Andrew Salway and Khurshid Ahmad, “Computational Semiotics: a framework for integrating multimedia information?”

In: Holmqvist, Kuhnlein and Rieser (eds.), Procs. Integrating Information from Different Channels in Multi-Media Contexts. Workshop at ESSLLI 2000.

Computational Semiotics: a framework for integrating multimedia information?

Andrew Salway and Khurshid Ahmad, Department of Computing, University of Surrey
(a.salway@surrey.ac.uk)

Introduction

Users can trigger reactions from machines through different channels, such as speech, hand gestures and facial expressions, such that keyboards and the mouse may be superseded. However multimodal integration is not only an important issue for the ephemeral interactions between humans and machines, but it is also important for dealing with digital archives, including those that exist as records of scientific and artistic endeavour. Consider collections of scientific images like medical X-rays and weather pictures from satellites, and formally constrained works of art like paintings and dance. When people, particularly experts, analyse these still and moving images they often produce verbal descriptions and interpretations which are sometimes complemented with diagrams, graphs and notations. Such information can be seen as 'collateral' to the images since it is related to them in various interesting ways: notably it serves to elucidate the contents of the images for less-expert viewers (which may include machines).

An everyday kind of collateral information is exemplified by considering someone who owns a video cassette of a movie, the novel on which it was based, the CD of the soundtrack and maybe some posters of the actors and actresses (Breakfast at Tiffany’s is a good example here). These items will probably be inconveniently spread around their home; not to mention other related information such as the biographies of the actors and the musical score which may reside in some distant library. However, though they are physically disparate, all these collateral media artefacts are in a sense integrated in the mind/brain of the individual who has seen/read/heard them. So, a few notes of the soundtrack may evoke a scene in the movie and reading the novel might help in understanding the movie. Whilst multimedia computing systems allow digital renderings of movies, books, soundtracks, etc. to be gathered within the same system, there is still much to be done in order to integrate these media artefacts at higher levels.

Dance is a more specialist example to consider here since it is clear that a dance expert attends to a range of visual information - primarily dancers' movements, but also costumes, stage sets and lighting, as well as auditory information - especially music. One kind of information can provide cues or a context for understanding information in another mode, for example a haunting piece of music might mean a dancer's movements were seen to be fearful. In the case of moving images information arrives at different times, such that, for example, a dancer's character as indicated by his/her costume is apparent from the start, but it is through the ongoing movements that the narrative unfolds. There is also a contribution made to understanding by the textual information in a dance programme read before, during or after a performance. For these reasons dance may be considered an interesting case of a ‘composite signal’.

Human cognition requires the combination of information from different sensory sources, and now machines have access to repositories of digital multimedia information comprising related texts, images and sounds to which users require integrated access. In order for 'intelligent' information access it will be necessary to understand the relationships between the different media artefacts so that appropriate surrogates can be attached for query-retrieval purposes, and so that structures of nodes and links can be laid over the information for hypermedia browsing. Another aim would be to automate the combined analysis of the different information sources. So, although visual, auditory, olfactory and tactile sensations can be transformed into digital data streams along with text, for the further development of information systems at least, there remains an important challenge in 'providing an integration of representations - a media interlingua - or at least integrated access to heterogeneous media representations' (Maybury 1997:xxi).

The integration of multimedia/multimodal information is relevant to scholars of computation who are concerned with the formal description of such information, with the development of information systems including digital libraries, and with the study from a computational perspective of human cognitive and communicative processes. The relationships between different modalities and media, for example vision/language and image/text have been of interest for an even longer time to philosophers of aesthetics, and to semioticians who have sought to develop a science of signs. This paper attempts to synthesise discussions from these areas to start developing a framework for investigating multimodal information systems. In particular we are concerned with modelling, or representing, dance in a computing system. On a more empirical note, we have been interested in how dance experts combine different strands of information in their analyses of dance, and how they articulate descriptions and interpretations of dances. This study has contributed to the development of an information system which integrates dance video data and collateral texts for retrieval and browsing.

Dance: a composite signal

The spectator of a dance deals with a rich and dense mass of visual information concerning the poses, gestures and actions of one or more dancers. The quantity of this information is tempered though by a degree of systematicity: the dance is the product of human minds and bodies and is therefore is some sense contrived, be it through the conventions of a dance genre, the intention of the choreographer, or physical constraints like gravity and the flexibility of the human body. But it is not just information about the movements of dancers that the spectator deals with: there is other visual information about the costumes and the set of the dance. Then there is auditory information, most notably music, occasionally speech, and if the spectator is close enough perhaps the sound of the dancers movements on the stage. Whilst a dance may invoke a response in anybody, a knowledgeable spectator might be able to understand more about a dance due to an analytic training, and access to prior information about the dance, the choreographer, etc. that puts the ephemeral movements and music in a rich context of multimedia information. Thus dance experts are able to highlight the intricacies of a dance to a lay viewer, and to explain the meanings that the dance seeks to convey.

It might be argued, whatever the media or modes of the artefact, that once a human expert begins to analyse and then to describe and interpret the artefact, it is natural language that comes to the fore of all the interacting sign systems. And so it is perhaps possible to study some aspects of the cognitive processes involved in analysing dances through the language spoken and written about them, and furthermore it may be that natural language and/or formal languages (logic and knowledge representations) provide suitable surrogates for storing and accessing multimedia data. Of course there are also movement notation systems to consider as surrogates for dance. Systems like Labanotation, Benesh and Eshkol-Wachman can all produce a record of a dance performance, or describe how a dance is to be performed (there is some similarity between these systems and musical notation). The degree of precision to which the dance is replicated depends on the needs and the skills of the annotator. However, the physical movement recorded by the dance annotator comprises but one strand of the information required for understanding a dance, and does not on its own equate with the meaning of the dance as intended by the choreographer or as understood by the viewer.

A method used by the linguist Wallace Chafe and his colleagues to study language production involved showing subjects a short film and then asking them to speak a narrative of the film after viewing (Chafe 1980). This technique has been related to the cognitive psychology method of Protocol Analysis in which subjects are asked to 'think aloud' and their 'verbal reports' are then taken as the data for developing theories about cognitive processes (Ericsson and Simon 1993). We have used a similar technique to observe how human experts combine the strands of information in a composite signal, like dance, and produce a collateral artefact in a single medium, e.g. natural language. The instructions given to subjects can focus their attention on particular strands in the composite signal and it is perhaps reasonable to assume that what they speak about is what is predominant in their thoughts at the time. When commentating on a moving image the speaker's words tends to follow the flow of the moving image allowing for some alignment between moving image and text (compare this with an expert commentary on a still image like a painting).

Commentaries were elicited from five dance experts as they watched a compilation of five dance sequences totalling 20 minutes. First the experts were asked to 'Describe' the sequences and then to 'Interpret' them. A linguistic analysis of the commentaries (21,000 words in all) highlighted systematic differences between descriptions and interpretations which might be taken to reflect differences in the kinds of information being attended to when performing the tasks. Descriptions tended to focus on one strand of information, most notably the dancers' movements, with occasional mentions of costume, stage and music. For the interpretations, on the other hand, the experts seemed to draw the information sources together in order to explain what the dance conveyed.

The results of this investigation, which also included the analysis of a 350,000 word corpus of extant dance texts, have contributed to and benefited from the development of an object-oriented multimedia information system which organises digital video data and collateral texts (e.g. experts' commentaries of dances) so that one can be used to access the other. Keywords are identified in time-coded text fragments and used to index video intervals: preliminary results also suggest that the time-coded commentaries can be segmented in a way that reflects significant sections in the dance. Once a user has retrieved a video sequence they can browse relevant texts and navigate through the video by them. For full details of our analysis of experts talking and writing about dance, and of the KAB video annotation system see Salway (1999).

Multimedia Information Processing

The integration of multimedia information is achieved in one respect through current systems which can handle digital images, videos, audio files as well as text: the fact that these media artefacts exist in the same systems (as sequences of 0's and 1's) constitutes a level of integration. However, the encoding of complex and composite signals, like dance, in computing systems does not currently recognise the contributions made by different kinds of information to communication. So, although a dance may be reproduced by a digital video data file (including audio data) via a monitor and speakers its digital encoding does not make the different strands of information in the dance and nor its semantics accessible to the machine. In order to develop systems for automatically analysing or generating dance, computer-based representations will need to make explicit how the parts of a composite signal contribute to the whole. Here we review a variety of research concerned with integrating multimedia information in computational systems: some of this work has been motivated by the need for intelligent information systems to store, retrieve, analyse and generate multimedia data and other work has been carried out to study the human mind/brain. Another contrast can be made between those approaches that deal with statistical (some might say sub-symbolic) features of multimedia information and other approaches that involve an intermediary (symbolic) language.

Multimedia data (image files, sound files, etc.) can be associated according to perceptual similarity which may be captured by statistical features (like colour, shape, texture metrics for images and frequency metrics for sound). The notion of perceptual thesauri that realise associations between visual, haptic, taste, olfactory and audio sensations using the same statistically-based computational model was presented by Picard (1995); she suggested that such associations within and between modes would constitute a representation of 'perceptual knowledge'. Such a representation would be useful for information retrieval purposes, e.g. to retrieve similar looking images from a database. Visual and auditory features can be used to recognise objects and events: in temporal media, like video, the timing of these features can be significant. Chang et al. (1999) describe multijects (multimedia objects) which are bundles of low-level features organised in interlinked Hidden Markov Models to address their multimodal and temporal nature; e.g. an explosion might be characterised by a thundering sound closely followed by a flash of red. The authors go on to suggest how further evidence for the identification of multijects might be gained from information about the co-occurrence of multijects: this co-occurrence information would be held in a probabilistic graph that is perhaps reminiscent of some models of semantic memory.

For some researchers interested in human cognition the awareness of sensory events, either experienced immediately or as a memory, can be equated with an 'iconic representation' – that is the neural activity normally caused by those events. From this point of view the process of acquiring knowledge is the process of linking iconic representations to one another over time, and crucially across sensory modalities. This is exemplified by the task of object naming that was learnt by artificial neural networks that maintained the physical characteristics of visual inputs and linked them to phonological inputs (Aleksander 1999). Other researchers have worked on the premise that 'basic language categories (e.g. concerning objects and actions) are 'grounded' in sensory/motor experiences, while more abstract concepts are learned through metaphorical extension' (Nenov and Dyer 1994). Their DETE system, comprising 50 ANN modules learnt to describe object names, adjectives and motions for visual inputs. The idea that abstract concepts are metaphorical extensions of (multimodal) physical experience has been elaborated under the Neural Theory of Language Paradigm. Jerome Feldman and colleagues have produced a series of neural network simulations that show how systems trained on information from an immediate physical environment can then cope with abstract reasoning tasks (Lakoff and Johnson 1999

Statistical and neural network approaches for multimedia integration may be contrasted with approaches that use an intermediary language. Halliday's functional systemic grammar was used to represent the 'ideational' content of texts about epidemics and graphical illustrations of related data (Cross and Matthiessen 1998). Another researcher used a logic-based representation as a common denominator from which the text of instruction leaflets could be generated and appropriate diagrams selected from a database (van Deemter 1998). In hypermedia systems, texts and images can be navigated via links which may include generic links from any instance of a given anchor to a fixed set of text fragments and images. The use of a thesaurus to provide a 'conceptual layer' allows links to be made from an image to a 'concept' such that the image is automatically linked to further texts and images associated with the concept (Dobie et al. 1999).

The use of a movement notation system was proposed as an intermediate representation for the generation of human movement animations from textual specifications and for the generation of descriptions from video sequences (Badler and Smoliar 1979). As noted previously, movement notations may provide excellent records of human movement but do not explicate the meanings that might be conveyed by the movement. A set of perceptually grounded movement primitives was used to represent movement verbs in such as way as to account for some aspects of their meaning, i.e. by facilitating inferencing (Siskind 1996. A range of knowledge representation schemes have been used to provide surrogates for video data. They have been used to give unambiguous descriptions of entities and actions in space and time (Davis 1995) and to represent causal relationships between video sequences, e.g. a sequence showing an explosion and a sequence showing the person who pressed the detonator (Roth 1999).

Two broad approaches to the integration of multimedia information can perhaps be characterised on the basis of this literature review; both approaches are realised in work that has the practical concerns of multimedia information systems at heart as well as in work that is concerned with developing theories of human cognition. The first kind of approach deals with multimedia information by characterising its physical properties, usually statistically, e.g. through artificial neural networks. The other kind of approach aims to explicate the meaning of multimedia information using intermediary formalisms bearing some relation to natural language.

Computational Semiotics

Current multimedia systems process various types of multimedia objects, including written texts, sounds and pictures at a physical level – as streams of binary data. However, it should not be forgotten that these multimedia objects are digital renderings of artefacts that owe their existence to various modes of human cognition and communication. In order to address the meaning-bearing potential of multimedia artefacts it might be important to consider the so-called science of signs, that is semiotics. Semiotics is concerned with the nature and use of signs in signification and communication involving humans, animals and machines (Eco 1976, Sebeok 1994).

Semiotic theories attempt to explain how sign systems organise knowledge and convey meaning. They do this be classifying different types of signs (according to how they convey meaning), by elaborating models of semiosis (the process by which signs are interpreted) and by the analysis of semiotic systems (both formal grammars and theoretical discussions about meaning in domains like art, cinema and dance). We believe that the classification of signs and theories of signification and communication will be valuable in developing computational systems that integrate multimodal information: computer-based systems could in turn contribute to the development of semiotic theories. Thus we have envisaged 'computational semiotics' (Ahmad, Salway and Lansdale submitted). There are parallels here for us with computational linguistics where natural language processing and theoretical linguistics enjoy a symbiotic relationship. Linguistic theories and empirical descriptions have guided the development of systems for speech recognition, information extraction and text classification. In return linguists have benefited from computational models which they may use to ground their theoretical discussions. The success of computational semiotics as we envisage it will depend on identifying aspects of semiotic scholarship that can contribute to the development of multimedia information systems. (Coincidentally, the term ‘computational semiotics’ was recently put forward as a less contentious name for artificial intelligence (Sowa 2000)).

Semioticians have studied multimedia communication as a form in its own right for some time now. In 1980 Elam wrote, 'it is not, clearly, a single-levelled and homogeneous series of signs or signals that emerges, but rather *a weave of radically differentiated modes of expression*' (Elam 1980). Although Elam was writing about communication in the context of theatre, his words are pertinent to computer-based multimedia information. He elaborated 29 sets of codes by which communication takes place in theatre: these include the movement and appearance of the actors, the stage, the language spoken, as well as cultural, psychological and aesthetic aspects. In discussing multimedia communication, semioticians distinguish *transmission channels* (light-waves, sound-waves, biochemical, thermodynamic, electromagnetic), *senses* (acoustic, olfactory, gustatory, haptical, optical), *modes* (icons, symbols, indices) and *codes* (verbal, paraverbal, non-verbal, socio-perceptive, psycho-physical) (Hess-Luttich 1991). Research in multimedia computing has dealt with the physical coding of multimedia signals and the detection of perceptual features, cf. 'Transmission Channels' and 'Senses', and perhaps in these areas the field could contribute to semiotics. Where multimedia computing might benefit from semiotics are in those aspects described as 'Modes' and 'Codes'. These aspects refer to the higher level semiotic descriptions of signs and the way they convey meaning, both individually and as part of systems, mediated by social and cultural factors and including the phenomenon of metaphor.

One aspect of semiotic study is the sign itself and the ways in which it can bear meaning. To this end, many scholars have sought to elaborate Peirce's trichotomy of *Icon*, *Index* and *Symbol*. Signs are typically characterised as some mixture of *icon* (a sign which conveys meaning by some perceived similarity to its referent); *index* (a sign connected to its referent by a causal relationship); and, *symbol* (a sign with an arbitrary, but socially mediated, relationship to its referent). Recently an article in IEEE Multimedia suggested that Peirce's sign types could be used as the basis for a taxonomy of representational systems for multimedia computing (Purchase 1998). The three dimensions of the taxonomy were the nature of the sign (concrete-iconic, abstract-iconic or symbolic), the arrangement of signs (individual, augmentation, temporal, linear, schematic) and the modality (visual or aural). Another researcher has described how the syntactic aspects of semiotic theory can contribute to the structured representation of hypermedia (Gonzalez 1997). As well as 'static' taxonomies of representation systems and data models, it is important to consider the processes by which media artefacts are created and understood: it has been suggested that 'such an inquiry is not well served by the intellectual foundations of computer science but must instead turn to principles laid down by disciplines such as semiotics and hermenuetics' (Minneman and Smoliar 19XX).

Summary

This paper has considered some of the challenges related to the modelling of composite signals, specifically dance (for analysis and for generation). A review of computational systems showed how different kinds of information can be integrated through statistical/numerical encodings and through intermediary languages. The further development of such systems may benefit from a framework based on semiotic theories that characterise different kinds of signs, the systems in which they are organised and the processes by which they convey meaning. The success of such an endeavour would perhaps lie in the strength of the mapping between theories about human communication and the data structures and algorithms of information processing systems.