Formalising Stories: Sequences of Events and State Changes

Andrew Vassiliou, Andrew Salway, David Pitt
Department of Computing, University of Surrey, Guildford, GU2 7XH, UK
+44 (0)1483 686058
a.vassiliou, a.salway, d.pitt@surrey.ac.uk

Abstract

An attempt is made here to synthesise ideas from theories of narrative and computer science in order to model high level semantic video content, especially for films. A notation is proposed for describing sequences of interrelated events and states in narratives. The investigation focuses on the idea of modelling video content as a sequence of states: sequences of characters' emotional states are considered as a case study. An existing method for extracting information about emotion in film is formalised and extended with a metric to compare the distribution of emotions in two films.

1. Introduction

Multimedia data is treated at different levels of abstraction: as a bit stream for storage and transmission, as low-level features such as colour, texture and shape for sketch-based retrieval or in terms of 'semantic content'. For video data, semantic content is typically considered to consist of an inventory of objects and events organised in space and time [1, 2]. Our suggestion is that high-level semantic content needs to include information about state changes: this idea seems important for multimedia data that tells a story.

Whether experiencing current affairs stories in newspapers and on television, or works of fiction in novels and films, humans seem to be interested in characters' motives and feelings in order to make sense of and anticipate what is happening, i.e. the sequence of events. Scholars of narratology study stories, or narratives, as well as the processes of story telling and story understanding. Narrative is commonly defined as a sequence of events, organized in space and time, where the events are linked by cause-effect relationships [3]. For films this definition has been extended so that the agents of cause-effect relationships are said to be characters [4]. When human beings watch a film they make sense of and anticipate the unfolding events that are depicted on-screen, based at least in part on what they think about characters' cognitive states, e.g. their goals, beliefs and emotions. In this way an audience can explain what happened in a film in terms of causal connections. States change as a story progresses and these changes help to make a story engaging and help the audience to make sense of what is happening.

States may be divided into two categories: physical and cognitive. Some events lead to changes in physical states; 'standing up' changes the state 'seated' to 'standing'. Other events cause, or result from, changes in cognitive states. A film's characters trigger and react to the events that occur: the characters' actions and reactions may be explained in terms of their cognitive states such as goals, beliefs and emotions.

Consider an example from the film "The English Patient" (1996). The following text summarises a scene from the film.

> Hana, an army nurse, is riding in the back of a truck. Jan, her friend, is driven up behind her in a jeep and asks Hana for money. The two exchange friendly pleasantries and Hana gives Jan some money. They laugh at the effort. Hana sits back down and watches the jeep drive away. The jeep then hits a landmine and explodes. The convoy halts and Hana runs desperately towards the wreckage.

A sequence of events and states based on this scene is sketched informally in Figure 1. This captures something about changes of each character's physical state, e.g. Hana changes from standing to seated, and their cognitive state, Hana changes from happy to distraught. Her being distraught can be understood as experiencing the emotion DISTRESS because an event (explosion) causes her to think that

0-7803-8603-5/04/$20.00 ©2004 IEEE

something bad may have happened (e.g. her friend Jan being harmed).

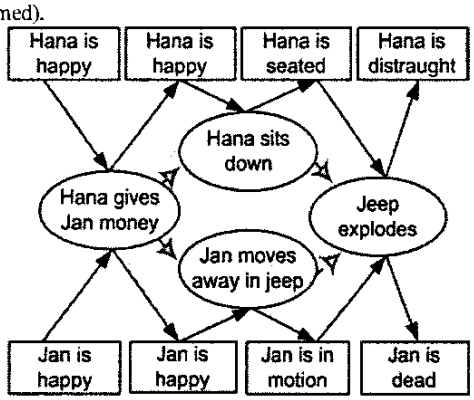

Figure 1. *Changes in state for Hana and Jan and the connecting events that bring about these changes in state.*

It seems that, for video retrieval and browsing applications based on semantic video content, information about changes of state may be as important as information about what is happening. Our proposal is to model the semantic content of multimedia as a sequence of events in space and time (as is common already), plus an interrelated sequence of state changes.

Talk of sequences of events and state changes is fundamental to computer science. These concepts have been utilized in most models of computation including automata theory and have formed the basis for most work on formal methods such as CSP [5], essentially based on sequences of events and B [6], modelling state changes. The object orientation paradigm is also built around objects with states and events or methods changing those states.

Previously, systems have been proposed for browsing video by narrative structures [7] and formalisms have been developed for representing and reasoning about narratives [8] and for the generation of narrative prose [9]. Systems to represent high-level semantics of stories through low-level video features have also been proposed, i.e. using motion in shots and shot length to determine a film's *tempo* or pace [10] and mapping colour, colour saturation and motion intensity to high-level emotional events of a film [11]. There has also been a recent growing interest in giving machines emotional intelligence [12], recognising facial expressions [13], recognising emotions from speech [14] and animating facial expressions [15].

Our research is much in the same spirit as [10, 11], representing high-level semantics of narratives using low-level textual features.

This paper presents the initial stages of an attempt to synthesise ideas from narrative theory and from computer science with the aim of applying them to the task of accessing and interacting with films stored in video databases. A formalism is introduced and discussed in the context of a case study in which sequences of cognitive states, specifically emotions, are considered as representations of semantic video content.

2. Towards Formalisation

Our interest is in modelling sequences of interrelated events and states, i.e. characters' emotions, depicted in films. At any time t, something is in some state and something *may* be happening. There is also an *order* of both events and states that is preserved in the viewer's mind. Thus the order and knowledge of what is occurring at t must be captured in our notion of change of state.

In order to formalise some of these observations let us consider a finite set **S** which encompasses all possible states, both physical and cognitive. There is a timeline **T** of time-points of the real numbers $\Re$ and a set **Ex** that denotes a set of entities or existents relating to the states and events. For any given state, the *existent* in that state and the *time* the state occurs must be known. At any point in time, t an existent, ex, is in a state $\mathbf{existSt}(t)(ex)$. Two functions accommodate this:

$$\mathbf{existSt}: \mathbf{T} \rightarrow (\mathbf{Ex} \rightarrow \mathbf{S}) \quad (fn\ \text{a})$$

$$\mathbf{timeSt}: \mathbf{Ex} \rightarrow (\mathbf{T} \rightarrow \mathbf{S}) \quad (fn\ \text{b})$$

Thus $\mathbf{existSt}(t)$ maps existents to their states at time t and $\mathbf{timeSt}(ex)$: $\mathbf{T} \rightarrow \mathbf{S}$ is the 'time state' line for existent ex. It must be noted that an existent may exist in both a physical and cognitive state. Note that:

$$\mathbf{existSt}(t)(ex) = \mathbf{timeSt}(ex)(t)$$

There is a finite set of events **Ev** and a function, **whenEv**, mapping events to the time at which they occur.

$$\mathbf{whenEV}: \mathbf{Ev} \rightarrow \mathbf{T} \quad (fn\ \text{c})$$

Events here are modelled as discrete atomic units with no duration and as thus can be considered to act at only *one* point in time. The concern is not the internal structure of the event but rather what happened.

A relation exists: **Impacts**: $\mathrm{Ev} \leftrightarrow \mathrm{Ex}$, linking events to involved existents, where an event ev impacts an existent ex, (written ev**impacts**ex). The existent may or may not change state. However, the notion that an existent may only change state as a consequence of an event that impacts it, is required.

I.e. if ex, $t_2 > t_1$ are such that $\mathbf{existSt}(t_1)(ex) \neq \mathbf{existSt}(t_2)(ex)$
Then there exists $ev \in \mathbf{Ev}$ such that ev impacts ex and $t_1 \leq \mathbf{whenEv}(ev) \leq t_2$

Issues of modelling time for changes of state exist. It has already been established that time is modelled discretely for events and this has been applied to states as well. This is done because there is the issue of duration for both an event and a state. Questions such as when exactly did something happen and how long were our characters in state **S** are not addressed here.

3. Sequences of States: Emotions

If information about characters' cognitive states, and how events impact upon them, is important for understanding a film's semantic content, then it is interesting to consider a cognitive theory of emotions that defines a character's emotions in terms of their beliefs and goals [16]. For example, FEAR is defined as the belief that something will happen with negative consequences for that character's goal; RELIEF is experienced if the feared event doesn't happen. Someone may be nervous, a kind of FEAR, because they think they have failed an exam, but experience RELIEF when they hear they have passed.

Information about emotional states can give insights into characters' goals and beliefs about the events taking place around them, and can also help to explain their actions. Audiences respond and relate to characters' emotions and criticize lack of emotion. Characters' emotions are manifested in different ways, like a change in body language, tone of voice or facial expression.

Previous work describes a method to extract information about characters' emotions in films from audio description [17].

A set of 22 emotion types, originally from [16], is denoted by: E_{type}.

$$E_{type} = \{JOY, FEAR, ANGER, SHAME...\}$$

Time is considered a timeline here composed of a series of time points. Thus let 'Timeline' denote a finite subset, **T** of time-points, of the real numbers $\Re$.

A proposed Emotion-Time frame can be applied to the elements E_{type} and **T**. Here the Emotion-Time frame is a triple (**E**, **T**, **OcAt**) where:

$\mathbf{E} \subseteq E_{type}$, **T** is a Timeline

OcAt (Occurs At) is a relation **E** to **T**

I.e. $\mathbf{OcAt} \subseteq \mathbf{E} \times \mathbf{T}$ (1)

Example *E1*

OcAt = {(JOY, 5'10"), (JOY, 25'34"), (JOY, 37'01"), (JOY, 55'19"), (JOY, 73'42"), (JOY, 121'22"), (FEAR, 12'42"), (FEAR, 14'), (FEAR, 17'54"), (FEAR, 50'41"), ..., (ANGER, 103'38")}

The relationship in (1) would be written as *x***OcA***t y* (Where *x* is an E_{type} and *y* is a time **T**), giving a representation 'JOY Occurs at 73 minutes 42 seconds' for example.

An increasing numbers of films (currently >500) are provided with audio description for the visually impaired – this is a description of important visual information that is spoken in the gaps between dialogue in scenes. Audio description potentially includes information about emotions that are manifested visually, e.g. 'She looks afraid' where 'afraid' is taken to be a token of the emotion type FEAR. Note, before it is recorded audio description is scripted as a time-coded text.

A number of tokens are assigned to an emotion-type and this association partitions E_{type}. That is there is a function:

$$\mathbf{typeOf}: E_{token} \rightarrow E_{type} \quad (fn\ 1)$$

For example:

typeOf(euphoria) = JOY and **typeOf**(afraid) = FEAR

Thus the set $\mathbf{typeOf}^{-1}(x)$, as *x* ranges over the set E_{type}, partitions E_{token}. *E.g.*

$$\mathbf{typeOf}^{-1}(JOY)^{\dagger\dagger} = \{euphoria,...pleased\} \subseteq E_{token}$$

E_{token} is taken to denote the set emotion-tokens. E_{token} has respective, numbered sets e.g. E_{token1} = {euphoria, elation, happy, jolly, pleased...} for JOY, $E_{token21}$={distraught, anguished, depressed...} for DISTRESS.

The function **OcAt** can then be constructed by scanning the time-coded audio description text for occurrences of emotion tokens, and the results displayed in a graph.

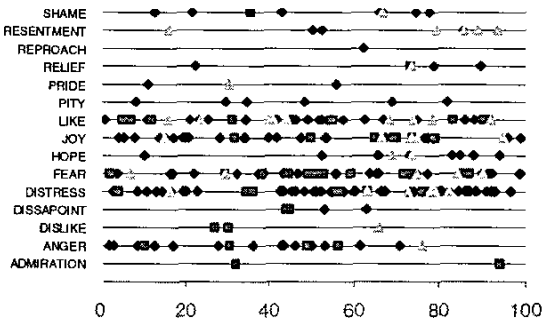

Figure 2. *A plot of emotion tokens found in 2 audio description scripts and 1 post-production script for:* The English Patient *(Minghella, A., 1996) against time (percentage of total film time).*

The plots in Figure 2 show that the most frequent emotions of the film were FEAR, DISTRESS, JOY and LIKE. Only about fifteen of the possible twenty-two emotions types [16] were found, some with only one instance, such as PRIDE and ADMIRATION. We can see from the plots that areas of the graph may relate to key events in the film where the points are 'densest'. In this case FEAR and DISTRESS are most compact around the middle of the film and this correlates directly to the film's 'war' scenes. This was verified by viewing the respective film segments.

By glancing over Figure 2 we can see that in areas of the graph the Audio description scripts and the actual script do correlate. There are however areas that do not correlate. This may be for several reasons. Negative words such as: 'not', 'don't look', that precede the emotion tokens are not taken into account. This may be giving false results, false 'positives'; for instance when "He was not happy!" is encountered it is recorded as JOY due to the word 'happy'. Also, there is no account for perspective, e.g. the overall atmosphere may be of a FEAR emotion type but the

[††] Here $\mathbf{f}^{-1}(x)$ is being used to denote a pre-image of x under a function $\mathbf{f}(x)$.

characters may be in DISTRESS and the villain may be in a state of JOY, in this case our system may detect: JOY, FEAR or DISTRESS.

Aside from these issues, it is hoped that our ideas can be used to locate patterns in different films and as a means of comparing distributions of emotions over time in different films, e.g. perhaps for genre classification or video retrieval by story similarity. For this purpose the following comparison metric has been developed.

4. Comparing Sequences of States

Let **M** be the metric space of the real numbers $\Re$ with metric distance $\mathbf{d}(x, y) = |x-y|$ The elements of **M** will be used as timepoints. Let $x \in \mathbf{M}$ and **U** be a finite subset of **M** i.e. a member of $\mathcal{F}(\mathbf{M})$. Then the distance **d** from x to **U** is defined as the minimum distance from point x to a point in **U**.

Definition: $\mathbf{d}: \mathbf{M} \times \mathcal{F}(\mathbf{M}) \rightarrow \Re^+$

$$\text{distance } \mathbf{d}(x, \mathbf{U}) = \min\{\mathbf{d}(x, u) \mid u \in \mathbf{U}\} \quad (2)$$

E.g. For $x = 3.1$, $\mathbf{U} = \{1, 2.5, 3.7, 4.1\}$,
distance $\mathbf{d}(x, \mathbf{U}) = 0.6$

If $x \in \mathbf{U}$ then this distance, **d**, is zero and if $x \notin \mathbf{U}$ then this distance is non-zero.

It may be the case that points in one timeline may not appear *at all* in another set **V** for example. For this metric distance comparison, if a time-point (say x) is not in the set **U** then the metric distance is not equal to zero. Thus to compare two time-point sets (timelines) of for the same emotion-type the following may be considered.
Let **V** be another finite subset of **M**.

Definition: Distance $\mathbf{d}(\mathbf{V}, \mathbf{U}) =$

$$\frac{\sum_{v \in \mathbf{V}} \mathbf{d}(v, \mathbf{U})}{\text{Cardinality } (\mathbf{V})} + \frac{\sum_{u \in \mathbf{U}} \mathbf{d}(u, \mathbf{V})}{\text{Cardinality } (\mathbf{U})} \quad (3)$$

In (3) the two time sets **U** and **V** are being compared within the metric space **M**. Basically (3) is the sum of the average distance from points in **V** to the set **U** and the average distance from points in **U** to the set **V**.

With distance metric (3), the only way that the distance between the subsequent time sets can be *zero* is if one set is identical to another or a subset of the other; ignoring the outliers (completely outlying points).

5. Ongoing Work

In order to investigate the extent to which a sequence of states alone can be used to represent semantic video content, we are currently using the distance metric to compare: (i) sequences of emotions extracted from audio description of different films; and, (ii) sequences of emotions extracted from different audio descriptions and a script of the same film. More generally we are continuing to explore ways in which formal languages can be applied to the events and states of filmed narratives.

6. Acknowledgements

This research is supported by EPSRC grant GR/R67194/01 – TIWO: Television in Words.

7. References

[1] Chen, S.-C., Kashyap, R. L. and Ghafoor, A., *Semantic Models for Multimedia Database Searching and Browsing.* Kluwer Academic Publishers, 2000.

[2] Agius, H.W. and Angelides, M.C., "Modelling Content for Semantic-Level Querying of Multimedia", *Multimedia Tools and Applications,* Volume 15 (1), 2001, Pages 5-37.

[3] Chatman, S., *Story and Discourse: narrative structure in fiction and film*, Ithaca: Cornell University Press, 1978.

[4] Bordwell, D. and Thomson, K. *Film Art: An Introduction*, McGraw-Hill, 5th edition, New York, 1997.

[5] Roscoe, A.W., *Theory and Practice of Concurrency*, Prentice Hall, 1997.

[6] Abrial, JR., *The B-Book*, Cambridge University Press, 1996.

[7] Allen, R. B. and Acheson, J., "Browsing the structure of multimedia stories." *In Procs. 5th ACM Conference on Digital Libraries*, ACM Press, New York, 2000, Pages 11-18.

[8] Baral, C., Gabaldon, A. and Provetti, A., "Formalizing narratives using nested circumscription.", *Artificial Intelligence*, Volume 104, Issues 1-2, September 1998, Pages 107-164.

[9] Callaway, C. B. and Lester, J., "Narrative prose Generation.", *Journal of Artificial Intelligence*, Volume 139, June 2002, Pages 213-252.

[10] Adams, B., Dorai C., and Venkatesh S., "Towards Automatic Extraction of Expressive Elements From Motion Pictures: Tempo", *IEEE Transaction on multimedia*, Volume 4, 2002, Pages 472-481.

[11] Kang, H, "Affective content detection using HMMs", *In procs. of the 11th ACM International Conference on Multimedia*, ACM press, Berkeley, November 4-6, 2003, Pages 259-262.

[12] Picard, R.W., Vyzas, E., Healey, J., "Toward machine emotional intelligence: analysis of affective physiological state" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Volume 23, Issue 10, October 2001, Pages 1175 –1191.

[13] Pantie, M. and Rothkrantz, L.J.M., "Automatic analysis of facial expressions: the state of the art", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Volume 22, Issue 12, Dec 2000, Pages 1424-1445.

[14] Devillers, L., Lamel, L., Vasilescu, I., "Emotion detection in task oriented spoken dialogs", *In Proceedings of ICME 2003*, IEEE, Pages 549-552.

[15] Raouzaiou, A., Karpouzis, K., Kollias, S., "Emotion representation for online gaming", *In Procs. of ICME 2003*, IEEE, Pages 417-420.

[16] Ortony, A., G. L. Clore and A. Collins, *The Cognitive structure of emotions*, Cambridge University Press, 1988.

[17] Salway, A. and Graham, M., "Extracting Information about emotions from film." *In procs. of the 11th ACM International Conference on Multimedia*, ACM press, Berkeley, November 4-6, 2003, Pages 299-302.