds. This is followed by an exploration of temporal cues as a further opportunity to construct narrative coherence.

## 4. Findings: Sequencing Cues and Issues

### 4.1 Character Identification and Tracking

Whether video descriptions are used as a means of making audiovisual content accessible for consumers experiencing visual or cognitive impairments, or as a basis for retrieval of material for use in future programme-making, identifying and labelling key protagonists or characters is essential. The human mind uses a wide range of cues for this purpose. In multimodal material, these cues can be characterised into two main groups; linguistic versus non-linguistic. The former includes elements that involve character naming, using a range of referential expressions, and character tracking to maintain the continuity of references to the same character between frames and scenes. The non-linguistic elements that contribute to continuous character identification and recognition between frames include visual cues, i.e. cues enabling *facial recognition* (recognising the face of the same character between the scenes), *body silhouette* (e.g. distinguishing between the body shape of a male character versus a female character), *clothing* (e.g. a character seen in the same clothing item), as well as vocal cues such as the pitch, tone and speech patterns in a character's voice.

Current AI models for video scene description use visual cues such as objects, characters, actions, locations and some facial expressions. They do not normally capture audio input yet (but see Jin & Liang, 2016). However, currently available machine-generated video descriptions, using even the most advanced neural network models, still lack granularity, accuracy and nuance, when compared to human-generated video descriptions. At the most elementary level of character identification, current AI models fail to allocate traditional binary gender labels to characters/protagonists with accuracy (through object recognition combined with NLP [natural language processing], introducing a serious confound that affects character descriptions). Whilst this is a relatively simple task for most human beings, our analysis suggests that training the computer in a way that allows it to detect multiple cues and assimilate this knowledge into a reasoned conclusion remains challenging. Curiously, the training data upon which machine-generated gender identification is based have been compiled by human volunteers and are thus likely to be largely accurate. Nevertheless, hair styles and other human characteristics do not always resonate with stereotypical gender assignments or may be ambivalent in an individual shot/frame, making rule-bound feature extraction difficult. For instance, as illustrated in Figure 2, a person with short hair—or what appears to be short hair—is generally labelled as a man, irrespective of facial features, mannerisms, and other cues implying gender.

Figure 2.

*An Education (2009)*



*A man* is talking and smiling at someone

As an alternative to “a woman” or “a man” the computer model also applies the gender neutral “a person” when it recognises the presence of a human being without there being a face in the frame. The frequency of “a person” in the MD corpus is 139 (1976.82/m) whereas in the training data this is a more frequent phenomenon (MS COCO: 3312.99/m; TGIF: 3240.83/m). In a random sample of fifty concordances retrieved from the MD corpus, 86% ( $N=43$ ) of these occurrences incorporated some part of a body that was featured in the image captioned, e.g. an extended hand, holding an object, as illustrated in Figure 3.

Figure 3.

*Bruce Almighty (2003)*





*A person* is holding a cell phone in their hand

Although the computer model generating this description clearly mislabels the held object as a “cell phone” (see also section 4.2), it appears to recognise a human being (by the hands), selects the gender-neutral description “a person” and applies the pronoun “their”, perhaps as an attempt at gender neutrality. Use of the neutral expression “person” will, of course, be helpful in storytelling where a character has not been identified yet and/or where using a neutral expression is motivated by keeping up the suspense, but the point here is that a human being would be able to detect the nuances of a pair of hands and use these as cues towards gender identification, whilst the machine-generated captions remain at a more generic level.

With regard to sequencing, the unreliable nature of gender detection, which impacts the choice of referring expressions in captions, will also affect the creation of coherence across frames and scenes when image sequencing becomes more sophisticated. In Figure 4 for example, the CD (our ground truth) achieves the sequencing and continuity of character appearances between frames and scenes through repeated mention of the characters’ names. The AD uses pronouns, which ease the natural flow of the narrative. By contrast, the MD not only features much more generic character descriptions (*a woman*), but also fails to mark the previously introduced characters as “given” (i.e. previously introduced), for example through the use of definite noun phrases (e.g. *the same woman*) or third-person pronouns.

Figure 4.

*Pretty Woman (1990)*

Frame	CD	AD	MD
	Vivian sits by Edward’s side and looks at him, smiling.	She sits down in front of him, smiling brightly.	A man is kissing a woman’s hand.
	Vivian presses one of her fingers onto her lips, and then presses it onto Edward’s lips. Edward keeps sleeping.	Tenderly, she presses a finger to her own lip and then to his.	A woman is looking at something and smiling.

As a matter of fact, third-person pronouns are under-represented in the MD corpus. The most frequently used pronouns in MD are “she”, “he”, “her”, “his” occurring 705 (10026.31/m), 613 (8,717.91/m), 195 (2773.23/m) and 198 (2815.9/m) times respectively. The occurrences of the same pronouns (i.e. *she, he, her, his*) in CD are respectively, 534 (10362.49/m), 639 (12400.06/m), 761(14767.52/m) and 670 (13001.63/m) and in the AD corpus the figures are 503 (17316.76/m), 590 (20311.91/m), 682 (23479.19/m) and 556 (19141.39/m).

Where pronouns occur in the MD corpus, they are often inaccurate and therefore fail to deliver cohesion in character identification between frames and scenes. In Figure 5, for instance, the machine has failed to create a correct link/agreement between the subject and object pronouns.

Figure 5.

*Sex and The City (2008)*







A man is talking to a woman while she is looking at *her*.

Other character cues such as clothes, hairstyle, and body shape, which are used by the human video describer to further aid character identification across frames and scenes, are rarely present in our MD corpus. The current computer model would therefore appear inadequate in the way it reads and applies these cues to the development of narrative cohesion. For comparison, the extract shown in Figure 6 illustrates how human describers use the available character cues such as clothes, hair and body silhouette to infer that the character appearing in the consecutive frames is the same person. The MD for this extract contains inaccurate descriptions: for instance, while the scene is correctly identified as a room containing tables and chairs, it lacks saliency in terms of the actions of the pre-known (by human describers annotating CD and AD) principal protagonist. Once the scene is established, a second MD labels the woman as a man, failing to identify the action pertaining to a head movement; this is followed by incorrect identification of the protagonist, object and actions (frames 3 and 4). Although each of these elements raises questions about computer vision competency, all of which are unfathomable given the “black box” nature of machine learning, the main point to note here is that even if description 2 (*A man is driving a car*) were correct, the expectation would be that description 3 presents the man (and the car) as “known”, using a definite article (i.e. *The man is now sitting in the car* instead of *A man is sitting in a car*) to recreate the visual continuity in the verbal description.

Figure 6.

*The Forgotten (2004)*

Frame	CD	AD	MD
	Telly is sitting at a long wooden table with a laptop.	A new day. Telly sits working at a laptop at one end of the large wooden dining table in the spacious, open-plan living area of the Paretta's home.	A room with a lot of tables and chairs
	Now standing up, Telly opens a drawer and takes out another box, opens it	On her feet, Telly opens a dresser drawer and takes out a small box.	A man is driving a car and <unk> his head
	and takes out a key.	She opens the box and finds a key.	A man is sitting in a car and smoking a cigarette
	She walks down a corridor holding the key and uses it to unlock a door.	She returns the box to the drawer and, taking the key with her, approaches the closed door of Sam's bedroom and unlocks it.	A man is looking at a woman and then she looks away

In summary, the human consumer of audiovisual material is adept at reading many different visual cues to detect character and narrative continuity between shots and scenes. Although the machine would appear to apply some rule-bound feature extraction measures to determining male and female, this is likely to include the blanket application of certain “rules” (e.g. short hair equals male, long hair equals female), and the level of human sophistication in this process cannot be matched. As humans, we know men can have long hair and wear skirts, and that women may have short hair and wear trousers, but we seldom make the wrong judgement call in a binary reading because we look beyond the obvious and seek affirmation or negation of our first hunch. The styling of clothes, facial hair, types and styles of jewellery and other adornments, and the wearing of make-up, are all

examples of visual cues which help with this task. Moreover, the nature of feature extraction and machine learning is such that the use of visual cues for character identification is also likely to apply to object recognition more broadly. This will be analysed further in the next section.

## 4.2 Object Recognition and Tracking

In addition to character tracking, object sequencing is perhaps the most obvious approach to creating coherence between shots and scenes in a multimodal narrative. Where plot and sub-plot are often interwoven in one or more scenes, viewers frequently decode visual cues related to narrative coherence via objects and locations in order to determine where plot strands continue, meet or diverge. In the scene illustrated in Figure 7 (*Click*, 2006) an object salient to the unfolding narrative is present in all three still images taken from three different shots. The object, a television remote control, is featured from different angles and in varying scale in each frame, and yet as human beings we are able to detect not only that it is the same object throughout, but that the younger man in the first image is likely to be the same individual as the man in the third image. We also conclude from the narrative significance of the remote control that the younger man is the owner of the disembodied hand in the second image. In reaching these conclusions, we draw on visual cues as well as cues from the film dialogue, combining them to formulate a storyline revolving around the remote control behaving unexpectedly, and backed up with the continuity supplied by the younger man's consistent vocal tone. For this reason, it is difficult to pinpoint whether narrative saliency is derived from the audio or visual prompts at any given moment, but even if the sound is omitted, there appear to be sufficient clues to object cohesion to infer that there is a continuity of plot and associated characters between frames.



Figure 7.

*Click (2006)*



A man is sitting in a chair and talking



A dog is playing with a toy



A man sitting in a chair with a laptop computer

However, human beings employ a highly complex layered approach to consuming narrative, beginning at the stage of establishing characters, objects and actions, and culminating in intra- and inter-textual referencing and inferential cue interpretation (Braun 2016; Braun et al. 2020; Herman, 2013). Current machine-based content descriptions, as can be seen from the above example, are created from single frame object and action recognition overlaid with natural language processing. The data output from this process is unreliable due to the nature of object recognition, which is often confounded not only by non-standard angles, variation in scale or size of the objects, but also small changes in the arrangement or light gradation of pixels. Frames displaying similar images with only minor differences elicit considerable variation in machine-generated captions, as illustrated in Figure 8. This cannot be explained simply by an absence of relevant images in the training data since in the example below, where a person appears to be performing a card trick or game, wide variations

in the detected objects and actions occur (*plates of food, couch, sitting, dancing*) even though the concept of “playing cards” is represented in both sets of data (MS COCO,  $n=4$ ; TGIF,  $n=8$ ).

Figure 8.

*Bright Young Things (2003)*



A person is putting a plate of food on a table



A man is sitting on a couch and is dancing

The current method of labelling objects is also often narratively inadequate in terms of the required level of granularity or specificity. As such, the machine may correctly identify *a car* but there may be strong narrative saliency in conveying greater detail. In Figure 9, we cannot see the driver of this car, but as human observers, seeking coherence, we are able to discern that it is not just any car—it is neither a standard family saloon, nor a flashy sports car, but is in fact a luxury executive vehicle. By the time we see the second image it is apparent that the car is a Chrysler or similar, and even though we cannot identify the driver yet, from the previous narrative we know by the type of car that it is likely to be “Jack”, the principal protagonist (*Firewall, 2006*).

Figure 9.

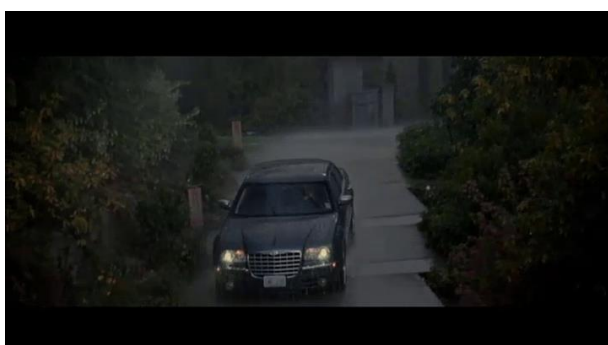
*Firewall (2006)*



A car is driving through a forest



A car is driving down a road in the rain



A car is driving down a road

Our human ability to make these inferences is reflected in the human description. The narrative relevance of the type of car is specifically targeted for exposition by the human audio describer since it suggests the identity of the driver: “Headlights glaring through the rain, the Chrysler speeds through the driveway and comes to a halt outside the house”. These descriptions, and those found in the CD corpus (“A car drives in the rain, then stops in front of a house. Jack is sitting in the car; he looks worried.”) define moments that a machine trained on still imagery is not readily capable of determining, e.g. *driving* and *coming to a halt* are generally indeterminable from still images, even if certain aspects suggest one or the other (a car in the middle of the road may have broken down, but is more likely to have been captured during movement; a car pictured outside a house is likely to have stopped rather than been captured in motion).



In practice, the volume of data available to train for image recognition will dictate how rapidly automatic image descriptions can be fine-tuned to this level of detail. When and where to provide the detail, in terms of narrative saliency, is a significantly bigger challenge, since it requires an assessment of what is missing from the scene (the driver's face) and which cues are most helpful in filling this gap.

One further point of note is that the MD represents the car as a new or unidentified object in each description, as indicated by the use of the indefinite article. This is characteristic of the MD corpus, and there are wide discrepancies in the numbers of definite and indefinite articles between the MD corpus and corpora including human descriptions, both AD and CD. In the case of MD the indefinite article ("a") is significantly over-represented (MD: 209,116.12/m; AD: 32,636.78/m; CD: 34,483.32/m), suggesting that each referent is treated as a new entity in the plot although, as the example in Figure 9 illustrates, this is often not the case. Once again, this is a consequence of each frame being treated as a discrete object, as opposed to part of a narratively linked sequence, but it is also a feature that could be remedied once objects are sequenced by similarity ranking (colour, shape, scale etc.) or other means. If automatic object recognition and continuity are improved, parsing of the corresponding descriptions may be adjusted to acknowledge recurring artefacts by incorporating the definite article ("the") more frequently in MDs. At present, this is far less frequently observed in the machine descriptions than in the human-generated corpora (MD: 3,740.31/m; AD 66,120.61/m; CD 69,451.66/m).

Similar observations can be made about location as a narratively relevant visual cue that can denote continuity between frames and scenes. Figure 10 exemplifies the use of this cue in human descriptions, illustrating the different ways in which the human describers have included narratively salient details about the location (the lift). The content describer highlights the physical characteristics of the lift (*narrow*), facilitating the inference that it may be uncomfortable inside, while the audio describer emphasises the lack of comfort, facilitating the inference that the lift is narrow. The content description also includes a mention of the metal door, which enables the audience to situate the lift in the early 20<sup>th</sup> century. By contrast, in this example the MD lacks reference to the location, highlighting the fact that the computer model has yet to be developed in a way that utilises location cues as a means of creating continuity and coherence between frames and scenes.

Figure 10.

*The King's Speech (2010)*

Frame	CD	AD	MD
	A man and a woman (Duke—Bertie—and Duchess of York) enter a narrow lift. The man closes the metal door behind them.	The Duke and Duchess squeeze uncomfortably into the lift in Logue's building.	A man is kissing a woman on the cheek
	The lift goes down.	They stand chest to chest.	A man is looking at a woman and then she looks away

### 4.3 Temporal Cues

An important aspect of audiovisual storytelling is its temporal structure. In other words, as noted by Kress (1998, p. 68), audiovisual content normally sequentialises and temporalises individual visual images. In the verbal description of audiovisual material, temporal words therefore provide a useful shorthand to the passage of time and can assist with the cognitive processing of narrative. In the case of AD, where audio hiatuses may be short, indications of time often explain a change in scene and may be accompanied by the act of naming or confirming the characters visible at the point of shot-change. This device is typically used in fast moving serial dramas (“soaps”) such as *EastEnders* where the programme is dialogue heavy and the time in which to describe fine details is somewhat limited. For example, where the scene moves from the sitting room of one character to an interior shot of a party at a different venue, the AD might state: “Later, in the pub. Bernadette and Tiffany arrive”. In film AD, if time allows, a more fulsome explanation is often given: “Later, fully dressed in a grey tweed suit and waistcoat, Adam trots down the stairs into the lobby” (*Being Julia*, 2004).

Temporal expressions of particular relevance in the storytelling context are position adverbs, i.e. adverbs that locate entities or actions temporally (Musn, 2002). They often depend on the context of an utterance or the wider discourse context (e.g. *yesterday*, *the day after*) (Altshuler, 2011). Temporal position words and phrases can be considered to fall into three categories: progressive (before, after, earlier, later, first, last, during); epochal (yesteryear, in times gone by, in the 1970s, last century, the Jurassic age); and diurnal/nocturnal (day, night, tonight, last night, this morning, yesterday). All of these elements and other time-related means of expression are encountered when

humans describe the unfolding of a series of connected audiovisual events, using them to anchor actions in time, even where the exposition of a story occurs non-sequentially or in a non-linear fashion (e.g. *500 Days of Summer*, 2009).

Machine-derived content descriptions, being based on non-sequential caption production, exhibit a curious relationship with the temporal lexicon: the isolated captioning of each frame in our MD corpus would suggest a lack of engagement with temporal concepts. However, these are not wholly absent from our MD corpus. Table 2 illustrates the relative frequencies of common temporal words and phrases in the MD corpus by comparison with both sets of training data (MS COCO and TGIF).

Table 2.

*Temporal Words and Phrases; Absolute Frequency and Relative Frequency Per Million Word Tokens*

WORD OR PHRASE	MD CORPUS		MS COCO		TGIF	
	<i>Rel.(f)</i>	<i>f</i>	<i>Rel.(f)</i>	<i>f</i>	<i>Rel.(f)</i>	
BEFORE	0.00	0.00	89.45	622	407.83	543
AFTER	0.00	0.00	133.45	928	398.81	531
DURING	14.22	1	506.05	3,519	249.35	332
NEXT	213.33	15	5,590.14	38,873	772.09	1028
NEXT TO	213.33	15	5,535.64	38,494	749.56	998
NEXT EXCL. NEXT TO	0.00	0.00	54.50	379	22.53	30
LATER	0.00	0.00	1.44	10	3.00	4
EARLIER	0.00	0.00	1.44	10	0.00	–
SINCE	0.00	0.00	1.58	11	0.75	1
WHILE	1,663.94	117	2,197.20	15,279	5,729.10	7628
UNTIL	0.00	0.00	1.29	9	54.08	72
MEANWHILE	0.00	0.00	0.43	3	3.76	5
AT THE SAME TIME	0.00	0.00	11.07	77	74.36	99
DAY	0.00	0.00	631.59	4,392	61.59	82
NIGHT	241.77	17	371.16	2,581	192.27	255
WHEN	0.00	0.00	17.83	124	413.84	551
START	56.89	4	38.11	265	490.44	653
FINISH	0.00	0.00	31.49	219	28.54	38
BEGINNING	0.00	0.00	11.50	80	8.26	11
END	0.00	0.00	160.20	1114	109.66	146
ENDING	0.00	0.00	1.44	10	3.00	4
BEFOREHAND	0.00	0.00	0.00	–	0.00	–
AFTERWARDS	0.00	0.00	0.00	–	16.52	22
ALREADY	0.00	0.00	8.77	61	3.76	5
THEN	15,515.89	1091	9.49	66	3,696.73	4922
FIRST	0.00	0.00	40.70	283	61.59	82
LAST	0.00	0.00	5.46	38	12.02	16
YESTERDAY	0.00	0.00	0.00	0.00	0.00	0.00
TOMORROW	0.00	0.00	0.14	1	0.00	0.00
TODAY	0.00	0.00	10.21	71	2.25	3

The first observation to be made in relation to these results is that temporal words are under-represented in the MD corpus by contrast with both of the datasets used to train the machine. Furthermore, all referenced entries except *beforehand* and *yesterday* are represented in one or other of the training datasets. This general sparsity of temporal words in relation to the MD corpus is to be expected, given the current lack of coherence between frames. However, there are notable exceptions where temporal words/phrases are present in machine descriptions, suggesting intercedence in the form of time or timescales, and these clearly require further exploration. *During* (14.22/m), *next* (213.33/m), *while* (1663.94), *night* (241.77/m), *start* (56.89/m), and *then* (15,515.89/m) are all present in the MD dataset, albeit with widely ranging frequency.

*During* is a lexico-grammatical device that suggests simultaneity, and its relative frequency in the MS COCO corpus, which is close to twice that observed in the TGIF corpus, is surprising given that MS COCO is derived from the captioning of still images, while TGIF comprises brief sequences of moving images (GIFs). Although *during* occurs only once in the MD corpus, collocated with *night*, its presence suggests capturing a particular moment in time: “A plane flies through the sky during the night” (extract 201117\_417). Both MS COCO and TGIF include the phrases *during the night* and *night sky*, but neither dataset contains the string *sky during the night*, suggesting that the machine has learnt to connect the notions of *sky*, *during* and *night* in a temporal fashion. Thus, if the next scene were to show a plane landing in daylight, it would not be too great a leap of narrative coherence to infer that the same aircraft had completed its flight the following day. Basic sequential events such as these are likely to provide the opportunity for early attempts at creating a joined-up narrative from theoretically discrete video frames. With additional information extracted from the image, in this case perhaps the colour and livery of the aircraft, it should be possible to establish continuity and ascertain coherence where none currently exists.

Another cue for plot continuity can be relayed by the use of *next*, either as a determiner preceding a noun (*next week*), as an adverb to qualify an action (*who sings next?*) or as an adjective to elaborate on a noun (*in the next carriage*). In each of these scenarios, *next* operates as a coherence marker, denoting forward projection or ordinal arrangement, and in both cases it situates a description within a specific sequence. However, in all instances within the MD corpus *next* is collocated with *to* (213.33/m), i.e. it is used as a prepositional phrase (*next to a fence*; *next to a woman*; *next to a red stop sign*). It cannot therefore be regarded as a temporal word in the MD context. Within the training data, most of the instances of *next* also occur within the prepositional phrase *next to*, yet a small number falls outside this category (MS COCO: 54.5/m; TGIF: 22.53/m). For example, in MS COCO, we find phrases such as *next train*, *next move*, *next pitch*, *next destination*, *next period*, *next trip*, *next player* and so forth; while in TGIF collocations include *next one*, *next scene*, *next task*, *next lane*, *next woman* and *next room*. Equipped with this data, the question arises as to why sequential or ordinal terms of this nature were not generated by the machine when creating the MD captions, most particularly since they were applied to single image examples in the MS COCO captioning exercise. Simultaneous actions within one video frame would offer an opportunity to test this concept as part of the feature extraction process.



As an alternative approach to concurrency of action, the use of *while* is more widespread across all three corpora (MD: 1,663.94/m; MS COCO: 2,197,20/m, TGIF: 5,729.10/m). Examples in the MD corpus include single-character simultaneity (*talking on a cell phone while holding a drink, smiling and laughing while looking down, walking down a corridor while holding a bag*) and two-person concurrent actions (*a man is playing the piano while a woman is playing a guitar, a man is holding a woman's head while they are both wearing a tie*). The same patterns can be identified in the human-captioned TGIF dataset (*a woman is eating while looking angry, a young girl is dancing while she listens to music, two men are dancing while showing hardly any emotion*) and MS COCO training data (*a cat that is looking up while sitting down, a man is sitting on the sidewalk while a police officer is doing something behind him*). The under-representation of *while* in the MD corpus requires further investigation to establish whether the computer fails to detect co-occurring actions as accurately as the human eye, or whether this outcome is a matter of weighting during the NLP parsing operation.

Diurnal and nocturnal considerations pose a particular subset of issues related to temporal markers within video narrative. One might expect that both bear equal significance in the storytelling rubric, since where they are mentioned, they are likely to be narratively salient, and therefore of relevance. In this context, it is especially notable that in the MD corpus there was no specific reference to *day*, yet 241.77/m references to *night*. A possible explanation is that daytime might be considered the default condition in feature films, the time when the majority of the plot unfolds. Events taking place at night are more likely to be iconographic, being employed as a shorthand to denote danger, clandestine activity or specific plot-defining episodes (party, wedding etc.) and for this reason the temporality is narratively salient. The computer does not yet detect saliency, but it detects a difference in colour saturation, and this difference is likely to correspond with captions in the training data where *night* is an iconographic element of the image. On closer inspection, all but one instance of *night* in the MD corpus correspond with *a road at night* or *a street at night*, collocations that are likely to have been derived from a combination of object + *at night* in the training data (MS COCO: 275.53/m; TGIF: 126.18/m). The collocation of *street* or *road* and *at night* is particularly common in MS COCO, and suggests that the combination of a street or road and a night sky prompted the machine to select the above combinations.

*Day* is used in a different manner, not to denote the time of day, but rather the type of day: (*cloudy day, windy day, sunny day, calm day, rainy day, foggy day, etc.*). The use of *day* is around three times less frequent than *night* in the TGIF corpus, although conversely, *day* is twice as common as *night* in the MS COCO dataset. It is difficult to say why this might be the case, but the moving image nature of TGIFs perhaps suggests to the viewer that the action occurring is more important than the mise-en-scène, whereas the static nature of the MS COCO images leaves the captioner with fewer actions to depict, and therefore more cognitive resource to describe the setting. Since it is unlikely that weather cues were incorporated into the feature extraction mechanisms used to create the MD corpus, this could explain the absence of *day*-related words and collocations.

*Beginning* and *end* are nebulous concepts in the audiovisual world with feature films frequently commencing a narrative in the middle or at the end of a storyline, and recounting plot in a non-

sequential manner. *500 Days of Summer* (2009), in our film corpus, employs an unconventional temporal technique, unravelling the various stages of a brief relationship between two lovers, numbering the episodes by the day of the relationship, as it jumps around the plot. On the other hand, *start* and *finish* are factually determined concepts, since a feature film or movie extract will always have a well-defined start and end. Likewise, actions have a determinable start and end, whether they occupy the narrative space of two frames (e.g. a blink), or stretch across a period of seconds, minutes or longer.

There are two applications of *starts* in the machine-generated (MD) captions: *a man is looking at a paper and then he starts to run away*; and, *a man is sitting on a bench and then he starts to run away*. Whilst both uses are correct, the presence of the phrase *starts to* is unexpected for a body of work based on individual frame descriptions, but also curious in the fact that it is only to be found on two occasions. Given the wide margin in relative frequencies between TGIF (490.44/m) and MS COCO (38.11/m) training data, it would seem most likely that the machine is drawing upon the former in its application of *start*. TGIF examples include *starts to shake*, *starts smashing a computer*, *start to kiss*, *start of a race*, *starts fighting*, *start to fall over*, *start hitting*, *start walking*, all of which might be expected in descriptions of short animated clips where there is a natural start and finish. In TGIF, *starts to run* occurs just three times, and is not collocated with *away* in any of these captions; the phrase is wholly absent from MS COCO. The question this finding poses then, is how the machine created the collocation *starts to run + away* in the MD corpus, which appears to be a combination of *starts to run* and *run away* (the latter occurs 25 times in TGIF, but is absent from MS COCO). By contrast, *finish* does not appear in the MD corpus, which could be explained by the relatively minor role it plays in the training data (MS COCO: 31.49/m; TGIF: 28.54/m).

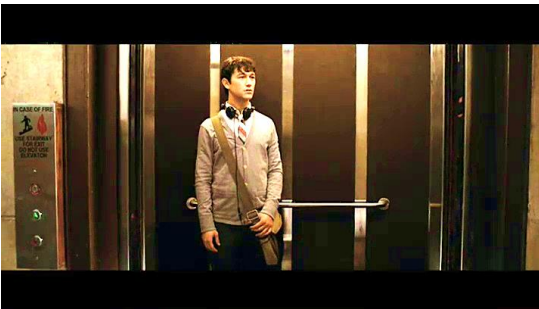
Perhaps the most interesting result from the analysis of temporal words is the high relative frequency of *then*, which is disproportionately represented in the MD corpus (MD: 15,515.89/m; MS COCO: 9.49/m; TGIF: 3,696.73/m). Upon further investigation it would seem that a particular computer vision phenomenon is at play in this case, which has resulted in an idiosyncrasy occurring in the training data being replicated in the MD corpus output. From the relative frequency statistics, it is clear that *then* occurs far more commonly in the TGIF than in the MS COCO dataset, where it is primarily used in error (*then* substituted for *than*). The relatively high frequency in the TGIF data has been explained as being the result of “split frame” captioning. This occurs where actions are represented across a split screen, effectively combining two or more images in one frame, causing the human captioner to link the frames in an inferred sequence. *And then* is therefore invoked as a means of connecting two images by some inferred link devolved from the human’s perception of an unfolding narrative portrayed in the GIF. Although it is not clear how neural networks derive solutions to visual problems, it can be surmised that in the case of the MD corpus, the machine may have interpreted any single image divided by a strong vertical line as comprising two interconnected images, and applied *then* as the “connector” when generating the caption.

This effect is illustrated in Figure 11, taken from the movie *500 Days of Summer*. The machine-generated caption for this image reads: *A man is looking through a door and then he opens the door*.

While it is not possible to discern with certainty the computer's steps in moving from analysing the pixels in the image to generating this caption, the above phenomenon would seem to suggest that the vertical lines forming the decoration on the door behind the young man have had the effect of "splitting" the image into two notional frames, causing *and then* to be invoked in the second half of the machine caption (shown in brackets). As such, it might be considered a forced sequencing process.

Figure 11.

*500 Days of Summer (2009)*



A man is looking through a door [and then he opens the door.]

While this may at first appear to be an anomaly in the creation of captions by machine, it also provides a window on the workings of the "black box" effect characteristic of computer vision neural networks. If the impact of vertical markers in film stills encourages the machine algorithm to return a linguistically sequenced result, however elementary, then it is conceivable that this might be a first step in generating fully sequenced narrative. Instructing the machine to use conjunctions in an intelligent and object-driven manner, seeking the principal source of action and attaching the conjoined action to this object, could feasibly be one future channel for investigation.

## 5. Discussion and Conclusions: Next Steps for Automated Video Description

The aim of our study was to gain an understanding of the key linguistic differences between human and machine-generated descriptions of (moving) images, focusing on the principal elements facilitating contextualisation and the creation of narrative coherence. Linked to this was an appraisal of the current ability of machine learning algorithms to emulate human comprehension to create coherent, human-like descriptions of the audiovisual content, while at the same time acknowledging that there is currently an absence of meaningful narrative sequencing.

Our analysis has highlighted some of the fundamental problems with the current state and quality of machine-generated description, i.e. the machine's failure to detect *visual cues* (e.g. face, object and gender recognition) with accuracy in the moving images and to use key linguistic elements (e.g. pronouns and temporal lexicon) as cohesive ties to contribute to narrative coherence and underpin caption sequencing. In image/video description, recognising salient visual cues from the "ground up"

(common scenes and scenarios providing contextual references of the type explored in Huang et al., 2016) and using cohesive prompts helps the audience identify and track a character and related objects, as well as make assumptions about the nature of the person at the centre of the narrative action, thereby establishing continuity. The examples analysed in this paper demonstrate that it is often the human's life experience and world knowledge that make the output of human description different from the machine-generated video captions, as they enable humans to disambiguate and analyse multimodal materials at a higher level of complexity, seeking meaning and relevance in small details and cues. On the other hand, these examples also illustrate how the computer model still "sees" images in isolation, as single frames, and deconstructs them into a collection of objects and actions. The model does not yet integrate techniques to connect objects and/or create character continuity markers as a way of creating coherent narrative.

The *identification of characters* remains rudimentary in the machine descriptions. Character descriptions are mostly restricted to the most generic level of description (*a man, a woman*), and details about age or appearance are generally absent, meaning that character descriptions in the MD corpus were generally found not to match the level of human sophistication. The machine's current failure to connect individual frames also means that characters are not tracked in the machine-generated captions. Face recognition techniques and other cues denoting continuity between frames and scenes (e.g. costume colour labelling) would, in theory, make it possible for the machine to achieve this and, at the caption generation stage, to use pronouns and other referential expressions in the captions for character referencing.

Closely linked to character identification, the concept of *gender identification* was one of the computer vision issues investigated in our corpus. It was discovered that the machine descriptions from our computer model were unreliable in detecting gender correctly in many instances, although the crowdsourced, i.e. human-generated training data from which they were derived are likely to have been relatively low-error in this regard. Gender detection is not simply a visual phenomenon, but is also aided by vocal cueing—pitch, tone and patterns of breathing can all provide audio prompts which confirm or negate first choices. Work undertaken in the field of vocal diarisation has produced encouraging results for the purposes of automatic gender detection (Doukhan, Carrive, Vallet, Larcher, & Meignier, 2018) although the results appear to be language specific. Consideration is therefore being given to combining video gender cues and audio voice profiling, with the latter providing material which could represent a new dimension for the feature extraction process.

Like gender recognition, *object recognition* in moving imagery requires substantial improvement before the machine will be able to produce meaningful narrative. At the current stage of machine description development, basic errors occur frequently, with objects labelled incorrectly as a consequence. Improving object recognition accuracy is a pre-requisite for advancing narrative sequencing which is largely dependent on improvements in the quality and quantity of data used to train the machine.

*Temporal linguistic cues* were studied as a means of determining the extent to which the computer model was able to extract sequential narrative from existing training data and apply the principles underpinning human temporal sequencing (or similar principles) to the creation of machine descriptions. Although temporal words were not widely present in the MD corpus, certain anomalies existed, both in terms of words present in the training data that were not applied to the MD captions (*day*) and in the extent to which certain collocations were combined to provide a sense of continuity (*starts*) where ostensibly none existed. The anomaly of “false frames” in the film corpus prompting the use of *and then* suggests that training the machine to treat individual sequential frames of film as “false frames” might prompt similar linking phrases (*and next, the next day* etc.). It is certainly a type of narrative sequencing, albeit forced and rudimentary, that might be explored further in the future.

In summary, current caption algorithms can be applied to moving images to construct the simplest of stories only as long as basic visual cues such as objects and gender are identified correctly. However, since there is no continuity and coherence across the temporal narrative, the result will often be incoherent or incomplete. A first step in improving machine descriptions is to resolve basic errors that occur due to computer vision problems (objects, actions, characters). Once these fundamental issues have been resolved, higher order problems such as determining saliency, establishing cohesive ties to sequence storylines, and incorporating story grammar to frame narrative, can be foregrounded.

## References

- Altshuler, D. G. (2011). Toward a more fine-grained theory of temporal adverbials. *Semantics and Linguistic Theory*, 21, 652–673.
- Aston, G. (2002). Getting one’s teeth into a corpus. In M. Tan (Ed.) *Corpus studies in language education* (pp. 131–144). Bangkok: IELE Press.
- Bai, S., & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, 311, 291–304.
- Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17(1), 47–64.
- Braun, S. (2011). Creating coherence in audio description. *Meta: Journal des Traducteurs*, 56(3), 645–662.
- Braun, S. (2016). The importance of being relevant? A cognitive-pragmatic framework for conceptualising audiovisual translation. *Target*, 28(2), 302–313.
- Braun, S., & Starr, K. (2019). Finding the right words: Investigating machine-generated video description quality using a human-derived corpus-based approach. *Journal of Audiovisual Translation*, 2(2), 11–35.
- Braun, S., Starr, K., & Laaksonen, J. (2020). Comparing human and automated approaches to visual storytelling. In S. Braun & K. Starr (Eds.) *Innovation in audio description research* (pp. 159–196). Abingdon: Routledge.

- Coraci, F. (Director), & Sandler, A., Giarraputo, J., Moritz, N. H., Koren, S., & O'Keefe, M. (Producers). (2006). *Click* [DVD]. USA: Sony Pictures Releasing.
- Dayton, J., & Faris, V. (Directors), & Turtletaub, M., Friendly, D. T., Saraf, P., Berger, A., & Yerxa, R. (Producers). (2006). *Little Miss Sunshine* [DVD]. USA: Fox Searchlight Pictures.
- Doukhan, D., Carrive, J., Vallet, F., Larcher, A., & Meignier, S. (2018). An open-source speaker gender detection framework for monitoring gender equality. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Retrieved from <https://hal.archives-ouvertes.fr/hal-01927560>
- Forceville, C. (2014). Relevance theory as a model for multimodal communication. In D. Machin (Ed.), *Visual communication* (pp. 51–70). Berlin: De Gruyter Mouton.
- Fry, S. (Director), & Carter, G., & Davis, M. (Producers). (2003). *Bright Young Things* [DVD]. UK: Film Four.
- Ghadessy, M., Henry, A., & Roseberry, R. (Eds.) (2001). *Small corpus studies and ELT: theory and practice*. Amsterdam: Benjamins.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*. Retrieved from <https://arxiv.org/abs/1512.03385>
- Herman, D. (2013). *Cognitive narratology*. Retrieved from <http://www.lhn.uni-hamburg.de/node/38.html>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short term memory. *Neural Computation*, 9(8), 1735–1780.
- Hooper, T. (Director), & Canning, I., Sherman, E., & Unwin, G. (Producers). (2010). *The King's Speech* [DVD]. UK: The Weinstein Company.
- Huang, T. H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Dhruv, B., Zitnick, C., Parikh, D., V., Vanderwende, L., Galley, M., & Mitchell, M. (2016). Visual storytelling. *Proceedings of NAACL-HLT*, San Diego, CA, USA.
- Husain, S. S., & Bober, M. (2016). Improving large-scale image retrieval through robust aggregation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9), 1783–1796.
- Jin, Q., & Liang, J. (2016). Video description generation using audio and visual cues. *Proceedings of 2016 ACM on International Conference on Multimedia Retrieval*. Retrieved from <https://dl.acm.org/doi/abs/10.1145/2911996.2912043>
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography* 1(1), 7–36.
- King, M. P. (Director), & Parker, S. J., King, M. P., Star, D., & Melfi, J. (Producers). (2008). *Sex and the City* [DVD]. USA: Warner Bros. Pictures.
- Kress, G. (1998). Visual and verbal modes of representation in electronically mediated communication. In I. Snyder, & M. Joyce (Eds.) *Page to screen*. Sydney: Allen & Unwin, 53–79.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D., Bernstein, M. S., & Li, F.-F. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International Journal of Computer Vision*, 123, 32–73.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.



- Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., & Luo, J. (2016). TGIF: A new dataset and benchmark on animated GIF description. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Retrieved from <https://ieeexplore.ieee.org/document/7780871>
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollar, P. (2015). Microsoft COCO: Common objects in context. *Computer Vision, ECCV 2014*, 740–755.
- Loncraine, R. (Director), & Bernstein, A., & Iwanyk, B. (Producers). (2006). *Firewall* [DVD]. USA: Warner Bros. Pictures.
- Mandler, J. (1978). A code in the node. *Discourse Processes*, 1(1), 14–35.
- Mandler, J. M., & Johnson, N. S. (1980). On throwing out the baby with the bathwater: A reply to Black and Wilensky's evaluation of story grammars. *Cognitive Science*, 4(3), 305–312.
- Marshall, G. (Director), & Milchan, A., Reuther, S., & Goldstein, G. W. (Producers). (1990). *Pretty Woman* [DVD]. USA: Buena Vista Pictures.
- Merkel, M. (1999). *Understanding and enhancing translation by parallel text processing* (Doctoral dissertation, Linköpings universitet).
- Musan, R. (2002). Types of temporal adverbials. In R. Musan (Ed.) *The German perfect, Studies in Linguistic and Philosophy*. Dordrecht: Kluwer.
- Park, J. S., Rohrbach, M., Darrell, T., & Rohrbach, A. (2019). Adversarial inference for multi-sentence video description. *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*. Retrieved from <https://arxiv.org/abs/1812.05634>
- Propp, V. (1958). *Morphology of the folktale*. Bloomington, Indiana: Research Center Indiana University.
- Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L.-J. (2017). Deep reinforcement learning-based image captioning with embedding reward. Retrieved from <https://arxiv.org/abs/1704.03899>
- Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Retrieved from [http://www.cv-foundation.org/openaccess/content\\_cvpr\\_2015/papers/Rohrbach\\_A\\_Dataset\\_for\\_2015\\_CVPR\\_paper.pdf](http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Rohrbach_A_Dataset_for_2015_CVPR_paper.pdf)
- Rohrbach, A., Rohrbach, M., Tang, S., Oh, S. J., & Schiele, B. (2017). Generating descriptions with grounded and co-referenced people. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Retrieved from <https://arxiv.org/abs/1704.01518>
- Ruben, J. (Director), & Cohen, B., Jinks, D., & Roth, J. (Producers). *The Forgotten* (2004). [DVD]. USA: Columbia Pictures.
- Scherfig, L. (Director), & Dwyer, F., & Posey, A. (Producers). (2009). *An Education* [DVD]. UK: Sony Pictures Classics.
- Shadyac, T. (Director), & Shadyac, T., Carrey, J., Brubaker, J. D., Bostick, M., Koren, S., & O'Keefe, M. (Producers). (2003). *Bruce Almighty* [DVD]. USA: Universal Pictures.
- Shklovsky, V. (1990). *Theory of Prose*. (B. Sher Trans.). Illinois: Dalkey Archive Press. (Original work published 1965).
- Sjöberg, M., Tavakoli, H. R., Xu, Z., Mantecón, H. L., & Laaksonen, J. (2018) PicSOM Experiments in TRECVID 2018. *Proceedings of the TRECVID 2018 Workshop*, Gaithersburg, MD, USA.
- Smilevski, M., Lalkovski, I., & Madjarov, G. (2018). Stories for images-in-sequence by using visual and narrative components. *Communications in Computer and Information Science*, 940, 148–159.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. 2nd ed. Oxford: Blackwell.



- Szabó, I. (Director), & Lantos, R. (Producer). (2004). *Being Julia* [DVD]. USA: Sony Pictures Classics.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D. Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference 2015 on Computer Vision and Pattern Recognition*. Retrieved from <https://arxiv.org/abs/1409.4842>
- Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence – Video to text. *Proceedings of 2015 IEEE International Conference on Computer Vision*. Retrieved from <https://arxiv.org/abs/1505.00487>
- Webb, M. (Director), & Novick, M., Tuchinsky, J., Waters, M., & Wolfe, S. J. (Producers). (2009). *500 Days of Summer* [DVD]. USA: Fox Searchlight Pictures.
- Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Retrieved from <https://ieeexplore.ieee.org/document/7780940>
- Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Retrieved from <https://www.cv-foundation.org/openaccess/ICCV2015.py>
- Yus, F. (2008). Inferring from comics: A multi-stage account. *Quaderns de Filologia. Estudis de Comunicació* [Philology Notebooks. Communication Studies], 3, 223–249.