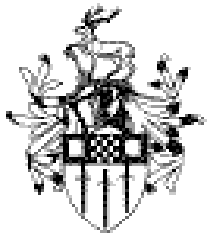
Video Annotation: the Role of Specialist Text

Andrew Salway

Ph.D. Thesis

Submitted to the University of Surrey in partial fulfilment of requirements for the degree of Doctor of Philosophy



UniS

Department of Computing
School of Electronic Engineering, Information Technology and Mathematics
University of Surrey
Guildford, GU2 5XH
United Kingdom

December 1999

Abstract

Digital video is among the most information-intensive modes of communication. The retrieval of video from digital libraries, along with sound and text, is a major challenge for the computing community in general and for the artificial intelligence community specifically. The advent of digital video has set some old questions in a new light. Questions relating to aesthetics and to the role of *surrogates* – image for reality and text for image, invariably touch upon the link between vision and language. Dealing with this link computationally is important for the artificial intelligence enterprise. Interesting images to consider both aesthetically and for research in video retrieval include those which are constrained and patterned, and which convey rich meanings; for example, dance. These are specialist images for us and require a special language for description and interpretation. Furthermore, they require specialist knowledge to be understood since there is usually more than meets the untrained eye: this knowledge may also be articulated in the language of the specialism.

In order to be retrieved effectively and efficiently, video has to be *annotated*; particularly so for specialist moving images. Annotation involves attaching keywords from the specialism along with, for us, commentaries produced by experts, including those written and spoken specifically for annotation and those obtained from a corpus of extant texts. A system that processes such collateral text for video annotation should perhaps be grounded in an understanding of the link between vision and language.

This thesis attempts to synthesise ideas from artificial intelligence, multimedia systems, linguistics, cognitive psychology and aesthetics. The link between vision and language is explored by focusing on moving images of dance and the special language used to describe and interpret them. We have developed an object-oriented system, KAB, which helps to annotate a digital video library with a collateral corpus of texts and terminology. User evaluation has been encouraging. The system is now available on the WWW.

Acknowledgements

My sincere thanks are due to Prof. Khurshid Ahmad whose tireless inspiration and guidance have brought out parts of me I never knew I had. I am also very grateful to the Department of Computing and the School of EEITM for giving me a scholarship, and for providing excellent facilities for research.

I have enjoyed a stimulating and friendly working atmosphere with the members of the Artificial Intelligence Group. Particular thanks go to Lee Gillam for advice about system design and implementation; to Tracey Bale for leading the way on the PhD path; and, to Caroline Jones for knowledgeable proof-reading well beyond the call of duty. Conversations with other members of the group have also helped to develop my thoughts, and made my time at Surrey an enjoyable one: at the risk of forgetting someone, I'd like to mention David Boulton, Grace Guo, Leon Miles, Jon Machtynger, Matthew Casey, Lena Tostevin, Shaikha Al-Jabir and Steve Collingham.

Prof. Janet Adshead-Lansdale made it possible for me to take dance as the subject area of research into moving images and collateral texts. Through her I met staff and postgraduates in the Department of Dance Studies at Surrey and because of her teaching I am left with an enhanced appreciation of dance. Others who gave their time to help me include John Duke, Jean Johnson-Jones, Carol Martin, Sherril Dodds and Helen Roberts. I am also grateful to all those who participated in verbal reporting sessions and in user evaluation.

The greatest thanks are due to my mum and dad who got me far enough in life to be able to start studying for a PhD and then continued to give immeasurable support and encouragement for three more years.

Neither words or pictures are enough, but for her love that keeps me going I will be forever thinking of Elaine.

Contents

Chapter 1

Introduction

The development of 'video-on-demand' systems requires new techniques for accessing libraries of video data. These libraries may contain, for example, films, news broadcasts, archive footage, and moving images that are of interest to specialist groups like meteorologists and dance scholars. The categorisation of video data, essentially a large number of still images each comprising a matrix of pixels, can be based on a labelling system, using for example the names of films and the people who made them, and the date and place of their production. The scheme could be refined, so that it referred to important events depicted in images, say, of changing weather systems, or of human movement in more aesthetic images like dance videos. In all cases, labels facilitate the storage and retrieval of video data: the question of whether a certain kind of label is up to the task can only be answered with reference to a specific retrieval scenario.

Labelling, be it eponymic or aesthetic, uses symbol systems other than images. The symbol system could be a ratings system used to classify films, or a notation system that can record dancers' movements. The current generation of video processing systems can be used to look at patterns in the video data itself to highlight changes in colour, texture and motion. However, in order to recognise arbitrary objects and actions in video data, it seems that research in computer vision has a way to go to automate the process for general application.

For the purposes of retrieval then, a surrogate of an image, whatever the symbol system, has to be placed side by side with the image: in the case of a moving image, the surrogate may be parallel in time. Crucially, though the surrogate may carry less information than is contained in the tens of thousands of still images that make up a video sequence, it conveys information about the moving image in a form that may be more meaningful to both a human, and to a machine. Surrogates that are placed side by side, are accompaniments to, and perhaps have an intrinsic relationship with, moving images, can be referred to as *collateral* – a word which encapsulates these different meanings.

We have explored the link between vision and language by focusing on the relationship between moving images of dance and the special language used by experts to describe and interpret such images. On the one hand the link between moving images and the collateral texts produced by dance experts can be used to access digital video data. On the other hand, perhaps more ambitiously, the link can provide a basis for investigating the dance experts' cognitive and communicative processes.

With regards to textual surrogates for images, it is perhaps interesting to note that language co-occurs with both still and moving images in various ways. There are, for example, the opening and closing credits of a film, running subtitles, and the spoken words of presenters and actors in news broadcasts and dramas. These types of text which are placed side by side with moving images may be called collateral texts. There are also the words written and spoken, often by experts, about images like dance sequences and weather patterns. So collateral texts combine with images to convey information: sometimes images illustrate texts, and in the case of experts' words, collateral texts elucidate the 'contents' of images.

Even when the argument is presented that an image or sequence of images is independent of any collateral text, such that we cannot define the humour of Charlie Chaplin in any text, and a book about a ballet might leave much to the imagination, the inadequacy of a collateral text still tells us something about the link between language and vision. Much as a collateral text depends on the existence of an image, the development and dissemination of knowledge about the image depends on collateral texts. A case in point here would be certain kinds of modern art whose practitioners strive to free their works of any kind of language, but who must resort to language to share their thoughts about their endeavours, and to receive the critical acclaim of their peers.

The relationship between an image and its collateral texts is important for storing and retrieving images, especially in the light of developments in automatic information extraction from text. Consider the information potentially available in the texts produced by experts when they analyse and discuss moving images: for example, dance experts who in their texts describe dancers' movements, delimit important sequences, elucidate the meanings of a dance

and evaluate it in relation to other dances. It might be argued that the texts are the artefacts of cognitive and communicative processes, such that the relationship between image and text corresponds in some sense to the link between vision and language.

This link is studied in different fields by scholars who are interested in human intelligence. They variously compare, contrast and discuss the relationships between verbal and visual thought; between poetry and painting; between words and pictures; and, between the mental faculties of vision and language. In artificial intelligence the link between vision and language is the subject of computational enquiry, with regards to information processing in humans and in machines. A general question here is how to correlate visual information with textual information, so that one can be converted to the other, or so that diverse information sources can be combined.

For Rohini Srihari this is the *correspondence problem*: that is, "how to correlate visual information with words" (1995:350). The use of the polysemous word 'correlate' indicates that Srihari, much as we do, thinks that the correspondence goes further than attaching keywords, i.e. nouns, to pictures; rather one has to "associate visual information with events, phrases or entire sentences" (ibid.). The correspondence problem is especially relevant with the advent of 'multimedia' computing systems that must integrate access to images, graphics and video data alongside text and speech (not to mention other kinds of data).

1.1 Content-based Indexing and Video Annotation

The widespread availability of still and moving images in digital forms has led to challenges and new opportunities for the developers of multimedia computing systems. Some of these challenges were faced previously by the curators of picture and film collections, whilst others have arisen with the opportunity to store and present image and video data, along with text and speech data, in a flexible and integrated fashion. The basic function remains that of retrieving information to meet a user's needs. The service provided by a traditional library, picture gallery or film collection ends with the location of a text, image or movie. However, in a digital library, the retrieved information can be used as the starting point for interactively

accessing further information through user interfaces that present texts alongside still and moving images.

To be able to meets users' information needs, both traditional and digital libraries must provide surrogates for the artefacts in their collections. The surrogates capture the important features of the artefact and can be used for indexing or classifying it: the question of what needs to be captured by the surrogate depends on the intended retrieval scenario, e.g. the kind of images and the intention of the user. Texts have traditionally been indexed by keywords that capture their 'aboutness', or have been organised by classification schemes such as Dewey's decimal classification. Computer-based systems for storing and retrieving texts have been able to index and classify texts automatically by generating keyword surrogates on the basis of statistical measures of word occurrence.

Picture and film collections have also been indexed by keywords and longer phrases, and have been organised by classification schemes such as Henri van de Waal's *ICONCLASS*. There are different kinds of information that can be attached to an image, be it still or moving, including details of the people involved in its production, places and dates, as well as details of what it depicts and what further significance it might have for a viewer. When indexing languages and classification schemes are implemented in computer-based systems it is feasible to associate multiple surrogates with an image, and in the case of moving images, they can be associated with particular intervals. Furthermore, once textual surrogates have been associated with an image or image sequence, techniques developed for text-based information retrieval and information extraction can be applied for the retrieval of images; for example, query expansion through thesauri. Such application of established language technologies to image and video retrieval was a major theme at a recent international workshop (Hiemstra, de Jong and Netter 1998).

Apart from the cost of manually providing surrogates for digital image and video collections, there is a problem of subjectivity which needs to be addressed. Images can depict and mean many different things to different people, and moving image sequences more so. To add to this, the choice of words to refer to the same thing in images may vary between people. Such drawbacks have led researchers to explore the idea of content-based indexing

for image and video data. In this case 'content' refers specifically to the properties of image/video data, rather than the meanings it conveys to a viewer: thus, content-based indexing includes techniques that capture the spatio-temporal distribution of pixels. Measures of colour, texture and shape can be computed for images, including key-frames in video sequences, so searches can be made in terms of *visual similarity*, as for example in IBM's *Q*uery *B*y *I*mage *C*ontent (*QBIC*) system (Flickner et al. 1997). Such features, along with measures of motion, can also be computed for *objects* which are identified across a series of frames, as in Columbia University's *VisualSEEK* (Chang et al. 1998): note that these unnamed objects are 2-dimensional regions of pixels, and may or may not correlate with an object recognisable to a viewer. The measurement of motion means that queries can specify spatio-temporal relationships between regions of pixels.

For retrieving some kinds of images, like those in trademark databases and fabric databases, visual features may do a good job of identifying appropriate images because visual similarity is of the essence. There are also restricted cases in which visual similarity will correspond with more conceptually significant similarity. However, a user will often want to retrieve images and video sequences according to semantic features, that is by naming the entities and actions depicted by the moving image and other meanings it might convey to a viewer. It must be noted that the state-of-the-art in computer vision technology cannot support the general recognition of entities and actions in images and video data, nor can it infer many of the meanings which are conveyed to a viewer.

If semantic features are to be generated automatically then, for the meantime at least, it will have to be from a source other than the visual component of video data. The notion of content-based indexing has been extended to exploit the speech and closed caption components of video data. In the case of news broadcasts and documentary programmes much of the information content is carried by the spoken words of the presenters, and the subjects on which they are speaking will reflect, albeit to varying degrees, the entities, actions and themes shown in the accompanying images. The *Informedia* system, developed at Carnegie Mellon University, indexes news broadcasts and documentary programmes by keywords that are extracted from speech and closed captions (Wactlar et al. 1999).

Other kinds of video do not contain 'integral' text, but they can be *annotated* with text that was produced specifically to describe or explain them: this collateral text is processed into machine-executable surrogates for video retrieval. The *WebSEEK* system, which has been used to index 500,000 images and videos on the WWW, selects keywords from the text of hyperlinks to images and videos on WWW-pages (Smith and Chang 1997). Another system, developed at NHK (Japan Broadcasting Corporation) laboratories, parses the notes which are kept in the production of documentary programmes, and which describe the entities and actions in the recorded footage and which are time-aligned with them. The user is then able to make queries in terms of the relationships between entities and actions (Kim and Shibata 1996).

The use, for video annotation purposes, of collateral text can be extended in scenarios where there is a rich variety of discourse about moving images, as is the case in dance scholarship. It could be argued that the spoken and written words of experts serve to explicate the semantics of moving images such as dance: that is the texts help their readers, and maybe machines, to better understand moving images. Furthermore, it could be argued that the texts produced by experts are grounded in, reflect, and in fact develop, knowledge about their domain.

This thesis presents a *K*nowledge-rich *A*nnotation and *B*rowsing system (KAB) that was developed to explore how a video database can be annotated with a collection of texts. KAB processes the text into machine-executable surrogates for retrieving video sequences and provides relevant texts to be read alongside the moving image. The term knowledge-rich is deliberate if ambitious. Keywords are *sparse* embodiments of knowledge; any elaboration of the keywords, for us, enriches the value of video surrogates.

The kinds of surrogates that are available to a retrieval system will determine how it can access digital libraries of moving images. A classification of such systems can distinguish between those which use content-based indices (features generated directly from video data); and, those which use video annotations (features generated from text that has been associated with video data), Table 1.1. It must be noted that a combination of content-based indexing and video annotation may be the ideal: in this regard the developers of the *WebSEEK* system

may be seen as pioneers. Text which has been used for video annotation includes hyperlink tags, the notes made by documentary makers, and in the KAB system a corpus comprising a variety of text types.

Video Retrieval Systems	**Surrogates for Retrieving Video Data**						**Reference**
	Content-based Indexing			**Video Annotation**			
	Key-frame Features	***Object Features***	***Keywords in Speech***	***Text in Hyperlinks***	***Production Notes***	***Text Corpus***	
QBIC (IBM)	√						Flickner et al. (1997)
VideoQ (Columbia)	√	√					Chang et al. (1998)
Informedia (CMU)	√		√				Wactlar et al. (1999)
WebSEEK (Columbia)	√			√			Smith and Chang (1997)
Dependency Structures (NHK)					√		Kim and Shibata (1996)
KAB (Surrey)				√	√	√	

Table 1.1: Six video retrieval systems organised according to the kinds of surrogates generated for video data. The content-based indexing techniques automatically generate features from video data, be they visual features (for key-frames and objects) or keywords extracted from speech or closed captions. In contrast, video annotation techniques exploit text that was produced specifically to elaborate the semantics of moving images.

1.2 Research Overview

If one accepts a link between an image and its collateral texts, the next step is to find out how the texts can be used as surrogates for accessing an image sequence. In this thesis the focus is on the use of specialist texts for video annotation. One can argue that texts like critical reviews of dances and *programmes* accompanying dance performances make good surrogates for the moving image. Between them these texts will tell us something about the movements in the dance, the technique of the dancers, the intent of the choreographer and also about the relationship of that dance with others. For labelling a video sequence one can take from the collateral texts important keywords provided by an expert writing about the sequence; for example, the expert may regard the dance as 'post-modern', or as conveying the theme of 'love'. It might be argued that the experts only give us keywords, but in fact they can also provide a wealth of other information.

When considering how experts analyse moving images and produce collateral texts it appears that the link between vision and language is actually a link mediated by *knowledge*. Specifically, this is the knowledge that well trained experts have about moving images (vision and knowledge), and the knowledge that allows them to articulate their descriptions and interpretations (knowledge and language). For us, the key issues are the processes involved in analysing and communicating about moving images. A better understanding of these issues can perhaps benefit the practical problem of annotating video data with collateral text so that it can be accessed more effectively. Figure 1.1 sketches these two potentially synergistic lines of investigation.

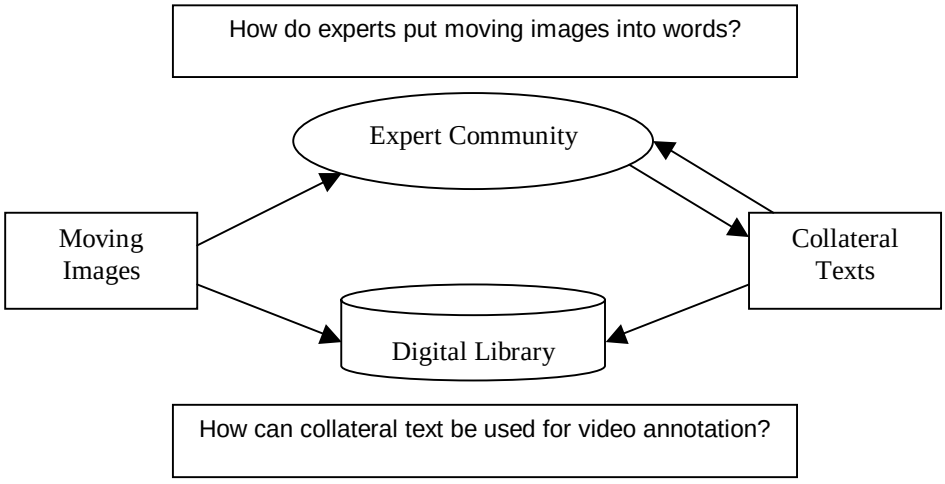


Figure 1.1: The figure shows that the link between a moving image and its collateral texts can be the basis of potentially synergistic investigations into how experts produce descriptions and interpretations of moving images, and into how the collateral texts they produce can be used as surrogates for accessing video data.

There is a literally inexhaustible category of images that one might choose to study in this context; even if one separates still and moving images. However, one category of images stands out here – images of dance. Here, as we elaborate throughout the thesis, the link between vision, language and knowledge is palpable. Dance is a highly specific form of movement, aesthetic and entertaining at the same time, and there are experts on dance who critique dances and develop theories, and who teach and train others. Such a group of people can be viewed as a specialist community; much like a community of engineers, or a community of hobbyists. These communities are distinguished from one another by a

number of facets. One facet is language: a specialist community can be distinguished from the rest of a language population by having its own vocabulary, by using that vocabulary in idiosyncratic sentential structures, and by communicating with particular intents in different types of text.

Based on the above, we may argue that there is possibly a 'language of dance' that differs from general language at lexical, syntactic, semantic and pragmatic levels. The notion of a special language is important because the use of experts' texts for video annotation requires engineering their knowledge (about moving images) on machines, and this will be helped by the structure provided by a special language. There is much discussion these days about *explicit* and *tacit* knowledge. Explicit knowledge is that available in books, journals, dance reviews and programmes, etc. Tacit knowledge is seldom articulated formally but this does not mean that it cannot be articulated at all. Of course it would not be in the polished form of a text, but it could be elicited nevertheless, for example by asking an expert to speak about moving images as they watched them. The resulting verbalizations would also be collateral texts. In our view of video annotation diverse collateral texts are associated with video data in a systematic fashion so that machine-executable surrogates can be generated from the texts, and so that the texts are made available to the viewers of moving images. The availability of a text corpus means that the system might elaborate the surrogates, for example with (lexical) semantic relationships for query expansion: in this case the system might be called 'knowledge-rich'.

This approach to video annotation may be construed by some as another example of the traditional model for the search and storage of images and video, both in human memory and for computerized search. Pentland suggests that the traditional approach "has been to create propositional annotations that describe the content of the image, and then enter these annotations into a standard database or semantic net" (1997:1283). The drawbacks of such an approach which concern Pentland have already been discussed, i.e. the ambiguity of language and the fact that images mean many things to different people. From our point of view these drawbacks are challenges to be overcome, and not, as they are for some researchers, the reasons for pursuing 'content-based' video retrieval approaches, to the neglect of annotation.

We hope to contribute to research in video annotation by dealing with collections of whole texts that capture more of the plurality of meanings than keywords or fragmented propositions alone, and by limiting our investigation, in the first instance at least, to domains where there is relative consensus amongst experts about how to describe and interpret images.

1.3 Contributions and Thesis Structure

This thesis aims to make four contributions which may be of interest to scholars of artificial intelligence, computer vision, terminology and semantics. The contributions may also be of practical value to the developers of video annotation systems, and of interest to dance scholars. The first contribution relates to the choice of a specialist moving image, that is dance, which is generally described and interpreted through a special language. The second contribution relates to the examination of what, for us, is a many-to-many mapping between a set of moving images and a corpus of texts. The third contribution relates to the methods by which knowledge about moving images can be elicited, for example through the application of verbal reporting. These theoretical contributions are complemented by a practical contribution, that of a system in which video data and collateral texts are organised and processed side-by-side. The system both depends on the theoretical investigations to highlight structure in the knowledge/language of experts and provides a computational environment for exploring the vision-language link.

This work is an attempt at a synthesis of key developments in the study of aesthetic images, in the study of special languages, in corpus linguistics, in cognitive psychology and in computing science – particularly multimedia systems and artificial intelligence. We believe that this kind of synthesis can give some understanding of the link between vision and language – two key human faculties. The structure of the thesis, comprising six further chapters, reflects this synthesis.

The characteristics of different kinds of images determine how they can, and should, be stored, analysed and accessed. Chapter 2 extends a classification of moving images that is based on the intent of their producers, their content, the manner of their production and their

usage. The class of specialist moving images and the characteristic of collateral text are added to the existing classification. It is argued that specialist moving images, which are produced with a specific intent and follow narrow conventions, may be more amenable to analysis than other kinds. A comparison is then made between the ways in which human experts analyse aesthetic images and the state-of-the-art in computer vision. The distinction between the description and the interpretation of images is elaborated with reference to aesthetic frameworks and the current limits of computer vision systems.

Video data is now widely available in computing systems but new techniques are required so that it can be used effectively. Chapter 3 considers the challenges in developing systems to access video data. A discussion of strategies for retrieving and browsing video data suggests that a range of surrogates, including both visual features and semantic features, need to be attached to intervals and regions within image sequences. As well as requiring a means for generating surrogates, this requires video data models and techniques for video segmentation. The emphasis in the chapter is on a number of systems that generate visual features and/or semantic features for video retrieval; proposals to use knowledge representation schemes for video surrogates are also considered.

In order to understand more of how experts describe and interpret moving images, methods from corpus linguistics and cognitive psychology were applied: the methods share the premise that texts can be analysed as data in investigations of cognitive and communicative processes. Chapter 4 reports the analysis of a collection, or corpus, of dance texts. A statistical analysis of the corpus showed systematic variance with a general language corpus at lexical and collocational levels: this is taken as evidence for a special language of dance. The potential range of collateral texts that might be used for video annotation is then elaborated by a manual classification of text types.

The study of a text corpus sheds light on the explicit knowledge of a domain, but to deal with the knowledge that an expert does not usually articulate requires the use of knowledge elicitation techniques. Chapter 5 presents a method for eliciting verbal reports about moving images. Results suggest that the instructions given to experts before they verbalize can guide the systematic elicitation of their thoughts about moving images so that, for example, the

distinction between description and interpretation is maintained, and so that the text refers to the moving image at a pre-determined temporal granularity.

A challenge commonly faced by the developers of modern computing systems is that of dealing with multimedia information. It has been argued that the object-oriented paradigm provides the means to handle different kinds of multimedia data in the same system. Chapter 6 presents the KAB system which is being developed in an object-oriented manner to store collateral texts alongside video data. The texts are processed to label video sequences with key terms, and appropriate texts are presented to the user as they are watching video sequences. The evaluation of the KAB system showed that users appreciated the exploitation of the video-text link for annotating, retrieving and browsing video data. Although the current implementation of KAB makes elaborations of surrogates available to a human reader, it does not incorporate the kinds of machine-executable elaborations that might properly warrant the label 'knowledge-rich'.

However, in Chapter 7 we suggest that through our theoretical and practical investigations we have laid some of the ground for the development of knowledge-rich systems to access image and video data. In considering the outlook for future work we present a specification for a knowledge-rich video annotation system and discuss potential research domains and research questions.

Chapter 2

The Analysis of Moving Images

Moving images are rich sources of information for their viewers and they now abound in popular media and in specialist areas of aesthetic and scientific interest. Recent technological developments now mean that digital video data can be received in real-time over communications networks and can be viewed at high quality on a personal computer. Whether broadcast, selected from archives, or recorded on-the-fly, moving images are becoming widespread in computer-based information systems for entertainment, for education, for keeping abreast of current affairs, and for scientific and aesthetic study. However, if these computing systems are to be more than conduits for streams of video data then they must assist humans to retrieve, browse and understand the moving images.

In order to provide such assistance, a system must in some sense understand the moving images that it stores so it is necessary for the moving image to be analysed. In some cases this may require no more than a textual surrogate giving the name of a film, its genre and perhaps its director and lead actors; this would be sufficient for a general user of a film library to select a film they wished to watch. Sometimes though a more in depth understanding of the moving image will be required, as in the case of the user who wishes to retrieve video sequences according to the objects and actions they depict, and the meanings they convey. It could be that it is easier to provide surrogates to meet these needs for the kinds of moving images that are created and used in aesthetic and scientific works, since these are produced with a specific intent and their content and structure is usually formally constrained.

One characteristic of moving images such as dance sequences and weather patterns is that they are only fully understood by trained experts who study video recordings and related information in the course of an analysis. In relation to aesthetic images, experts also reflect on the processes by which visual information is communicated and understood in theories about visual signs, their combination with other modes, and the respective roles of the producer and the receiver of visual information. Such discussions have led to the proposal of frameworks for the analysis of images, so that despite the complexity of aesthetic (and also

scientific) images, and the diverse information sources that bear on their full understanding, the claim can be made that experts are able to describe and interpret them in a systematic fashion.

The distinction between *description* and *interpretation* is important for this thesis. In their general language senses the distinction is that between "detailed account of a person, thing, scene, or event; a verbal portrait" and "the action of explaining the meaning of something" (The New Shorter Oxford English Dictionary, 4th Edition). These definitions suggest that a description provides a record of someone or something which could be agreed upon by different experts. In contrast, the nature of an interpretation, in which somebody explains the meaning of something, is more likely to be affected by subjectivity, that is to say the interpretation will be determined to some extent by the interpreter's own knowledge and experience. This distinction between a relatively objective description and a more subjective interpretation is emphasised in a discussion on the philosophy of art in which description is considered to imply a "stable, public, relatively well-defined object" such that differences in descriptions can be "reconciled by further examination". However, interpretation "requires the contribution of the interpreter" and so there is the possibility of "alternative interpretations" (Margolis 1980:111).

This chapter continues by discussing different kinds of moving image in order to highlight characteristics which might aid systematic analysis, and to note the surrogates which might be needed for retrieval (Section 2.1). A close look at the theory and practice of dance analysis then shows how a systematic analysis of visual information can go beyond a literal recording of apparent objects and movements, by incorporating related information sources, and by applying domain expertise. In these respects, the analysis of dance can be compared with approaches to the analysis of other specialist images like paintings and films (Section 2.2). Since human involvement in providing an image or video collection with surrogates is expensive, and sometimes unreliable, it is important to consider the potential for providing video surrogates automatically with computer vision technology (Section 2.3). A discussion of the analysis of specialist images by humans and by machines, notes the common use of languages and schemes which involve different levels of processing (Section 2.4).

2.1 Kinds of Moving Image

Moving images are created to inform, to entertain, to record events, and to be aesthetic works. Generally they depict multiple entities and actions, organised in space and time, and they are often accompanied by a soundtrack which may include speech and music. Sequences of moving images may be further structured in the editing process. They are understood by viewers with reference to a context which includes the users' expectations and background knowledge, and related information sources which might add to their appreciation of the moving image.

As well as 'everyday' moving images from home video, television and cinema, there are moving images that are studied by specialists working in scientific and in aesthetic domains. Consider moving images recorded by a weather satellite that will be analysed by meteorological experts who can identify key weather patterns, and make predictions; or recordings of dance which are studied by experts who can highlight significant movement sequences and explain their meanings. For storing and retrieving digital video data, it is important to recognise the characteristics of different kinds of moving images, in order that they are given appropriate surrogates for retrieval and browsing alongside relevant information.

Everyday Moving Images

Television and cinema provide moving images to large audiences: the subject matter of programmes and films is wide-ranging, and in the case of live events is to some extent unpredictable. However, for a given type of programme, such as a news broadcast or a sports broadcast, there will be certain conventions followed for the organisation of visual information. Furthermore, in television dramas and crime shows, and their feature length equivalents in the cinema, a sophisticated set of camera techniques and editing effects are used to thrill, shock and surprise the audience.

These kinds of moving image each include a speech track: though the relationship between what is said and what is seen varies between different types of moving image. A sports commentator may accurately describe the action as it unfolds, so that their speech is tightly-

coupled with the moving image; at other times the commentary might digress to trivia about the sportspeople and the sporting occasion. There is perhaps a weaker link between the spoken word and moving image in the case of news broadcasts in which the stories are illustrated, only in places, by still and moving images that show the people and places involved in the story, or its general theme. There is a weaker link still in dramatic works, where the spoken words of the actors may say little about what can be seen on the screen.

Specialist Moving Images

More stylised and patterned visual information, like aesthetic works and scientific data, convey meanings which are heavily dependent on the contexts in which the images are produced and viewed. Some of these meanings are only understood by trained experts who study video recordings and related information in the course of a systematic analysis. The term *specialist (moving) image* is used here to refer to visual information which contains meanings in its patterned and formal aspects – meanings which are confounded for a layperson but which can be recovered systematically by an expert. Specialist images proliferate in the realms of the sciences and the arts. In the arts, meanings arise in the communication that takes place between the producers of artistic works and their audiences. In the case of scientific data there is no 'sender' of visual information (unless we count a deity), but it is still understood to convey meanings: these meanings will be determined in part by the situation in which the data is gathered.

Scientists gather still and moving images from microscopes, telescopes and satellites, and use X-rays and infra-red light to capture visual information that is not available to the human eye. Such visual data is important for scientific endeavours, such as the testing of theories and the development of technologies. Scientists also use images to illustrate abstract concepts and to organise classifications of domain artefacts: for a history of the scientific image see Robin (1992). Humans have been painting pictures and engaging in drama and dance for thousands of years. In more recent times, scholars have begun to investigate the ways in which fine art, film and dance can express meanings and convey messages. The writings of art theorists like Erwin Panofsky and Ernst Gombrich, film theorists like Christian

Metz, and dance theorists like Susan Leigh Foster and Janet Adshead-Lansdale, describe the elements of aesthetic works, explain their meanings and give insights into the historical and social situations in which they were produced.

Consider dance as an exemplar specialist moving image. A dance performance may be viewed as a sequence of stylised and patterned movements which are rhythmic and set to music. One can further argue that through the use of movement, music, costumes and sets, dance can convey emotions, tell stories, and make social comment and cultural statements. Dance is generally performed to a present audience; or performed as a social activity, in leisure and in rituals. In contrast, the study of dance is often based on film and video recordings which allow for repeated viewings of multiple dances: the dance scholar also has access to various sources of related information, including others' writings about dance and dances, and biographical details of the choreographer and dancers. The analysis will be guided by the theoretical persuasion of the scholar who may be influenced by the dominant philosophical trends of their times.

A Synthesis

Moving images are produced and viewed for different purposes, and vary in the kinds of things they depict and the ways in which they are made. When storing and retrieving digital video data it is important to recognise such differences in order to capture information about the moving image that is relevant for a particular use. Hampapur and Jain (1998) provide a classification of moving images according to their *intent, content, production* and *usage*. This classification is summarised, and extended, in Table 2.1, where the kinds of moving images have been ordered from the 'everyday' to the 'specialist': note that this is our distinction. Four further kinds of specialist moving image have been added to the original classification in order to supplement the everyday moving images that were the main focus of Hampapur and Jain's interest.

These new examples highlight the contrast between the everyday and the specialist moving images, according to the degrees of specificity and patterning in their content, of contrivance and formalisation in their intent and production, and of specialism in their usage.

A further characteristic of moving images has also been added to Hampapur and Jain's original classification: that is, related textual information. This characteristic is relevant for video annotation systems that may use such text. In going from everyday to specialist moving images, the related textual information becomes increasingly informative about the moving images. It is perhaps the high degree of specificity in the intent, content, production and usage of specialist moving images that allow experts to analyse them in a systematic fashion.

	MOVING IMAGES								
	EVERYDAY				SPECIALIST				
					SCIENTIFIC			AESTHETIC	
	Surveillance Cameras	**Feature Films / TV Drama**	**News / Documentary**	**Sporting Events**	**Biomechanical**	**Medical Images**	**Satellite Images**	**Art-house Films**	**Dance**
INTENT	Recording all activity in field of vision.	Entertainment.	Information.	A record / Entertainment	Analysis of biomechanics.	Gathering data for analysis.		Conveying intent of filmmaker	Conveying intent of choreographer
CONTENT	Determined by what happens in area of surveillance. No structure.	Range of subjects, restricted within genres.	Unrestricted, but ordered into segments: politics; social stories; science; etc.	Structure comes from the nature of the events being recorded.	Highly constrained.	Area of interest for medical specialist.	Weather systems; geographic information systems.	Tightly scripted dialogue. Content chosen to reflect intent of film-maker.	Movements from a restricted set. Maybe accompanied by a musical soundtrack.
PRODUCTION	None beyond siting of camera.	High control over screenplay, filming and editing.	Greater control over prepared stories and news presented in the studio. Less control over the coverage of live, breaking news.	Camera strategically placed; no control over live event but some degree of predictability.	High degree of control over what images are filmed though they are unscripted and editing is unusual.	High degree of control over what images are filmed though they are unscripted and editing is unusual.		High degree of control over the scripting, filming and editing of the moving images. Drawing on conventional meanings of entities and actions.	
USAGE	Security agencies to monitor activity live and check for evidence after criminal activity.	Film goer for enjoyment. Film critic. Film database managers.	News viewer for information. News producers for reuse.	Sports viewer for enjoyment. Sports coaches for information.	Athletes and coaches. Biomechanics researchers.	Medical specialists.	Meteorologists; compilers of geographic information systems.	'Entertainment' Academic study by film scholars.	'Entertainment' Academic study by dance scholars.
RELATED TEXTUAL INFORMATION	Speech that might be captured by the camera.	Film script. Press previews and reviews.	Spoken words of the newscaster. Coverage of the stories in the press.	Spoken words of the commentator. Coverage of the events in the press.	***The spoken and written words of experts who analyse these moving images.***				

Table 2.1: A classification of moving images, summarised from Hampapur and Jain (1998) and extended with four further kinds of moving image and one further characteristic: these additions are shown with a grey background. The kinds of moving image have been ordered here from the 'everyday' to the 'specialist', with increasing degrees of specificity and patterning in their content, of contrivance and formalisation in their intent and production, and of restriction in their usage. Note the additional characteristic, that of 'Related Textual Information', is important for systems that process such text to annotate video. For us, it is the spoken and written words of experts analysing moving images that are of particular interest.

2.2 Beyond the Image

With regards to the human body, muscular-skeletal movements can be recorded on paper, measured biomechanically, and in limited cases can be recognised by computer vision systems. For example, a *slow* movement can be recorded, measured and recognised: however, depending on contextual factors, it might be understood as a *tender* movement or a *tired* movement. At this point the analysis of movement shifts from examining isolated movements to an integrated understanding of these movements along with other information. This crossing from the relatively objective recording of movements, to the attribution of meanings to movements may be viewed as the crux of dance analysis.

A dance scholar may be concerned with a particular performance of a dance, the choreography of a dance, the work of a choreographer or with dances of a certain genre. The dancing body may be described in terms of its parts, or itself may be described as part of a group of dancers. In each case, the scholar attempts to go beyond a description of human movements by making interpretations. Their interpretation will take account of the music, setting and costumes associated with the dance; as well as other related information and theories about dance analysis. Thus dance analysis is more than 'looking at people' – a phrase used in reference to computer vision systems that analyse human movement: dance scholars deal with both the perceptual and conceptual aspects of dance, and discuss them in historical, social and cultural contexts.

Dances can be organised according to genres and styles, in part following historical and geographical distinctions. For instance, the genre of ballet has seen the developments of different styles over two centuries, including *pre-romantic, classical, romantic* and *neoclassical*. Finer-grained distinctions can be made for each style, in retrospect, according to ground breaking individuals and companies in different countries at different times, for example the influential choreographer Marius Petipa, and the pioneering Royal Danish Ballet.

Over the course of the last century alternatives to the ballet genre have developed in Western dance. In contrast with ballet, which follows conventions for selecting movements and the form of dances in order to present narratives, modern dance (c. 1900-) is concerned

with the formal properties of movement, and post-modern (c. 1960-) dance enjoys mixing and breaking the conventions of earlier genres and styles, and interacting with other media such as film and television. There has perhaps been an accompanying change in the ways people theorise and write about dance: that is, a change from reviews that praised the beauty and technical virtuosity of ballerinas, e.g. Beaumont (1949), to learned articles exploring dance from different theoretical perspectives. A recent collection of dance studies literature shows the diversity of dances studied and theoretical positions taken (Carter 1998). Dances studied include ballet, modern dance, dance theatre, vernacular dance, dance in pop videos and dance in virtual reality. The perspectives from which these dances are considered include those of choreographers and dancers, as well as the audience; and references are made to theories from psychology, anthropology, sociology and philosophy.

In some dance writing, individual dances and choreographers are considered and a view of dance genres and styles is developed on a case-by-case basis. Thus, Judith Mackrell, a prominent dance scholar, has noted that classical ballet tries to create "the illusion of flight", whilst some classical Indian dances are "grounded on earth" (Mackrell 1997:116). The motivations for some modern American dance are given in comments on Martha Graham – that her dancing "was based on the pull of gravity" (ibid.); and on Merce Cunningham – that he wanted dance to "reflect the dense information overload that we're used to processing every day in the modern world" (ibid.). In some cases the history of a dance is tracked and the different productions are discussed, for example a history of Petipa and Ivanov's *Swan Lake* (Beaumont 1952). In other cases different versions of the 'same' dance are reported and the contrasting intentions and achievements of their choreographers are discussed (Mackrell 1997:23-24).

Whilst an analysis might start with the expert watching a recording of one performance of a dance and focusing on the human movement, many other information sources and theoretical constructs are introduced as the dance scholar analyses the dance. It is these sources, beyond the moving image itself, which allow experts to elucidate the meanings of a dance, beyond literal descriptions of dancers and their movements.

2.2.1 Levels of Dance Analysis

Contemporary discourses on dance deal with the *description of components* of a dance, including individual movements, and with the *discernment* of significant spatial arrangements and movement sequences such as recurrent motifs. Furthermore, these discourses deal with the *interpretation* and the *evaluation* of dances. The systematic analysis of dance has been discussed by Janet Adshead-Lansdale and colleagues, who drew on a framework for the teaching of art appreciation (Smith and Smith 1977) to argue for a four level approach to dance analysis (Adshead 1988). These four levels are elaborated here under two headings to highlight the contrast between those aspects of an analysis that deal with what can be seen in the moving dance image (i.e. description of components and discernment of form), and those aspects which go 'beyond the image' (i.e. interpretation and evaluation): this division may be broadly labelled as 'description' and 'interpretation' – using the words in their general senses.

Description of components and the discernment of form

The components of a dance which are to be described include the dancers, the stage or setting, any accompanying music and words, as well as the dancers' movements. Movements may be described in a notation system, such as *Labanotation*, or through vocabularies associated with certain dance genres and styles that name codified actions and positions; for example ballet's *plié*, *relevé* and *attitude*. For the purpose of dance analysis a description, be it movement notation or natural language, is expected to provide a record of the ephemeral dance that can be used as a common point of reference for further debate.

Notation systems provide a means for recording human, and in some cases animal, movement at fine levels of detail. They are used by dance analysts, as well as by scholars in other domains such as anthropology and zoology. An example of Labanotation, one of three commonly used systems, is shown in Figure 2.1. The notation is read from bottom to top, i.e. along a vertical temporal axis delimited by bars akin to those of a musical score. Symbols to the left of the centre refer to movements made by left-hand limbs, and *vice versa*. Moving out from the centre the columns of symbols indicate movements of the *feet*, *legs*, *torso*, *arms*, *hands* and *fingers* – and to the extreme right the head. The excerpt above records a dancer

performing a movement scale – stretching to reach the corners of an imaginary cube that they are standing inside, with the head following in the direction of the movements. The symbols' points, shadings and size capture the movement dynamics of direction, level of extension and duration, whilst the diacritics capture more subtle aspects of the movement.

Two other widely used movement notation systems are *Benesh* notation, which is particularly adapted for recording balletic movements symbolically, and *Eshkol-Wachman* notation which is geometrically grounded and can be applied generically to any jointed movement (both humans and animals): see Hutchinson Guest (1984) for a review and history of movement notation systems.

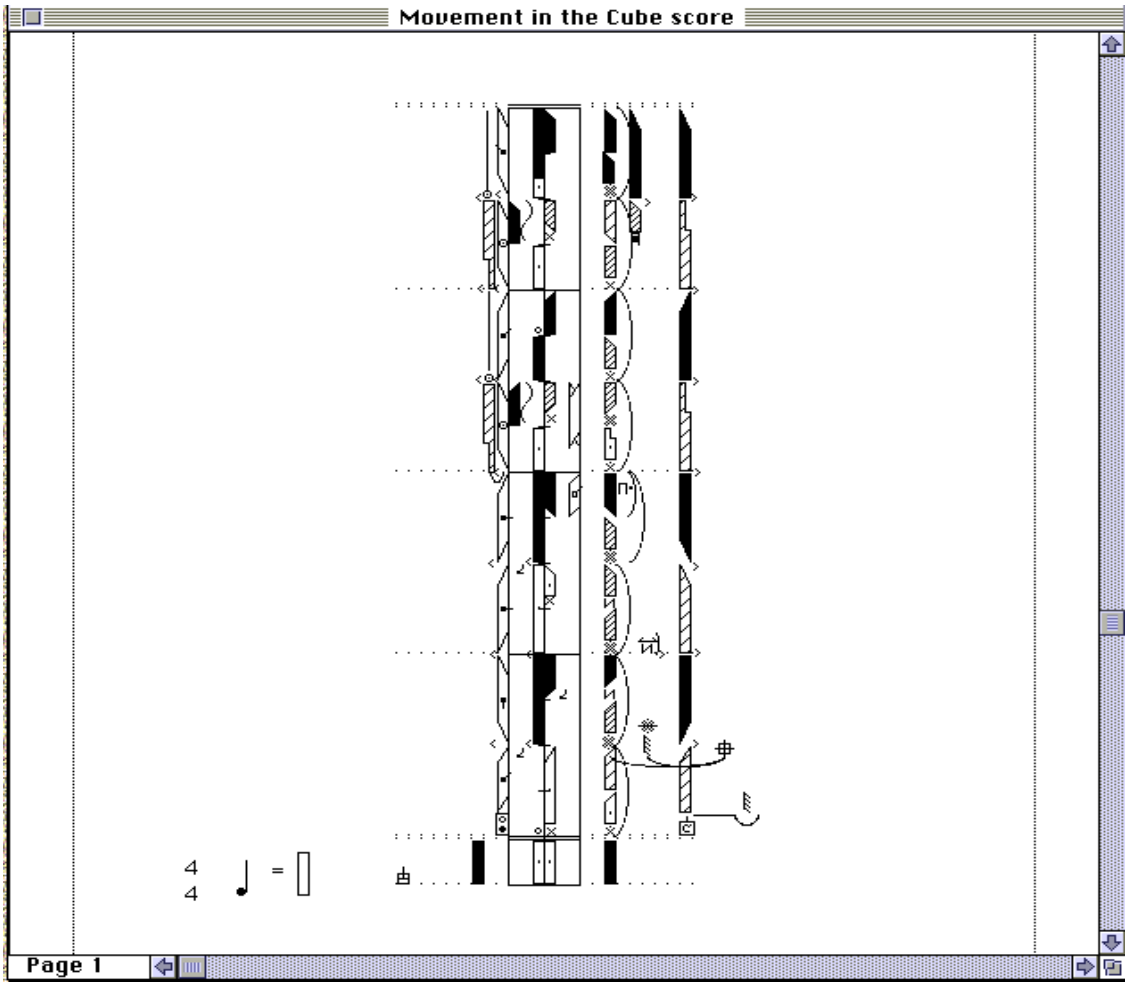


Figure 2.1: An example of Labanotation which records the movement of a dancer as they reach forward to four vertices of an imaginary cube that they are standing inside.

Whilst notation systems and specialist movement vocabularies can record movements at a fine grain of detail, such detail will not always be important for the analysis of a dance sequence. In order to start making sense of a mass of complex visual information, the dance analyst must discern salient spatial and temporal combinations of movements. These combinations include motifs which may be distinguished by their recurrence throughout the dance or by their simultaneous occurrence across the stage. Movement components also combine through the interactions between dancers, for instance in lifting, or when dancing in unison.

Interpretation and evaluation of dance

Whilst 'description' and the 'discernment of form' are concerned with capturing the dance in as objective terms as possible, the process of interpretation opens up the possibilities of subjectivity – though this is constrained by knowledge of what is to be expected from a particular kind of dance. In an interpretation the dancers may be referred to as their characters, e.g. 'the white swan', rather than the neutral 'second dancer', and the significance of movement qualities will be elaborated, e.g. what might be described as a slow movement may be interpreted as a tired movement. Such ascription of *quality* to movement is often the starting point of interpretation. Interpretation might include an elaboration of the dance's narrative and the artistic intent of its choreographer.

The analysis of dance is informed by, and contributes to, debates about the nature of non-verbal communication and meaning. Aesthetic and semiotic theories about how visual information is understood appear to deal with the ways in which signs convey meanings in non-verbal communication; the ways in which multiple modes of communication combine and interact; and, the varying contributions to the production of meaning made by the addresser, addressee and the 'text' itself. As such, these approaches complement research in visual cognition and in computer vision which has often concentrated on the computational processes required for tasks such as the recognition of objects and actions in visual information, usually with no further information available – for example, work inspired by David Marr (1982).

The question of how dance, as a form of non-verbal communication, can have meaning has been discussed in the subject of semiotics, leading one researcher to propose six ways in which movement can convey meanings. These are: *concretization* – that is, by the imitation of another's movement; *icon* – by reproducing the formal characteristics of a movement; *stylization* – by arbitrary, conventional movements; *metonym* – by associated movements; *metaphor* – by resemblance or analogy; and, *actualization* – by an individual according to their role (Hanna 1994). The viewer of a dance performance has to integrate not only movement and musical information, but also a range of other potentially meaning bearing signs, for example costumes and stage sets which can evoke cultural icons and shape the viewer's understanding of the dance. The question of how this information combines has been addressed by semioticians, particularly those interested in the semiotics of performance, who study "the interweaving of several expressive elements, organised into various codes and subcodes" (de Marinis 1993:1).

The use of the word 'codes' may suggest a shared understanding of signs, such that messages can be communicated reliably from a sender to a receiver. This notion has been challenged by literary and critical theorists who question the idea that a text gains all its meaning from its author and that it should therefore have a fixed meaning for all its readers (see Lodge 1988 for a range of such views). Dance scholars have adopted, and adapted, theoretical positions from both sides of this argument.

One semiotic perspective of dance which focuses on how it can achieve communication through a system of signs is based on Roman Jakobson's model of communication (Foster 1986:xvii). According to this view, communication through dance is achieved by the selection and combination of movements and by studying how these movements can refer to, or represent, worldly events. Meaning is coded by the addresser (choreographer) and decoded by the addressee (viewer), so that a dance is reliably understood with reference to a set of conventional codes and the context in which it is produced and viewed, Figure 2.2.

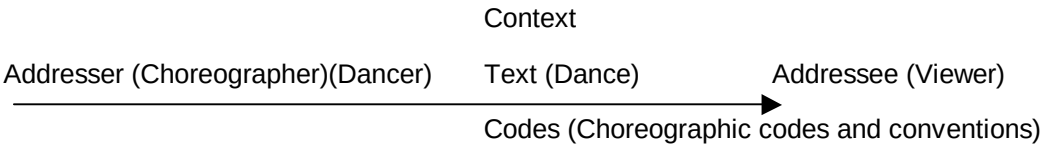


Figure 2.2: A communication model of dance, after Roman Jakobson (Foster 1986:xvii).

Such a model of conventional communication is perhaps suited to dance genres like classical ballet which have strictly encoded meanings for certain gestures and mimetic sequences, see for example those discussed by Beaumont (1952:78-79). Contrasting approaches emphasize the contexts in which communication takes place, and in particular the "role of the reader" in understanding texts (Eco 1979). This leads to the idea of multiple interpretations and possibly permanently unstable texts: these ideas are developed in the theory of intertextuality which considers that "a text … cannot exist as a hermetic or self-sufficient whole, and so does not function as a closed system" (Still and Worton 1990:1). 'Texts' bearing on the interpretation of a dance might include other dances (by the same choreographer, of the same theme, or otherwise related), and writing about the dance, as well as sources from a wider context. The study of dance and the practice of dance analysis from the standpoint of *intertextuality* has been studied by Adshead-Lansdale (1999).

Having understood a dance by interpreting it, the dance scholar is in a position to evaluate its aesthetic contribution: evaluation proceeds according to a set of values. Different types of artistic criticism may emphasise different values for judging works, for example the five types delimited by Stolnitz (1960): *criticism by rules* – such as the neoclassical criteria for evaluating art which took work from the Graeco-Roman period as the standard; *contextual criticism* – which seeks to understand the work with references to the social, historical and psychological factors that influenced it; *impressionist criticism* – which reports the subjective reaction of the critic to the work; *intentionalist criticism* – that seeks to identify the artist's purpose in creating the work; and, *intrinsic criticism* – which is only concerned with the formal properties of the work.

Summary

The levels of dance analysis discussed here, and summarised in Table 2.2, of course overlap and are not meant to imply a strictly linear progression in analysis. These provisos aside, it seems that there is an important divide between the first two levels, in which the analysis is concerned with what can be seen 'within the image', and the second two levels, in which the analysis goes 'beyond the image'. The analysis goes beyond the image in two senses, firstly in that it is based on more information than just the moving image, and as a consequence, secondly, the analysis says more about the 'semantics' of the dance than just naming the literally apparent entities and actions.

STAGE OF ANALYSIS	AIM
Description of components	To produce a record of the dance as a point of common reference for analysis. Components include movements, dancers, the stage or setting, music and possibly spoken words.
Discernment of form	To identify salient combinations of components, e.g. movement sequences, recurrent motifs and spatial arrangements of dancers.
Interpretation	To explicate the significance of the dance with reference to movement qualities, for example its narrative and subject matter, and to attribute meanings.
Evaluation	To judge the aesthetic worth of the work, with reference to other dances and the content in which it was produced.

Table 2.2: A summary of the four levels of dance analysis outlined by Adshead (1988).

2.2.2 Describing and Interpreting Images in Fine Art and in Film

One can argue that dance and dance analysis are archetypes of specialist images and expert analysis respectively. Like dance, other specialist images are stylised and patterned, and have some conventional meanings which are only properly understood in relation to the context in which they are produced and with reference to diverse information sources. And like dance scholars, experts who analyse other kinds of specialist images are able to draw on specialist symbol systems and domain expertise in order to describe and interpret complex visual information.

One aspect in particular can be observed across a range of specialisms: that is the distinction between the analysis of what can be seen within an image (a description) and the analysis of the meanings beyond the image (an interpretation). Frameworks for the analysis of moving images, like films and dances, have discussed these images both in terms of

generic descriptions of their contents, and in terms of specific interpretations which go beyond the image by invoking background knowledge and related information sources.

For instance, Metz (1974) discussed how films could be analysed at the *physical* level – where visual phenomena become perceptually meaningful as generic contents; the *diegetic* level – which comprises the 4D spatio-temporal world shown by the moving image; and the *connotative* and *sub-textual* levels in which the 'aboutness' of moving images is conveyed by metaphorical meanings and specialised symbols. With regard to the storage and retrieval of moving and still images, it should be noted that Lindley and Srinivasan (1998) have presented the work of Metz in some detail as the basis of a digital video generation system.

The work of Metz, and also that of Adshead on dance analysis, can be related to earlier work on meaning in images, albeit still images, by the art scholar Erwin Panofsky (1939). For Panofsky, the three levels of meaning in a painting were: the *pre-iconographic*, comprising the generic contents, e.g. a man in a city; the *iconographic*, forming the specific contents, e.g. a particular man in a particular city; and the *iconological*, that is what the painting is about, e.g. politics, war, peace. Panofsky's framework has been proposed as the basis for organising picture collections (Shatford 1986) – this idea was more recently discussed by Enser (1995).

The frameworks for the analysis of specialist images provided by Panofsky, Metz and Adshead, are compared and contrasted in Table 2.3. Though they were produced at different times, and to deal with different kinds of images, the three frameworks can perhaps be seen to share a common distinction between the description and the interpretation of images. The degree to which experts' analyses of images appear to be systematic suggests that the development and articulation of their knowledge will also be systematic to a degree, perhaps enough for a machine to compute some aspects of it. Though the discussion here has focused on aesthetic images, it is reasonable to suppose that those experts who gather and analyse scientific images also follow well established procedures and are guided in their analyses, especially interpretations, by the conventional wisdom of their domain.

	Analytic Levels for Specialist Images		
	Film: Metz (1974)	**Art: Panofsky (1939)**	**Dance: Adshead (1988)**
Levels of analysis concerned with what can be seen within the image: broadly 'Description'	Physical visual phenomena become perceptually meaningful Diegetic the 4-D spatio-temporal world posited by the film	Pre-iconographic generic who, what, where, when Iconographic specific who, what, where, when	Description of movements a recording of movements Discernment of form the meaningful grouping of movements in space and time
Levels of analysis concerned with the significance of the image, with reference to related information and background knowledge: broadly 'Interpretation'	Connotative metaphorical, analogical and associative meanings Sub-textual specialised meanings of symbols and signifiers	Iconological abstract aboutness	Interpretation genre and subject matter Evaluation aesthetic qualities and artistic statement

Table 2.3: A common contrast is noted in three frameworks for the analysis of specialist images between the 'description' and the 'interpretation' of images. The presentation of Metz's and Panofsky's frameworks here is based on the expositions by Lindley and Srinivasan (1998) and Shatford (1986) respectively. (There is a fifth level discussed by Metz which relates to cinematic features such as cuts, pans and zooms. It is not included in the table as it has no analogue in the realms of painting or dance – except in the case of dance made specifically for film).

2.3 Computer Vision

Generic video retrieval systems use surrogates that capture the changing patterns of pixels in a moving image regardless of its domain or context. However, the capture of semantic information currently relies on a human annotator or on textual information already associated with the video. In the future, the automated description, and perhaps interpretation, of semantic video content will depend upon the development of domain- and task- specific methods like those researched in the field of computer vision. As well as dealing with images of the immediate environment, computer vision systems are also used to analyse visual information that is not normally available to the human eye. In medical

practice and research X-rays, microscopes and a variety of scanning devices provide images of the human body to be analysed for symptoms of disease or as scientific data. Satellite-mounted cameras provide images of the earth's surface which can augment geographical information systems and images of weather systems which can be incorporated into meteorological information systems.

Video images of human movement provide challenging examples for computer vision research. The many degrees of freedom of the human body and its propensity for self-occlusion make the task of movement recognition a difficult one. State-of-the-art systems are able to recognise, say dance steps, only under certain constraints, such that for example the body is filmed from a fixed angle with nothing obstructing the view, or recognition is limited to previously seen individuals. Recognition in current systems is of short, relatively simple, sequences and produces literal descriptions of simple movements, compared with the complex descriptions and interpretations produced by humans.

2.3.1 Levels of Processing in Computer Vision Systems

Computer vision has been defined as "the information-processing task of understanding a scene from its projected images" (Cohen and Feigenbaum 1982:127). Computer vision systems take low-level codings of still and moving images as input and produce higher-level, symbolic classifications and descriptions of the objects and actions depicted by the pixels.

Thus, research in computer vision has been characterised as following a "signals-to-symbols paradigm" in which three broad levels of processing can be delimited (though this need not imply uni-directional processing): low-level – to extract local image properties; intermediate-level – to identify generic scene attributes; and, high-level – to produce a description of a scene (Fischler and Firschein 1987). These levels were elaborated in a set of functional requirements for a general purpose vision system proposed by the same authors. At the lower levels are functions such as geometric modelling and photometric modelling which deal with how light reflects off the three-dimensional objects in the scene. The intermediate levels include scene segmentation and 'naming and labelling'.

The naming and labelling of objects and actions is crucial for high-level image understanding in which the relations between objects and actions are explicated and reasoned over with reference to domain knowledge stores to make inferences or to update current knowledge. It is at the naming and labelling stage that the processing of visual information requires some kind of language to describe entities and actions.

Most approaches seem to stop once a mapping has been made from mathematical features of image data to a description, or at least a classification (Sonka, Hlavac and Boyle 1993) and it is a similar story for image sequences. In the case of moving images, information must be extracted about changes over a sequence of images to describe not only objects and spatial relations, but movements and complex co-ordinated events. A review of movement analysis techniques showed an emphasis on low-level motion information such as trajectory features and optical flow, and the systems discussed perform statistically-based classification of moving images – in effect treating the video signal as time-varying data (Cédras and Shah 1995).

However in some cases it may be important to have a structured description of moving images in order to facilitate further information processing. For systems to produce conceptual descriptions of moving images it was suggested that complex visual events need to be formed from simpler ones. Nagel (1988) proposed four levels of conceptual description to link signal level encodings of moving images to higher-level concepts: a *change* – any deviation in a sensory signal which is significantly different from noise; an *event* – any change which has been previously defined as a primitive for the construction of more complex descriptions; a *verb* – which describes some activity; a *history* – which describes an extended sequence of related activities. In later work the same author showed how this scheme could be used to recognise relatively complicated traffic manoeuvres: vehicle trajectories were extracted from traffic scenes and associated with a German language verb, by way of a transition grammar, such that the verb 'to overtake' comprises 'changing lanes' and 'accelerating' (in that order). Both of the 'primitives' have visual correlates in terms of vehicle trajectories (Nagel 1994). This building up of complex movements from simpler movements is particularly relevant when considering human movement.

2.3.2 Recognising Human Movements

Systems for recognising human movement must deal with moving images of a non-rigid, articulated object that has a propensity for self-occlusion – these three characteristics of the human body each increase the difficulty of human movement recognition. It is not only a matter of tracking a body in 3-dimensional space, but also capturing the movement of body parts relative to one another. The task is simplified by placing bright spots on the joints of the subject or by having them wear specially patterned clothing. As well as visual information, data from other kinds of sensors is used to detect and analyse human movement, for example, infra-red beams that are broken and reflected; transmitters on body parts; and, detectors of muscle contractions and neural impulses. State-of-the-art systems for the recognition of human movement are limited by the range of movements they can recognise and by the constrained conditions under which they can operate.

A variety of approaches have been explored to analyse human movement in video data. Both 2-dimensional and 3-dimensional image data have been analysed, following both contour-based and model-based approaches. The results have included the recognition of gestures such as simulated computer mouse movements and a subset of American sign language, lip motions, dance steps and 'everyday actions'. A recent review of the visual analysis of human movement defined action recognition as a 'classification problem involving time-varying feature data' and continued by suggesting the applicability of *Hidden Markov Models* and *neural networks* for such a task (Gavrila 1999:91).

Research in human action recognition is exemplified in work done at the MIT Media Lab where researchers developed a system that recognised nine ballet moves performed by a particular dancer (Campbell and Bobick 1995). The system was trained with tracking data that recorded the 3-dimensional location of 14 points on the body every one-sixtieth of a second. This data corresponded with the movements of 10 joints, six of which had one degree of freedom and four of which could move in three directions. As well as dealing with the mass of data, the system also had to cope with 'co-articulation effects' where a movement is changed by preceding and following movements, and also by varying degrees of limb extension. The nine ballet moves chosen for recognition were considered to be 'atomic' – in

that they could be composed into more complex moves. Other researchers have presented a system for recognising everyday 'atomic' human actions, comprising different kinds of walk (normal walking, line walking, marching and walking to kick). The discrimination of four kinds of walking is made by a system trained on features relating to the translation of five body parts (Yacoob and Black 1999).

Whilst some approaches are successful in treating the task of action recognition as a statistical problem of classifying time-varying feature data, there is perhaps a need for intermediate conceptual descriptions of human motion to deal with complex events; especially if the recognition of the more complex events requires drawing on information and knowledge beyond the immediate image sequence. Aaron Bobick acknowledged the work of Nagel when he proposed three levels of description for human motion, these are: *movements* – which do not require any contextual knowledge to recognise; *activities* – sequences of movements which are recoverable from the image data; and, *actions* – larger scale events which typically require background knowledge to recognise (1997). Examples are given for each level with reference to moving images of baseball: the swinging of a bat is a movement which can be recognised independently of any context; the pitching of a ball is an activity comprising a series of movements; and, an action would be the tagging out of a player which to recognise would require knowledge about the game of baseball.

Both notation systems and derivatives of natural language have been proposed as intermediate descriptions in the literature, not just for the 'signal-to-symbol' recognition of movement but also for generating movement from linguistic descriptions (consider graphical animations, and even robotics). It has been argued by computing scientists that Labanotation would make a suitable medium between language-based descriptions of human movement and moving images. Its use was demonstrated in a system for generating graphical animations of human movement (Badler and Smoliar 1979).

Arguments that movement notation systems are too complex and lack the structure necessary to be good for computer-based representations of human movement led to an approach which borrowed the notions of 'deep structure' and 'surface structure' from Chomskyan linguistics (Calvert 1986). This approach is based on the idea of 'macros' for

movements, like walking, which take parameters such as distance and step size. These would be used in transforming linguistic specifications of animations into moving images described at a physical level (surface structure) via an unambiguous semantic representation (deep structure). More recently, a researcher who wished to introduce a 'symbolic component' to image processing in order to facilitate inferencing about scenes, asked "whether a set of generic human actions can be defined which can be applied to a variety of applications?" (Gavrila 1999:95): his suggestions for candidate actions are listed in Table 2.4.

Stand-alone actions	*walking, running, jumping, turning around, bending over, looking around, squatting, falling, sitting (down), standing (up), climbing, pointing, waving, clapping*
Interactions with objects	*grasping, carrying, putting down, examining, transferring (from one hand to another), throwing, dropping, pushing, hitting, shaking, drinking, eating, writing, typing*
Interactions with people	*shaking hands, embracing, kissing, pushing, hitting*

Table 2.4: Candidates for a set of generic human actions proposed by Gavrila (1999).

2.4 Discussion

In the context of this thesis at least, it is important to note that some kinds of moving image seem to be more restricted in their intent, content, production and usage. It is such restrictions that perhaps allow experts to develop theoretical frameworks for the analysis of such images. These frameworks bring a degree of order to the understanding of images, which might be reflected in the language used to produce collateral texts that elucidate them: if so then the texts will be more amenable for processing as surrogates in a video annotation system. One aspect of this order is the distinction made between 'description' and 'interpretation': the former concerns what literally appears within an image or image sequence, the latter draws on related information sources to go beyond the image and expound meanings. Maybe the distinction will be reflected in the language used to articulate analyses?

Whereas scholars in aesthetics are concerned with the understanding of images at the level of description and above, most research in computer vision has concentrated on the mapping from image data to a level of description. The state-of-the-art for recognising objects and events cannot yet support the generation of surrogates for general video annotation purposes,

however recent proposals for recognising complex events in terms of simpler ones point to ways in which the performance of systems could improve. A prerequisite for such approaches is some kind of language for 'conceptual descriptions' at each level of processing, such that higher-level descriptions can be generated out of simpler ones, and so that other information sources can be integrated. For several authors discussed here, this suggests a need for 'primitives', the mention of which brings to mind research in knowledge representation, especially Schank's work in the context of movement; the applicability of knowledge representation schemes to moving images is considered in the next chapter.

It may be that computer vision systems meet with more success when analysing restricted specialist moving images, like dance. However, it seems that the challenge of generally recognising even simple, isolated movements like a *plié*, never mind combinations of movements, will be beyond the state-of-the-art for some time. When the need for interpretations is also considered, then the best source from which to generate video surrogates at least semi-automatically is perhaps a collateral text.

Chapter 3

Storing and Accessing Digital Video Data

Digital video sequences can be recorded with cameras, captured from non-digital recordings, and generated by systems for video editing and graphics animation. In each case the video data is optimised for efficient and reliable storage and transmission, and for the high quality reproduction of moving images, possibly accompanied by sounds and speech. Video sequences give the appearance of movement through the rapid presentation of still images (frames), each of which comprises contrasts between light/dark and between red/green/blue.

The same video sequence might convey different information to different viewers. For example a mass of pixels that indicates to an expert meteorologist an abrupt change in the weather, to a layperson appears only marginally different from the rest of a weather scan. A viewer's interests, as well as their level of expertise, will also determine what they see in a video sequence: where one dance scholar sees a series of *pirouettes*, another sees an archetypal example of classical ballet. Since a machine's immediate view of moving images is as streams of pixels, the successful storage and retrieval of digital moving images requires that a range of surrogates are attached to the raw video data.

This chapter discusses the need for systems that attach surrogates to video data, and reviews the state-of-the-art. An overview of the problem emphasises that the ways in which people would like to access video data mean that a range of surrogates is required: it is noted that this need has directed the recent development of digital video standards whose predecessors dealt more with the compression of video data (Section 3.1). Surrogates need to be attached to video data at different temporal and spatial granularities: for this reason, data models and automatic segmentation techniques have been developed to structure video data (Section 3.2). Perhaps the biggest challenge is dealing with the content of video data: current systems include those that automatically generate surrogates for 'visual content' and those that tackle 'semantic content' (Section 3.3). When considered with regard to specialist moving images, these issues suggest a need to use experts' analyses as collateral texts for video annotation (Section 3.4).

3.1 Overview of the Problem

Depending upon a user's needs, a system storing video data may supply a specifically requested video sequence, or may return a set of video sequences that match a query according to some similarity criteria. Once video data has been retrieved, a system should assist the user to browse the moving image interactively along with any available related information. If the user is accessing a set of moving images in order to analyse them or to edit them, then the system might also be expected to assist in this. Systems for retrieving, browsing and analysing video data rely upon surrogates that are attached to video sequences. The nature of specific retrieval/browsing/analysis tasks will determine what kind of information needs to be conveyed by the surrogates.

Discussions about the processes by which surrogates are attached to video data emphasise different aspects of the task, and use a range of terminology. The common feature of all discussions is the expressed need for surrogates to facilitate access to video data collections. Some research then emphasises the data models that are required to hold information about video sequences, other research emphasises how information about video sequences can be 'extracted' from video data and from text with which it is associated. Commonly used terms in these discussions are 'indexing', particularly 'content-based indexing', and 'video annotation'.

Though there are no dictionary definitions available for these words in this context, it is possible to characterise their usage in the video retrieval literature. 'Content-based indexing' tends to be used to refer to approaches that automatically generate statistically-based visual features directly from video data: in contrast, 'video annotation' tends to refer to approaches in which linguistic labels are attached (either manually or automatically). In its general language sense 'annotation' implies the addition of explanatory information: 'video annotation' thus suggests attaching surrogates which elucidate the contents of video data, and which may also be used to index video sequences.

The ideal system would produce video surrogates automatically and these surrogates would capture all the information about a moving image that was significant for a given use, in a machine-executable form. Given the state-of-the-art in computer vision technology, this

system is likely to remain an ideal for some time yet: it has been argued that 'in the near term, it is computer-supported human annotation that will enable video to become a richly structured data type' (Davis 1995:854). Though Davis is implying direct human involvement in the annotation process, it might be that an indirect contribution is sometimes sufficient: if someone has already spoken or written about a moving image, then these words can be used as a source of video surrogates for no extra cost in human labour. Progress in the development of systems that attach surrogates to video data can be evaluated both in terms of the degree to which surrogates are produced and understood by both humans and machines – in colloquial form, the progression can be stated as:

> "At least, Pat should be able to use Pat's annotations.
>
> Slightly better, Chris should be able to use Pat's annotations.
>
> Even better, Chris's computer should be able to use Pat's annotations.
>
> At best, Chris's computer and Chris should be able to use Pat's and Pat's computer's annotations." (Davis 1995:854-855).

3.1.1 Strategies for Accessing Video Data

A user's interaction with a digital video collection might proceed with a series of *queries*, for which the system returns matching video sequences, and/or, with the user *browsing* through and between related 'documents', including video sequences. Thus, discussions about accessing video data borrow concepts from the established paradigms of information retrieval (Sparck Jones and Willett 1997) and hypertext/hypermedia (Baecker et al. 1995). It is also possible to imagine that other established information management technologies, like data mining and expert systems, could be applied to analyse sets of moving images, or at least to assist a human analyst. In each case, the point to reiterate is that it is the surrogates attached to video data that determine the success of the operation: it is surrogates that are matched against queries; it is the surrogates that structure video data and link it to related information; and potentially, it is surrogates that will be analysed by data mining systems, and reasoned over by expert systems.

Queries to video databases take various forms. A home-user may simply request to see a film by its known title, or may request films that feature certain actresses or that were made in particular years. By contrast, advertising executives using video footage libraries will be concerned with the precise lengths of shots in a video sequence and may require certain colour schemes to match a campaign – such needs will lead to queries which combine information about video structure with low-level characterisations of pixel patterns. Furthermore, specialist users will want to request video sequences depicting certain objects, events and actions, perhaps with particular spatial and temporal arrangements; or they may request sequences with more abstract qualities and meanings. For example, a cell biologist might request sequences showing certain cell behaviours, or a dance scholar might want dance sequences on the theme of war.

Queries can be phrased in a language or they can made by drawing a sketch to indicate desired colour, shape and motion properties. A system will then match the query against an index of video sequences, either exactly or according to similarity criteria which may be adjusted by the user. The index must capture the aspects of the moving image which are significant for the user, and capture them in a form which can be matched against the user's query, either directly, or through a process of query expansion to generate further queries from the original.

Most systems follow such a query-based paradigm for accessing video data. More speculatively, consider how annotations could contribute to the browsing of video data in hypermedia systems. Once a video sequence has been retrieved, the viewing of it can be enhanced with structures by which it can be browsed: for example, a list of sections, or a hierarchy of shots and scenes. Viewing can be enhanced further if the system makes related information available as appropriate for the user's needs. Such information can include texts about the moving images, and further video sequences. Thus, if someone is watching the dancer Margot Fonteyn in *Swan Lake*, then any available textual information about Fonteyn and Swan Lake should be brought to their attention, as should any video material of other Swan Lake performances. At a finer granularity, a mouse click on a dancer in a moving image could bring up biographical details, and links to their other works. Likewise, if the

user is reading a text about dance there could be links from the written word to relevant moving images.

The *nodes* and *links* of hypermedia systems are sometimes hand-coded, but perhaps video surrogates could be used to dynamically generate links, taking into account a user's interests and experience; for more details of 'hypervideo' see Sawhney, Balcom and Smith who discuss how 'hypertextual commentary flows around the moving image, offering deeper associations' (1997:30). When moving images are the object of specialist study then systems should be able to assist in analysing them. In the first instance this assistance could be in the form of providing related information and explanatory material (as in hypermedia systems). Further assistance might be provided if systems were able to detect patterns in, or make inferences about, (the surrogates of) video data.

3.1.2 Video Data Standards

Early research about how moving images could be incorporated into computing systems concentrated on the processes of capturing and compressing video data. The amount of storage space required for uncompressed video data and the limited bandwidth available for its transport both within and between devices means that the development of compression schemes has been a major research theme for digital video systems. However, as powerful compression schemes become available and hardware performance for data storage, transmission and coding improves, there has been a shift in research interests to focus on surrogates (sometimes 'metadata') for accessing video data: this progression can be seen in the work of the Moving Pictures Expert Group (MPEG). MPEG is an organisation comprising several hundred representatives from academia and industry who, under the auspices of the International Standards Organisation (ISO), produce standards for digital video. Successive MPEG standards have built on earlier ones to increase the utility of digital video data in computing and communications systems.[1]

[1] The MPEG WWW-site has details of the standards discussed here - http://drogo.cselt.stet.it/mpeg/

The first standard, MPEG-1 which was finalised in 1993, dealt with the compression of video data for storage on a CD-ROM and playback on a local computer. The compression of video data works by removing both spatial and temporal redundancy from the visual part of the video signal on a probabilistic basis. Spatial redundancy occurs within a video frame, as it does in a still image, where regions of pixels have similar colour values. Temporal redundancy occurs between frames when the colour values for some pixels remain the same over time, as when an object moves across an unchanging background: in these cases only the changes in pixel values between frames need to be stored. MPEG-2 (1996) took a similar approach to the compression of video data but enhancements were made to allow for higher quality images to be transported in noisy environments: for example, transmission over the Internet and broadcasting in digital television systems.

The next development, MPEG-4 (1998), introduced structure to the video data stream which had been previously considered as frames of pixels and digitised sound waves. In contrast, MPEG-4 is an object-based coding scheme which explicitly encodes visual objects either as 2-dimensional regions of pixels or as virtual reality models. The scheme also includes auditory objects like multiple speech tracks and music, and text objects such as subtitles. These audio-visual objects are composed for viewing in ways that may be specified by the viewer who can for example select the face of a speaker and the language they speak. As well as making the presentation of video data flexible, the structure of MPEG-4 facilitates both the compression of video data and the indexing of its contents by making important objects explicit in the coding. It should be noted that the MPEG standards only specify the format that a video data file should have: they do not say how the data file should be generated for a moving image. With regards to MPEG-4, there are many open questions about how to automatically generate object-based video codings.

Work started on the latest MPEG standard in 1996: MPEG-7 (there was no –3, -5 or –6) has a target for completion of the year 2000. This standard in the making addresses the need for metadata to index the content of video sequences with the specification of a “Multimedia Content Description Interface”. A test content set has been compiled which includes 13 hours of video data, comprising moving images like news programmes, films, TV dramas,

documentaries, sport and shorter examples of other types like home video, cartoons, commercials and surveillance recordings. The aim is a standardised, but extensible, description scheme for such material, as well as for audio recordings and still images. It may be noted that with reference to our discussion of 'everyday' and 'specialist' moving images, the MPEG-7 test set concentrates on the 'everyday'; that is, moving images with relatively unconstrained intent, content and production (cf. Table 2.1).

The MPEG-7 working papers distinguish between video descriptors at low 'levels of abstraction', such as measures of colour and shape for image regions, and descriptors at high 'levels of abstraction', that refer to objects, events and whole video sequences as they are recognised and understood by the human viewer. They also note the importance of recording other information about video sequences such as their coding format, duration, viewing permissions and production details, and the importance of maintaining links to related documents.

Other organisations have been involved in the production of standards which may complement MPEG-7 developments. The Dublin Core Metadata set has been developed to index electronic documents that are stored on the Internet[2]. The initial aim was to produce an extensible scheme for the indexing of texts which offered more than unstructured, full-text indexing but that was also more general than schemes that implemented highly-structured record formats. Many of the features used to index texts are applicable to some still and moving images, for example *title*, *author*, *publisher*, *subject* and *date*. These features would apply to a video sequence as a whole without taking into account its internal structure.

The convergence of computing, communications and television technologies prompted a recent joint report by the Society of Motion Picture and Television Engineers (SMPTE) and the European Broadcasting Union (EBU), on the exchange of programme materials as bit streams. The report considered what kinds of metadata should accompany programme material through the stages of production, storage and transmission. Seven categories of metadata were identified, including information required to transmit, receive and playback the video data (i.e., the metadata categories labelled 'Essential', 'Access', 'Composition' and

[2] The Dublin Core Metadata WWW-site contains details of the standard - http://purl.oclc.org/dc/

'Relational'); information about the production and storage of the video data ('Parametric' and 'Geospatial' metadata); and information by which the video data could be retrieved ('Descriptive' metadata) (Schachlbauer and Weiss 1998:678).

The ways in which video data is described vary between different people using it for different purposes – such variation may exist within an organisation. The British Broadcasting Corporation (BBC) has addressed this issue with the development of a data model that tracks video data from its inception in the minds of programme creators to its marketing and its distribution by means of broadcast, the Internet and physical recordings. The developers of the Standard Media Exchange Framework (SMEF) discuss the challenges that arise in handling the diverse perspectives and vocabularies used throughout the organisation to refer to the same video data (Hopper, Owens and Croll 1999).

3.1.3 Classes of Video Surrogates

It is apparent from the preceding discussions that there is a range of information about video data that needs to be captured to assist in its retrieval, browsing and analysis by different users for different purposes. Depending on particular circumstances, the relevant information might be about the video's production – that is its makers, time and place of creation, and its digital coding format; its structure – that is its shots and scenes, and how they are delineated with special effects and other editing techniques; and, its content – which, variously, refers to the spatio-temporal distribution of pixels, the entities and actions depicted in the moving image and their meanings for viewers.

The different kinds of information about video data which need to be dealt with can be grouped under the headings *bibliographic*, *structural* and *content* (Rowe, Boreczky and Eads 1994). Bibliographic information comprises details about the title of a video and its production; structural information refers to the composition of temporal segments in the video like shots and scenes; and, content information relates to what is seen and understood in the moving image. Whilst bibliographic information tends to apply to the video as a whole,

information about structure and content may relate to the video at different levels of spatial and temporal granularity.

On a second axis, video surrogates can be classed as *media features*, *visual features* and *semantic features* (Chang et al. 1999): media features relate to the coding format of video data, and its resolution and frame rate; visual features relate to the spatio-temporal distribution of pixels; and, semantic features deal with what is seen and understood in a moving image by a human viewer. In discussions of video content a distinction is often made between *visual content*, which pertains to the spatio-temporal distribution of pixels, and *semantic content*, which has been defined as "the message or information conveyed by the video" (Hampapur and Jain 1998:250).

These two ways of classifying video surrogates – as bibliographic/structural/content-related and as media/visual/semantic features – are combined in Table 3.1. Whilst the automatic generation of visual features tend to be 'low cost', it will sometimes be necessary to attach surrogates for semantic features: the question of what is 'good enough' is specific to a user's information needs.

	Media / Visual features	**Semantic features**
Bibliographic	Coding format, file size, video duration. (*Media* features)	The people and places involved in the production of the moving image.
Structural	Sudden changes in video signal levels. (*Visual* features)	Shot and scene structures, and camera actions, which convey meaning.
Content	Measures of colour, texture and motion in the pixels. (*Visual* features)	Entities and actions, and meanings, as seen and understood by humans.

Table 3.1: Examples of the information captured by classes of video surrogates. The distinction between 'bibliographic', 'structural' and 'content' indices is due to Rowe, Boreczky and Eads (1994); the distinction between 'media', 'visual' and 'semantic' features is due to Chang et al. (1999).

3.2 Adding Structure to Video Data

Information about video data needs to be organised in a structured fashion to allow different users to have different views of the same video sequences. As well as organising diverse features of video data, a video data model must also specify to what parts of the video data the features relate, i.e. to the whole sequence, or to sub-intervals and sub-regions of frames. The process of virtually segmenting video data into sub-intervals and sub-regions can be

automated by algorithms which detect temporal and spatial discontinuities and other cues in video data, including textual information that may be part of it.

3.2.1 Video Data Models

Data models provide multiple viewpoints of moving images for users with different querying, browsing and analysis needs. Some information pertains to a video sequence as a whole, for example, a film's title and release date, other information is only applicable to certain temporal intervals or spatial regions within the video data, for example, the presence of an object or a movement. Depending on the type of video, meaningful sequences may range in length from fractions of seconds to hours.

Video data files can be treated as Binary Large Objects (BLOb's): data modelling then involves making decisions about the attributes of the video data file, and its relationships with other entities in its domain. The resulting data structure holds information that relates to the whole video file, along with a reference to the physical location of the video data in the file system. This approach is suitable for recording bibliographic information about a video sequence and for giving a broad classification of its content.

A video data model can facilitate the attachment of features to discrete sequences of video data within a file. For a given video interval which is delineated by start and end times, any number of features of different types can be attached to provide different kinds of information. The modelling of these features will be specific for a video type and will be determined by the function of a particular system. A series of intervals may follow on, one from another, in a strictly linear fashion: for example as when a news broadcast is broken into separate stories, Figure 3.1a.

It is not always appropriate to impose a fixed segmentation on video data such that information about video content can only be attached to a pre-determined set of intervals. In complex visual information many actions and events may be co-occurring and overlapping; as, for example, when many dancers are performing independently on stage. Thus surrogates might only relate to a portion of a fixed interval or may extend across several fixed intervals.

Furthermore it may be desirable to add descriptions of content incrementally, as and when they become available. Such considerations led to the specification of a Video-Object data model (Oomoto and Tanaka 1993). Each 'video-object' comprises a start time, an end time and an attribute-value pair: for flexibility, the temporal span of video objects may overlap, and their permissible attributes and values need not be specified in advance, Figure 3.1b. Video-objects were used in the OVID system to store descriptions about people and their activities as seen in news video. The temporal relationships between video-objects were reasoned over using Allen's interval logic (1983), so that for example an interval contained within a longer one could inherit its annotations.

The fact that there are many ways in which video content can be described simultaneously was addressed in the proposal of a *stratification* model for organising information about video data. This model allowed for multiple descriptions of the same video sequence at different temporal granularities (Davenport, Aguierre Smith and Pincever 1991). In some cases visual events can be seen as being composed of shorter events, for example the scenes and acts that make up a play. In order that descriptions of video content with fine granularities could be related to descriptions at coarser granularities, the *nested stratification* model was proposed (Weiss, Duda and Gifford 1995), Figure 3.1c.

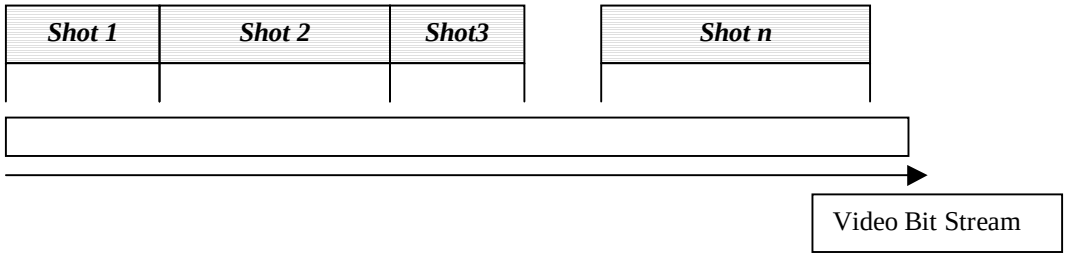


Figure 3.1a: An illustration of a simple video data model in which the video data stream is virtually segmented into discrete sequences that follow on one from another. This could be used to attach surrogates to separate stories within a news broadcast.

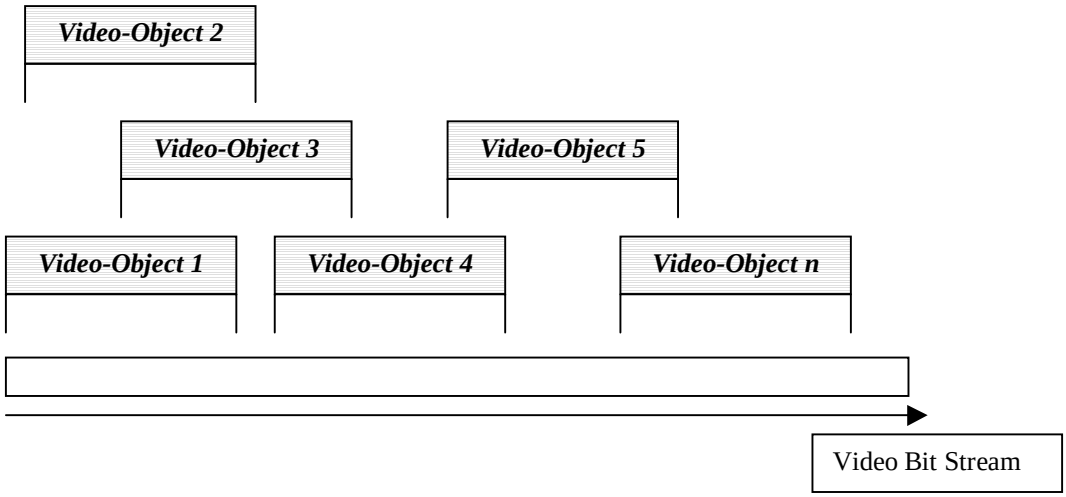


Figure 3.1b: An illustration of a video-object data model (Oomoto and Tanaka 1993). Each video object associates attribute-value pairs with a temporal interval to index objects and events. This could be used to associate surrogates with entities and events that overlap in the video sequence, as for example in a dance involving several independent performers.

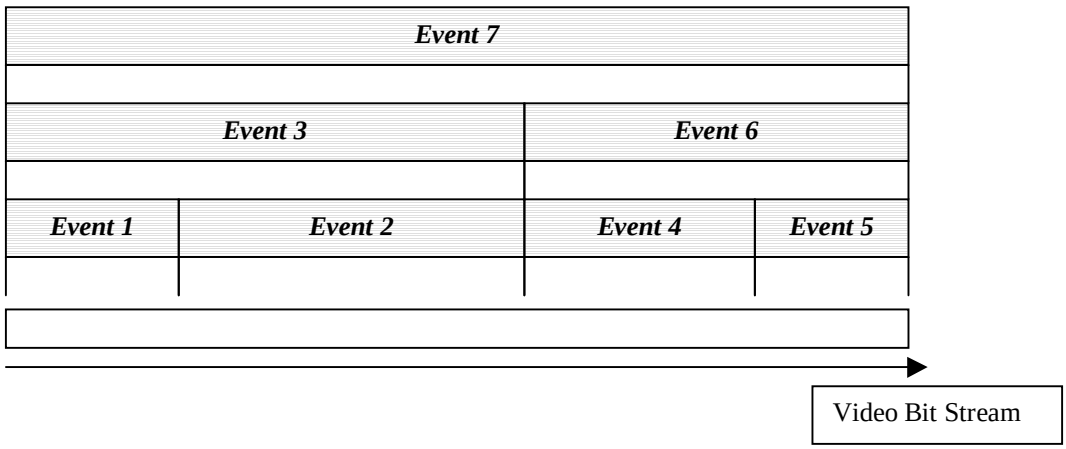


Figure 3.1c: An illustration of a hierarchical video data model in which complex visual events are composed of simpler ones (Weiss, Duda and Gifford 1995). This model captures the inherent structure of some kinds of moving image, like a recording of a play that comprises scenes and acts.

Until recently the nature of video coding, i.e. as frame sequences, has meant that video data models have organised information about video data in relation to temporal intervals. The advent of object-based video coding, as part of the MPEG-4 standard, will perhaps open the door for data models that attach descriptions to the objects portrayed in video sequences rather than to temporal intervals, Figure 3.2. The object-based coding of video data grants extra flexibility for indexing as well as presenting and manipulating moving images in computing and communications systems (Bove 1996).

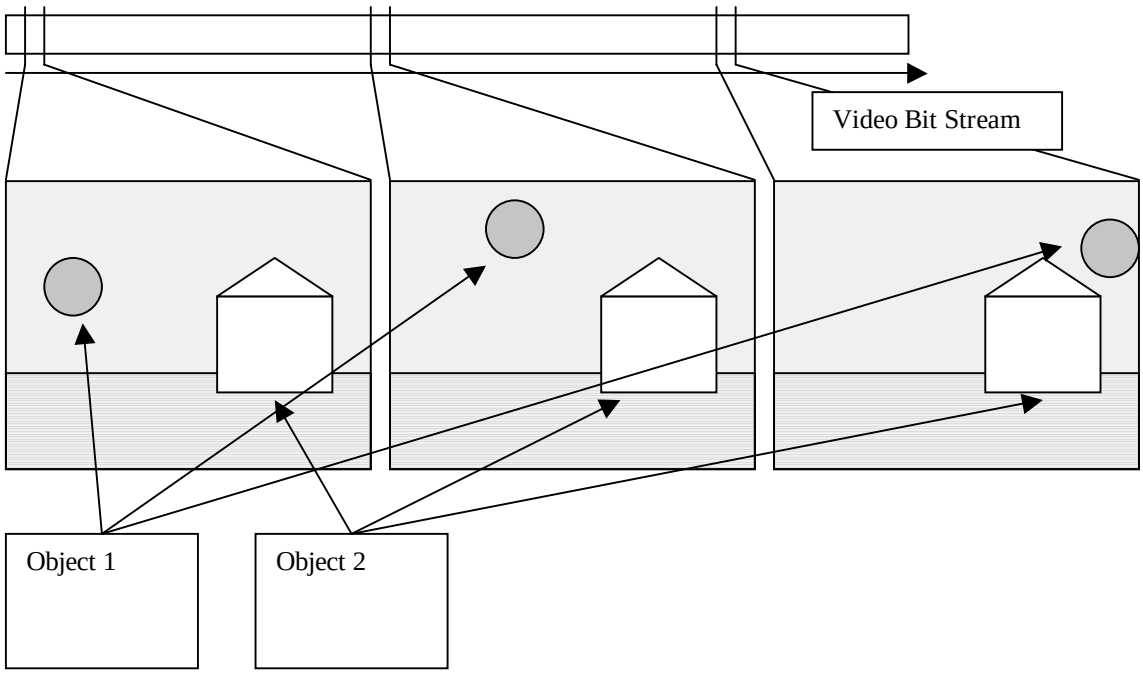


Figure 3.2: An illustration of an object-based video data model in which objects are identified across a sequence of frames – here three frames are shown. Annotations can then refer to these objects, rather than to the sequences of frames in which they appear.

In some types of moving image the composition of video sequences can play an important part in how their content is understood. Film theorists have elaborated the ways in which the ordering of shots, combined with editing effects and camera actions, can contribute to the messages conveyed by film (Bordwell and Thompson 1997). Thus, a video data model for cinematic film needs to incorporate information not only about content, like people and their actions, but also about the significant aspects of its structure which arise from the use of conventional camera techniques and editing effects.

A video data model intended specifically for movies and which is grounded in film theory was reported by Corridoni et al. (1996). This model facilitates the querying and browsing of digital video data based on combinations of technical film features and structure, as well as descriptions of content. A film is modelled as comprising shots, scenes and episodes. A film may have attributes like title and director; a scene may have a textual description and a shot may be described in terms of the objects and actions it contains and the camera techniques used to record them. These components of the film are joined by editing effects which are also described in the model; for example, sharp and gradual transitions. This data model allows a user to retrieve movie sequences according to a mixture of criteria, such as the name of an actor who appears in the scene and the kind of camera techniques used to film it. It was implemented in a system with a graphical user interface that allowed queries to be built up incrementally, to become increasingly specific.

3.2.2 Video Segmentation

Video data models specify meaningful units of video data like temporal intervals and spatial regions, to which surrogates can be attached. The task attaching surrogates to digital video libraries can be assisted by systems that virtually segment video data into such intervals and regions using algorithms that exploit either generic or domain specific features. Discontinuities of image features and textual features can both be generic cues to interval boundaries; spatial discontinuities of image features delimit regions within frames. In specific cases image features and textual features can indicate particular kinds of interval boundary.

Video Segmentation Using Image Features

Sudden changes between frames of image features, such as colour and texture, can indicate a change of scene and are therefore a cue for segmentation. Other image features may be derived from an analysis of optical flow to determine a change in camera motion, another cue for segmentation. A review of several generic techniques for identifying such discontinuities

in video data is presented by Hampapur, Jain and Weymouth (1996): this kind of approach exploits changes in low-level image features as cues to structure in moving images. The authors go on to propose a method for video segmentation which uses statistical models of image features to recognise special effects like *fades* and *dissolves*, as well as more straightforward cuts and changes of scene. Related techniques can also be used to recognise camera actions, like *pan* and *zoom*. For object-based coding schemes (like MPEG-4) it is necessary to segment video data into spatially contiguous regions of pixels that share colour, texture or motion features across a sequence of frames: for a range of such techniques see Torres and Kunt (1996).

Video Segmentation Using Textual Features

When video data contains textual information, such as the speech of a newsreader or the narrator of a documentary who both read from scripts, this can be analysed using linguistic techniques to segment the video data into meaningful intervals. Research in natural language processing has shown how cohesion patterns in text can be used to detect changes in topic. One such approach, Hearst's TextTiling technique, has been extended and used to segment news programmes into different topics (Mani et al. 1997). The technique is based on the comparison of terms in neighbouring fragments of text – a change from high similarity of terms to low similarity is taken to indicate a change of topic: the extension reported by Mani et al., is to compare the terms in the text fragments via a subject classification thesaurus.

In specific types of digital video linguistic phrases may mark the beginnings and ends of potential segments. The system presented by Mani et al. for segmenting news video uses phrases like 'And now over to X…', and 'Still to come…' as cues to supplement the cohesion based segmentation. The use of key phrases, or discourse cues, for highlighting the different stages of monologues in video data was explored by Takeshita, Inoue and Tanaka (1997).

The Potential for Automatic Segmentation

Automatic techniques have been shown to work well for certain kinds of video data, especially news video broadcasts which can be segmented by both changing shots (visual features) and changing storylines (textual features). However, in other kinds of moving

images there will not always be a correlation between important structural features and visual/textual features. In the case of a dance, for example, the entrance of the lead dancer onto a stage full of dancers is important for the structure of the dance, but it would not register as a discernible change in visual features (unless the director switched to, say, a close up of the dancer). This suggests the need for human involvement to segment some kinds of moving image.

3.3 Surrogates for the Content of Moving Images

Once bibliographic details about the production and distribution of a digital video sequence have been recorded, and once it has been decided what its meaningful segments are and how they are structured, the next task is to generate and organise information about the video's content. The term 'content' has been used, since long before digital video data, to refer variously to different aspects of communicative processes. In the context of digital video, a distinction is often made between 'visual content' that relates to the spatio-temporal distribution of pixels, and 'semantic content' which relates to the meaning conveyed/understood by the video/viewer.

3.3.1 Automatically Generated Visual Features

Visual features are computed from the colour values of pixels: such features might apply to a single frame or to a sequence. The review here discusses systems which use such features as the basis for indexing and retrieving video data. These kinds of approaches are currently widespread in the literature: for more details see the reviews by Mandal, Idris and Panchanathan (1999) and Aigrain, Zhang and Petkovic (1996).

Global Features

For a single frame, i.e. a still image, it is possible to compute global metrics of colour distribution and texture. The colour distribution within an image is thought to be a good

global property since it is invariant to image translation and rotation, and changes only slowly under the effects of different viewpoints, scale and occlusion. The red, green and blue values of pixels across an image can be collated in a histogram which is then taken as an index of the image. In order to cluster images or to match an image against a query, there are a range of mathematical techniques for assessing the similarity of colour histograms (Furht, Smoliar and Zhang 1995: 230-3).

Compared with metrics for image colour, there is less agreement about what constitutes a measurement of image texture. Two popular kinds of texture features are *Tamura features*, which seek to indicate the contrast, directionality and coarseness that a human viewer will see in an image; and, *Wold features* which capture periodicity, directionality and randomness. Both kinds of feature take the form of vectors which can be the basis for inter-image comparison and clustering (ibid.: 235-243).

In order that colour and texture features can be used to annotate digital video it is necessary to select representative key-frames from video sequences: the image features of the key-frames are used to index the sequence. This approach was used in a video annotation system reported by Zhang et al. (1997). The automatic selection of key-frames was made by picking the first frame in a shot along with any frame when there was a sharp change in colour within the shot that might indicate new visual content: the first and last frames of camera actions such as panning and zooming were also selected. The system also indexed features that applied to the whole shot, such as how brightness and colour levels changed throughout. Shots were then clustered by the image features of their key-frames, and by their global features, for subsequent retrieval and browsing.

It has been argued that the extraction of image features for the indexing of visual information is not only practical from an engineering perspective but may have a physiological basis. Researchers have suggested that global image features may contribute to the formation of perceptual categories, and that these categories in turn contribute to conceptual categories (Furht, Smoliar and Zhang 1995:227). It is also argued that the usefulness of image features for visual information retrieval can be enhanced by their weighting and combination in learning algorithms with human supervision (Minka and Picard

1997). Image features such as colour and texture have been the basis of prominent image retrieval systems which have been extended for video retrieval: these include IBM's *QBIC - Query By Image Content* (Flickner et al. 1997); MIT's *Photobook* and *FourEyes* systems (Minka and Picard 1997); and, Columbia University's *VisualSEEK* (Smith and Chang 1997).

The indexing of still images with low-level features has been applied in WWW-search engines, such as *AltaVista*. This system allows the user to make a query with an example image and ask for further images which are 'visually similar'. An example of this is shown in Figure 3.3 where the top-left image of a dancer is the example to which the remaining images have been matched according to visual features. It can be seen that whilst some of the returned images are also of dancers, most would normally be considered false matches: six of the returned images do not even include human beings, and of those that do only two depict dancers; if it was the particular dance position that was of interest then none of the returned images match the query. Although it is not applicable in this instance, visual similarity may nevertheless be useful for certain kinds of visual databases, like those storing images of trademarks and images of fabric samples.

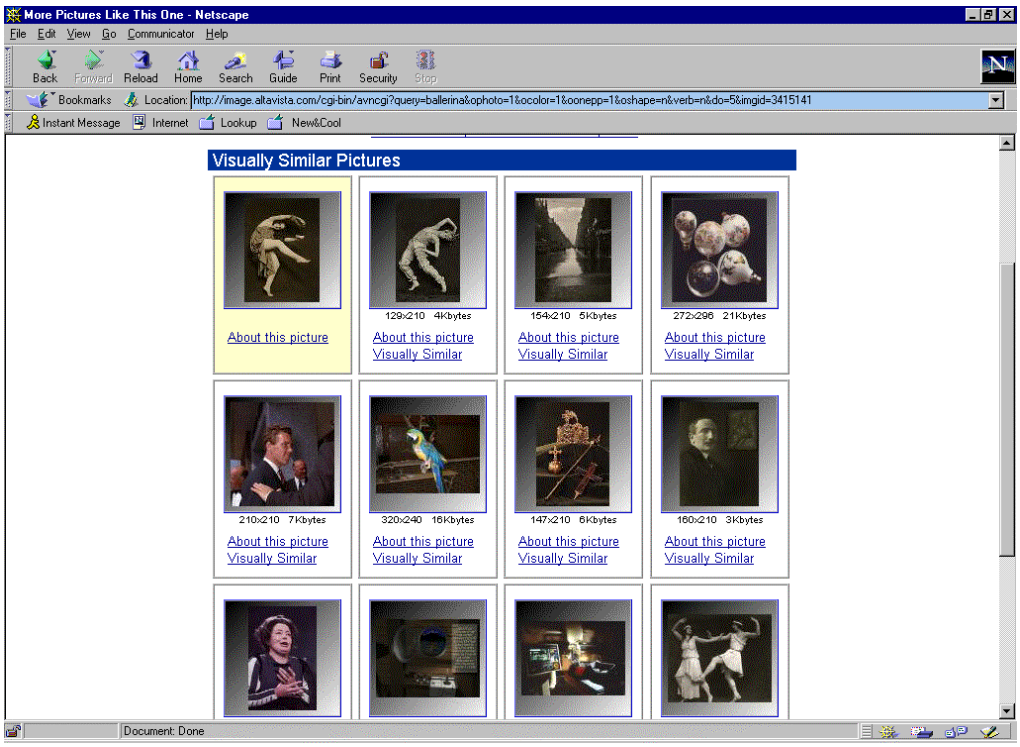


Figure 3.3: The results of a query by *visual similarity* using AltaVista's WWW-search engine for still images. The image of the dancer in the top-left corner was the search image: most of the returned images would generally be considered false matches.

Spatio-temporal Descriptions of 'Objects'

As well as extracting image features that apply to single frames, it is possible to segment and index 'objects' in video sequences. In this case an 'object' is a 2-dimensional region of pixels sharing colour, texture or motion features. The *VideoQ* system, implemented on the WWW, allows users to query a video database by drawing and positioning coloured and textured shapes, and indicating their motion with lines (Chang et al. 1998), Figure 3.4. The query is matched against an index of objects, automatically extracted from video sequences, which holds the colour, texture, shape and motion features of the objects. The shape of an object is recorded in terms of its area and its moments about its axes. Its motion is stored as a list of vectors which record the average translation of the centroid of the object between successive frames, compensating for global motion – such as that caused by a camera panning. This approach has been extended by the authors so that, through user-interaction, aggregated object features are associated with 'semantic' labels like *slalom*, *sunset* and *high-jump* (Chang, Chen and Sundaram 1998). Like computer vision systems, this approach depends upon a strong correlation between visual features and the depiction of concepts: though it works reasonably well for the examples given, there is no evidence yet that it will fare any better in general use than extant computer vision systems.

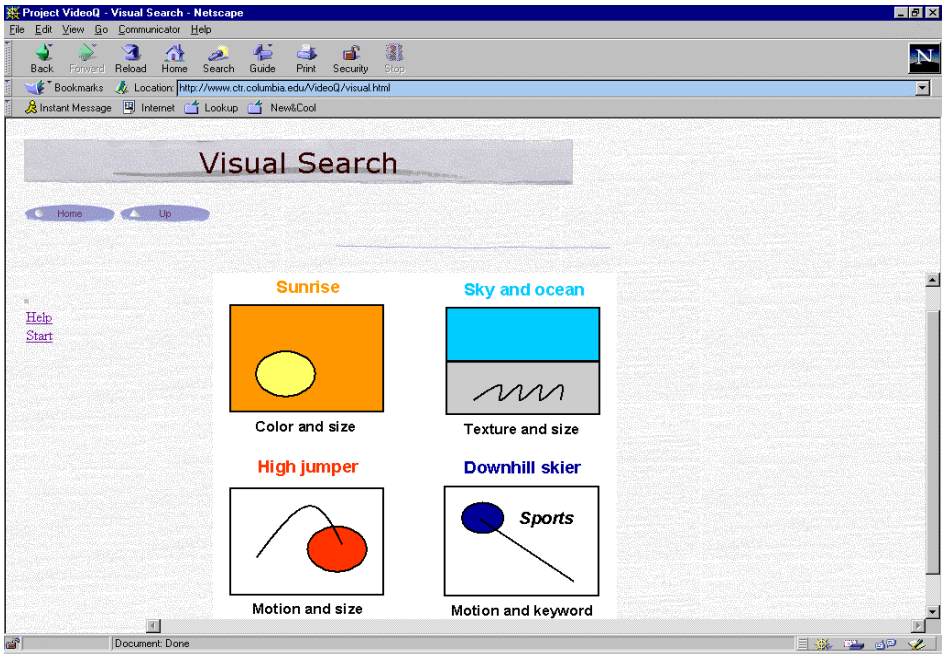


Figure 3.4: Examples of sketch-based queries which can be made to the *VideoQ* system for video retrieval over the WWW (Chang et al. 1998). The system generates an index of 2-dimensional objects extracted from video sequences to record colour, texture, shape and motion features.

In order to formally describe video sequences in generic terms of colour regions and their relative positions and motion, researchers proposed a *scene description language* (Gong et al. 1996). This language comprises 14 major keywords which are combined to describe the properties of objects and their spatial and temporal relations in a video sequence: *Colour, Region, Object, Event, Scene, Position, Shape, Area, Relation, Motion, Call, CameraOperation, ConsistOf, MatchMode*. A scene is defined in terms of its objects and four possible events which may apply to the objects over time: *Enlarge, Shrink, Appear* and *Disappear*. Each object comprises spatial regions and can have motion in one of eight directions: spatial regions are characterised by their shape and colour. The authors show how this scheme can be used to describe a moving image in which a train passes from left-to-right across the scene against a background of grass and sky: the train and the background are described as blocks of pixels with certain size and movement properties. Whether visual features are captured statistically, or in more formal schemes, their applicability depends upon an often weak correlation with conceptual features (except in examples where the visual properties of images themselves are important, e.g. trademarks and fabric samples).

3.3.2 Annotating Video with Related Text

It is not always convenient, nor indeed possible, to specify a query to a video database in terms of coloured and textured regions and their spatial and temporal relations. Furthermore, low-level image features do not reliably capture conceptual similarities between video sequences and queries. In fact, it may be more intuitive for a user to enter linguistic descriptions of the entities, actions or themes they want retrieved images to depict; it has been argued that language is closer to the concepts underlying a user's information need (de Jong 1998). From the machine point of view, it can be argued that some kinds of information, e.g. linguistic information, are better understood computationally than are other kinds, e.g. visual information: this further supports the notion of 'cross-modal information retrieval' such as querying video databases through language (Owen and Makedon 1999).

The simplest form of annotation is manually attaching keywords; possibly following the kinds of schemes discussed by researchers in pictorial information science – see Rorvig (1990) and Enser (1995). Keyword indexing schemes can exploit lexical relations, for example through the use of a thesaurus, to expand and to focus queries. Recognised problems with keyword indexing arise from human subjectivity and linguistic ambiguity, and from the cost of human resources. These problems can be exacerbated by the fact that images can simultaneously convey many meanings: if all the meanings are indexed there is a risk of low precision in retrieval, if too few are indexed then the retrieval will have a low recall rate. An alternative to manual annotation is to use textual information that is already associated with moving images. Video data sometimes includes a textual component in the form of speech and closed-captions. Other textual information arises in the making and distribution of videos, for example scripts and production notes. It is perhaps useful to distinguish between the use, as video surrogates, of 'integral' text and 'external' text.

Video Annotation with Integral Text

The linguistic channel in video data (i.e. speech and subtitles) can be processed in order to give a finer grained index of video data than keywords and subject classifications which apply to the video as a whole. The *Informedia* project, one of six major US digital library initiatives, has investigated how speech analysis and text processing technologies can be applied for video retrieval (Wactlar et al. 1999, Christel et al. 1996). The speech components of news broadcasts and documentary programmes are processed by a continuous speech recognition system. The resulting transcripts are processed to give keyword indices for video segments which have been delimited by image analysis techniques. The Term Frequency – Inverse Document Frequency (TFIDF) statistic developed in text-based information retrieval is used to rank the contents of the video segments: this means that a segment will be indexed with the terms and names that are spoken particularly frequently in its duration (compared with other segments).

Related questions about how to integrate speech recognition and information retrieval technologies for video access has also been explored in the context of video mail – email

messages comprising a head and shoulders shot of the sender and their speech. Speech recognition was used for 'spotting' keywords that were then used to index the video mail messages (Jones et al. 1997). Information retrieval techniques were also used in the development of a system for composing news stories (Ahanger 1999). The approach has been extended into a multi-lingual context by researchers and broadcasting agencies working in two major European research projects (Netter 1998). This work introduces translation technology so that textual information in one language which is associated with a video sequence can be used for cross-lingual access. The transcribed speech of news presenters has also been exploited in a system which allows the user to browse through transcripts of news broadcasts, following hypertext links between terms, and viewing associated video sequences (Shahraray 1999).

The processing of integral text may reap further benefits if it is combined with visual processing of video content. As part of the Informedia project, researchers combined a face detection algorithm with a linguistic processor for spotting names, in order to suggest candidate names for faces in video sequences (Satoh, Nakamura and Kanade 1999). Related work sought to establish correspondences between visual and linguistic cues to distinguish types of video sequence: for example, the detection of a facial close-up accompanied by direct narration suggests that someone is making a speech (Nakamura and Kanade 1997).

The use of the linguistic channel for indexing digital video is now a key feature of several commercial video retrieval systems particularly for use in archiving television and film material. These systems all use speech recognition and information retrieval technologies to index video data, particularly news broadcasts: examples include ISLIP's *MediaKey* packages, Excalibur's *VideoEngine* and Virage's *VideoLogger*: the latter was also implemented on the WWW as part of a trial video retrieval system tested by *AltaVista* which indexed video sequences of a politician speaking to camera.

Whilst terms in the text stream of video data may be good indicators of the overall message carried by the video, they will not necessarily describe the moving images which they accompany. This is in part because the text-image relationship varies between different types of video: for example the commentary in a sports broadcast may describe the action

quite closely; in contrast, the words spoken by a newscaster may have a rather ad-hoc relationship with moving images that are often selected from stock video footage libraries to illustrate the news story. Note that in the former case, although the images convey important information, the commentary can compensate for an absence of images; in the latter case, the important information is generally conveyed by the speech stream.

Video Annotation with External Text

A more informative, and perhaps more reliable, source of information about moving images might be found in texts, or at least text fragments, that are produced specifically to be informative about moving images. The two systems discussed here, which use text external to the video data, also extend the technique of keyword indexing. The first system adds a manually created subject taxonomy; the second parses text fragments so that the relationships between entities and actions depicted in videos are captured.

Classifying Video Sequences with Text Surrounding Hyperlinks: Smith and Chang (1997)

In addition to the low-level image features discussed previously, the *VideoQ* system uses keyword indexing and subject-based classification for video retrieval from the WWW. The user can form queries with keywords and Boolean operators, for example 'basketball game' and 'nature AND sunset'. Alternatively the user can browse down a hierarchy, following a subject taxonomy, for example from the category 'Nature' to the category 'Glacier' where there are 26 video clips of glaciers.

One part of the *VideoQ* system trawls the WWW to find HTML pages with links to digital video files. The HTML text in which the link to a video is embedded is used for indexing and classifying the video. All text strings occurring between non-alpha characters in the URL of the video and in surrounding HTML tags are considered as 'terms' for indexing (Smith and Chang 1997). The authors report the indexing of over 500,000 images and videos using 11,500 terms[3]. Terms are mapped to subject classes in order to place the video in one or

[3] Note that the *AltaVista* WWW search-engine now offers a similar video retrieval service based on the text found near hyperlinks to video data files.

more categories of a subject taxonomy, so that for example, a search for music videos returns videos classified as 'Rock Music' and also those classified as 'The Beatles'.

The subject taxonomy is created manually from a list of the most frequent terms in the index. A human annotator must decide on the selection and arrangement of the nodes in the taxonomy, and on the entries for a key term dictionary which maps terms to subject classes. A portion of the taxonomy, which totals 2,128 classes, is shown in Figure 3.5: note that the nouns used as labels for classes include concrete nouns (fruit and animal names), abstract nouns (scientific disciplines and music genres) and proper nouns (the names of places and people). The automatic classification of images and videos in nine classes was compared with human judgements and found to be 92% accurate. The authors reported that problems arose from ambiguous terms and from taking images out of context: for example, images of forests were classified as 'Possum' because they appeared on a WWW-page about possums to illustrate their habitat.

Animals (Dogs, Cats (Lions, Tigers), Whales)
Art (Paintings (Renoir, Da Vinci (Mona Lisa), van Gogh), Sketching, Sculpture)
Astronomy (Comets, Planets (Earth (Moon), Jupiter, Saturn), NASA)
Entertainment (Humour, Music (Pop, Rock (Elvis, Beatles)), Movies, Television)
Food (Vegetables, Fruits (Apples, Bananas), Drinks)
Plants (Cactus, Flowers (Roses, Sunflowers))
Science (Biology (Neurons), Chemistry (Proteins))
Sports (Baseball, Soccer)
Travel (Asia, Europe (France (Paris), Germany (Berlin)))

Figure 3.5: A portion of a taxonomy for image and video classification (Smith and Chang 1997) showing, as nested lists, nine of the top-level headings and some of their subclasses.

Smith and Chang's work is being extended as part of Columbia University's Digital News Service system where keyword indexing and a subject taxonomy are used to automatically provide images and videos to accompany textual summaries of news events. In this system researchers are currently exploring how longer passages of text associated with still and moving images, such as a news story, can be used to generate vectors of key terms for image and video classification (McKeown et al. 1998).

Annotating Video with Natural Language Descriptions: Kim and Shibata (1996)

Single keywords used for indexing digital video cannot capture the spatial and temporal relationships between the objects, actions and events portrayed by moving images. In longer textual descriptions of video content these relationships are encoded syntactically. The makers of television programmes and films may keep written records of the shots they take. A video annotation system has been reported which uses the descriptions of shot content written by film directors (in Japanese) during filming. In order to give more than just keyword matching the system uses a natural language analyser to parse the time-coded sentences into 'subject-centred dependency structures' which form an index to the video. These structures make explicit the subject of a sentence, the subject's actions in relation to other entities, and spatial and temporal information. Queries to the system are similarly parsed so that the user can request video sequences in terms of both entities and actions, as well as spatial and temporal constraints (Kim and Shibata 1996).

The system was tested with 34 natural language sentences describing 26 segments of video from a 40 minute long wildlife programme about a bird called a *saicho*; the sentences (translated from the Japanese) for the first three segments were:

1: A saicho flies from left to right.

2: A male saicho gives a fruit to a female saicho. There is a female saicho in the nest.

3: A saicho flies from the nest.

This approach is based on the assumptions that, in each sentence of video description: (i) the main object is encoded as the subject or topic; (ii) minor-objects are encoded as noun complements or adjuncts; (iii) actions and states are encoded as verbs or adjectives; and, (iv) relations between objects appear as grammatical dependency relations (ibid.: 697). The authors note that some linguistic processing is required prior to parsing in order to ensure that sentences have a single verb, a proper subject, minimal ambiguity and use syntactic structures consistently.

3.3.3 Knowledge Representation and Video Annotation

The ambiguities of natural language present problems for video annotation, with keywords having different senses and, as discussed above, with syntactic structures not relating consistently to participant roles. Such ambiguity can lead to a breakdown in communication between a video retrieval system and its user. Ways to represent the meanings of propositions unambiguously have been explored in the development of logics and knowledge representation schemes. Another aim of such endeavours is to facilitate inferencing about propositions. The promised lack of ambiguity and the potential for inferencing suggests the use of representations in a formal language as video surrogates.

The notion of *representation* is important to many aspects of computing: in the subject of artificial intelligence it is a central theme. A popular textbook in the subject defines representation as "a set of conventions about how to describe a class of things" (Winston 1992:16): by this definition, it seems reasonable to say that the practice of video annotation should at least aspire to be the 'representation of moving images'. It would perhaps be premature to say that video annotation systems are already representing moving images, since the term carries stringent connotations. For Winston a representation comprises four parts: the *lexical* – that is the symbols in the vocabulary; the *structural* – that is how the symbols can be arranged; the *procedural* – that is how to create and modify descriptions, and how to reason with them; and, the *semantic* – that is how to associate meanings with the descriptions. Although no current video annotation system meets these criteria, proposals have been made for using various kinds of formal languages for video surrogates.

Representations can explicate the relations that hold between entities and actions, and can locate them in space and time. A language based on a logical framework (Del Bimbo, Vicario and Zingoni 1995), and another based on an algebraic framework (Golshani and Dimitrova 1998) have been proposed for capturing the spatio-temporal relationships between objects and events in video sequences. Video surrogates that explicate, for example, 'at a given time X is behind Y, which is behind Z' allow the inference to be made that 'X is behind Z'. Herein lies the attraction of using a formal language for video annotation, rather than

natural language. However, it should be noted that these proposals, like the others discussed in this sub-section, only specify a language – not the means for generating surrogates for particular moving images; this is left as a (laborious) manual task.

As well as describing objects and events in space and time, knowledge representation formalisms can deal with participant roles (who is doing what to whom) and with causal relationships. Semantic networks, specifically propositional networks, have been used to represent the action of a feature film, not only in terms of entities and actions, but also in terms of the causal links between events. Links are made between regions of pixels and concept instances in the network, so that when the user clicks on an object depicted in the video, the instance is activated. This activation then spreads to associated concept instances, so that their associated video data is returned as a candidate match to the query – thus clicking on the scene of an explosion might return a video sequence depicting the character who pressed the detonator (Roth 1999).

Other knowledge representation schemes that have been used as the basis of video annotation include Minsky's frames (Davis 1995); Schank's conceptual dependency scheme which provided a 'context-free representation' that complemented a 'story grammar' in a video retrieval system (Tanaka, Ariki and Uehara 1999); and, Sowa's conceptual graphs that were applied in a system for creating novel video sequences (Nack and Parkes 1997, see also Parkes 1989). This interest, stemming perhaps from the apparent intersection of goals between knowledge representation and video annotation, suggests that representation schemes will play an increasingly important part in accessing video data. However, the subject of knowledge representation is a broad and long established one, so for now our review focuses on two examples: (i) frames which facilitate the inheritance of properties in hierarchies; and, (ii) conceptual dependency grammar which facilitate inferencing about actions.

Ahead of this, it is perhaps worthwhile drawing attention to the fact that there are some open questions which would impact on the use of knowledge representation schemes for video annotation. These include *quantification*, for example (how) could a scheme describe the fact that some of the dancers on stage are dancing, and some, but not necessarily the same

ones, are wearing skirts? Problems might also arise if a scheme had to deal both with a literal description of a moving image, and its interpretation. Consider the need to say that: (i) a film was made in Texas, and another film depicts Texas (though it may be filmed elsewhere); (ii) a dancer is the mother of another only in the context of the dance; (iii) the Prince, the lead dancer, and the dancer John Smith are the same person; and, (iv) the dancers are moving their arms as if they are swimming.

An Iconic Visual Language for Video Representation: Media Streams (Davis 1995)

As well as being a means for expressing unambiguous propositions about the content of moving images, knowledge representation schemes can also be used to develop hierarchies that facilitate inferencing about the content. Perhaps the most extensive hierarchy developed specifically for video retrieval is that reported by Marc Davis (1995). Davis' *Media Streams* system is grounded in knowledge representation theory: it was implemented in LISP and a language developed at MIT called *FRAMER* which is designed for cross platform media annotation, knowledge representation and database functionality (Haase 1996); the *FRAMER* language itself is grounded in Marvin Minsky's idea of frames (1975).

Davis argues for 'context-independent' video annotation on the grounds that video sequences might be reused in new contexts. Further, Davis is keen to remove any linguistic bias and ambiguity from the annotation-retrieval process. These considerations motivated the development of an 'iconic visual language for video representation' which aims to provide a general means of representing video content, irrespective of users' backgrounds and languages and irrespective of the video sequence's context. The language comprises some 3500 'iconic primitives' to represent entities and actions, as well as the settings of films and filming techniques: some of the icons for actions are animated. These primitives can be joined to represent the relations between subjects, actions and objects for both annotating and querying a video database.

Davis has organised the primitives hierarchically so that they can be selected, combined and attached to a timeline through a graphical user interface, allowing the user to make layered annotations at varying levels of detail. The top level categories of primitives are

space, *time*, *weather*, *characters*, *objects*, *character actions*, *object actions*, *relative position*, *screen position*, *recording medium*, *cinematography*, *shot transitions* and *subjective thoughts*. The hierarchy is seven or eight levels deep in places, for example the following steps lead to an icon for indicating that a video sequence is set in Texas: space → geographical space → land → continent → North America → USA → South Mid-Western States → Texas. Icons may be located in more than one place in the hierarchy, for example 'blow-dryer' falls under 'hand-held device', 'heat-producing device' and 'personal device'.

Davis suggests that the primary level of video content representation should be 'semantically invariant' and 'sequence-independent', and as such should capture what the viewer sees rather than what the viewer infers when watching a video sequence (cf. our earlier discussion of description and interpretation). This distinction is elaborated with reference to representations for human actions like hand-shaking which, according to Davis, should be represented in a decontextualised way, rather than in terms of contextually determined meanings such as 'greeting' or 'agreeing a deal'. In *Media Streams* there are icons for representing body parts and animated icons for actions which can be combined to represent more complex actions. These animated icons are organised according to whether they apply to the whole body or to a body part (head, arm, leg), and according to whether they are 'conventional' (walking, sitting, eating) or 'abstract' actions.

Proposals for 'universal languages' tend to be controversial, and perhaps Davis' work in the modern context of digital video libraries is no exception. The history of the 'search for a perfect language' has been discussed by the semiotician Umberto Eco who covers a period of several thousand years, towards the end of which comes work on knowledge representation languages (Eco 1995). It seems that in proposing a set of iconic primitives for video annotation Davis faces problems that have been faced (and not necessarily overcome) by scholars of many ilks. Firstly, like all those attempting to produce an *a priori* language, Davis would benefit from a principled approach for deciding what constitutes a primitive and how they should be arranged – this he seems to be lacking. Secondly, there are problems to do with the choice of visual icons for the primitives – recall that the motivation for using visual icons, as opposed to words, was to overcome cultural barriers (in the form of linguistic

barriers). However, it cannot be assumed that visual icons are free from culture-specific connotations any less than words are: for example, Davis suggests a 'thumbs-up' icon for representing the rating of a film, i.e. the more thumbs-up the better the film. But, the thumbs-up gesture is not universally understood with positive connotations.

Schank's Conceptual Dependency Grammar

Roger Schank has contributed to the representation of natural language utterances, in particular where some kind of movement is involved. His view of language understanding was as follows: "there exists a conceptual base into which utterances in natural language are mapped during understanding" (Schank 1975:188). Although the philosophical and linguistic import of this statement may be criticised, Schank's approach has been widely acclaimed, including by those developing virtual reality systems (Krueger 1991:163), and those interested in video retrieval. Schank is keen to emphasise the so-called conceptual dependencies which involve the definition of four primary 'conceptual categories' including: the 'picture producers (PP)' – largely objects in the real world; and, the 'acts (ACTS)' – including those of physical and mental transfer of objects and ideas, and acts of ingestion: then there are 'picture aiders (PA)' which modify nominals, and 'action aiders (AA)' which modify actions.

The key feature of his scheme is that it manipulates a set of primitives in order to make inferences about natural language utterances. Our interest here in recalling his work is to record the fact that Schank's scheme emphasises action, including physical action. Dance involves actions and reactions and an understanding of dance involves access to descriptions of dance. However, to attempt to reduce the simplest of rhythmic movements in a dance to a description in Schank's, or any other, primitives is some undertaking. Our purpose here is to draw the attention of the reader to this important work which may have some relevance to multimedia systems, especially since it deals with a language of actions, i.e. a set of primitives (vocabulary) and the grammar by which they are manipulated, according to the conceptual categories and their dependencies.

3.4 Discussion

Surrogates for a video sequence, or for a region within a frame, are required to capture details about its making, its structure and the entities and actions it depicts, along with the meanings these features convey. The exact demands placed on surrogates are dependent on their intended use in a system for accessing video data. In some cases a simple label of a film's title, or a dance's genre will suffice. In other cases, where the colour, texture and motion properties of video data are of interest to the user, or where these properties coincide with other, more meaningful, features, then statistically-based measures of these properties are appropriate. However, there will be instances, perhaps especially with systems storing specialist moving images, when a user wishes to make a detailed query in terms of a description of entities and actions, or even in terms of an interpretation.

Given the vast quantities of video data to be dealt with, it is important that the generation of video surrogates be as automated as possible. So-called visual features, i.e. properties of raw video data like colour, texture and motion can be automatically computed. Unfortunately, these features are not always 'meaningful' with regards to a user's information needs and so 'semantic' features are required. If a system is to successfully process queries that are made in terms of complex descriptions and interpretations, then video surrogates must convey this information. As noted previously, the state-of-the-art in computer vision precludes the automatic generation of such surrogates from video data, however manual effort on the task can be reduced if a system processes collateral text.

Research has already shown how collateral text can be processed for accessing video data, but the emphasis has been on integral text, like the speech of a newsreader, which may not have a strong relationship to the accompanying moving images. The use of text which was produced specifically to refer to the contents of moving images has been explored, but to date the text considered has been very limited, especially compared with the collateral texts produced by domain experts. Furthermore, the developers of these systems appear to take a rather ad-hoc approach to language, for example in their creation of taxonomies.

In the case of specialist images there is a range of collateral texts which describe and interpret moving images and provide other kinds of information about them, but that these

texts are intended for a human readership. If collateral texts are to be exploited for accessing video data then they must be processed into machine-executable forms. It might be that the ideal form would be an unambiguous representation of the meaning of the text that facilitated inferencing about the information it conveyed. In order to approach this ideal it will be important to understand more about the language used to produce collateral texts.

Chapter 4

Special Language and Moving Images

The analysis of moving images, especially for storing and accessing video data, can be carried in two ways. First, a given set of specialist moving images can be characterised in terms of variations in light, shade and colour, resulting in statistically-based or formally described visual features. The second way involves description and interpretation, and results in key terms and phrases which refer to the content of the moving image as it is understood by a viewer. Thus, visual features are complemented by 'semantic' features.

Howsoever 'semantic' is defined in this context, it is hard to imagine that semantic features could be consistently related to visual features. This statement should not be construed as a dismissal of visual features: recall that we suggested that the two levels are complementary. It is possible for machines to recognise some kinds of movement, such that they are automatically labelled with a linguistic descriptor, and thereby reducing the manual overhead in annotation. It is not clear however, and this is despite interesting developments in computer vision, that such automatic feature attribution could be scaled up so that muscular movements, like a *plié*, could be described in unconstrained examples.

Having clarified our position vis-à-vis visual features and semantic features, we will now focus on the latter. So what of linguistically-based semantic features? A sequence of images can be described in many different ways. To a neophyte, a prima ballerina's performance could be truly mesmerising, and the neophyte will be mesmerised without understanding the complex movements the ballerina used to convey the themes of the ballet. A critic asked to write two columns in a newspaper, or a magazine, will describe the historical context and previous performances of the ballet. The critic may focus on particular scenes or acts, while describing the virtuosity of the dancers, their individual and combined movements, and the ways in which their dancing conveys the ballet's narrative or other meanings.

A textbook of dance does not discuss specific performances in terms of location and time, but rather, relays an overall impression of a dancer, or a ballet, over many performances. The point about textbooks is that they try to enunciate general principles, for example that

classical ballet is based on an upright body and aspires to give the impression of weightlessness. A dance could also be described in a programme which refers to a specific staging of the dance, and will give a brief overview of the dance and a biographical sketch of the principal dancers. A journal paper on dance will look at the same dance from a specific theoretical perspective, attempting to situate the work of the paper's author in a network of other authors' papers.

This spectrum of people who write and talk about dance is broad, and indeed overlapping. The key point however is that the neophyte's articulation of dance, particularly in writing, will be similar in many respects to other people's everyday communication. For instance, the neophyte's vocabulary of dance will be limited and embedded in a variety of sentence types, and the semantics of the articulation will not be very rich. In contrast, the dance specialists will have a richer vocabulary of dance, but their articulation will be restricted to a limited set of sentence types, and at the semantic level there will be a premium on avoiding ambiguity. These lexical, sentential and semantic characteristics, that are seen in the specialists' choice of words, sentences, phrases and indeed whole texts, can be used to divide a given natural language into special language and general language.

For us this distinction is very important. If there is a special language of dance which is constrained at the syntactic level and idiosyncratic at the lexical level, then we have a language which can be manipulated by a computer program with a view to annotating moving images. And this annotation, because it is embedded in a language, will provide a more reliable basis for subsequently retrieving video sequences. This hypothesis, that the existence of a language constrained in meaning and grammar, and avoiding the problems associated with ambiguity in general language, will enable a more robust annotation of moving images is just that: a hypothesis. In this chapter we will attempt to support this hypothesis by providing evidence for a 'language of dance'.

We take the view that, in the course of their work experts may write about moving images for a variety of purposes resulting in different types of texts, but using a common language. Each text can be viewed as an artefact of an expert's analysis and as such it is informative about one or many moving images. Within a text, both descriptions and interpretations of

moving images may be interwoven with other information such as details about their production and their structure. Texts may also articulate concepts that are important for the analysis of moving images in a specialist domain. Thus, there may be a wide variety of textual information that describes, interprets and otherwise relates to moving images in a particular domain. In the case of dance, texts include programmes, critical reviews, textbooks and learned journal articles which are informative about dance and dances in different ways.

A parallel can be drawn between the potential use of such texts for video annotation purposes, and the established practice of knowledge acquisition which contributes to the development of expert systems. Some developers of expert systems make the assumption that domain expertise is encoded in the texts produced by experts, i.e. language reflects knowledge. Expertise comprises both *declarative* knowledge ('knowing that') and *procedural* knowledge ('knowing how to') and may be available in extant domain texts (*explicit* knowledge) or in the transcripts of knowledge elicitation sessions (to access *tacit* knowledge). The degree to which expertise is structured by the language of a domain determines the extent to which a knowledge-base can be (semi-automatically) built from a collection of domain texts: the knowledge-base will comprise *objects*, *facts* and *rules*, and when a reasoning strategy is invoked it will solve problems in a limited domain.

The development of a system to 'solve the problem' of analysing a dance is beyond our current ambitions, so we are not so interested in the procedural knowledge of dance experts. Rather, we are concerned with their declarative knowledge, i.e. what they know about the entities in their domain, and in particular, what they know about specific moving images. If this knowledge is structured by a special language then it is not only of practical import for annotating video data with collateral texts, but it may also say something about a three-way link between language, vision and knowledge.

This chapter considers the case for a 'language of dance' with reference to a systematically gathered collection of texts written by dance experts. An automated statistical analysis of word frequencies and distribution suggests significant contrasts with a general language sample (Section 4.1). The relationships between moving images and the different types of texts in the collection are then elaborated through a manual analysis which highlights ways in

which texts can be informative about moving images, apparently in accordance with the intent of their authors (Section 4.2). One particular text type, the dictionary, is then considered in more detail. Definitions of specialist movement terms demonstrate how language can 'go beyond the image', and the structure of definitions in general make them amenable to automatic processing for the generation of machine-executable relationships between terms (Section 4.3). The chapter closes by discussing the potential, given the evidence for a special language, to study the language-vision link, and for video annotation (Section 4.4).

4.1 Notes on a 'Language of Dance'

The grounds for supposing that a special language of dance might exist are observations made by linguists who delimit language varieties by different kinds of criteria. For example, variations in language registers have been attributed to regional, social, medial, field (of discourse) and attitudinal factors (Quirk et al. 1985). A special language is considered to be a distinct language register by virtue of its subject matter, i.e. field of discourse; other registers like dialects and sociolects are distinguished by the geographical and social groupings of their speakers. Such 'macro level' factors can be associated with variations in linguistic features at a 'micro level' – for example, patterns of vocabulary usage and syntactic features.

One scholar noted seven different approaches to the study of a special language, or a Language for Special Purposes (LSP) (Hoffman 1984): the first approach is the lexicological or terminological study of vocabularies for special purposes. Other approaches include: functionally orientated studies which seek to correlate linguistic phenomena with communicative needs in a teleological relationship; research concerned with the use of language in science and philosophy, and in scientific and technical translation; and, theories of sublanguages. Functionally orientated approaches tend to focus on the grammatical features of an LSP and note a preponderance of certain structures in an apparent response to a communicative need. In the terminological approach to the study of LSP, the relationship between vocabulary (often nominal groups) and the concepts of a domain is made explicit.

Thus it is this kind of approach that is perhaps most appropriate when the concern is with the knowledge of an expert community.

Language may be seen as a means to idealise the world and share thoughts about it with others. In this view, vocabulary develops in tandem with concepts about the world, and grammatical relations pertain to relationships between these concepts: "Language is used to isolate units of experience and knowledge and to order them in various ways" (Sager, Dungworth and McDonald 1980:19). Thus, a special language, serving a group of domain experts who are constantly developing new ideas, will be expected to exhibit a dynamic vocabulary. On the other hand, the importance given to maintaining clear communication in specialist discourse means that the special language tends to use only a limited set of the grammatical features of its associated general language.

Much research in linguistics since the 1960s has been dominated by structures derived from introspective grammaticality judgements that are elicited for decontextualised, and often contrived, language fragments. Since the 1980s there has been a return to favour of an earlier method of linguistic study based on large samples of real-world text: the collection and analysis of these text corpora has been greatly aided by the modern computer and the Internet. The importance of corpus linguistics was emphasised by several eminent linguists of different theoretical persuasions at a 1992 Nobel symposium (Svartvik 1992).

In the context of LSP research, a corpus-based approach allows a language register (i.e. a special language) to be characterised by a combination of statistically significant linguistic features which appear in a collection of texts of a certain register, but do not appear in a general language sample. The language register may be determined by a unique subject matter, a unique type of text or other 'situations of use' (Biber, Conrad and Reppen 1998:135).

4.1.1 Method: measuring linguistic variance in the Surrey Dance Corpus

One corpus-based methodology to test for a special language that has been proposed is based on the notion of *linguistic variance*, that is statistically significant and systematic differences between linguistic features observed in corpora of specialist texts and in a corpus of general language texts. The tests for linguistic variance consider lexical, morphological and collocational features (Ahmad 1999). The approach can also be used to systematically identify and elaborate the terminology of a domain. In the current work this methodology was applied to the study of a corpus of dance texts; the analysis was performed using *System Quirk* – the University of Surrey's corpus management and text analysis package.

The Surrey Dance Corpus was compiled to include diverse types of texts including: chapters from textbooks and theses, and journal articles – some of which were scanned from print, others of which were acquired in electronic format from their authors; and, newspaper articles, reviews, and publicity previews of dances – which were gathered from the WWW. The selection of these texts was also guided by the need to include different varieties of English, e.g. British and American; the majority of texts were contemporary, i.e. written in the 1990s. A breakdown of the corpus, which totalled 322 texts and 346,263 tokens, is given in Table 4.1. The general language sample used for comparison was the British National Corpus (BNC) which totals 100,000,000 tokens, and includes a wide variety of everyday text types.

Texts	**Breakdown by Numbers of Texts and Tokens**	
	Language Variety	**Text Type**
British English	140 texts / 245,847 tokens (71%)	
American English	182 texts / 100,416 tokens (29%)	
TOTAL	322 texts / 346,263 tokens (100%)	
Press		301 texts / 180,056 tokens (52%)
Theses		10 texts / 86,566 tokens (25%)
Book		6 texts / 55,402 tokens (16%)
Journal		5 texts / 24,239 tokens (7%)
TOTAL		322 texts / 346,263 tokens (100%)

Table 4.1: Breakdown of the Surrey Dance Corpus which totalled 322 texts and 346,263 tokens: note the selection of text types and language varieties to improve the 'representativeness' of the corpus.

The analysis presented here compares the occurrence and distribution of lexical items in the specialist corpus of dance texts with their occurrence and distribution in the general language corpus. The aim is to highlight linguistic variance between the special language and the general language as evidenced by the two corpora. Contrasts are noted between the frequencies of lexical items, morphological endings and collocation patterns.

The method of linguistic variance has been previously used to study special language with text corpora from the domains of Nuclear Physics, Linguistics and Automotive Engineering. It should be noted that in comparison with these fields, the subject of dance analysis is in its relative infancy and as such it is less likely to have developed such a prevalent terminology as some more established disciplines. One feature that might be expected in the terminology of a nascent field is a preponderance of borrowings, that is terms taken and possibly adapted from other fields: one dance scholar has noted an influx into their subject of terms from anthropology, psychology, philosophy and sociology (Adshead 1988:16).

4.1.2 Lexical Variance

Experts use some of their own general language when they speak and write: in particular they use grammatical constructions, and *closed-class* words that would be familiar to a non-specialist. A distinguishing feature of an LSP arises from its use to communicate about a limited subject matter which means that certain *open-class words* are repeated frequently in specialist texts. These words may, at first glance, appear familiar to the non-specialist but they may be used in restricted or modified senses by the experts: furthermore, they may be used as carrier words for the coining of new compound terms.

The proliferation of domain specific open-class words is observed by comparing the 100 most frequent words in the Surrey Dance Corpus with the top 100 from the BNC: in the general language sample only two of the top 100 words are open-class - *time* and *people*; the rest are closed-class - *the, of, and, a*, etcetera. In contrast, the nature of specialist discourse has caused a number of open-class words to rise up in the frequency list for the specialist dance corpus: here 21 of the top 100 words are open-class, Table 4.2a. These 21 open-class

words on their own account for about 5% of the corpus so they contribute significantly to the 'flavour' of experts' writings about dance.

The preponderant open-class words in the dance texts include words to refer to a dance as a whole - *work, works, performance, piece*; to refer to the entities and actions of a dance - *dance, dancer, dancers, dancing, movement, body*; to refer to a genre of dance – *ballet*; and, to refer to aspects of the dance other than human movement – *music* and *stage*. As well as appearing in different morphological forms, the word *dance* appears in the corpus as part of terms like *modern dance* and *dance analysis*. The preponderance of open-class words, to refer both to the practice of dance and to the practice of dance analysis, points to common concepts running through the discourses of dance experts.

Position	Token	Relative Frequency	No. of OCW
1-10	the, of, and, a, in, to, is, s, as, that	24.24%	0
11-20	**dance**, with, it, by, for, on, this, her, at, was	6.58%	1
21-30	from, are, his, an, but, I, which, she, their, be	3.95%	0
31-40	**dancers**, or, **music**, he, not, they, one, **movement**, **work**, **ballet**	3.15%	5
41-50	has, its, all, other, more, have, **new**, into, **body**, there	2.17%	2
51-60	**davies**, can, **company**, two, also, were, we, up, through, out	1.84%	2
61-70	when, **like**, so, what, these, **performance**, you, who, between, **stage**	1.63%	3
71-80	first, both, about, each, some, only, **piece**, **back**, than, **dancer**	1.42%	3
81-90	then, them, **time**, had, **see,** if, been, own, three, no	1.27%	2
91-100	**theatre**, very, **works**, **dancing**, much, way, just, most, my, over	1.08%	3
	CUMULATIVE RELATIVE FREQUENCY	**47.33%**	**21**

Table 4.2a: The 100 most frequent words in the Surrey Dance Corpus which account for almost half the corpus (47.33%). Open-class words (OCWs) are indicated through the use of bold type face.

The open-class words are 'diluted' by the amount of text in the corpus from WWW-based press reviews: these are not necessarily written for an expert audience and will therefore be less technical in nature. An analysis of a reduced corpus – 174,298 words without the press reviews – gave 25 open-class words in the top 100: these now include words that emphasise the practice of dance scholars, Table 4.2b. Note the use of *text, phrase* and *section* to refer to the dance (sometimes 'performance text') and its parts. The analysis of the reduced corpus also shows more words referring to the practice of dance itself, for example, *arms, position, movements*.

Position	Token	Relative Freq.	No. of OCW
1-10	the, of, and, in, a, to, is, as, that, **dance**	26.98%	1
11-20	s, with, for, on, it, this, her, by, which, from	6.50%	0
21-30	are, **movement**, was, his, an, be, their, or, **music**, at	4.23%	2
31-40	**dancers**, but, **davies**, **work**, not, one, she, they, **body**, he	3.19%	4
41-50	its, has, also, through, more, into, can, other, both, between	2.32%	0
51-60	I, two, all, up, these, **ballet**, when, out, have, first	1.88%	1
61-70	there, each, were, only, **time**, than, **new**, **stage**, so, **back**	1.51%	4
71-80	own, who, then, some, **performance**, **dancer**, **arms**, **bar**, same, what	1.26%	4
81-90	three, had, them, such, **analysis**, **text**, **phrase**, **section**, if, made	1.15%	4
91-100	**works**, **movements**, over, **position**, been, like, no, we, **different**, **company**	1.09%	5
	CUMULATIVE RELATIVE FREQUENCY	**50.11%**	**25**

Table 4.2b: The 100 most frequent words in the Surrey Dance Corpus, excluding press reviews. The more technical nature of the restricted corpus is reflected by the appearance of words like *text, phrase* and *section* to refer to dances; and, *arms, position* and *movements* that suggest more detailed descriptions of movement.

The case for a 'language of dance' is supported by earlier results that show a relative abundance of open-class words in other special language corpora (Ahmad 1999) - Linguistics (688,000 words), Nuclear Physics (472,000), Automotive Engineering (325,000). The frequency of open- and closed-class words in each of these corpora is shown in Table 4.3a, alongside the frequencies in the Dance corpus (excluding press reviews). The observed frequency of open-class words can be compared against the expected frequency, i.e. the average across the specialist corpora, to show that the corpora are statistically the 'same' with respect to frequently occurring open-class words: this is shown with the χ^2 statistic. By way of contrast, the comparison of the Dance corpus with the general language BNC gives a χ^2 value of 39.184, from the data in Table 4.3b: this provides evidence that the two corpora are not the same.

Word Class	Surrey Corpora				
	Automotive Engineering	Nuclear Physics	Linguistics	Dance	EXPECTED
Open	39	33	30	25	**31.75**
Closed	61	67	70	75	**68.25**

Table 4.3a: The make-up of the 100 most frequent words in four special language corpora. By using the χ^2 statistic, value 6.472, it is possible to say that there is no statistically significant evidence of difference between the corpora. (For 95% certainty, the χ^2 value for three degrees of freedom is 7.81).

Word Class	Surrey Dance Corpus	BNC	EXPECTED
Open	25	2	13.5
Closed	75	98	86.5

Table 4.3b: The make-up of the 100 most frequent words in the Surrey Dance Corpus and the British National Corpus. Here the χ-square value is 39.184, suggesting there is a difference between the corpora with respect to the occurrence of open- and closed-class words. (For 95% certainty, the χ^2 value for one degree of freedom is 3.842).

By looking at the most frequent words in a corpus of specialist texts it is possible to observe the general themes of experts' communications, but a different technique is required to look at more specific concepts that are only discussed in a few texts. Even if a term's absolute frequency is low in a special language corpus, its relative frequency may be high compared with its occurrence in general language. A 'coefficient of weirdness' can be calculated by dividing the relative frequency of a lexical item in a special language corpus by its relative frequency in a general language corpus. Lexical items with high coefficients of weirdness in the Surrey Dance Corpus include movement terms like *arabesque* and *pirouette*, and theoretical constructs borrowed from other domains like *post-modern, structuralism, formalist* and *minimalism*, Table 4.4. (It should be noted of course that many specialist ballet terms are borrowed/adapted from French general language).

	BNC		Surrey Dance Corpus		
	Absolute Frequency	**Relative Frequency**	**Absolute Frequency**	**Relative Frequency**	**Weirdness Coefficient (SL ratio/GL ratio)**
arabesque	32	3.20 x 10-7	35	1.01 x 10-4	316.22
post-modern	159	1.59 x 10-6	41	1.18 x 10-4	74.55
minimalism	35	3.50 x 10-7	8	2.3 x 10-5	66.08
pirouette	35	3.50 x 10-7	5	1.44 x 10-5	41.30
structuralism	185	1.85 x 10-6	24	6.93 x 10-5	37.51
formalist	109	1.09 x 10-6	12	3.46 x 10-5	31.83

Table 4.4: Example of 'weird' words in the Surrey Dance Corpus; weirdness is measured by dividing the relative frequency of a word in the special language corpus by its relative frequency in the general language sample.

4.1.3 Morphological Variance

Lexical items appear in different forms which are generated by both *inflectional* morphology and *derivational* morphology. Inflectional morphology signals grammatical relationships such as number and time, using in English, for example, suffixes like –s and –ed to mark

plurals and tenses. Derivational morphology changes a lexical item's grammatical class so that, in the example of nominalisation, a verb becomes a noun, e.g. *move* becomes *movement*. The morphological behaviour of lexical items in a special language corpus can indicate the ways that experts are using terms in their discourse. The use of singular or inflected plural forms may indicate the way in which the experts make generalisations: a preponderance of derived nominalisations may suggest that the domain experts prefer to communicate about processes as static entities.

The preponderant lexical items in the corpus show a preference for the base form in some cases and the plural in others. Experts are much more likely to write about *movement*, rather than *movement*s, but will write more about *dancer*s than *dancer* – see columns 1 and 2 in Table 4.5. The ratio between plural form and base form can be compared with the equivalent ratio in a general language sample: values of this ratio that are greater than one suggest a word is 'more plural' in the special language, and conversely, values less than one suggest it is 'less plural'. Divergence in either direction from the 'norm', i.e. 1, is indicative of linguistic variance – see column 7. Such variance may be partially explained by the ways in which experts make general statements about, for example, *movement* on the one hand, and *dancers* on the other.

	Surrey Dance Corpus			British National Corpus			Dance / BNC
	Frequency		Plural / Base	Frequency		Plural / Base	(3) / (6)
Token	Plural Form (1)	Base Word (2)	(3)	Plural Form (4)	Base Word (5)	(6)	(7)
performance(s)	219	522	0.42	1627	12,998	0.13	3.23
dancer(s)	1174	444	2.64	824	598	1.38	1.91
performer(s)	157	69	2.28	623	482	1.29	1.77
art(s)	165	277	0.59	5354	15,587	0.34	1.74
arm(s)	281	141	1.99	11,143	9194	1.21	1.64
hand(s)	151	191	0.79	18,978	35,352	0.54	1.46
movement(s)	278	1045	0.27	4,388	13,504	0.32	0.84
choreographer(s)	163	288	0.57	160	179	0.89	0.64

Table 4.5: Frequently occurring plurals and their base forms in the Surrey Dance Corpus, and the equivalent values in the BNC. Some lexical items tend to occur more in the singular, others more in the plural. The right-hand-most column divides the specialist corpus ratio of plural/base by the general language ratio: values away from the 'norm', which would be 1, indicate linguistic variance. Note, since our corpus is untagged we are unable to deal with lexical items that are both nouns and verbs, e.g. *dance(s)* and *work(s)*.

It has been argued that, in the eyes of scientists (and so perhaps other specialists) "the world is a world of things, rather than of happening; of product rather than process; of being rather than of becoming" (Halliday and Martin 1993:116). In an analysis of a Nuclear Physics corpus it was noted that several key lexical items were nominalised more than 90% of the time (Ahmad 1999): this might indicate the specialists' preference for talking about 'things' rather than about 'happening'. An analysis of frequently occurring nominals in the Surrey Dance Corpus shows that they are used many times more often than their base forms – see columns 1 and 2 in Table 4.6. However, with one exception – *move(ment*) – they are no more 'nominal' than they are in a general language sample, and, in some cases – *construct(ion), interact(ion)* and *situat(ion)* – they are significantly less nominal, see column 7. So, although this data shows linguistic variance in the dance corpus, it is not in the direction predicted by the earlier analysis of scientific texts; perhaps the 'happening' nature of the dance experts' subject of study has an effect here?

	Surrey Dance Corpus			**British National Corpus**			**Dance / BNC**
	Frequency		**Nominal/ Base**	**Frequency**		**Nominal/ Base**	
Token	**Nominal Form (1)**	**Base Word (2)**	**(3)**	**Nominal Form (4)**	**Base Word (5)**	**(6)**	**(3) / (6)**
move(ment)	1323	160	8.27	17,892	20,427	0.88	9.40
collaborate(ion)	68	7	9.71	1401	221	6.34	1.53
express(ion)	81	34	2.38	8782	4959	1.77	1.34
relate(ion)	65	11	5.91	14,007	2602	5.38	1.10
associate(ion)	67	9	7.44	13,538	1207	11.22	0.66
develop(ment)	90	34	2.65	37,523	8636	4.34	0.61
construct(ion)	45	26	1.73	6643	1412	4.70	0.37
interact(ion)	38	12	3.17	3388	198	17.11	0.18
situate(ion)	47	3	15.67	19,856	48	413.67	0.04

Table 4.6: Frequently occurring nominalisations and their base forms in the Surrey Dance Corpus, and the equivalent values in the BNC. Whilst all the lexical items tend to occur more as nominals in the dance corpus, the ratio of nominal/base tends to be less than in the general language sample. The right-hand-most column divides the dance corpus ratio of nominal/base by the BNC ratio.

4.1.4 Co-occurrence Variance

When innovating concepts, experts sometimes combine lexical items to form compound terms with which to refer to them. Thus, one characteristic of a special language is idiosyncratic sequences of open-class words, which individually may or may not be common

in the general language. The data available to us about the BNC only related to the frequency of individual words and not to co-occurrence information, so this sub-section does not measure linguistic variance statistically. However, it is still possible to identify idiosyncratic sequences, or *collocations*, of lexical items within a corpus.

Compound terms, in English at least, tend to be formed by a series of open-class words particular to a domain: thus, one heuristic for identifying compound terms in a special language corpus is to look for sequences of open-class words that are uninterrupted by closed-class words and other common general language words. This technique provides further evidence for a 'language of dance' by producing a number of candidate terms from our corpus, Table 4.7. These candidates were the most frequently occurring sequences of lexical items that were not found in stored lists of closed-class and common general language words. Many of these candidates are not terms, but are proper nouns referring to the names of choreographers and dance companies. The list has been refined by checking for the occurrence of plural forms in the corpus to highlight candidates such as *performance text*, *Graham technique* and *hybrid site*.

Compound Term: Singular		**Frequency Singular**	**Compound Term: Plural**	**Frequency Plural**
Candidate Terms with plural forms in the corpus				
ticket	price	74	ticket prices	5
swan	lake	34	swan lakes	1
world	premiere	18	world premieres	7
performance	text	15	performance texts	1
ballet	school	14	ballet schools	2
choreographic	style	11	choreographic styles	1
music	phrase	9	music phrases	3
violin	line	8	violin lines	1
graham	technique	6	graham techniques	1
hybrid	site	4	hybrid sites	1
diagonal	line	4	diagonal lines	1
central	character	4	central characters	1
Proper nouns – no plural forms found				
siobhan	davies	39	-	-
merce	cunningham	22	-	-

Table 4.7: Candidate compound terms extracted from the Surrey Dance Corpus by taking sequences of lexical items not found in lists of closed-class and other common general language words. The list is refined by separating out those candidates that have plural forms in the corpus: these are considered more likely to be terms.

Some compound terms are formed from lexical items that are individually common in general language so would not be extracted using the technique outlined above. However, such compounds might be identified by a technique that analyses statistically significant co-occurrences of lexical items in a corpus. Table 4.8 shows the distribution of lexical items that frequently occur in the Surrey Dance Corpus within a neighbourhood of five words either side of the nucleate *movement*. Whilst these frequent co-occurrences of lexical items close to the nucleate may suggest collocation patterns, further discrimination is required to establish statistically significant collocations.

Two statistics were proposed by Smadja (1994) to determine the significance of collocation patterns of lexical items around a nucleate. The first statistic highlights lexical items that appear mostly in the same position relative to the nucleate: this 'U-score' is calculated as the variance of the frequencies for lexical items across the 10 positions. The lexical items co-occurring with the nucleate *movement* are ordered by their U-scores in Table 4.8: prominent are candidate terms such as *movement vocabulary*, *movement material* and *movement analysis*. The second statistic, the 'k-score', normalises the frequencies of the collocations by taking into account the expected co-occurrence of the lexical item and the nucleate according to their respective frequencies in the corpus. A high 'k-score' strengthens the candidacy for *dance movement* and *movement vocabulary* to be terms.

		P_{-5}	P_{-4}	P_{-3}	P_{-2}	P_{-1}	P_{+1}	P_{+2}	P_{+3}	P_{+4}	P_{+5}	**U-score**	**k-score**
movement	vocabulary	0	0	1	3	0	45	1	1	0	0	177.69	12.86
movement	material	2	0	2	0	0	22	0	0	0	1	42.01	6.55
movement	human	0	1	0	0	20	0	1	0	1	0	35.01	5.50
movement	dance	6	11	6	3	17	0	13	12	7	8	22.81	21.26
movement	analysis	0	0	2	5	0	13	0	1	0	0	15.49	4.98
movement	own	1	5	0	0	13	0	0	2	3	2	14.44	6.29
movement	pedestrian	0	1	3	0	13	0	1	1	0	1	14.2	4.71
movement	style	1	0	1	4	0	11	0	0	0	2	10.69	4.45
movement	laban	0	1	5	0	10	0	0	0	1	0	9.81	3.93
movement	quality	0	0	10	2	0	6	1	1	0	2	9.76	5.24

Table 4.8: The frequency of collocations with the 'nucleate' *movement*, in positions from five words before (P_{-5}) to five words after (P_{+5}). Smadja's (1994) statistics for measuring the strength of collocations are used here to highlight candidate terms. High U-scores and k-scores indicate strong collocations and here suggest *dance movement* and *movement vocabulary* as terms.

4.2 Text Types Related to Moving Images

The previous section presented statistical evidence for a 'language of dance', at lexical, morphological and collocational levels across a corpus of dance texts. As well as considering features of the corpus as a whole, it is important to remember that each text, which may be collateral to one or many moving images, is produced in a particular communicative scenario. The intention of the text's author and the expertise and expectations of its readers will determine the selection of information about moving images, and how this information is expressed in the text. Common communicative scenarios give rise to text types, some of which are seen across subject domains, for example textbooks and journal papers; other types may arise within a domain. With regard to the language-vision link, it is interesting to examine the relationships different types of text can have with moving images: with a view to video annotation, it is particularly interesting to see how these texts are informative about moving images.

The delimitation of text types can draw on many criteria relating either to the use of texts, or to their inherent features. An exhaustive typology can be very detailed: when documenting text forms in the domains of science and technology Sager, Dungworth and McDonald listed over 100, starting "address, agenda, aide-mémoire, announcement, calendar, catalogue…" (1980:147-181). One scholar proposed five groups of criteria for the classification of special language texts which included features inherent to the text and features relating to its use: the groups of criteria fell under the headings of *situational*, *social*, *pragmatic*, *semantic* and *syntactic-stylistic* (Weise 1993). Another proposal distinguished LSP communications at three levels (Gläser 1993): first, texts were divided according to their intended receiver, so they were either 'expert-to-expert' or 'expert-to-nonexpert'; then they were grouped according to their purpose – that is whether they were intended to convey information or to be directive, didactic or interpersonal; thirdly, texts were organised as primary forms (research articles, theses, experimental reports), derived forms (abstracts and reviews of primary texts), pre-text forms (outlines of proposed texts) and quasi-text forms (invoices, certificates and agendas).

The classification of text types needs to address many criteria and might only be resolved in relation to a particular need. For studying the language-vision link, particularly with a view towards video annotation, it is appropriate to characterise text types by how they are informative about moving images. That is to say, text types should be delimited in terms of the kinds of information about moving images they convey, and how this information is organised in the text. The examples with which this point is elaborated here are listed in Table 4.9; note the distinction between texts that refer mainly to a specific dance, and those which refer to many dances. Though the classification made here is based in part on intuition, it is supported in part by the existence of some of the text types in other domains. Further support is provided in the following discussion which suggests links between the communicative scenarios that text types are produced in, and some of their linguistic features. It is noted that a more reliable method for justifying the classification of text types would perhaps be to measure linguistic variance *within* a special language corpus; at the time of writing, the Surrey Dance Corpus does not contain sufficient examples of each text type to make such an analysis possible.

Text Type	**Size (words)**	**Examples**
About an individual dance		
Choreographer's Notes/ Choreographic Script	100's – 1000's / 10,000's	Lea Anderson's notes for making *Flesh and Blood*; in a teaching pack for the dance, from the National Resource Centre for Dance. Cyril Beaumont (1952/1982), *The Ballet Called Swan Lake*. New York: Dance Horizons, pp. 80-143.
Advertisement/Preview/ Catalogue Entry	10's – 100's	Leaflets for Derek Deane's *The Nutcracker*, London Coliseum Theatre, December 1998; and, for Lea Anderson's *Flesh and Blood*, part of the Spring Loaded festival.
Programme/Sleevenotes	100's – 1000's	Programmes for Christopher Bruce's *Cruel Garden*, Sadler's Wells, November 1998; for, William Forsythe's *Hypothetical Stream 2*, Sadler's Wells, November 1998; and, for Derek Deane's *The Nutcracker*, London Coliseum Theatre, December 1998.
Critical Review	100's	Camille Hardy reviewing Merce Cunningham's *Beach Birds*. *Dance Magazine*, July 1992, p. 59.
About many dances		
Journal article/Learned book	1000's – 10,000's / 10,000's – 100,000's	Roger Copeland (1979), 'Merce Cunningham and the Politics of Perception.' in: Cohen and Copeland (eds.), *What is Dance?* Oxford: OUP. Susan Foster (1986), *Reading Dance*. Berkeley: University of California Press.
Textbook/Teaching Pack	10,000's – 100,000's	Judith Mackrell (1997), *Reading Dance*. London: Michael Joseph.
Dictionary/Encyclopaedia	100,000's / 1,000,000's	Horst Koegler (1987), *The Concise Oxford Dictionary of Ballet*. 2nd Edition, Oxford and New York: OUP. Selma Cohen (1998), *International Encyclopaedia of Dance*. New York and Oxford: Oxford University Press.

Table 4.9: The classification of texts produced by dance scholars: note the distinction between texts which refer to one dance in particular, and those which refer to many dances.

4.2.1 Texts About an Individual Moving Image

In terms used to classify video surrogates, texts that are collateral to specific dances appear to contain what might be regarded as a mix of *bibliographic*, *structural* and *content*-related information: that is, a mix of information relating to the production of the dance, to the arrangement of sequences within it, and to its dancers and movements. Importantly, the information relating to the content of dances in these texts is not restricted to literal descriptions of dancers and their movements: some of these texts go beyond the image to expound the meanings of dances. A range of grammatical devices are used to convey this information, including short imperative sentences to instruct a dancer, and long sentences with several clauses to explicate a movement sequence. The texts can also be characterised in terms of their temporal relationship with the moving image, which ranges from a close correlation at a fine granularity, to texts which refer to dances only as wholes.

Texts to Create a Moving Image: Choreographer's Notes and Choreographic Scripts

A textual description of dance sequences, complemented by sketches and movement notations, might exist before a dance is even performed. Depending on their mode of work, choreographers may make written notes to keep track of developing ideas from which they and the dancers then work. The choreographer may use idiosyncratic shorthand codes in their, often handwritten, notes which might make the text difficult for others to comprehend, and the changing descriptions of ideas may not match the final dance. However the authoritative status of such notes means that they are an important source of information about the content of a dance especially with regard to its inspiration (1a), its intention (1b) and the genesis of its movement sequences (1c).

These examples are taken from notes written by the choreographer Lea Anderson during the making of her *Flesh and Blood*. The fact that the notes are (primarily) for personal use means that they rarely comprise well-formed sentences that could be automatically parsed. However they do contain keywords that pick out the themes of the dance, for example 'Joan of Arc' and 'silent film'. Of course there is always a danger in taking keywords out of

context, as in (1b) where the dance is not 'violent, obsessive, etc.' The third example refers to a movement sequence but it is not clear from the text whereabouts in the dance this takes place, however it does give insight into the choreographer's thought processes at work.

(1a) Joan. Mostly stands still. Based on the silent film – The Passion of Joan of Arc which deals mostly in close-ups of the faces of Joan + her tormentors.

(1b) A lot of European work is described as being violent, obsessive, etc. I don't want to describe this in the same way.

(1c) Landing people on one leg? Vibrate arms, perhaps.

Some dances have been performed over many decades or centuries, changing with every performance as they are passed down the generations. In some cases a choreographic script is made to record a 'definitive' version of a dance, as was the aim of the ballet historian Cyril Beaumont when he wrote 'The Ballet Called Swan Lake' in 1952. The core of this book is a choreographic script which describes the on-stage action of *Swan Lake* as it should be performed according to Beaumont. An excerpt of this script shows how the description of movement combines general language with specialist dance terms and abbreviations.

The order of the text follows closely the order of the dance at a fine level of detail, and the layout of the text and headings are used to break complex movement sequences into simpler ones. Here we see instructions for the first part of a *Pas de Trois* and the action leading up to it (2). The text contains sufficient information for the enactment of the dance, in a fairly well formalised style – enough that a computer program might be able to extract key descriptive features of the dance and maintain their temporal order; for example, the sequence of ballet moves 'temps levé, développé, chassé…'. Not only is there information to attribute these movement features to particular intervals of the dance, but maybe also to particular dancers.

(2) The Prince hands all the flowers to Benno, and with him crosses stage to 2, when Benno deposits the flowers on the table, while Siegfried takes a cup of wine and talks with his friends.

Pas de Trois.
Commence at 3 two village girls and man (G., M., G.), the leading girl does the first solo.

I. *Temps levé* on L.F., *développé à la 2de* with R. leg, *chassé* on to R.F. and *pas de bourrée dessous* towards R. Repeat the whole three times in all, travelling in straight line across back to 8, then Man lifts 1st Girl in *temps levé en arabesque* on R.F., 2nd Girl does same *temps levé* without lift.
Repeat the whole travelling to L. to 2 (2nd Girl is lifted on this side)

The choreographic script is presented in summary form elsewhere in Beaumont's book as an overview of acts and scenes that shows the structure of the ballet (3a), and as a list that classifies the ballet's movement elements (3b). This information could be used for labelling the dance with keywords at a coarse granularity to pick out the movements and characters associated with each scene or act.

(3a)
```
ACT ONE, SCENE ONE
Short mimed scene (Siegfried and Benno).
Entrée of Peasants.
Pas de Trois.
Mimed scene (Princess Mother, Siegfried; then Siegfried, Benno).
Humorous Mimed scene (Tutor and Peasant Girl).
Ensemble for Peasant Girls and Youths.
Short mimed scene.
```

(3b)
```
Movements: développé, retiré, rond de jambe, grand rond de jambe
Poses: arabesque, attitude
Beating steps: cabriole, entrechat, petit battement
Cutting steps: coupé
Gliding steps: chassé, glissade, pas de basque, pas de bourrée
Hopping steps: ballonné, temps levé
Raising steps: relevé
Springing steps (horizontal): échappé, grand jeté, petit jeté
Springing steps (vertical): sissonne
Turning steps: petit tour, pirouette, renversé
Whipping step: fouetté
```

As well as describing the movements that constitute *Swan Lake*, Beaumont also provides a version of its story – that is the narrative being conveyed by the ballet. Like the choreographic script this 'book of the dance' has a close temporal alignment with the dance due to its straightforward story-telling style (4). Rather than detailing the movements of individual dancers, this text tends to refer to larger events such as the 'arrival of Prince Siegfried' and the 'start of the festival'. By elaborating the narrative of the dance, this text is more interpretative than the strictly descriptive choreographic script.

(4)
```
(I) Benno and his friends await the arrival of Prince Siegfried,
preparatory to celebrating with him his coming-of-age.  Enter Prince
Siegfried accompanied by his tutor Wolfgang.  The festival begins.
Village girls and youths arrive to congratulate the Prince who offers
wine to the youths and favours of coloured ribbons to the girls.
Wolfgang, already slightly tipsy, gives effect to his pupil's commands.
Peasant dances.
```

The narrative of the dance is elaborated further in a chapter about the mimetic sequences in the dance which contribute to maintaining its storyline. An explication of the meanings that are supposed to be conveyed by a dance sequence shows that understanding these meanings

requires knowledge of the conventions of the ballet genre (5). This information allows the reader of the text to go 'beyond the image'. Note that there are cues to the explicative process, such as 'to convey this' and 'is conveyed by': such cues may be used by a machine to help a human access interpretative information.

(5) *Odette declares that she is doomed to die*. To convey this, she mimes: I – here – must die. That is, she places both finger-tips to her breast, points to the ground, then raises her arms above her head, clenching the fists, crosses her wrists, lowers her arms straight in front of her, then, just as the arms fall vertically downwards, she unclenches her hands and sharply separates them. Death is conveyed by a gesture of strength which is suddenly snapped in two.

The multi-faceted and extensive description of the Swan Lake story and the balletic movements of the dance are complemented by further textual information which advises on the settings, costumes and music for the dance; and which discusses the roles of the different characters. A book like 'The Ballet Called Swan Lake' shows the range and extent of textual information that can be associated with moving images of dance, although it should be noted that its prescriptive intent means that it does not necessarily relate to any particular instance of the dance. Furthermore, such books are available only for the few most popular dances.

Texts to Select a Moving Image: Advertisements, Previews and Catalogue Entries

A new performance of a dance is brought to the public's attention through press releases which appear in newspapers and specialist magazines as previews, and through advertisements placed in conjunction with the performance venue. These short texts are intended to catch attention by promoting famous names associated with the dance – its choreographer and lead dancers for instance, as well as providing information about the time and place of the performance and the purchasing of tickets. A synopsis of the dance may also be given in a way calculated to attract a certain audience: compare the adverts for a popular ballet (6a) – 'the world's favourite'; and, for an avant-garde dance (6b) – 'idiosyncratic and highly original'. Current information extraction techniques could identify some of the proper nouns in these texts and maybe assign them roles with respect to the production of a dance, for example attributing Tchaikovsky as the composer of The Nutcracker's score.

(6a) The Nutcracker, 15 December – 9 January 1999, London Coliseum Theatre. Evenings at 7.30pm, Matinees at 2.30pm.
With enchanting designs by Sue Blane, Tchaikovsky's gloriously familiar score … and the talents of a world class company of dancers … promises to be the theatrical event of the Christmas season. The production has everything you would expect to find in the world's favourite ballet: a magical Christmas tree that grows and grows; snowflakes that come spectacularly to life…

(6b) Flesh and Blood, 25-29 March, The Place Theatre 8pm.
First staged in 1989, Flesh and Blood draws the audience into a secret and disturbing world. Seven women, swathed in silver, appear as saints, martyrs and fallen angels. … The all-female Cholmondeleys are renowned for their idiosyncratic and highly original style.

As well as live performances, the dance viewer also selects dance recordings to watch. Video recordings of dances are catalogued, and organised on shelves, both by merchandisers and by librarians using more or less ad hoc schemes. A catalogue entry may use a system of attributes to organise information about the dance's title, choreographer, running time and genre, as well as information about the production and distribution of the video cassette. These well-structured details, that might be accompanied by a summary of the dance's content and themes, would map to a computer-based index in a relatively straightforward way.

Texts to Assist the Viewing of a Moving Image: Programme and Video Sleevenotes

Programmes available to the audience of a dance performance give information about the people involved in making the dance such as the choreographer, dancers, stage personnel and lighting crew. They also provide background information about the story and themes of the dance, and about the music and its composer. For example, a programme for a recent production of *The Nutcracker* tells the original fairy tale by Hoffman, and gives biographical details of Tchaikovsky. A programme for a ballet inspired by the life of the poet and dramatist Federico García Lorca, *Cruel Garden*, tells about his life and times, and prints some of his poetry. In some very interesting sense these texts are all 'collateral' to the dance but the links are perhaps not yet clear enough to be explored computationally.

Perhaps the most important function of a programme is to explain something about the dance to the audience (a function it shares with the sleevenotes that accompany a video recording of a dance). But what might it 'explain' to a machine? A programme usually

contains a synopsis of the dance which outlines its narrative or suggests other ways in which it might be interpreted depending on the dance genre. Compare the excerpts of synopses for three dances: a traditional ballet with a set narrative that is explained as an ordered sequence of events (7a); a less narrative, but still 'symbolic' contemporary ballet which is explained in terms of its inspirations and themes (7b); and an 'abstract' dance which is described in terms of its movement and lighting qualities (7c).

Keywords could be taken from each of these texts: in the first example they could be attached precisely to the acts and scenes of the dance. In the second example there is some temporal order in the text, and cues like 'Cruel Garden begins…' might allow keywords to be attributed to sub-intervals of the whole dance. The final example refers to the dance mainly as a whole, though it alludes to some change in the nature of the dance as it progresses.

(7a) The Nutcracker
Act 1
The Party
It is Christmas Eve and a lavish Christmas party is underway. Drosselmeyer, Clara's mysterious and eccentric Godfather, arrives with a gift for Clara. He has brought a magnificent new Nutcracker doll as her Christmas present…
The Battle
The clock strikes midnight. Giant rats and mice appear in the living room which has undergone a strange transformation. The Nutcracker doll becomes a life-size Nutcracker Soldier and a great battle ensues…

(7b) The ballet, *Cruel Garden*, is not an account of Lorca's life, though there are echoes of what he was and did; his theatre work, puppet plays and his love of music. It has its own story, drawing on images of Lorca's poems and plays to evoke the character of a man alone; a poet, a bullfighter, even a Christ figure.
Cruel Garden begins in a bullring, with the entrance of the Moon, whose role is to reveal and betray…
The last bull dance, the Poet's final death, symbolises Lorca's murder at the hands of the Fascists.

(7c) Hypothetical Stream 2 is a masterfully realised piece for 9 dancers, its grey light and ambient score creating a sense of shadowy rapture. Entangled bodies and angular shapes gradually unfold into lines of open, more joyful movement as the journey begins.

Texts Expressing Opinions of Moving Images: Critical Reviews

In the days following the performance of a dance it is reviewed by critics working for newspapers and specialist magazines. Reviews serve several purposes: they report information about the people responsible for the production of the dance; they elucidate and interpret the content of the dance for potential audiences; and, they pass judgement on the

dance – advising the reader if the dance is worth seeing. In expressing their opinions of a dance, critics will draw attention to particular aspects that support their point of view. Reviews also comment on issues pertaining to a particular performance of a dance such as the virtuosity of the dancers.

Often afforded limited column inches, reviews are condensed texts with sharp changes of topic. For example, a review of Merce Cunningham's *Beach Birds* takes just one paragraph to inform the reader of how the dance was commissioned and who composed the music, as well as commenting on its movement quality and giving an appraisal of the dance (8a). A second paragraph reports on the dancers' costumes and the images evoked by their movements (8b). Such condensed information, though richly informative to a human reader, would be difficult to deal with in a machine; it would require being able to separate the different kinds of information and attribute them to appropriate sub-intervals of the dance, or to people associated with making it.

(8a) At the opposite end of the energy spectrum, *Beach Birds*, is almost elegiac in its stillness and silences. Set alongside John Cage's Four[3], this is one of Cunningham's most exquisite works. The piece was commissioned in Zurich for last June's James Joyce and John Cage festival.

(8b) Skinner has given black gloves to the cast of eleven and a stripe that extends across the shoulders of their white unitards from fingertip to fingertip… The dancers flock and soar or stand quizzically on one leg. One trio recalls a gaggle of sand pipers. Another group evokes a pattern of gnarled sea grasses blowing in the wind…

4.2.2 Texts Which Refer to many Moving Images

As well as producing texts that pertain to particular moving images, domain experts also produce texts that make generalizations about sets of moving images. These texts articulate analytic procedures, theoretical constructs and background information that contribute to experts' analyses. Experts draw on different dances to support theoretical positions, and to compare and contrast categories of dances such as genres and choreographers' oeuvres. These texts develop and establish the domain knowledge that is explicit and is shared amongst experts. Texts written by experts for experts, such as learned journal articles and monographs, as well as texts written by experts for novices such as textbooks, and reference

works such as dictionaries and encyclopaedias, can all be seen to embody the knowledge of a domain. By articulating important concepts these texts provide a foundation for texts like those discussed previously that pertain to specific moving images.

Texts that Debate the Subject: Journal Articles and Learned Books

In expert-to-expert discourse the writer of a text will be trying to make a case for a particular point of view with regard to a theoretical issue. Their position will determine the particular aspects of a dance or set of dances that they focus on to maintain their argument. The writer will also draw on established theories and previous findings in order to support their ideas.

A learned journal article by Roger Copeland about the work of Merce Cunningham extensively discussed the works of artists and choreographers associated with the formalist and modernist movements in an attempt to situate the work of the article's main subject. Copeland used his theoretical position to draw parallels and make contrasts between the work of Cunningham and his predecessors and contemporaries. The main point of contention discussed in the article is whether or not Cunningham's work is 'escapist' because it does not seek to convey a message. Discussions about the oeuvre of a choreographer may be used to annotate a set of dances directly, or may be used to add information about a choreographer to a system, so that any dance known to be by that choreographer 'inherits' appropriate properties.

In writing a book an expert may be able to develop their ideas and arguments more extensively and cover a broader range of material. Susan Foster cites the work of Roland Barthes and Michel Foucault as the theoretical backdrop to her book 'Reading Dance' in which she compares the work of four twentieth-century American choreographers, before addressing philosophical questions about the nature of representation in relation to dance. Such a text deals with fundamental concepts that underpin the practice of dance analysis.

Texts that Introduce the Subject: Textbooks

Another book, also titled 'Reading Dance', written by Judith Mackrell is intended to introduce key concepts for the analysis of dance. Such textbooks and other teaching materials

combine a wide range of information, selected for its prime importance, and organise it for presentation to students according to conceptual and methodological principles. Mackrell's 'Reading Dance' is divided into two parts: the first part comprises three chapters, covering ballet, modern dance and post-modern dance respectively. Each chapter is organised around historical periods and key figures to review the work of major choreographers and give synopses of defining works. The second part of the book comprises 13 chapters: each deals with a different aspect of a dance that can be analysed, for example the narrative, the music and the steps. It would appear that the key concepts, which have been developed in learned discourses, are better explicated for the novice in textbooks, and are perhaps more accessible to a machine.

Texts for Reference: Dictionaries and Encyclopaedias

Key concepts are still more rigorously explicated in reference works like dictionaries and encyclopaedias; though the distinction between the two can be blurred. As well as defining movement terms and dance genres, Koegler's 'Concise Oxford Dictionary of Ballet' contains more encyclopaedic information about important people and institutions, and synopses of dances. In a dictionary this information is organised in relatively short entries that are organised alphabetically by terms and names. In an encyclopaedia the same material is covered more extensively and may be organised differently. In Cohen's 'International Encyclopaedia of Dance' entries for important choreographers cover several pages. The entry for Merce Cunningham chronicles his dances and collaborators in some detail and comments on the significant moments in his career. Other kinds of entries cover the dance traditions of individual countries, and discuss the latest theoretical debates in the field of dance analysis. Dictionaries have been targeted as important sources for automated knowledge acquisition methods and are discussed as such in the next section.

4.2.3 Summary

The different types of texts outlined here realise different viewpoints on dances, including the choreographer's intentions, the audience's appreciation and the critic's reactions. The variety of factors determining the production of a text about a dance means that text types vary in the kinds of information they convey, and in the ways in which they present this information, Table 4.10. From the point of view of developing a video annotation system, it is useful to observe how texts written about individual dances can be seen to comprise a mix of bibliographic, structural and content-related information.

Some text types comprise predominantly one kind of information, for example, adverts and previews convey bibliographic details about the people involved in the production of a dance, and about the time and place of its performance; and a choreographic script concentrates on describing the content of a dance at the level of every dancer's movement - movements are also explained in terms of a narrative or other meanings. Other text types, like critical reviews, combine and condense these kinds of information.

Text types also vary in their temporal relationship with the moving image. The order of information in some texts that describe or interpret a dance follows the moving image sequence closely. For example, a choreographic script describes the action of a dance at a fine temporal granularity and a programme synopsis of a dance explains its narrative at a coarser level. Other texts may only select parts of the dance to focus on and may rearrange their order to draw attention to a particular sequence. For example, a critical review might start by commenting on the finale of a dance and continue by discussing the dance as a whole. In such cases the interval of the dance being referred to might be indicated with textual cues.

<table>
<tr><th rowspan="2">Temporal Relationship to the Moving Image</th><th colspan="4">Type of Information Conveyed about the Moving Image</th></tr>
<tr><th>Bibliographic</th><th>Structural</th><th>Content-descriptive</th><th>Content-interpretive</th></tr>
<tr><td rowspan="2">Refers to the dance as a whole.</td><td>ADVERT/PREVIEW</td><td></td><td></td><td></td></tr>
<tr><td>PROGRAMME CAST LIST</td><td colspan="3" rowspan="2">PROGRAMME SYNOPSIS</td></tr>
<tr><td>Refers to the dance in sequence: coarse-grained.</td><td rowspan="2"></td></tr>
<tr><td>Refers to the dance in sequence: fine-grained.</td><td rowspan="2"></td><td colspan="3">CHOREOGRAPHER’S NOTES</td></tr>
<tr><td></td><td>CHOREOGRAPHIC SCRIPT</td><td>BOOK OF THE DANCE</td></tr>
</table>

Table 4.10: Types of text written about an individual dance, organised here to show how each type is informative about the moving image, i.e. what kind of information it conveys and how it is temporally related to the dance. Note the text type *critical review*, not shown here, could possibly fit across the whole table.

Not all of the text types considered here are collateral to an individual dance. Some texts refer to many dances in order to develop, or to record, concepts for the practice of dance and dance analysis. Journal articles, textbooks and reference works can be viewed as a foundation for the production of other texts in the domain. These texts make abstractions, so that rather than describing or interpreting specific sequences of moving images, they elaborate concepts that may be applied to analyse various sequences.

4.3 Defining Specialist Movement Terms

Dictionaries and glossaries produced for human use record the terminology of a domain, and in their definitions they articulate some domain knowledge. As such they, perhaps more than any other text type, explicate the conceptual framework within which experts work. Such knowledge is important for computer systems that process the language used by experts. Representations that compose the meanings of words from primitives, and representations that link words with different kinds of relationships can both facilitate inferencing in

computer-based systems. A range of methods have been proposed for generating the necessary representations from texts, in particular dictionaries: for examples and a discussion of related computational and theoretical issues, see Wilks, Slator and Guthrie (1996).

When defining specialist movement terms, a dictionary might complement written definitions with images to provide examples of movements: such visual information could perhaps be used by a computing system to assist in the recognition of movements. For dance, notation systems can also be used to record exemplar movements. Whilst images and notations can give precise examples of movements, they do not highlight their distinguishing features, nor can they convey other information associated with the movement. Thus, examples of definitions from dance dictionaries, including ones that incorporate images and notations, can indicate ways in which language can 'go beyond the image'. These definitions also suggest a high level of complexity, with subtle distinctions between a myriad of movements (Section 4.3.1). It has been argued that the syntactic structure of definitions can be exploited in order to process them into machine-executable forms: here we consider how such an approach could benefit the annotation-retrieval of dance sequences (Section 4.3.2).

4.3.1 Dance Dictionaries

Dance experts share specialist vocabularies with which they can both describe dances in terms of movements sequences and discuss theoretical issues pertaining to the interpretation and evaluation of aesthetic works: here we consider the definition of terms used to refer to dance movements. In the case of ballet vocabulary, which includes many terms borrowed and adapted from French general language, terms refer to the positions held by dancers, such as *arabesque* and *attitude*; some to general kinds of movement, *adagio* – slow, and *batterie* – with beating legs; and others to particular movement sequences, *pirouette* – turning on one leg with one foot touching a knee, and *promenade* – turning slowly with body held in a set position.

In Kersley and Sinclair's (1997) *Dictionary of Ballet Terms*, some written definitions of movements and positions are complemented by sketches, and the reader is also pointed to

exemplary sequences from well known ballets, Figure 4.1. Some sketches convey movement through the way in which dancers' clothes hang and through their apparent centre of gravity. Most of the sketches however are used to illustrate key positions; 9 for the feet, 10 for the legs, 9 for the body and 18 for the arms.

entrechat. The dancer jumps straight up, beating and changing the legs in the air beneath him … The male dancer in the Blue Bird *pas de deux* finishes his solo with six *entrechats six* and a DOUBLE TOUR EN L'AIR.

Figure 4.1: The entry for *entrechat* from Kersley and Sinclair's (1997) *Dictionary of Ballet Terms.* Note how the definition is supplemented with a sketch and a reference to a famous work.

The use of illustrations in dictionaries has been discussed by Landau who noted that though they are useful in some cases, they are limited since they can not elucidate the 'defining qualities' of a word (1989:111-115). Regarding the selection of illustrations for dictionary entries he suggested that drawings were generally to be favoured over photographs since a drawing can combine and highlight the important features of the thing being defined, whereas any particular instance of the thing in a photograph may not have all the important features. If a digital dictionary of dance were to be produced, these observations would suggest favouring handcrafted movement animations over video recordings of actual movements.

A record of movement can be produced on the printed page with a movement notation system. In the case of ballet, notation can capture subtle distinctions between numerous varieties of positions and movements. For example, Miles' (1976) *The Gail Grant Dictionary of Classical Ballet in Labanotation* records 26 kinds of *arabesque* which are grouped by

nationality (French and Russian), and by individual dance masters (Cecchetti). There are 13 kinds of *entrechat*, three of which are shown in Figure 4.2.

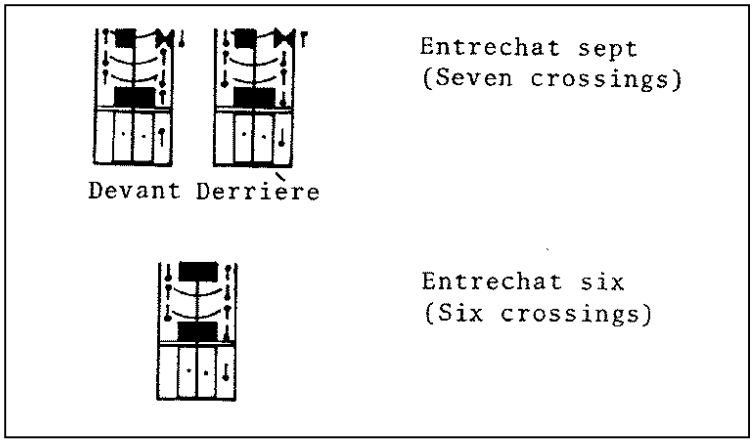


Figure 4.2: Three varieties of *entrechat* recorded in Labanotation, from Miles' (1976) *The Gail Grant Dictionary of Labanotation*.

Images and notations provide ostensive definitions of movements; that is to say, they show an example movement and say 'that is an X'. In contrast, written definitions are somewhat less precise but are able to specify characteristic features of a movement. For example, a definition of *entrechat* from Koegler's (1982) *The Concise Oxford Dictionary of Ballet* focuses on the dancer's legs and says nothing about the rest of the body (9). This written definition also supplies information about the history of the movement that would be missing from an image or movement notation.

(9) **entrechat** (Fr. caper). Designates a criss-crossing of the legs before and behind each other while the dancer is in the air. They are numbered from deux to dix according to the number of movements performed – with each crossing counted as 2 movements (one out, one in)… Camargo is said to have first executed an entrechat quatre on stage, for the better execution of which she slightly shortened her skirt.

In Koegler's dictionary specialist dance terms are explained in general language and sometimes with reference to further specialist terms: in this way complex movements are broken into simpler ones. For example, the entry for *pirouette* (10a) states a number of positions in which the movement can be performed, one of which, *attitude* (10b), is itself defined both in general language words and with reference to specialist terms that refer to its varieties. Thus, there is a richly interconnected terminology to describe an established set of dance movements and to express their interrelationships and historical background. These

specialist movements can be defined in general language such that a *pirouette* is a particular type of *turn* (or to *pirouette* is to *turn* in a particular kind of way). Doing a *pirouette* involves a combination of actions – a leg *supporting* the body and the *touching* of a knee with a foot. The *pirouette* can be executed in a number of ways, including *en attitude* – in which case one arm will be *raised* above the head and the other *extended* to the side.

(10a) **Pirouette** (Fr., spinning-top). Designates in ballet one or more turns of the body on one leg (on half or full point), with the point of the working leg generally touching the knee of the supporting leg. It can be executed à la seconde, en attitude, en arabesque, or sur le coup-de-pied.

(10b) **Attitude** (Fr.). A position of the body inspired by Giovanni da Bologna's statue of Mercury, and codified by Blasis. The body is supported on one leg with the other behind, the knee bent at an angle of 90°, turned out, with the knee higher than the foot. The corresponding arm is raised above the head, while the other arm is extended to the side. The various national schools have developed different variants. It can be executed in several ways: relevée, sautée, en tournant en avant (or en arrière), effacée, croisée, etc.

4.3.2 The Structure of Definitions

It has been argued by Barnbrook and Sinclair that definitions have discernible structures, such that there are 'repetitive patterns' in definitions (1995). It appears that when lexicographers and terminologists define things, they bring to bear their training in the enterprise of definition. This factor, perhaps along with the limited space available for each definition, leads to repeated kinds of definitions across both general language dictionaries and a range of specialist terminologies. Definitions are sometimes further constrained in that they only use words from a restricted set. Thus, ideally, the user of a dictionary is presented with clear and concise explications of words' meanings. Viewed as a text then, dictionaries would appear to be a good source from which to acquire knowledge about the relationships between terms: their simple form means the process can be automated.

Specific repetitive patterns are found in the definitions of words belonging to a particular grammatical category, e.g. nouns and verbs. Typically, when a noun is defined it is usually associated with an object, and the tendency is to organise objects in hierarchies. So, for instance, much of modern-day science is concerned with the search for genera of objects, new instances of the genera, and new differentiae between instances. The effect of the effort to

create such hierarchies can be seen in typical dictionaries, where if possible the definitions will refer to a superordinate and to the specific attributes of each word.

Barnbrook and Sinclair have suggested eight definition classes (A-H), each with an idiosyncratic structure. It was later observed (Al-Jabir 1999) that in specialist terminologies, including *law, mathematics* and *health*, more than 85% of the definitions were of type A, which uses an idiosyncratic syntactic pattern to present a superordinate and differentiae. This observation was exploited in the development of a *De*finition Analysis and *R*epresentation *S*ystem (*DEARSys*), which parsed the tagged text of definitions to give first-order logic representations, which in turn were processed into frame-based representations.

Our purpose in drawing attention to work on the 'language of definitions' in this thesis is that if time had permitted, we would have investigated how relationships between dance terms could be extracted from a dictionary, and so support the annotation-retrieval process. The dance dictionaries discussed previously treat words referring to movement as nouns, so they deal with 'a pirouette', rather than 'to pirouette'. This means that the pattern of superordinates and differentiae can be seen in these definitions too; for example, a pirouette is a 'turn' which is characterised by being performed on one leg. This piece of knowledge could be used to expand the query of a user in a video retrieval system, such that, for example, a request for pirouettes returned sequences labelled 'turns' as approximate matches, or so that a request for all movements performed on one leg returned pirouettes.

The discussion of definition structures and relationships between words ought to perhaps consider the electronic lexical database *WordNet* (Fellbaum 1998). The developers of this freely available database[4] claim that it was inspired by current theories of human lexical memory, both psycholinguistic and computational. Four major categories of English words – nouns, verbs, adjectives and adverbs – have been organised into so-called synonym sets, each of which stands for an underlying lexicalised concept.

This database, which is a useful resource for research in general language and is widely referenced, at first glance seems relevant to the current discussion of a language of dance. It has entries for some words used in dance, like *pirouette*, for which it has two senses, as a

[4] WordNet is available from http://www.cogsci.princeton.edu/~wn

noun and as a verb. The set of synonyms provided for the noun are 'spin, twirl, twist, twisting, whirl', and for the verb (which incidentally is defined as 'do a pirouette') the set is 'pivot, swivel'. WordNet will also provide hypernyms, i.e. superordinates, and at the first level or two these seem interesting, but they quickly reach levels of such generality as to be rather meaningless: for example, moving up the hierarchy from the noun 'pirouette' we have five successive hypernyms – 'spin-rotation-motion / movement / move-change-action-act / human_action / human_activity'. This extensive network is impressive in its scale, but as it spirals out, links are made which may quickly lead to entries far removed in meaning from the starting point: this may well be a feature of general language, but as such it does not suit domain specific enquiries like ours.

4.4 Discussion

Studying a collection of texts that are related to a set of specialist moving images highlights the interdependence between text and image, and hence between language and vision. Texts can take images as their subject matter, and images may be illustrated by texts. Moving images are appreciated better when their viewers have access to other people's thoughts about them; and, texts are sometimes clearer when they are illustrated.

A special language can be viewed as structuring, and thus reflecting, the knowledge of domain experts: by extending this view to domains where experts deal with moving images, the hypothesis is that special language will structure experts' knowledge about moving images, thus realising a language-vision link. As artefacts of human cognition and communication, collateral texts that are used to create, select and view moving images are grounded in the structures of experts' knowledge/language. These structures make the information about moving images conveyed by the texts accessible to human readers, and potentially to machines.

Special languages are distinguished by a profusion of specialist terms that experts introduce, discuss, modify and sometimes discard in their discourse. If specialist terminology is taken to reflect the shifting conceptual organisation of a domain, then the elicitation and

elaboration of terminology is important for maintaining accurate and reliable communication not only between humans, but between humans and machines. Candidate terms may be semi-automatically extracted from a corpus of domain texts, and the definitions of terms in specialist dictionaries can be processed to elaborate relationships between the terms. Such a resource as a terminology database could support the video annotation-retrieval process, which could also exploit texts that elucidate specific image sequences.

The preceding discussion suggests that a system for annotating a set of specialist moving images should, at least, incorporate the means to manage a terminology database so that video sequences are labelled with currently relevant terms (and names). The labelling process might be automated by selecting key terms from the various texts associated with moving images: the level of temporal granularity at which terms can be automatically attached will depend upon how explicit the temporal relationship is between text and moving image. In principle, texts could be exploited much further in video annotation systems by utilising language technologies to extract knowledge from dictionary definitions; to identify events and participant roles, along with spatio-temporal and causal relationships in collateral texts; and, to gather and classify the texts themselves. Such language engineering predicates the existence of a systematic language: in this chapter we have gone some way to establishing such a language for moving images of dance.

Chapter 5

Eliciting Verbal Reports About Moving Images

Video annotation requires the comprehension of moving images: the subsequent articulation of that understanding can then be associated with video sequences and processed into machine-executable surrogates to facilitate video retrieval. The aim in video annotation is to refer to groups of sequences in a unique manner so that they can be matched against queries (or perhaps processed in other ways). As we have seen in the previous chapter, knowledge about a set of moving images may exist in a range of texts which can perhaps provide a basis for video annotation. But, in the case of dance at least, most extant texts refer to long sequences, e.g. whole dances, so it is not possible to associate text fragments with shorter intervals in the moving image. Furthermore, the texts mix information about people, costumes and scenes with movement descriptions and interpretations of meaning, and so on. This makes them interesting for their intended readership, but makes it difficult to process these texts reliably into machine-executable forms.

For a more flexible and reliable source of video annotations, it is perhaps necessary to request specific analyses of images and image sequences directly from an expert (that is given the current limits of automatic techniques for describing objects and actions, and for interpreting visual information). So, should we get an expert to sit at a machine annotating video data? This would require the expert to have computing expertise and would take up more of their time than might be necessary. Furthermore, it would be difficult to control the kinds of information they provided. What is needed is a method for gaining, from experts, specific information about video sequences at pre-determined temporal granularities.

The challenge faced here perhaps parallels the challenge of knowledge acquisition whereby experts' knowledge about a subject is transferred onto a machine; we might say that, for specialist moving images, video annotation involves the transfer of an experts' knowledge about an image sequence. That is the practical perspective of the vision-language link in this context: from a more theoretical perspective, the controlled elicitation of verbalizations about images has been used as the basis for investigations of language production.

Research in knowledge acquisition has borrowed methods from psychology for gaining access to human expertise: one such method is *verbal reporting*. Scholars in cognitive psychology have asked subjects to 'think aloud' as they perform a cognitive task. The transcribed words of the subjects, known as *verbal reports*, are taken to reflect the cognitive structures and processes that were utilised to perform the set task: it is argued that the cognitive structures and processes can be recovered from the verbal report by following the method of *protocol analysis*. Some tasks are well-defined, like the resolution of anagrams and mental multiplication: for these cases it is possible to posit a sequence of discrete information processing steps. The study of less well-defined or less well-encapsulated tasks can share the approach of verbal report elicitation, though the reports may not provide enough evidence for detailed elaborations of how the task was performed. Nevertheless, it has been noted that the collection of verbal reports has become a standard method of research in many applied areas, including knowledge acquisition (Ericsson and Simon 1993).

The early expectation of cognitive psychologists was that verbal reports would allow the same insights into cognitive processes as was the case with more physical probes like EEG. It may be that the validity of such claims is not clear, and subsequent experimentation has yet to substantiate them. Nevertheless, the verbal report literature has been systematically organised and has undergone extensive peer review. Whether or not verbal reports can make cognitive processes fully explicit, we would like to argue that a suitably elicited verbal report will make an expert's thoughts about an image explicit enough for the purposes of video annotation, in much the same way as knowledge acquisition methods make knowledge explicit enough to be manipulated by a knowledge engineer for use in an expert system.

In our experience, details of which are discussed in this chapter, dance experts can be guided to produce verbal reports that explicate a variety of information about moving images at different levels of temporal and spatial detail. An expert may summarise intervals in video sequences lasting tens of seconds, or maybe tens of minutes: thus keywords can be associated with many hundreds or many thousands of frames in a digital video sequence. The experts might emphasise the physical features of the location of the dance, like the colour of the background, or conceptually more complex features, like describing a dance studio. They can

also list sources of other information which would be relevant when analysing a given sequence, for example, books about the choreographer and writings about the dance's genres.

In order to refer to moving images at a finer temporal granularity, experts can be guided to focus their attention on limbic movement and provide detailed descriptions of individual limbs, conventional movements and interactions between dancers. Such *description* of movement may be distinguished from the *interpretation* of moving images. When asked for the latter, experts discussed the 'imagery' in dance sequences, for example talking about religious iconography, and concomitant notions of *kneeling*, *praying* and *clasping*. The descriptions and interpretations of moving images provided by experts have the status of texts, in that the experts' accounts use typical devices of cohesion like repetition and pronominal referencing.

Based on the work of the cognitive psychologists Ericsson and Simon, we have attempted to formulate a method for the elicitation of verbal reports about moving images. The aim of this method is to collect verbal reports that make information about moving images explicit enough for subsequent video annotation. We want to be able to control the kinds of information provided by the experts in order to expedite the automatic processing of the verbal reports for video annotation. The method, comprising four related scenarios, was used to elicit verbal reports from nine dance experts (university lecturers and postgraduate students) as they watched a variety of dance sequences totalling 20 minutes (excerpts from a neoclassical ballet, a modern dance and two post-modern dances). The rest of this chapter comprises an overview of the four scenarios used in our method (Section 5.1) and a discussion of how each scenario was used to elicit dance experts' thoughts about moving images (Sections 5.2-5.4). The use of the resulting verbal reports for video annotation purposes, and the ways in which they may also give insight into how experts analyse moving images, are then discussed (Section 5.5).

5.1 Method

As an expert analyses a specialist moving image, like a dance sequence, they combine information from different sources in order to articulate new information. As well as the moving image, the sources of information include their own knowledge about their domain, complemented by texts, and possibly other moving images. As we have seen this process leads to the generation of texts which are richly informative about moving images, and which embed them in the knowledge of a domain. However, for video annotation purposes we would like to isolate different aspects of experts' analyses in order to produce texts with a consistent information content in relation to moving images. There are a variety of verbal reporting techniques discussed in the literature: here we report on the adaptation of some of these for isolating aspects of an experts' knowledge about a moving image.

The key to eliciting suitable verbal reports is in the instructions given to the expert. The instructions might ask them to speak as they are watching the moving image, resulting in what are known as *concurrent verbalizations*, or the expert might speak after viewing, giving a *retrospective verbal report*. It is suggested that concurrent verbalizations can reflect the rapidly changing thought content dynamically – and so, for us, can be considered to follow the contents of moving images. Retrospective reports give the subject more chance to compose their thoughts, but also more chance to forget important information. Cognitive psychologists discuss reasons for using certain kinds of verbal reports rather than others, in particular they are concerned about the effects of instructions on the completeness and validity of verbal reports as data. However, we are encouraged by Ericsson and Simon's comment that "concurrent verbalizations and retrospective reports will produce verbalizations of at least a subset of the thoughts heeded while completing a task" (ibid:xxxv).

Instructions may also specify that a subject is to give either a description or an interpretation of what they see. Although this distinction is motivated for us by our previous discussion of aesthetic frameworks for analysing images, it can perhaps be related to what Ericsson and Simon have called *context-free verbalization* (including the description of images) and the *verbalization of complex thoughts* (which we suggest could include the

interpretation of images). In the first case, the information to be verbalized is immediately available (e.g. in Short Term Memory or as sensory stimuli) and so the question is how to select the words with which to speak it: issues here include the selection of lexical items and the selection of syntactical forms. This assumes a fairly straightforward mapping between thoughts and propositions that "denote one or more attributes of an object or conceptual entity, or denote relations between entities" (ibid.:228). The verbalization of complex thoughts requires the selection and combination of different information sources which may include previously known information and information that has just been generated.

In their discussions of concurrent and retrospective verbalizations, and of context-free verbalizations and the verbalization of complex thoughts, Ericsson and Simon make a link with research in language production, which has used still and moving images as 'controls' for eliciting language samples (ibid.:221-259). An important point made in their discussion is that the description (and hence, interpretation) of an image differs from the direct verbalization of thoughts, in that the subject is more careful to select and organise information for the intended recipient of the description; Ericsson and Simon make this distinction in terms of 'inner speech' (direct verbalization) and 'outer speech' (for a recipient). Nevertheless, they argue that both kinds of verbalizations correspond with the speaker's thoughts. The link with language production research is useful because scholars in that area have observed how pauses and intonation information which delimit units of articulation, may also delimit significant cognitive processes. There is also an overlapping interest between verbal reporting techniques and language production with regard to the subject's selection of linguistic devices, e.g. lexical items and syntactic structures.

5.1.1 Four Verbal Reporting Scenarios for Moving Images

Four related verbal reporting scenarios were used to elicit experts' thoughts about a set of specialist moving images. These scenarios varied in how the experts were asked to make their verbalizations. As mentioned above, whether an expert speaks as they are watching the moving image, or after, determines whether they produce a concurrent, or a retrospective, verbalization. Furthermore, if they are asked to describe what they see, then they may give a context-free verbalization but if they are asked to interpret, then this may involve the verbalization of complex thoughts. In the first scenario the expert was free to talk about whatever aspects of the moving image they thought appropriate, and whenever they thought appropriate: in this sense it was akin to an *informal* interview, to use a term from the knowledge acquisition literature. The three remaining scenarios were more *formal* in that they required the experts to follow precise instructions.

In the first scenario (elaborated in Section 5.2) an expert was asked to break a dance into important sequences and to pick out the characteristic features of the sequences: they were also asked to discuss other information sources that they would consult when analysing this dance. The expert was free to stop, start, rewind and fast-forward the video, and to comment on aspects of their analysis in any order. This meant that a high degree of involvement was required of the investigator, firstly to makes notes to be validated later by the expert, and secondly to make use of the 'verbal report' in a prototype video retrieval system.

To get finer-grained analyses of dance sequences with more homogeneous information about the moving image, the next scenario required experts to speak as they watched continuously playing video sequences (Section 5.3). The experts were first asked to describe the sequences, and then to interpret them. The resulting verbal reports, which were recorded, transcribed and time-coded, had a strong temporal relationship with the moving image, and referred to intervals as short as single seconds. In particular, the descriptions picked out the dancers' movement and positions, and related them to one another, whilst the interpretations tended to refer to longer sequences with more abstract statements.

The descriptions and interpretations were analysed to see how the experts selected and organised information about the moving images: systematic differences were apparent between the descriptions and the interpretations in the choice of lexical items, syntactic structures and cohesive devices. These results assisted in the later use of the verbal reports for video annotation. Given the possibility of using speech recognition for automating the transcription of verbal reports, and the fact that they are inherently related to the image sequences in time, then this scenario seems to be the most promising for video annotation: it is this scenario that we concentrate on, and which was the basis for our final video retrieval system.

The time constraint imposed on the experts as they spoke while watching continuously playing video sequences may have encouraged them to focus on the most important aspects. However, they may not have had time to mention all the relevant details; nor did they have the opportunity to explain their descriptions and interpretations. Furthermore, the scenario was unnatural in that it required the experts to analyse complex visual information having, in some cases, seen it only once or twice. Perhaps then it is important to examine the homogeneous descriptions and interpretations produced when an expert can watch a sequence an unlimited number of times. It may also be interesting to compare the spoken verbalizations with their written equivalents. For the third and fourth scenarios experts were asked to speak, and to write, respectively, after watching image sequences (Section 5.4).

In the third scenario an expert was asked to spend as long as they could to elaborate the transcripts of a previously spoken description and of an interpretation. They were asked to add details to, as well as to correct and to explain, their original verbalizations. The elaborated description referred to dancers' movements at a still finer detail, and the expert also provided partial definitions for some movement terms. In elaborating their interpretation, the expert discussed the theoretical position taken, and highlighted aspects of the movement which supported their analysis. These elaborations combine the advantage of a retrospective report, i.e. that the expert can give in-depth explanations, with the advantage of the concurrent reports on which they are based, i.e. an inherent temporal relationship with the moving image.

When experts, in the fourth scenario, were asked to write an interpretation of a sequence immediately after viewing it, they each adopted different strategies for selecting and presenting information; though all of them appeared to benefit from the opportunity to 'compose their thoughts'. Two discussed the dance as a whole, another relayed it as a narrative: all discussed its meanings, and one went further and explicated the devices used to convey the meanings. Compared with their spoken equivalents, the written verbal reports were better organised and more concise - at least to the human reader, and compared with extant written texts, they were more clearly focused on one aspect of an analysis.

The instructions given to the experts in each of the four scenarios serve to elicit different kinds of verbal report, and hence may make explicit different aspects of an expert's knowledge about a moving image at predetermined temporal granularities. Whilst the informal character of the first scenario makes it difficult to classify, the other scenarios can be more clearly seen as leading to either concurrent or retrospective verbalizations, and as either involving context-free verbalization, or the verbalization of complex thoughts, Table 5.1. The characteristics of the different verbal reports mean that varying levels of human involvement will potentially be required in order to 'animate' them in a video retrieval system. Concurrent reports have an inherent temporal relationship with the moving image, so it will be perhaps easier to associate text fragments with specific video sequences. With regards to the kinds of knowledge they elicit, context-free reports will perhaps be simpler to deal with than ones that involve the verbalization of complex thoughts.

Scenario	Instruction	Concurrent / Retrospective	Context-free / Complex thought
1	Comment on important sequences (informal).	Both	Both
2	Speak while watching moving images: either 'describe' or 'interpret' (formal).	Concurrent	'Describe' – context-free 'Interpret' – complex thoughts
3	Elaborate descriptions and interpretations (formal).	Retrospective (but based on a concurrent report)	'Describe' – context-free 'Interpret' – complex thoughts
4	Write an interpretation (formal).	Retrospective	Complex thoughts

Table 5.1: The four verbal reporting scenarios that are presented in this thesis for eliciting different aspects of an expert's knowledge about a moving image. The characteristics of the resulting verbal reports will determine to some extent their potential as a basis for video annotation.

5.2.2 Details of Dance Sequences and Experts

The co-ordinating expert, Prof. Janet Adshead-Lansdale, advised on the selection of the dance sequences used in this investigation. The selection was made to include a variety of dance genres: excerpts comprising movement sequences of medium complexity (typically duets) and ranging in length between approximately three to five minutes, were taken from four dances, Table 5.2a. (Note that in the first verbal reporting scenario, the expert watched the whole of one of these dances).

Dance	***Beach Birds for Camera*** ***(NB. Two sequences)***	***Perfect Moment***	***Steptext***	***Swan Lake***
Choreographer	Merce Cunningham	Lea Anderson	William Forsythe	Matthew Bourne
Genre	Modern Dance	Post-modern Dance	Neoclassical Ballet	Post-modern treatment of Classical Ballet
Music	John Cage	Steve Blake	Johann Sebastian Bach	Piotr Ilyich Tchaikovsky
Date	1993	1992	1984	1995
Length of sequence	Duet: 2m 58s Ensemble: 4m 50s	Duet (4 pairs): 3m 37s	Duet: 2m 44s	Duet: 5m 34s

Table 5.2a: The selection of five dance excerpts provided by the co-ordinating expert to give a range of dance genres and sequences of medium complexity. Note that the dances from which these excerpts are taken range in length from tens of minutes to hours.

The co-ordinating expert also gave the investigator the opportunity to talk to groups of lecturers and postgraduate students in the Department of Dance Studies at the University of Surrey. Two lecturers, three PhD students and four MA students then volunteered to act as experts in this study, Table 5.2b.

Expert Code	**Level of Dance Expertise**	**Native Language**
α	University lecturer	English
β	University lecturer	English
γ	MA student	Dutch
δ	MA student	English
ε	MA student and choreographer	English
ζ	MA student	English
η	PhD student and performer / choreographer	English
θ	PhD student and college lecturer	Finnish
ι	PhD student, dancer, lecturer and dance critic	English

Table 5.2b: The nine dance experts who gave verbal reports: note that all the students had taken postgraduate courses in Dance Analysis at the University of Surrey. All but two experts were native English speakers.

5.2 Elaborating Important Sequences within a Moving Image

For the first, informal, verbal reporting scenario, five sessions were conducted with a university lecturer (expert α) over a two month period, lasting a total of eight hours. (Some time in these sessions was spent discussing the design and implementation of a prototype video retrieval system). During these sessions the expert and the investigator watched one dance a number of times (Merce Cunningham's *Beach Birds for Camera*) and the expert spoke about a preliminary analysis of it. In particular, the expert talked about ways in which the video could be broken into important sequences and how they could be distinguished; the expert also spoke about the information sources they would use to assist their analysis. Some viewings ran through the whole dance without stopping, other viewings repeated short sequences several times. Notes taken during these sessions were written up by the investigator and later validated by the expert.

The expert, who was already familiar with the 25-minute long dance, provided two structural overviews of it. The first was in terms of its three main sections which were delimited according to the number of dancers on screen; the expert also noted changes in the background of the dance, Table 5.3a. In the second overview sections are delimited with reference to camera shots. For each shot the expert picked out characteristic features of the dancers' movements, Table 5.3b.

Section Time	Summary
0:00 – 11:58	Up to 12 dancers in what looks like a dance studio.
11:57 – 14:55	A man and a woman in a duet against a blue background.
14:56 – 25:24	Back to the group of dancers, now against the blue background.

Table 5.3a: A coarse overview of a 25 minute long dance with sections delimited according to the number of dancers on screen. The expert has also noted the changing background of the dance.

Shot Time	Camera Action	Summary
0:00-0:05	Still	Arms and torsos only; rocking slightly. Crossed-arms motif.
0:06-0:16	Slow merging transition to still	New shot is more torso and more torsos. Crossed-arms motif.
0:17-0:29	Quicker merge to still	Shot comprises images of arms crossing over – barely any torso visible.
0:30-1:49	Merge to wide shot	Group shot. 11 dancers on screen standing with knees together and slightly bent, arms spread wing-like. All still then gradual spreading of slight movements: hands twitch, arms flap, torsos rotate, then bending legs.
1:50-2:03	Camera travelling left to right, then backwards	The camera movement centres two female dancers, who take the first steps in the dance. There is a contrast in tempo between slow and quick.

Table 5.3b: A finer-grained overview of the first two minutes of the same dance, here delimited by camera shots. For each shot the expert has noted camera actions and picked out characteristic features of the dancers' movements.

The expert went on to identify and elaborate four *motifs* in the dance: a motif is a gesture, or a sequence of movements, that is prominent in a particular dance due to its innovation, its frequent repetition, or its coincidence with key moments in the musical score. In *Beach Birds for Camera*, our expert noted for example a 'shaking leg' action that was performed by several dancers at five points between 8:00-12:00 and that returned towards the end of the dance. Another motif identified by the expert was an *attitude penché* position in which the dancer's arms are spread like wings - this occurs in a variety of forms throughout the dance, Figure 5.1.

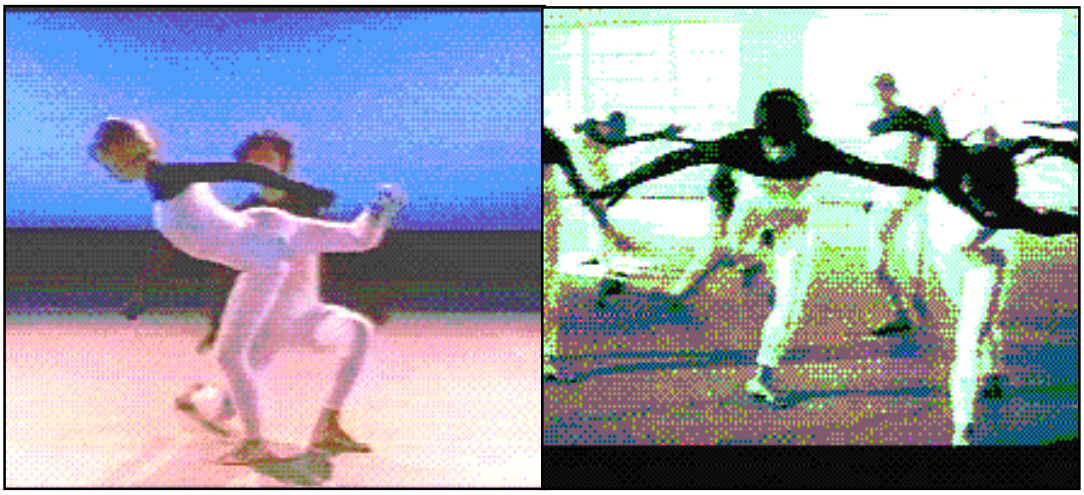


Figure 5.1: Two instances of the attitude penché motif identified by the expert watching Merce Cunningham's *Beach Birds for Camera*. The position involves the dancer standing on one leg with the other stretched behind – the motif is characterised by the dancer's arms being spread like wings.

Automatic techniques can segment digital video by measuring changes in image features such as colour and motion which are associated with certain kinds of editing techniques and camera actions. However, such techniques would not necessarily be able to recover other important sequences, such as the sections of a dance and its motifs as they are identified by an expert. And, whatever the means of delimiting sequences, an expert is currently required to label them with meaningful characteristic features.

With regards to the related information sources that should be referred to when analysing the dance, the expert listed texts including critical reviews of the dance, journal articles about the choreographer, the choreographer's own writings, and a book about the choreographer's ideas on dance - the 'Cunningham Technique'. Relevant non-textual information that was discussed in the interviews included the musical score, an earlier stage version of the dance, and computer animations of key movement sequences. It seems important then that a system for accessing specialist moving images like dance, provides the user not only with digital video data, but also with links to a variety of other information sources; as is the way in *hypermedia* systems. Such a system would benefit from organising the labelling of video sequences, and related information, at a range of granularities.

5.3 Speaking about Moving Images in Real-time

The second verbal reporting scenario aimed to elicit verbalizations that were temporally-aligned with the moving image. One lecturer and four postgraduate students (experts $\beta,\gamma,\delta,\varepsilon,\zeta$) were recorded speaking as they twice watched a video compilation of all five dance sequences lasting about 20 minutes – the sequences were separated with 45 seconds of blank screen followed by a message giving 10 seconds warning of the next dance. Before the first recording each expert was read an instruction to 'Describe' the dances, speaking as they watched. For the second recording the instruction was to 'Interpret'. These instructions were elaborated with definitions, provided by the co-ordinating expert, which the subjects read before recording began:

- Describe – ‘by describe we mean, focus particularly on the detail of the movement, its use of space and its dynamic emphasis’;
- Interpret – ‘by interpret we mean, outline one or many kinds of significance you might attribute to the interaction in this section’.

The experts’ verbalizations were recorded onto one of the sound tracks of the video cassette that they were watching in order to maintain the temporal relationship between the moving image and the spoken word.[5] At the end of each recording, details about the subject’s dance background and education were gathered, and the subject was asked to note any of the dances they had seen before.

The recorded speech was transcribed orthographically (prosodic information was not transcribed); speech fragments were delimited on the basis of pauses and the perceived completeness of each fragment; contracted forms were expanded, e.g. *there’s -> there is*; and, false starts and unfinished utterances were prefixed with ‘~’. The high quality of the recordings meant that more than 99% of the utterances could be transcribed. Guidelines for the transcription of spoken linguistic data[6] estimate that an orthographic transcription of spontaneous speech takes 10 times the length of the speech: this was found to be the case in this instance. During checking, each speech fragment was given a time-code to record the time from the start of the dance sequence that the fragment was spoken.

The five experts had similar dance backgrounds; all had taken postgraduate courses in Dance Analysis at Surrey and one was then a lecturer; all had seen at least two of the dances before, and in two cases had studied them extensively. The transcripts were analysed to compare and contrast the information heeded and the strategies, vocabulary and linguistic structures used when experts describe and interpret moving images; Surrey’s text analysis package *System Quirk* was used to assist this part of the investigation.

The 25 descriptions (five experts each describing five dance sequences) totalled 11,300 words in 1600 speech fragments, from approximately 100 minutes of speech: the 25

[5] This was possible using the ‘Audio Dub’ facility on a conventional video cassette recorder.
[6] The Expert Advisory Groups on Language Engineering Standards (EAGLES) produced a handbook for spoken language systems which is available on the WWW: for guidelines on the transcription of speech see http://coral.lili.uni-bielefeld.de/EAGLES/eagbook/node151.html#4918

interpretations totalled 9754 words in 987 speech fragments, Table 5.4. There was no difference in average word length between the descriptions and interpretations but the speech fragments in the interpretations were 25% longer. For the 'Describe' task the most fluent speaker averaged a rate of about 150 words per minute, the least fluent about 75: there was no significant variation in the rate of speech for descriptions of different dances. For the 'Interpret' task the most fluent speaker spoke at about 140 words per minute, the least fluent about 50: again there was no significant variation between different dances.

	Descriptions	**Interpretations**
Total words	11,300	9754
Total speech fragments	1,600	987
Average word length (letters)	5.63	5.45
Average fragment length (words)	8.04	9.96
Most fluent speaker (words per minute)	150	140
Least fluent speaker (words per minute)	75	50

Table 5.4: Statistics pertaining to the *descriptions* and the *interpretations* spoken by five dance experts whilst watching a 20 minute compilation of dance sequences.

5.3.1 Strategies for describing and interpreting

A manual analysis of the descriptions and interpretations highlighted contrasting strategies used by the experts when performing the two tasks. When describing the moving images the experts were able to rapidly identify movements: at times they made single word utterances to name movements in quick succession (1).

(1)
```
0:47  hugging
0:47  releasing
0:49  wide open stance
0:52  rolling
0:53  pushing
0:54  turning
```

At other times the experts used fuller utterances in their descriptions to give temporal and spatial details associated with the movement; to relate the roles of the dancers; and, to comment on the movement's quality (2).

(2)
```
0:40  she does some adage steps
0:43  he is balancing
0:45  she is now moving behind him in a locomotion pattern
0:51  he leans and balances
0:53  she counter-balances with him
```

When describing, the experts usually kept up with the dance and spoke only about events that they had just seen. (It should be noted that the time-codes refer to the time of the utterances which will be some time later than the time of the movements they describe – this discrepancy presents a serious challenge when using these texts for video annotation). In contrast, when they were interpreting, the experts sometimes talked in abstract terms about longer sequences or about the dance as a whole (3a-c).

(3a) `it is in the ballet genre`
(3b) `it is also the formal movement relationships which are important`
(3c) `religious iconography of people looking after the sick or caring for the dead or the ill`

One indication of the more general statements made by the experts when interpreting is the relative frequency of the phrase *there is*: in the descriptions it occurs once per 283 words; in the interpretations it is once per 91 words. It seems to be that the phrase was often used in the interpretations to introduce comments about the abstract qualities of dance (4a-4b).

(4a) `there is a general sense of exploration`
(4b) `there is a certain amount of tension between them`

Two descriptions and a corresponding interpretation of the same image sequence are shown in Table 5.5. When describing the same visual events, the experts' verbalizations differ in several respects. Expert β chooses to detail where the man is walking - 'across the stage', whilst expert γ says how - 'at a very slow pace'; expert β describes outstretched arms as a process - 'extending their arms outwards', expert γ describes them as a state - 'locked behind their backs'. The interpretation of the same sequence by expert δ says less about individual movements and positions, and more about the mood of the characters – 'a sense of loneliness', 'looking longingly'; and, about their relationship – the man is 'separated from the rest of the characters'.

Description by Expert β	Description by Expert γ	Interpretation by Expert δ
[0:05] a single man walks across the stage area [0:10] his back is to the audience [0:11] he hugs himself [0:13] he is surrounded by a group of dancers who are bent over from the waist [0:18] extending their arms outwards and behind their backs into a cross shape [0:25] they are looking upwards [0:27] the central character who is a male who has walked across the stage wanders towards the audience looking around [0:36] meanwhile the male group of dancers of about twelve are continuing to spread their arms and they are running around, arms undulating	[0:00] we see a big scene full of blue figures [0:05] a man is walking at a very slow pace [0:08] see a lot of backs of people [0:10] their arms are locked behind their backs [0:16] they are actually higher than their backs [0:17] and they gradually move their torsos up and they are standing [0:22] they are all men with the ~left ~foot, left leg bent [0:26] we see the man who was walking turning round and his face is looking upwards in a sort of romantic pose [0:36] we see the arms of the men who wear the white pants	[0:07] the male character seems to be out looking at the moon, searching for something, he has a sense of loneliness and isolation about him [0:14] the swan characters in the chorus are very earthy [0:17] seem very still and calm in comparison to the man [0:19] he is looking longingly at the moon [0:25] he is unhappy, distressed, a bit soulful about something [0:32] his mood is accentuated by the swans – we see them at one with their environment [0:35] whereas he somehow seems dislocated [0:38] being separated from the rest of the characters

Table 5.5: Two descriptions and a corresponding interpretation of the same dance sequence (from Matthew Bourne's *Swan Lake*). Note that the two descriptions pick out different aspects of the same sequence and sometimes refer to the same movement or position in different ways. In contrast, the interpretation deals with more abstract features of the sequence.

It is clear from a manual analysis that the verbal reports elicited with the instructions to 'describe', and then to 'interpret', are richly informative about image sequences, at least to a human reader: but what of their use for video annotation? For video annotation purposes it would be necessary to (automatically) identify keywords and use them to label intervals within moving images. Further processing might pick out relationships between entities and different kinds of actions, or segment video sequences by noting breaks in the cohesion of a collateral text. The rest of this section reports a linguistic analysis of the verbal reports to show how experts selected and organised information as they spoke: this analysis points to systematic differences between descriptions and interpretations.

5.3.2 Vocabularies for describing and interpreting

The differences between descriptions which concentrate on the positions and movements of dancers, and interpretations which make more abstract statements about dance sequences, are reflected in the words frequently used in the experts' commentaries. These words include both specialist movement terms and general language words to describe movement, as well as more abstract words used for interpretation.

Table 5.6 shows the 100 most frequent open-class words used in the descriptions and interpretations respectively: 51 of these words appear in both lists. The words are grouped according to whether they refer to the dancers and their surroundings; to their movements and positions; or to movement qualities. Of the most frequent words in the descriptions, 88% refer to dancers, their surroundings and their movements and positions; for the interpretations the figure is only 68%.

The descriptions of dance sequences frequently use both general language words, which may have restricted meanings in relation to dance, like *turns*, *jumps*, *spins* and *leaps*, and specialist ballet terms like *arabesque* and *plié*. General language words and specialist terms for referring to movements also feature in the interpretations of the same sequences but here there are also words whose use is perhaps less literal when referring to entities – *prince*, *wings* and *moon*; to movements – *flight* and *manipulating*; and to movement qualities – *caring* and *underwater*. Other words appearing frequently in the interpretations reflect the more theoretical nature of the task, for example *significance* and *references*.

The frequency of words referring to body parts perhaps says something about how the experts attend to movement – it might also say something about the genres of the dances being described. In the descriptions, body parts appeared in the following order of descending frequency (merging singular and plural forms): *arm, leg, hand, head, foot, body, back, shoulder, chest, neck, torso, waist, face, elbow, hair, knee, palm, spine.*

Words that refer to …	In Top 100 Words of Descriptions and Interpretations	In Top 100 Words of Descriptions Only	In Top 100 Words of Interpretations Only
Dancers, bodies and surroundings	arms, background, bodies, body, character, dancer, dancers, feet, female, floor, hand, hands, head, leg, male, man, men, people, person, space, stage, swan, swans, woman, women (***25 words***)	arm, centre, chest, couple, couples, figure, foot, heads, legs, neck, partner, shoulder, weight (***13 words***)	birds, characters, earth, moon, prince, wings (***6 words***)
Movements and positions	action, balancing, circle, duet, gesture, gestures, holding, lifting, looking, lying, movement, movements, moves, moving, position, shaking, standing, supporting, touching, turning, turns, walks (***22 words***)	arabesque, attitude, balance, bent, circle, contact, facing, following, goes, going, holds, jumping, jumps, kneeling, leaps, lifted, motif, move, plié, point, positions, sitting, spins, steps, stroking, turn, twisting, walking (***28 words***)	coming, curve, curved, dance, dancing, falling, flight, leaving, manipulating, pushing, shape, showing, support (***13 words***)
Movement qualities	big, little, slowly, unison (***4 words***)	extended, flexed, gestural, lyrical, quick, slow, small, sustained (***8 words***)	caring, formal, great, independently, jerky, physical, sensual, stillness, strange, strength, strong, underwater, undulating (***13 words***)
Other			control, dead, element, enjoying, fantasy, idea, images, important, life, presence, references, relationship, ritual, sense, significance, suggesting, type (***17 words***)
TOTAL	**51**	**49**	**49**

Table 5.6: The 100 most frequent open-class words used in the 'Describe' and the 'Interpret' tasks. Note that 51 words fall in the intersection of the descriptions and the interpretations.

Recall that a statistic which divides the relative frequency of a word in a set of specialist texts with its relative frequency in a general language sample gives a word list in which words peculiar to that set of texts rise to the top: this has been termed a 'weirdness coefficient'. Among the 127 words most peculiar to the descriptions, i.e. words occurring 100 times relatively more often than in the general language sample, were: 14 specialist dance terms, e.g. *adagio, arabesque, attitude*; 38 general language words referring to quite particular movements e.g. *arching, balancing, clasping*; and, 18 general language words referring to movement quality e.g. *animalistic, dynamic, flexed*. In the mid-ranges of the list were more general movement words e.g. *bend, come, hold, roll, kneel, walk*, and words that locate movements in space e.g. *forward, left, right, across*, and time e.g. *continuing, occasional, sporadic, while.*

The most peculiar words in the interpretations included fewer for referring to movements and spatial and temporal details, but did include 18 words for movement qualities e.g. *animalistic*, *sculptural* and *dreamlike*; as well as 11 abstract nouns e.g. *ecstasy*, *ethereality*, *recalcitrance*. These lists reinforce the vocabulary differences noted in the absolute frequency lists of words used to describe and to interpret moving images of dance. The descriptions are full of words that refer literally to movements and to locate them in space and time, whilst the interpretations include more words for referring to movements non-literally and for making more abstract statements about longer dance sequences.

It is also interesting to note the co-occurrence, or collocation, of certain words in the experts' commentaries. When describing, the experts used certain words as nucleates to improvise collocations, as well as using established collocations, for referring to poses and movements. Distinct collocation patterns are apparent for the words *position*, *gesture* and *action* as they appear in the descriptions: for each word a particular type of preceding word is seen to collocate in about half of its occurrences, Table 5.7. The collocations of *position* are established terms used in classical ballet, e.g. *first position*, *second position*, etc.; the collocations of *gesture* tend to refer to body parts and are thus constrained; the collocations of *action* refer to mimetic movements and are thus potentially the most productive. None of these collocation patterns appear in the interpretations.

	Descriptions		
Nucleate	**Frequency**	**Typical Preceding Word (%)**	**Examples**
position	68	*first-fifth* (44%)	`first position…fifth position`
gesture	19	[*BODY_PART*] (53%)	`leg gesture, head gesture`
action	18	[*MIMETIC*] (56%)	`pendulum action, sawing action`

Table 5.7: Collocations of *position, gesture* and *action* in the descriptions; such collocations were not found in the interpretations. The frequency given is the number of occurrences of the nucleate. Note that a typical preceding word occurs in about half the instances of the three nucleates.

For video retrieval purposes, keywords (and collocations) need to be associated with particular intervals within a moving image. The keywords spoken in the experts' concurrent descriptions and interpretations have a temporal relationship with the moving image: thus, the time at which a keyword is spoken can be a guide to the interval that it should label. However, two factors lead to complications here. Firstly, the keyword will be spoken some

time after the expert has seen whatever it is that the word refers to. Secondly, some keywords, like *jump*, will refer to a very short interval, whilst others, say *dreamlike*, may refer to an entire dance. There is also a danger in taking keywords out of context, as for example if an expert said 'this part of the dance is not dreamlike as it was before': perhaps then more reliable keywords will come from the 'context-free' descriptions, rather than the interpretations which also refer to temporal intervals less precisely.

Whilst acknowledging these issues, it would perhaps be reasonable for a video annotation system to identify keywords and use them to label an arbitrary interval either side of the time they were spoken. This heuristic approach would then be supported by expert validation and editing. Improved performance would perhaps be gained if the system had access to lexical knowledge about the keywords, such as whether verbs referred to short finite actions, or long continuous ones.

5.3.3 Describing and interpreting at the clausal level

It was at the clausal level in their descriptions that experts attributed movements to dancers and located them in space and time, and it was often by conjoining clauses that they articulated their interpretations of individual movement sequences.

Observation of the descriptions suggested that clauses bore mainly three types of information. This categorisation was validated by the co-ordinating expert: *Static-Pose* clauses describe dancers' locations on stage, their held positions and their gazes; *Movement-Gesture* clauses describe a spatial reconfiguration of body parts in relation to one another; and, *Movement-Action* clauses describe a spatial relocation of the whole body along spatial pathways. Understanding how such information about moving images is selected and organised in clauses will perhaps be useful for developing systems that extract this information from texts into machine-executable forms.

A manual analysis of one description (80 speech fragments describing 333 seconds of dance) showed an even distribution of information about positions, gestures and actions, Table 5.8: note that a speech fragment may contain more than one clause. These clauses

sometimes refer to one dancer and sometimes to more than one when the dancers might be dancing in unison, or might be taking different roles as in *supporting* and *lifting*.

Information Content of Clause	Freq.	Examples
Static-Pose	27	there is a group of four of them in the background .. .. they are looking upwards .. ..in an arabesque..
Movement-Gesture	27	he hugs himself .. .. his arms are undulating ..
Movement-Action	26	.. a single man walks across the stage area .. .. they circle around each other ..

Table 5.8: The classification of clauses by information content shows an even spread between poses, gestures and actions in one description of 333 seconds of dance.

There were 12 clauses which did not fit this scheme because they referred to aspects of the dances apart from the dancers' movements. Although the task instructions had not requested it, some of the experts commented on theatrical features such as the stage set and costumes, cinematic features such as shots and camera actions, and the music accompanying the dances.

Like the descriptions, the interpretations included clauses which described poses and movements but in this case they were sometimes conjoined with second phrases and clauses to comment on the meaning of the pose or movement. Words and phrases used in the formation of these interpretative statements included *seem*, *sense of*, *suggest*, *as if*, *like* and *appears to be*, Table 5.9: these words and phrases may be useful as cues in a video annotation system.

Linking word/ phrase	Frequency in all interpretations	Example
seem*	46	gestural sequence by the two men seems quite aggressive
as if	40	moving faster as if something is driving him
like	33	the stretching of the neck, like a swan
sense (of)	19	holding the wrists, a sense of being bound
suggest*	17	aerial steps, which could also suggest flight
appear to be	3	the constriction also appears to be a support

Table 5.9: Words and phrases used to make interpretations by linking clauses referring to movements with other phrases or clauses. The frequencies refer to the occurrences of these words and phrases in 9,754 words of spoken dance interpretation.

5.3.4 How the cohesion of verbal reports can reflect the moving image

The nature of the 'Describe' task meant that the resultant commentaries followed the dance closely and it is possible to argue that a description indicates a dance's focus on particular dancers and clusters of related movements. In the experts' interpretations clusters of related words reflected a chain of argument rather than the dancers and their movements. The phenomena noted here resemble two kinds of cohesion defined by Halliday (1994:309-310): *cohesion by reference*, and *lexical cohesion*. The notion of cohesion is interesting here because it can be exploited for automatic text segmentation, which in the case of texts aligned in time with video sequences, entails video segmentation.

In an example passage of text from a description that exemplifies cohesion by reference there are references to four nominal entities – two single dancers and two groups of dancers, Table 5.10. The first mention of each entity uses an indefinite article, *a* or *another*; subsequent mentions in close temporal proximity use pronouns or ellipse the nominal, whereas subsequent mentions after mentions of other entities use a noun phrase with a definite article, *the*. Here the switch of focus in the dance to a new dancer or group of dancers is mirrored in the text by the use of a noun phrase with an article; if it is an indefinite article then the dancer or group of dancers has not been mentioned before.

Some passages of text are marked by the repetition of semantically-related words, exemplifying lexical cohesion. In an example found in the descriptions, the semantically-related words refer to movements and thus may delimit a dance sequence characterised by a certain kind of movement, e.g. 'aerial steps' (5a). In the interpretations there is a less direct link between the text and the moving image so that rather than reflecting sections of the dance, clusters of semantically-related words arise when the expert elaborates a point of discussion over several speech fragments. In this case the word clusters point to the themes of the expert's interpretation (5b).

(5a)
```
[2:44]  there are a lot of leaps
[2:45]  hops
[2:48]  jumps
[2:48]  with legs usually extended
[2:51]  there are different qualities of aerial steps
```

(5b) [0:17] and it is also a certain amount of **religious** imagery that you see in some medieval and renaissance painting of the ***sick*** and the needy and people caring
[0:23] the idea of touching, clasping hands
[0:25] images of **praying**
[0:29] also people lying down, there is images of the ***dead***
[0:31] and also certain **religious** iconography of people looking after the ***sick*** or caring for the ***dead*** or the ***ill***

Time	Dancer 1	Dancer 2	Group of 12	Group of 4
0:05	a single man			
0:10	his back			
0:11	he			
0:13	he		a group of dancers	
0:18			-	
0:22			their hands	
0:25			they	
0:27	the central character			
0:36			the group of dancers of about 12	
0:55			they	
0:58				a group of 4 of them
1:00	the central character			
1:05		another character		
1:07		who		
1:14		-		
1:15		-		
1:16		his arms		
1:18		he		
1:20		he		
1:22	the central character			

Table 5.10: References to four entities in a description which shows cohesion by reference; note that '-' stands for ellipsis, i.e. no explicit mention. All the first mentions use indefinite articles; subsequent mentions with no other entities intervening use pronouns and ellipted references; and subsequent mentions after an intervention use a definite article.

5.4 Speaking and Writing after Watching Moving Images

Because they appear to be the most immediately promising for video annotation, our main interest is in the concurrent verbal reports discussed in the previous section. However, it is interesting to consider how they may be complemented by other kinds of verbalizations. In this section we discuss a third verbal reporting scenario in which an expert elaborates their previously spoken descriptions and interpretations (5.4.1), and a fourth scenario in which experts produce written verbalizations, according to precise instructions, after watching an image sequence (5.4.2).

5.4.1 Elaborating the Transcripts of Spoken Commentaries

The lecturer (expert β) who had produced spoken descriptions and interpretations of the five dance sequences was asked to revise the transcripts of one description and one interpretation. The expert was encouraged to make corrections, insert addenda and give explanations to supplement the transcripts whilst watching the dance sequences again, but this time being allowed to stop, start and rewind the video player. First, the expert took 30 minutes to elaborate a description of one minute of dance (from Matthew Bourne's *Swan Lake*), stopping the video every second or so to talk about the movements in detail. Then she spent 15 minutes elaborating her interpretation of a dance (Lea Anderson's *Perfect Moment*) in which she talked about the dance as a whole, then commented on particular sequences ranging in length from 10 to 20 seconds. The expert also discussed issues concerning the appropriate choice of vocabulary and theoretical framework for describing and interpreting different genres of dance.

The expert started by discussing the issue of the vocabulary to be used for the description, noting that this kind of dance (Matthew Bourne's *Swan Lake*) called for a mix of classical ballet terms and other kinds of movement description, like terms from modern dance. A switch between two movement vocabularies was evident as the expert described the sequence – first using classical ballet terms to describe a series of aerial steps, then using contemporary dance terms to describe some contact work.

The elaborated description gave more information about various aspects of movement in the dance by, for example: adding details about the positions of arms and legs during steps; commenting on whether the back was curved or straight; noting the degree of turns and the extent of movements; and by referring to positions as relaxed or held. The expert was also able to give more information about the dancers' movement across the stage, in particular making frequent references to the diagonal pathways they took, and later to circular floor patterns as they moved around one another. The examples below show how much more information can be added to the 'live commentary' (6a) when the expert is given the opportunity to elaborate (6b). The elaborated commentary is still tightly time-aligned with the moving image, so is still a good source of video annotations.

(6a) 2:14 the second swan character comes towards him
2:18 [the swan] leans against him

(6b) 2:14 the swan takes two steps towards the prince, he has his arms to his sides
2:18 now the prince gently leans forward and the swan is leaning with his thigh against the prince's back and flexes the knee so that the prince takes his weight

The expert was also able to explain the movement terms used in the description, highlighting some of their defining characteristics. In talking about one minute of dance the expert elaborated, partially at least, a dozen candidate terms in relation to specific instances in the moving image, e.g. (7a-c). Such information could be used to elaborate the machine-based annotation language, or terminology.

(7a) a leap and a kind of flick kick with the legs to the front rather than the back, sometimes known as a **barrel leap** in contemporary dance terms

(7b) **aerial leap**, leaving the ground off two feet and landing on two feet as well, **one of five different kinds of aerial steps**

(7c) he **pitches** over, in contemporary terms, reminiscent of an **arabesque** position, he is on a supporting leg, he has one leg extended behind

When elaborating her interpretation (of Lea Anderson's *Perfect Moment*) the expert started by discussing the theoretical position to be taken and noted that dances of a post-modern genre, like this one, comprise a variety of imagery. The expert then went through the dance and noted movements that were cues for different kinds of imagery, e.g. undulating fluid movements that evoked imagery of the sea, and other gestures that had religious connotations. Although the expert had commented on this imagery in her original commentary, she was able, in their elaboration, to explain in more detail what it was about the dance that created the imagery. The further explanation of the interpretation made it more informative for a non-expert human, and perhaps for a machine.

5.4.2 Written Interpretations

Three further dance experts (η,θ,ι) watched a sequence (from Matthew Bourne's *Swan Lake*) with an instruction to speak a description; then they watched it again having been told that they would be asked to spend up to 10 minutes writing an interpretation of the excerpt after viewing.[7] The same definitions of 'Describe' and 'Interpret' were used as in the elicitation of spoken commentaries. The exercise was repeated for each expert with a second dance (Merce Cunningham's *Beach Birds for Camera*: the duet section).

Given ten minutes to write their interpretations for each dance sequence the three experts (who were PhD students, compared with the one lecturer and four Masters students who provided spoken interpretations) produced a total of 756 words at an average rate of some 30 words per expert per minute of dance viewed: this is in contrast to an average fluency of about 100 words per minute for the spoken interpretations. The average word length in the written interpretations, 5.95, was similar to their spoken equivalents but the average sentence length, 13.95, was 40% greater than the spoken speech fragments.

Though the three experts picked up on similar themes as those experts who provided spoken interpretations, they had the time to organise their verbalizations, rather than being directed by the unfolding moving image. Each displayed a distinct strategy in writing their interpretations. One expert (η) produced a narrative, written in the present tense, as if they were replaying the dance in their mind as they wrote. This narrative refers to movements and actions; as well as making interpretations about the emotional state of the main character (8).

(8) Night. A man, lonely and cold, enters the scene. He is apparently blind to his surroundings; lost in his own thoughts. He is, however, looking for something or someone. He searches the night sky, while around him men as "birds" come to life. They move away from him, leaving him isolated.

Another expert (θ) chose to abstract away from the movements in the dance and wrote about the dance as a whole, commenting on its meanings and themes (9).

---

[7] At this stage the interest was in the written interpretations; the spoken description was included to match the conditions for the elicitation of spoken interpretations.

(9) The work is about relationships. First and foremost this can be looked at as spatial relationships. However the duet - and the fact that we have two human beings on the stage - immediately speaks also about personal relationships. Particularly evident were reciprocity and support. Also independence.

The third expert (ι) took a viewpoint from which they could comment on the devices being used to create impressions in the viewer's mind, for example the lighting effects and the music, as well as the movements (10).

(10) The time is set visually and aurally by the muted blues of the setting, and by the gentle harmonies in the harp. We see the backs of men in 'feather' skirts, positioned like a conventional corps de ballet. Their slow movements bending forwards, carrying their arms behind their (naked) torsos and joining their hands (at the wrists) can be seen in tandem with the atmosphere setting music as having a dream-like quality. (A point emphasised by a 'moon' in the 'sky').

The kinds of information provided by these verbal reports is similar to that which might be found in extant written texts about the same dances. However, a possible motivation for eliciting verbal reports, rather than using the extant texts, is that here the experts' interpretations are free from the 'noise' of other information that is often interwoven into extant texts: the suggestion is that the elicited texts will be more easily turned into machine-executable surrogates (whether manually or automatically).

5.5 Discussion

The elicitation of verbal reports about moving images was motivated by the observation that extant collateral texts in a domain mix different kinds of information and refer to moving images at variable levels of temporal detail. The method presented in this section appears to elicit verbal reports that contain specific kinds of information about moving images, at pre-determined temporal granularities. Thus the verbal reports may be considered as more reliable collateral texts for the purposes of video annotation. The kind of informal reports elicited in the first scenario may be used to segment video data at a coarse level, and to associate with each segment characteristic features, as well as related information: however,

the process of ‘animating’ the verbal report would involve a high degree of human involvement.

In contrast, the concurrent verbal reports elicited in the second scenario appear to be more amenable for automatically generating video surrogates: keywords in these descriptions and interpretations could be used to label corresponding temporal intervals; information concerning the relationships between dancers and movements, and between movements and meanings, might be recovered with information extraction techniques; and, breaks in the cohesion of the verbal reports might be used for automatic segmentation. The identification of keywords in collateral texts is the basis of the video retrieval system presented in the next chapter. More speculatively, we have investigated the use of neural networks for clustering and segmenting verbal reports (and hence their corresponding moving images) on the basis of keyword vectors produced for each verbal report: see Appendix A.

As well as possibly providing a basis for video annotation, verbal reports may give some insight into how experts analyse moving images. On the one hand, verbal reports indicate the information ‘heeded’ by experts as they perform the task, for example, the part of the dance, or dancer, they are focusing on, or the related information they are considering. On the other hand, verbal reports give a clearer view (than normal texts) of how the experts ‘chunk’ the output of their analyses, that is how they select and organise information about the moving images. Systematic differences in both the information heeded, and in information selection and organisation, were observed between the processes of description and interpretation.

Research at the intersection of verbal reporting and language production may contribute to understanding more about how experts articulate their thoughts about moving images, and may thus assist the application of verbal reports for video annotation. This chapter closes by considering how ideas explored in one famous study of language production, the so-called ‘Pear Stories’ study, may cast some light on the verbal reports discussed in this chapter.

Wallace Chafe (1980) led a group of researchers in an investigation of language production using a method in which subjects were asked to speak a narration of a film that they had just been shown: the film was about seven minutes long and showed scenes such as a man picking pears and a boy riding and falling off a bicycle. The film served as a

'constant' in the investigation of cultural variance in language production: it was shown to different groups of speakers and the resulting narratives were compared to discuss differences in the choice of words and structures. Some of the ideas discussed by Chafe's colleagues, notably Pamela Downing and Deborah Tannen, are perhaps relevant in understanding our dance experts' production of verbal reports.

When describing moving images, the experts used both general language words, some with restricted meanings in relation to dance, along with specialist terms to identify the entities and actions in the moving image, and to situate them in space and time. In the spoken commentaries the most frequently occurring words to describe movements were common general language words with what might be called a *broad referential scope* (Downing 1980). When the experts were not under a time constraint they were better able to use more specialist terms to identify movements more specifically. Another factor discussed by Downing which may be pertinent when considering the experts' lexical choices is the degree of *codability* of movements: a well recognised movement would be more consistently labelled with a specialist term than a recently innovated and unfamiliar movement.

In their interpretations of moving images the experts assigned non-literal qualities to particular movements and to longer sequences. The interpretations of particular movements were sometimes cued by words such as *like, as if* and *seems*, such that the utterances resembled similes: the interpretation of longer sequences was sometimes cued by the phrase *there is*. Other interpretations were not so explicitly cued and perhaps exemplify what has been termed *interpretive naming*, whereby a speaker chooses a lexical item in order to "create an interpretation by triggering a series of associations" (Tannen 1980:70). An example would be the choice of the word *wander* over the word *walk* in order to convey a feeling of aimlessness in a character's movement. There is an acknowledged tension between description and interpretation in the field of dance analysis, which may be replicated in the analysis of other specialist moving images, so it is perhaps interesting to be able to study the distinction from cognitive and linguistic viewpoints.

Chapter 6

The KAB System: design, implementation and evaluation

Recall the idea of a knowledge-rich video annotation and browsing system that was proposed in the first chapter: it was suggested that collateral texts might be gathered and processed alongside a video database. The previous chapters discussed how collateral texts may be used to describe and interpret images. This relationship is important for digital libraries because it means that systems can annotate still and moving images with these texts, such that machine-executable surrogates are generated from the texts for retrieval purposes, and so that users can browse between video data and explanatory information. In this context, sequences of video frames and their collateral texts can be regarded as *objects*: that is, identifiable components within the system which are in some sense self-contained and have identifiable boundaries. For us, this *object-oriented* view of moving images and collateral texts is helpful in the design and implementation of a system to store and access video data.

Computationally, objects are components comprising data structures and procedures which manipulate the structures: objects may include attributes to refer to data files containing, say, text or video data. The procedures, or *methods*, which manipulate the data structures can be invoked by sending messages to the objects. Two important properties of objects are *inheritance* and *encapsulation*. Objects can be thought of as instances of a template and collectively objects can be said to form a class: new classes may be defined in terms of existing ones, inheriting some or all of the properties of the existing class. Computation proceeds by virtue of message passing between objects, in such a way that the objects sending and receiving messages do not need information about each other's internal structures: in this way they are encapsulated.

The twin notions of inheritance and encapsulation are important for developing systems that deal with different kinds of interrelated data, such as video data and text. Collections of video sequences and texts may be arranged in separate hierarchies, and sequences of images may be related to many texts, and *vice versa*. Here inheritance will play a key role in allowing classes of texts and videos to be defined in terms of existing classes. The notion of

encapsulation makes it easier to deal with the multiple standards for encoding text, and in particular video data. For example, an object sending a message requesting an action on some video data does not need to be concerned with the particular coding format of the video data file: matters such as decoding and displaying the video data will be dealt with internally by the recipient object. Object-oriented design thus focuses on defining classes of objects and the procedures (methods) that act on them, and on capturing the relationships between classes of objects and the messages to be passed between them. In a system that uses collateral text to annotate video data, the most important aspect of the design will be capturing the properties of moving images and texts, and their interrelationships.

This chapter presents the KAB system which is being developed in an object-oriented manner. The system has been used to access a collection of digital dance videos with experts' descriptions and interpretations. The chapter starts with a discussion of requirements specification, prototyping and development issues (Section 6.1). The object-oriented design of the current system is shown using the Unified Modelling Language (Section 6.2) and its implementation in Java is reported with a series of screenshots and explanations of the objects used to deal with video and text data (Section 6.3). The results of system evaluation (a questionnaire-based user study) were encouraging both for the KAB system, and for the general exploitation of the video-text link in digital libraries (Section 6.4). The chapter ends with a consideration of further developments for the integrated processing of video data and collateral information (Section 6.5).

6.1 Overview of System Development

Two overlapping sets of factors, or requirements, guided the development of KAB. On the one hand, there were considerations to do with how a user might wish to access video data and related information interactively in a digital library. On the other hand, there was the wish to develop a computational environment in which to explore the link between moving images and collateral texts.

To gain a potential user's perspective, time was spent with a dance expert discussing how a digital library could be used both as an educational resource, e.g. for students and for lecturers, and as a resource for research, e.g. for experts. Requirements common to all potential uses that were identified included the capabilities to locate relevant video sequences in response to queries, and to provide further information related to the sequences, for example texts, musical scores, movement notations, animations, and other video sequences. It was also specified that users should be able to build their own collections of videos and texts, and make their own annotations.

Two prototypes were implemented with the 'multimedia authoring system' Macromedia Director. The final KAB system was implemented in the object-oriented Java programming language, supported by the Java Media Framework: the program comprises approximately 4000 lines of code. KAB also makes use of a commercially available text corpus management and analysis package, *System Quirk* from the University of Surrey. Dance was used as an exemplar domain for development: 20 minutes of dance videos were digitised (at a resolution of 320x240 pixels in MPEG-1 format which required 400 MB of disc space) and stored alongside the transcribed words of experts' concurrent verbal reports (21,000 words) and the Surrey Dance Corpus (350,000 words).

6.1.1 System Prototyping

The key characteristic of the KAB system that crystallised during prototyping is that arbitrary intervals of video sequences can be given a linguistic label, and can have further information associated with them, e.g. text files. The progression in prototyping was in the reduction of effort required to delimit, label and attach relevant information to intervals of video data. In the first prototype handcrafted sections, labels and links were hardwired into the system, Figure 6.1a (Ahmad, Salway and Adshead-Lansdale 1998). In the second prototype decisions about sections and labelling were still left to a human, but the system could store different sets of annotations, and these could be entered through a user interface, Figure 6.1b (Ahmad, Salway and Adshead-Lansdale forthcoming).

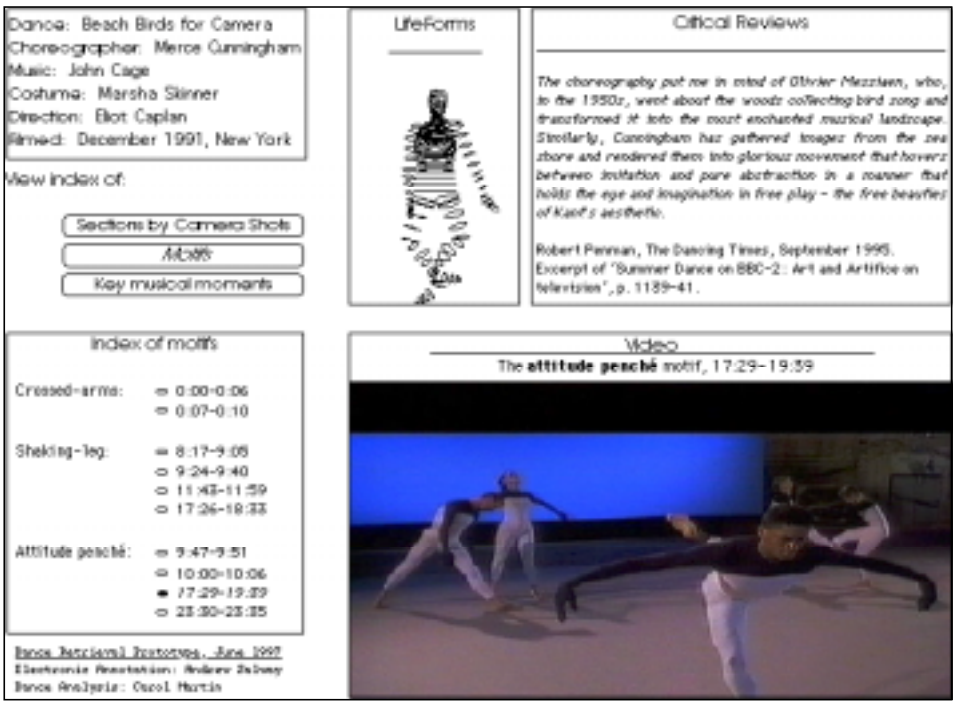


Figure 6.1a: The first prototype in which sections and labels, and related information, were hardwired into the system. The moving image can be navigated by a list of sections, e.g. corresponding to the occurrence of motifs. Further information made available includes bibliographic details, a graphical animation of an important movement, and a critical review.

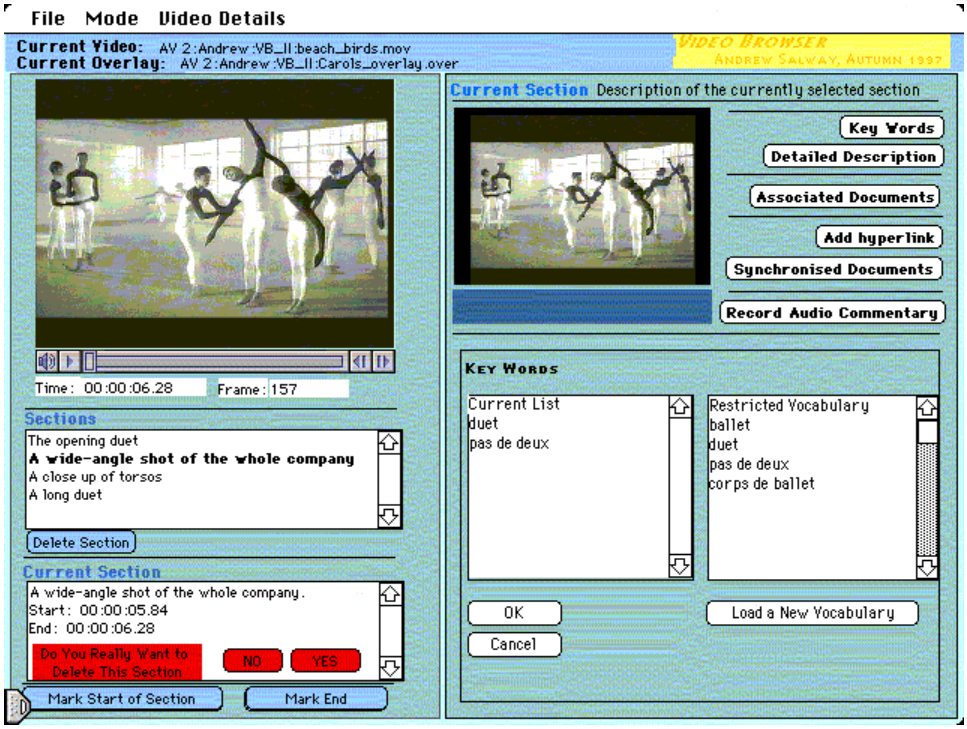


Figure 6.1.b: The second prototype in 'Annotate' mode whereby the user can delimit and label sections, and attach other kinds of information to them. Details of the sections, labels and links are stored as 'overlays' which are separate from the video data, so that different overlays can be used to access the same video data when the system is switched to 'Browse' mode.

The interest during prototyping was in how the user could interact with video sequences and a range of related information, that is not only texts but notations, graphical animations, etc. For the design of the final system, the focus was limited to just collateral text, with the aim of producing labelled intervals of video data from the text.

6.1.2 Design and Implementation Issues

The KAB system was developed in an object-oriented manner. It has been suggested that object-orientation "provides better concepts and tools with which to model and represent the real world" (Khoshafian and Abnous 1995:7). The same authors predicted that object-orientation would be exploited to "integrate diverse sets of objects ranging from text and voice to images and video" (ibid.:30). A recent discussion about the storage and retrieval of diverse data types like voice, images and video suggests that this prediction is already being realised (Subrahmanian 1998).

A number of object-oriented modelling and development paradigms have converged in the Unified Modelling Language (UML) which has wide-ranging support from theorists and systems developers, coordinated through the UML Partners consortium. It has also been accepted as the standard for modelling languages by the Object Management Group (Eriksson and Penker 1998). UML can be used to model all aspects of system development: for us it provided a convenient notation for designing the KAB system.

As mentioned above prototyping was done with Macromedia Director, but Java was used for the final system implementation: the reasons for this approach are discussed now. Macromedia Director is a package for combining texts, images, video, audio, etc. into interactive presentations, and it is widely used as such by graphic designers and others who may not have a computing training. It provides a drag-and-drop interface to co-ordinate the presentation of these different data types and a programming language, 'Lingo', which can be used to enable interaction with a user. Director uses cinematic 'metaphors' like *cast*, *stage*, *score* and *script* to order the interactive presentation of multimedia information. A 'cast' is a set of objects (text, image, sound and video files) that are organised spatially on the 'stage'

and temporally along the 'score'. Each object may have an associated 'script', written in Lingo, that further specifies its behaviour. Director addresses the needs of its target users very well: through its 'metaphors' it shields the non-computing specialist from the conceptual difficulties of object-orientation, and makes it easy for them to develop one-off solutions. For the current work Director facilitated the rapid development of prototypes which helped to maintain the interest of our dance experts and to elicit further system requirements.

However, it was felt that for a systematic exploration of the link between moving images and collateral texts, the idiosyncratic nature of Director was not appropriate. Although its developers have acknowledged the importance of object-orientation for systems dealing with multimedia data, their proprietary approach led the authors of a programming guide for Lingo to note that they "faced a different task than authors writing about HTML, JavaScript, Java, C++, or indeed any number of open standard languages" (Plant and Smith 1997:5). It also seems that there is confusion, or at least vagueness, about object-orientation with respect to Director: "In the context of the Lingo programming environment, the definition [of an object] could be: anything you want to put into the computer's memory" (Small 1996:7).

A deciding factor in favour of Java as the 'open standard language' for implementing KAB was the availability of the Java Media Framework (JMF) which offers high-level operations for video presentation and manipulation. The JMF abstracts from the physical layer of video data so that programs can be written independently of video coding formats. The JMF Application Programming Interface provides high-level operations for controlling the display of video data, i.e. in effect 'stop', 'start' and 'go to time X'. With regards to handling text, Java has class libraries which facilitate the implementation of language processing techniques. Java also has the technical virtues of being independent of hardware architectures and operating system platforms. Finally, Java is a modern language in widespread use, thanks in part to its free distribution, and perhaps now most importantly, the fact that Java programs can be run in a Web-browser over the Internet.

Further advantages of Java became apparent during system implementation. Its 'object serialization' functionality relieves the need to connect with an external database for saving complex data structures – this saved time when structures were changed during development.

Instead of storing data in an external database, it is possible to declare a Java class to be 'Serializable': this means that an instance of this class (which may contain all the data used by a system) can be written to a file as easily as a simple data type. The multithreading functionality offered by Java makes it suitable for an application with several windows each running processes together, for example while one is showing a video, another can be analysing texts.

6.2 System Design

Whilst the prototypes attempted to explore how different kinds of collateral information could be presented alongside video data, the emphasis in the development of the final KAB system was on processing, and presenting, collateral texts. The use of collateral text for annotating digital collections of moving images requires a system to store, retrieve, process and display video data and text files in an integrated and flexible manner. The design of the KAB system addresses these requirements with data structures that capture important features of videos and texts, and the link between them. The user is given a variety of ways to access the stored information with retrieval and browsing combined in a graphical user interface. The system is geared towards collateral texts that follow the temporal order of moving images, such that they can be time-coded; like the concurrent verbal reports elicited from dance experts.

Details about a moving image, and hence a video data file, that might need to be stored include its name, when and where it was produced, and the people involved in making it. This information can be held in a series of fields, along with a reference to the location of the video data in a local file system or on a network. The necessary fields can only be fully specified with reference to a particular kind of moving image. Since KAB has not been tailored to a specific domain, it uses two broad fields – *name_of_video* and *name_of_person*; for a collection of dance videos these fields record the name of a dance and its choreographer. For storing texts it will be important to record details such as *text_type* and *author*, and crucially a reference to a related video. The type of a text may determine how it can be used by the system since different types are informative about moving images in different ways.

To annotate sub-intervals of a video data file, KAB uses a data structure based on Oomoto and Tanaka's (1993) 'video-object' (see also Tanaka, Ariki and Uehara 1999). This structure supports flexible annotation, and hence retrieval, since annotations can refer to sub-intervals of any length and they may overlap (in contrast with the lists of non-overlapping sections in the prototypes). In KAB an annotation comprises: a reference to a video; a start time and an end time to delimit an interval within the video; and, a text string which refers to some aspect of the interval's content. The text string is selected from lists of terms which are stored by KAB – these lists can be used to group terms that refer to the same aspects of video content.

The creation of term lists can be semi-automated through text analysis to elicit candidate terms from a text corpus, e.g. using the method of linguistic variance reported earlier. The term lists then guide the automatic generation of annotations from time-coded texts related to specific video data files. The user selects a term list to indicate the kind of content to be annotated, then the system finds occurrences of the terms in the texts. For each occurrence an annotation is created: its start time, and end time, are calculated by subtracting, and adding, fixed amounts from, and to, the time-code of the text fragment in which the term occurs. The annotation's text string is set to equal the term.

The user's view of moving images and texts in KAB is through a graphical user interface that combines the processes of annotation with strategies for accessing video data and related information. The annotation and the accessing of video data are combined so that the annotator can consult related information and the person retrieving and browsing video data can make personal annotations. KAB has not yet been tailored for a particular domain, nor for a particular kind of user because at this stage the interest is in exploring the potential of the video-text link: however, this means that, for the uninitiated, the user-interface is not as 'friendly' as it could be.

The principal data structures and processes in the KAB system can be shown in a class diagram, Figure 6.2. (User interaction with the system is best explicated by the series of screenshots in the next section). The first class is KAB_Database, of which there is only ever one instance in the system at a time. As the aggregation links show, a KAB_Database comprises an unlimited number of Videos, Texts and TermLists. The Video class has attributes

to hold details such as the name of the video and the person who made it, along with the file location where the video data is physically stored. The Text class has similar attributes, but it also has an attribute to associate a particular text with a particular video: the link in the figure shows that many instances of the Text class can be associated with one instance of the Video class.

The methods of the Video class serve to control the playback of video data, i.e. starting, stopping and moving to a particular point in the sequence. The object's property of encapsulation means that these methods can be invoked by other objects without their needing to know the idiosyncrasies of different video data formats. The methods of the Text class are concerned with the presentation of texts to the user who can modify them by highlighting or filtering terms from stored lists.

Many instances of the Annotation class, possibly overlapping (in time), can be associated with one instance of the Video class. An instance of the Annotation class is generated either manually through user-interaction, or automatically by the system processing a time-coded text (as described previously): note that the method generateAnnotations belongs specifically to the TimecodedText class; this class inherits further attributes and methods from the more general Text class.

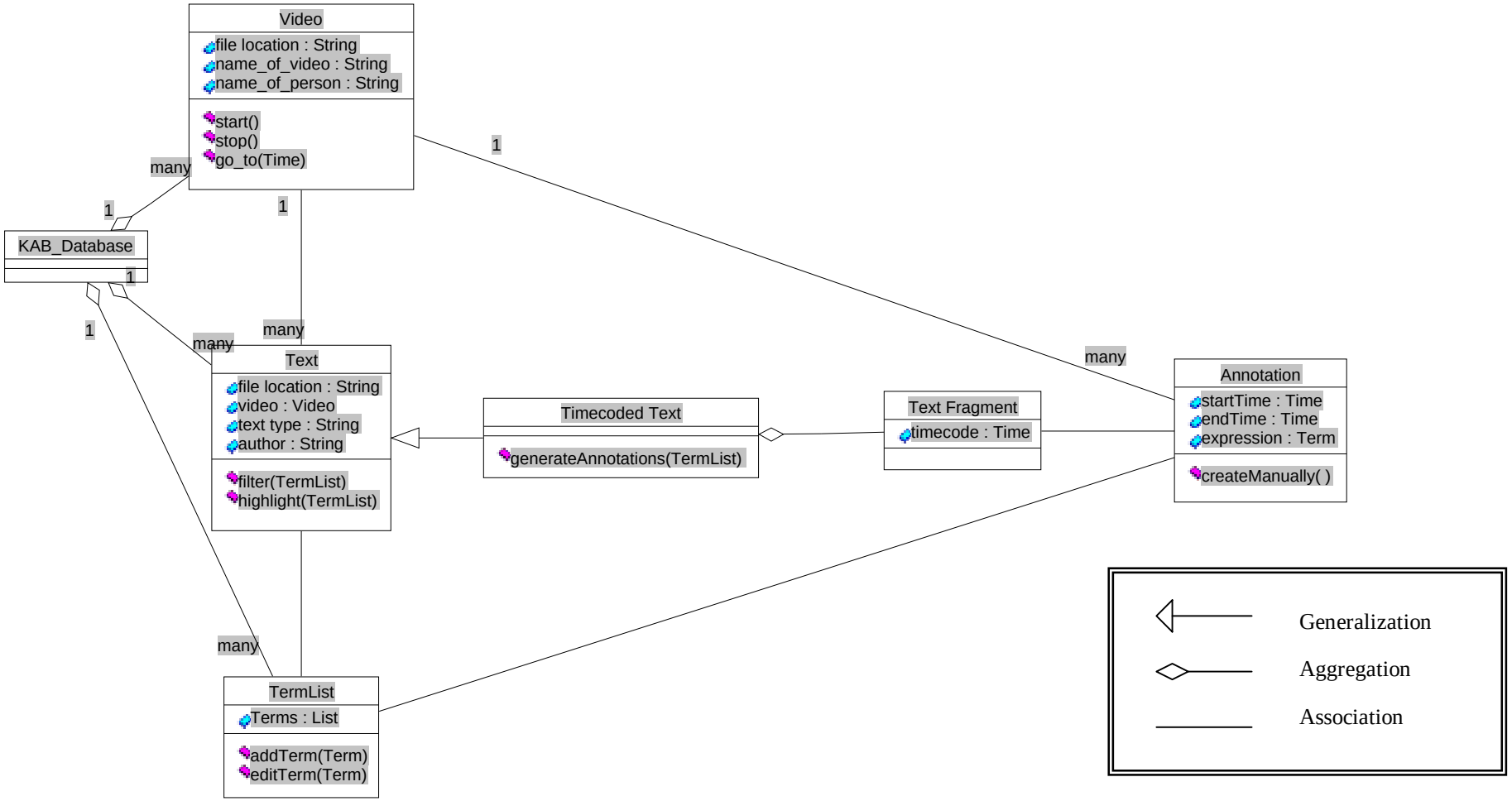


Figure 6.2: A class diagram showing the main data structures and processes in KAB. The notation shows the attributes and methods of each class, and the relationships between them, i.e. *generalization*, *aggregation* and *association*; for details of the Unified Modelling Language (UML) see Eriksson and Penker (1998).

6.3 System Implementation

The main KAB window co-ordinates the processes for storing and accessing video data and collateral texts. The system's functions are organised in five horizontal bands, Figure 6.3. The first band displays the name of the current instance of the KAB_Database class: there are also buttons to load and to create a new instance, and to save the current one. As the previous class diagram showed, an instance of the KAB_Database class holds information about a set of videos, texts, term lists and annotations. Java's object serialization facility means that loading and saving this data is a simple operation, requiring only one object (i.e. an instance of KAB_Database) to be read from and written to disc.

Details about the videos and texts in the current KAB_Database are listed in the second and third bands of the window respectively: there are also buttons for maintaining information about texts and videos, and for viewing the highlighted choices. The bottom two bands comprise buttons for annotation and for browsing. All these functions are presented in the rest of this section.

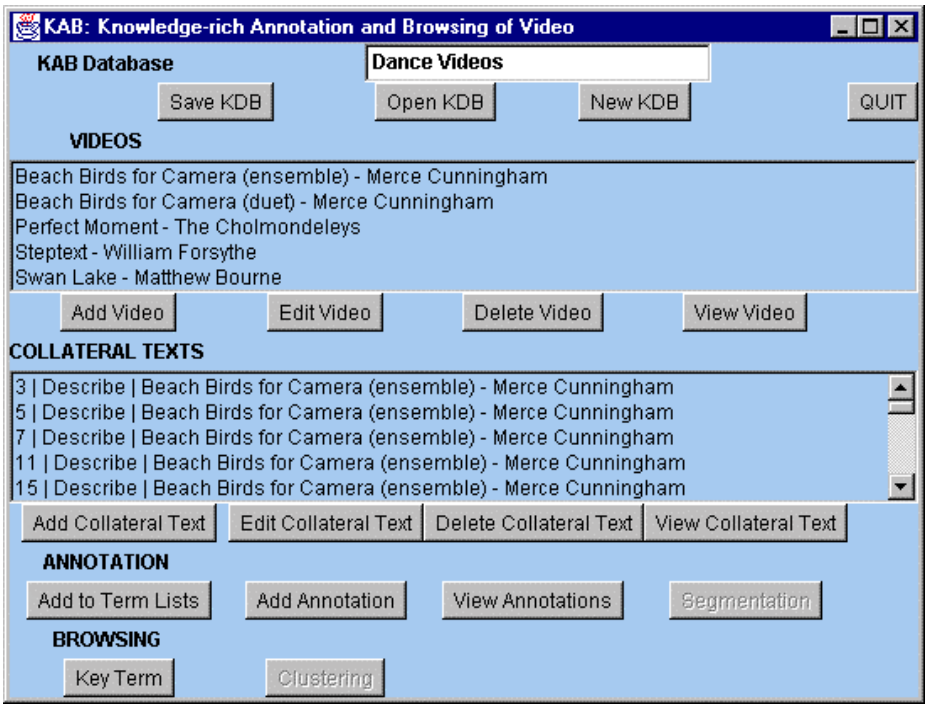


Figure 6.3: The main KAB window comprising five horizontal bands of functions grouped under 'KAB Database', 'Videos', 'Collateral Texts', 'Annotation' and 'Browsing'. Note that the details of the listed texts include an author given as a number (for experts' anonymity) and a text type which is either 'Describe' or 'Interpret' – the third column is the video to which the text relates.

6.3.1 Viewing Videos and Texts

When the 'View Video' button is pressed, the system identifies the object corresponding to the selected video in the list, and 'reads-off' the file location of the video data. The file location is then sent in a message to an object which creates a window to display the moving image, Figure 6.4: when 'View Collateral Text' is pressed on the main KAB window, a similar sequence of events leads to a text being displayed. The video window includes standard controls that start and stop the video, move it back and forward, and adjust its volume: the text window has standard scrollbars for moving through the text. The KAB-specific functions of the buttons on these windows are explained in following sub-sections.

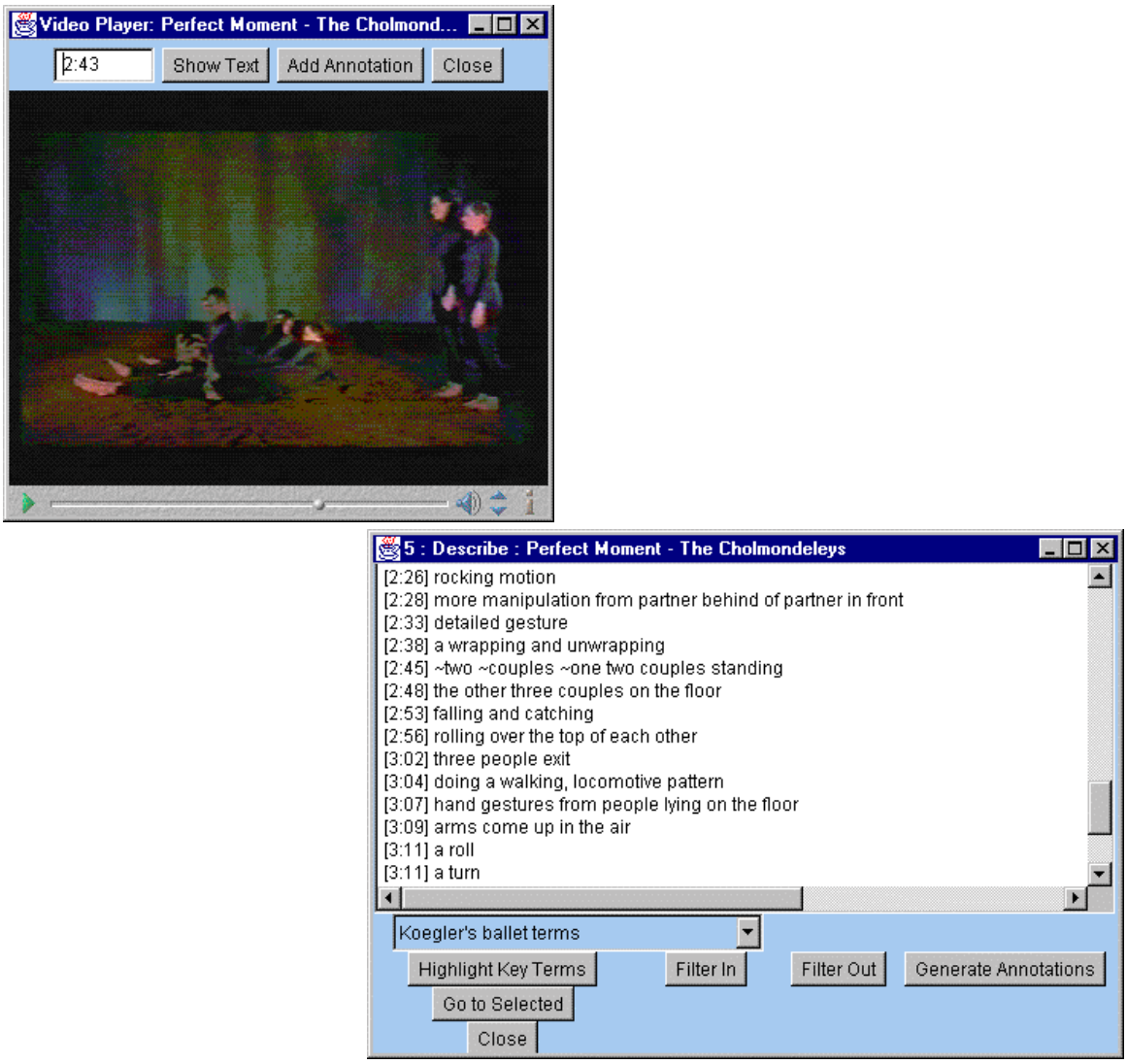


Figure 6.4: The windows for viewing videos and texts in KAB. Note there are standard controls for controlling the playback of a video, and for scrolling through a text. The other buttons on the window relate to KAB-specific functions.

6.3.2 Retrieving Video Sequences with Key Terms

Rather than choosing a video by its title, the user can search the stored videos by their annotations. The 'Key Term' button on the main KAB window brings up a window from which the user selects a term (from the stored lists). This term is then matched against the list of annotations in the current KAB_Database, and those that match are listed in a results window. When one of the returned results is chosen, the corresponding video interval is played, starting at the start time of the annotation (and returning there to play again when the end time is reached): further details about the annotation are also displayed, Figure 6.5.

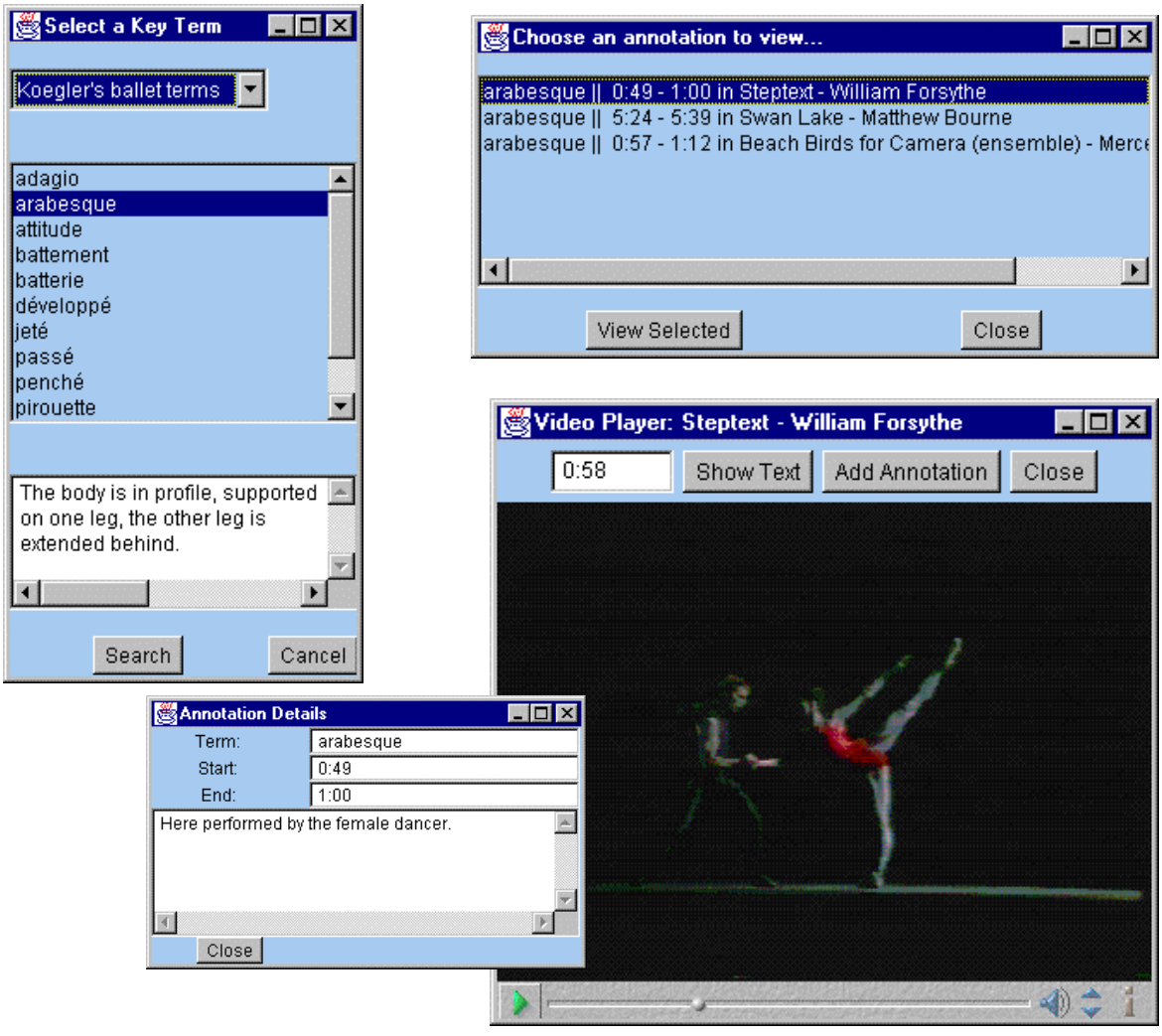


Figure 6.5: Retrieving a video sequence with a 'Key Term'. Note that a definition of the currently selected term is shown in the selection window. The result of the search is a list of all annotations in the current KAB_Database which match the term. The interval corresponding to the chosen annotation from the list is shown, along with further details.

6.3.3 Navigating Video Sequences through Collateral Texts

We previously saw how windows displaying a video and a text could be viewed next to one another (from the main KAB window), but nothing was said about the interaction between them. In fact, as the user watches a video sequence, be they annotating or browsing, they are able to call up a list of all the texts in the current KAB_Database that are related to the moving image. As well as being read by the user as a source of background information, time-coded texts can be used to navigate through the video data. When the user highlights a line in the text and presses 'Go to Selected', the system takes the time-code from that line and sends it to the video player, which then moves the video to the appropriate time, Figure 6.6. Note that with the concurrent verbal reports used here a time-code is the time of utterance, so the video should be set to an earlier time in order to show the action to which the utterance refers.

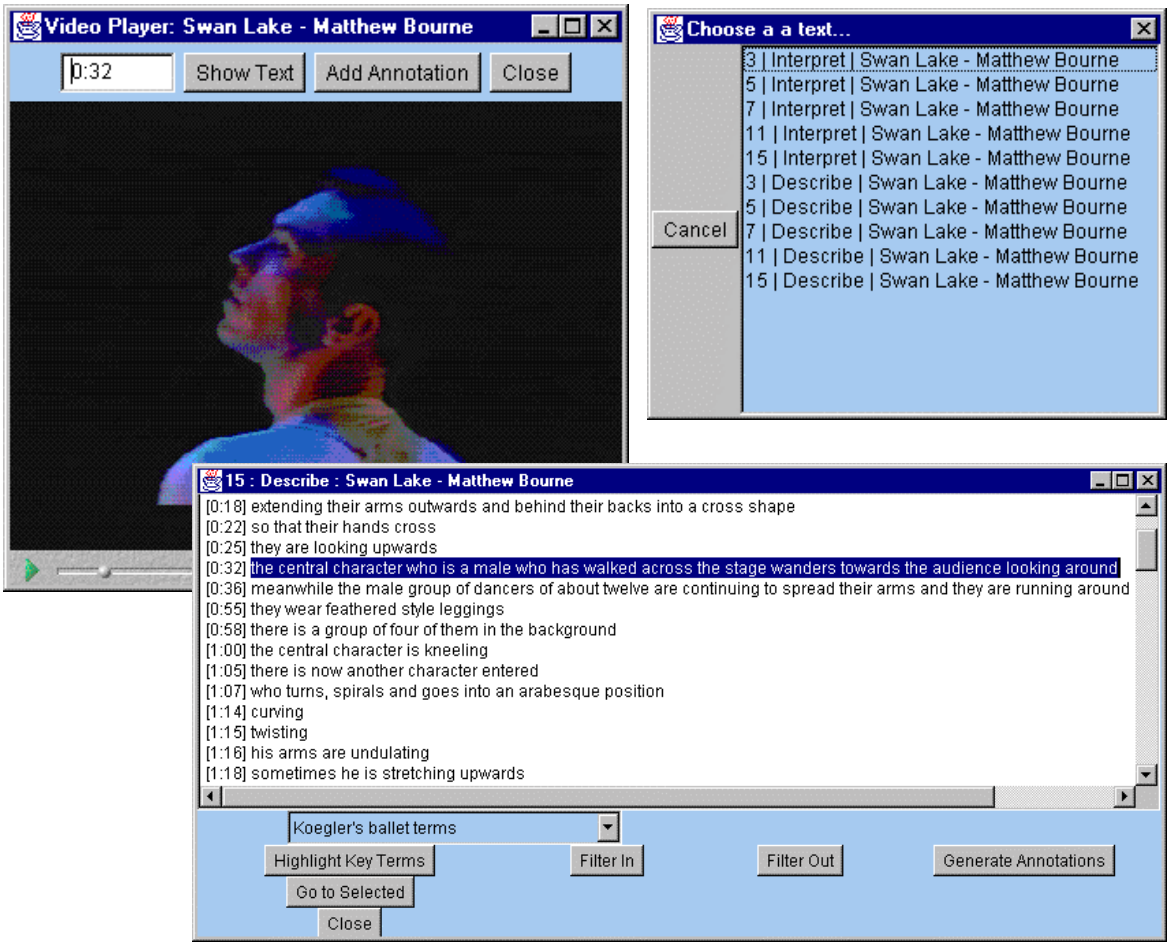


Figure 6.6: A list of texts related to the moving image is displayed by pressing the 'Show Text' button in the video window. The selected text is shown here below the video. When the 'Go to Selected' button is pressed, the video moves to the time of the highlighted text fragment.

This text-to-video navigation can perhaps be improved if the user's attention is drawn to key terms, or other significant lexical items in the text. For example, if all words but those in a list of terms are removed, then it is easy to spot certain patterns of movement, Figure 6.7a.

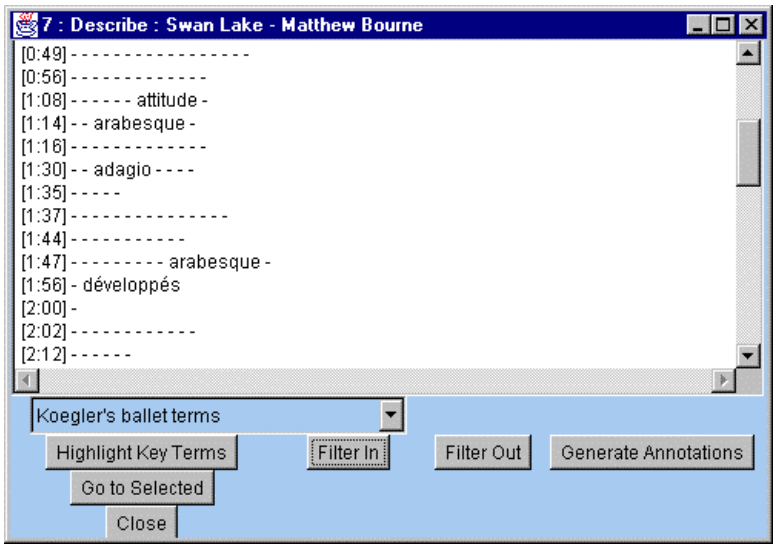


Figure 6.7a: By 'Filtering In' terms from a selected list (here ballet terms), KAB draws attention to sequences of specific movements in a moving image. Note that '-' stands for a word in the text that has been left out.

Other textual features might also assist the user to get a better understanding of the moving image. As noted previously, experts sometimes use one of a limited set of phrases to link a movement description with an explication of its meaning. If such phrases, including 'as if', 'like' and 'suggesting', are entered into a list in KAB, then specific acts of interpretation can be highlighted in a text, Figure 6.7b.

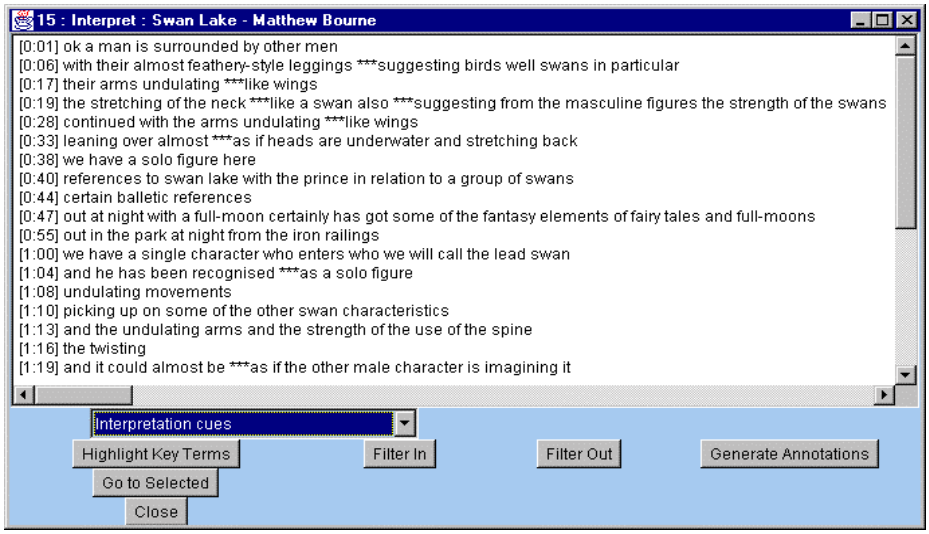


Figure 6.7b: 'Highlighting' interpretation cues (with '*') helps the viewer by drawing their attention to where the expert may be explicating the meaning of the moving image.**

6.3.4 Adding Videos, Texts and Terms

To build a collection of videos and collateral texts, the user enters bibliographical details for each video and each text, along with the physical location of the corresponding data file. Crucially, when entering details about a text they must also specify the video in the current collection with which it is to be associated. The windows for adding details about videos and texts to a KAB_Database are shown in Figure 6.8: when the user presses 'Submit' in either window, a new instance of a Video or Text class is generated with the specified attributes.

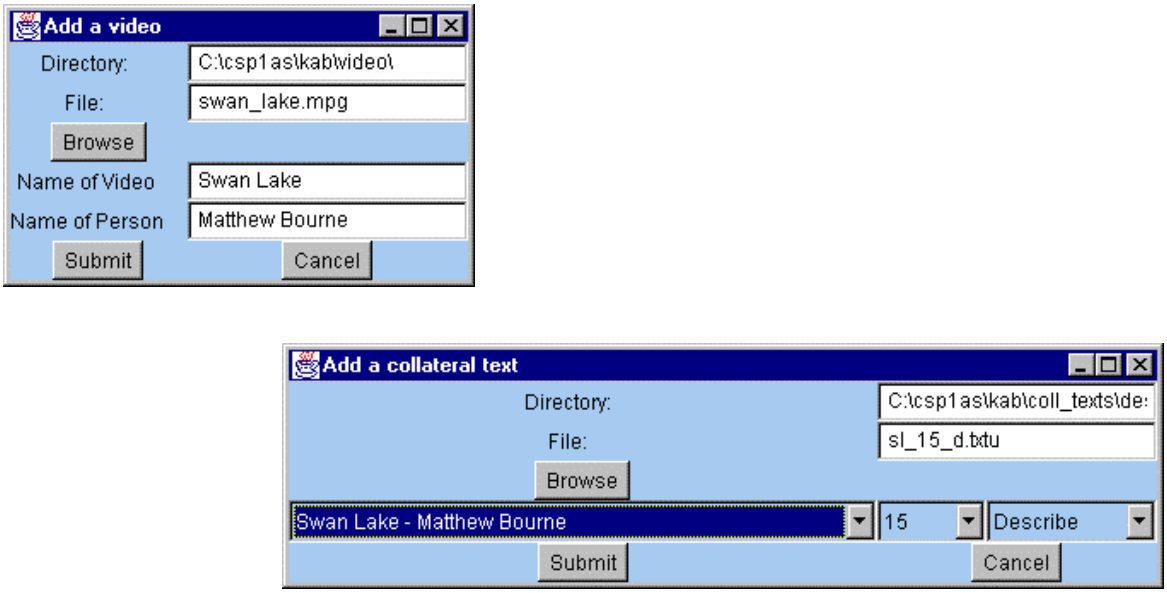


Figure 6.8: When the user presses 'Add Video' or 'Add Collateral Text' on the main KAB window, the appropriate window opens for entering details about a video or a text. Note that the system is geared towards the use of verbal reports provided by experts, hence the 'author' of a text is given as a number (for expert anonymity) and the text type is given as 'Describe' or 'Interpret'.

Term lists play an important role for annotating, retrieving and browsing videos and related texts in KAB. Since the choice of an 'annotation language' might be considered to bias a system, it is important that a system can handle different 'languages'. In KAB the user creates named term lists, and adds terms with definitions and other information as required, Figure 6.9. Note that the lists can also be used for other kinds of lexical items or phrases, like 'interpretation cues' or the names of the characters in a dance.

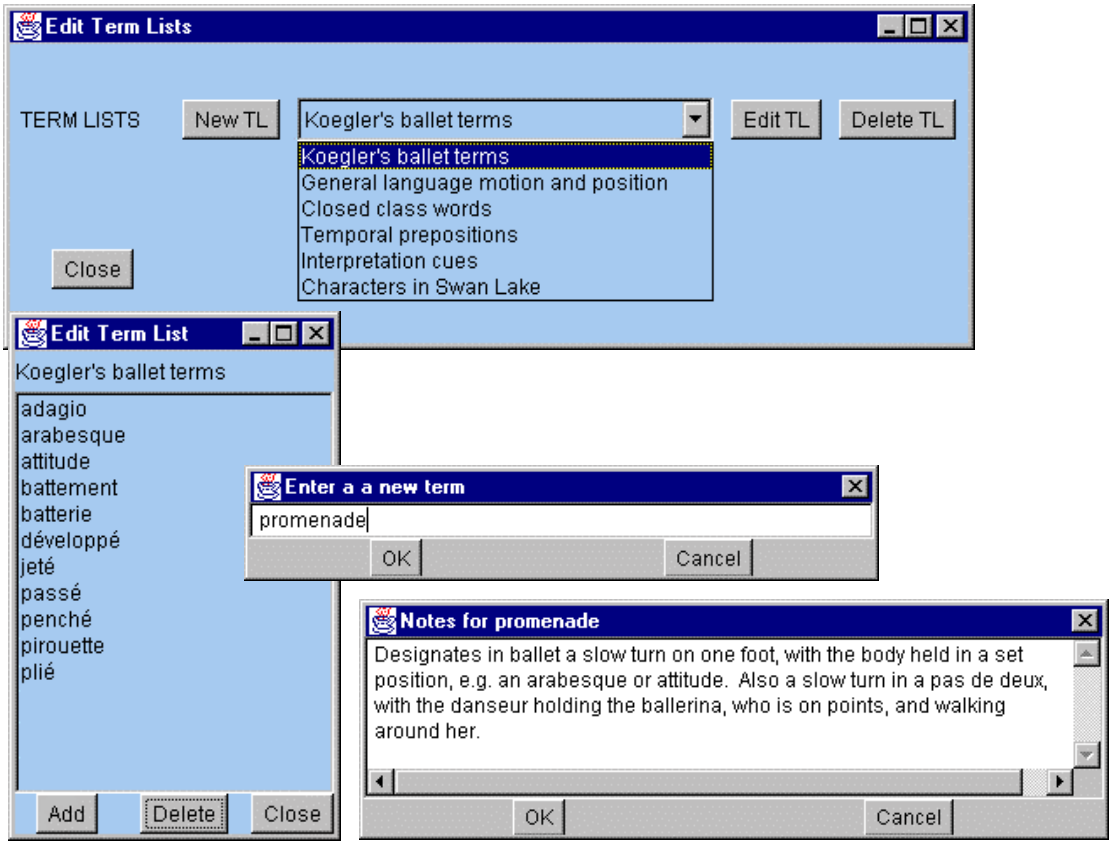


Figure 6.9: Modifying a term list. The user chooses which list to modify (or creates a new one), then adds or deletes terms. When adding a new term, a definition or other information can be associated with it as 'notes'.

6.3.5 Annotating Video Sequences with Key Terms

The annotation process in KAB, whether manual or automatic, involves creating an instance of an Annotation object with a start time, an end time and a key term. The user may annotate a video sequence by marking the start and the end of an interval as a video is playing, and then selecting a term from the lists stored in the system: the user may also add notes to give further information about the annotation, Figure 6.10.

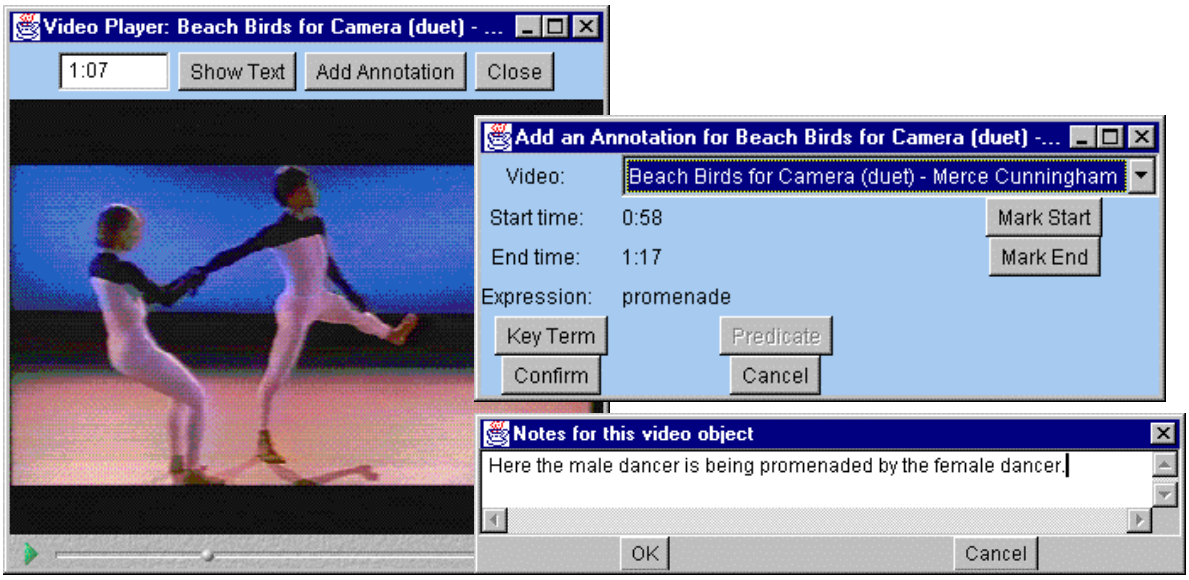


Figure 6.10: Manually annotating a video sequence. The 'Mark Start' and 'Mark End' buttons are used to delimit the interval as the video plays, then a term, here *promenade*, is selected from the lists stored in the system. Further information about the sequence can be added as notes.

Alternatively, candidate annotations are generated automatically from time-coded texts that have been associated with video sequences. The user specifies a term list according to which aspects of video content are to be annotated, then whenever a term from the list is found in a text a new instance of the Annotation class is made. It is given a start time and an end time equal to the time-code of the text fragment containing the term, minus and plus a fixed number of seconds. In this case, the annotation's notes record the fact that it was automatically generated. A list of candidate annotations is presented to the user who can view them before accepting, editing or deleting them, Figure 6.11.

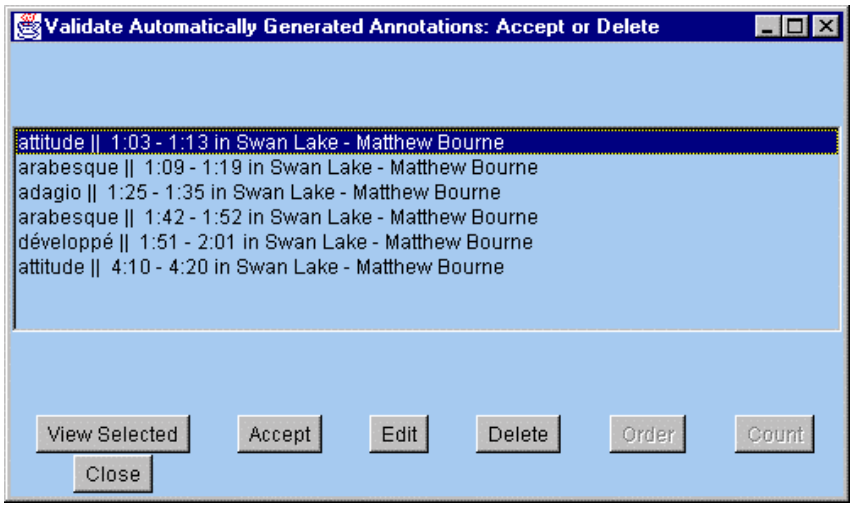


Figure 6.11: KAB presents automatically generated candidate annotations for the user to accept, edit or delete.

6.4 System Evaluation

Traditionally, information retrieval systems have been evaluated with statistical measures of *precision* and *recall*. There appears to be little sign yet in the video retrieval literature of such statistically rigorous approaches being used to evaluate systems there; this may be because the notion of 'document relevance' becomes (more) problematic in relation to video data, and also because video collections tend to be too small. For the evaluation of hypermedia-like systems, in which users interactively browse between information sources, it is more difficult to determine performance metrics because of the potentially unlimited ways in which information can be usefully provided.

The KAB system combines video retrieval with the browsing of related documents, but the size of the data set available, and the system's interactive nature, meant that an empirical evaluation of its performance was not feasible. Rather, it was decided that the most appropriate evaluation would be by means of a questionnaire-based user study. The baseline for evaluating KAB was the situation in which somebody is confronted with a room full of video cassettes and printed material relating to them; or the electronic equivalent of being presented with lists of files and standard video players and text viewers. That is, we are interested in how users judge the features that KAB offers for accessing visual and textual information, beyond simple video playback controls (start/stop/rewind/fast-forward) and the scrolling of text.

6.4.1 Evaluation Method

The KAB system was evaluated by six users filling in a questionnaire that comprised two parts. In the first part the users rated the system's ease of use for set tasks in terms of fixed grades: this gave a 'formal' evaluation with quantifiable results. For the second part the users answered open-ended questions about how they thought the system could be developed and applied further: this 'informal' evaluation allowed users to decide upon which aspects of the system to comment.

The users included three academics from the performing arts, all of whom were computer literate and had expressed an interest in KAB; and, three computing professionals with different levels of experience, Table 6.1. Each user was given a one-to-one demonstration of the system, lasting 20 minutes. They were given the questionnaire and encouraged to use the system for themselves as they answered the questions. The users spent between 30-50 minutes using the system and filling in their responses.

User	Occupation	Education / Training	Computing Experience	Dance Experience
1	***University Professor of Dance Studies***	Degrees -> PhD	Word processing, spreadsheets	Plenty – over many years, as performer and spectator
2	***Senior Lecturer (Sound Recording)***	Bmus Tonmeister PhD	Considerable operational and some programming	None
3	***Lecturer in Dance Studies***	BA Creative Arts / MA Dance Studies / PhD – Dance on Screen	Basic WP	See Education / Training
4	***Senior Lecturer (Computing)***	PhD in Physics.	Extensive Research in Computer Assisted Learning and Computer Modelling & Simulation	None
5	***Systems Engineer***	Business / Computer Science / Training / Consultancy	20 years	None
6	***Computing undergraduate***	'O' and 'A' levels, and City & Guilds.	Final year honours degree Computing student. One year industrial experience in an IT department.	None.

Table 6.1: The profiles of users in the evaluation study – note this information is reproduced here as it was entered on the questionnaires. The three non-specialists in computing were all at least computer-literate and had an interest in using video retrieval systems: the specialists had different levels of professional experience, including familiarity with video retrieval systems.

The questionnaire and the users' full responses are reproduced in Appendix B.

6.4.2 Ease of Use

The users were asked to rate the set tasks in terms of five grades from 'Very Difficult' to 'Very Easy'; the tasks encompassed the system's functionality as it is presented in Section 6.3. The results suggested that the users were at least reasonably happy with storing and accessing video data in KAB, Table 6.2. Only twice did a user mark a task as 'Difficult', and

none were considered to be 'Very Difficult'. The grades of 'Very Easy' tended to be given by the computing specialists who perhaps had a better appreciation of the underlying system, and so could 'see through' the interface to some extent. Nevertheless, the non-specialists in the main gave grades of 'Easy' or 'OK'.

Task (corresponding section)	Very Easy	Easy	OK	Difficult	Very Difficult
***Select and view videos and collateral texts* (6.3.1)**	4	1	1	-	-
***Search for video sequences with key terms* (6.3.2)**	1	2	2	1	-
***Navigate through a video from a collateral text* (6.3.3)**	2	2	2	-	-
***Highlight and filter key terms in a collateral text* (6.3.3)**	1	3	2	-	-
***Add a video and a text to the system* (6.3.4)**	-	2	4	-	-
***Add terms to the system* (6.3.4)**	2	3	1	-	-
***Manually annotate a video sequence with a key word* (6.3.5)**	-	2	4	-	-
***Generate annotations from a collateral text* (6.3.5)**	-	2	3	1	-

Table 6.2: The number of users giving each grade for the eight set tasks. Only two tasks were considered to be difficult, and then only by one user. Though it is not shown by the table, most of 'Very Easy' grades were given by the computing specialists: nevertheless, the results suggest that all the users found the system reasonably intuitive.

6.4.3 Users' Views on Further Development and Application

The open-ended questions were intended to elicit users' opinions and ideas about the potential of the system, and also to gauge, albeit subjectively, the degree to which the users appreciated the ways in which the system exploited the relationship between videos and collateral texts. The users' comments suggested that they felt KAB was useful as it was, and that the idea of annotating videos with collateral text was worth extending further.

In their answers to the open-ended questions, three users (1, 2 and 6) explicitly mentioned the link between texts and videos as being a good feature of the system. With regards to its functions for annotating and retrieving video data, two users (1 and 5) mentioned that it had good features, and another (4) noted the ease with which a video sequence could be found. The other user (3) commented on the interactive nature of the system as a plus point.

The criticisms of the system tended to concern the appearance of its interface which does not have a standard 'look and feel'. As well as following a standard for interface design, the system's user-friendliness would be improved by it being tailored to a specific domain, and to specific users, so that, for example, videos were referred to as 'dances' and the makers of moving images as 'choreographers'. The other repeated criticism concerned the discrepancy between the displayed time-codes of text fragments, and the visual action to which they related. This discrepancy is an artefact of the concurrent verbal reports used in the system because the expert necessarily speaks about what they have seen after it has passed by. Because there was no reliable way of determining the length of the lag in every case, it was decided to show the time at which each utterance was spoken: however, more care needs to be taken to explain this to users.

With regards to the potential use of the system, the Professor of Dance Studies thought that it could be used in teaching for developing on-line resources, and also as a means for pursuing research by producing personal analyses of moving images: it was even suggested that a collection of annotated videos and related texts (i.e. a KAB database) might constitute a new form of publication. Other users saw opportunities for applying the system in their own fields, e.g. to annotate sounds, and to annotate the graphical results of computer-based simulations. However, one user noted that since the system did not deal with filming and editing techniques its applicability was limited for certain kinds of moving image. Some other interesting suggestions for additional functionality included 'speech-to-text' and 'text-to-speech' for inputting and outputting collateral texts in time with video sequences; and, 'gesture input' whereby a user would perform a movement by way of making a query.

6.5 Discussion

At one level different media, like moving images and texts, appear as digital streams. For humans, mixing these media is fundamental to communication but it is currently a significant challenge for researchers in computing to characterise different kinds of media, and the relationships between them. In order to deal with moving images and texts and the link between them in a computational environment, we turned to the paradigm of object-orientation for the development of a knowledge-rich video annotation and browsing system.

Object-orientation contributed at each stage of KAB's development cycle. To develop programs in open standard languages, like Java, requires a considerable amount of effort, so for rapid prototyping Macromedia Director was used. This multimedia authoring system, which has an object-oriented basis (albeit proprietary and idiosyncratic), made it easy to present potential users with graphical user interfaces to encourage their feedback and further participation. The use of the Unified Modelling Language for designing the final system helped to state the attributes of video data files and text files, the link between them, and the operations that could be applied to them.

When implementing a system in an area with ever evolving standards, e.g. video coding standards, it is important to guard against obsolescence. The choice of Java for implementing KAB was considered a good way of making the system future-proof: the producers of Java are quick to add to its capabilities in response to new technologies, and successive versions of Java are backwardly-compatible. For example, the fact that the Java Media Framework deals with new video coding standards means that systems like KAB do not have to be changed; keeping up-to-date requires only the installation of the latest JMF.

The evaluation of KAB demonstrated its robustness at the hands of potential users and computing specialists, both groups appreciated its features for accessing visual and textual information. KAB demonstrates how the link between moving images and collateral texts can be realised in an object-oriented system by classes that associate text files with video data files. This link can then be exploited for accessing video data: firstly, by processing the text into machine executable surrogates for retrieval purposes; and secondly, by allowing the user

to browse between the moving image and explanatory texts. It was shown here how key terms identified in concurrent verbal reports can be used to label an interval of a video data file. More advanced processing of collateral texts might produce richer surrogates from a range of text types. As for browsing between moving images and collateral texts, KAB allows the user to jump from a time-coded text fragment to the corresponding point in the video. Further developments could allow the user to click on entities in the moving image and jump to related text, and to make texts scroll in time with the moving image.

Chapter 7

Closing Remarks

Textual information guides us through a world of sights, sounds and smells with, for example, photograph captions, television schedules, compact disc covers, deodorant labels, and more expansive texts like museum and art gallery catalogues, and press reviews of films, plays, concerts, restaurants and wines. We have concentrated on the specialist texts which are produced to elucidate moving images of dance. Such collateral texts are important for video annotation because there is information about moving images that is more readily accessible from the text than from video data. The phenomenon of moving images and their collateral texts is also an interesting focus for examining the link between vision and language.

We have considered theoretical questions to do with how experts put moving images into words. The application of methods from cognitive psychology and from linguistics allowed the characterisation of some cognitive and communicative processes involved in articulating descriptions and interpretations of moving images; these methods are based on the principled collection and analysis of texts. From a practical point of view, we have been concerned with how collateral texts can be used to access video data. The object-oriented KAB system allows users to build collections of moving images and texts: the texts are processed in order to label intervals of video data and are also made available to the viewer of moving images.

The development of the system to date, along with evidence about special language and verbal reporting, allows us to specify a knowledge-rich system. This specification is characterised by more extensive processing of collateral texts into machine-executable surrogates, and a terminology/knowledge-base to facilitate inferencing. Speculation about the development of the system in different application domains suggests a variety of interesting research questions.

This chapter elaborates the results of our theoretical and practical investigations, and their potential synergy (Section 7.1) and discusses the outlook for future research (Section 7.2).

7.1 Conclusions

Recent developments in hardware and software technologies mean that video data is widely available for use on computers but, paradoxically perhaps, the coding standards that enable the reliable transmission and high quality reproduction of moving images obscure many of their important features as understood by a human viewer. Surrogates must be associated with video data for it to be accessed effectively in digital libraries: on the one hand, surrogates are matched against users' queries to retrieve appropriate video sequences, and on the other hand, surrogates, like collateral texts that elucidate the moving image, may be read by the user. For retrieval purposes, mathematically-based visual features that capture colour, texture, shape and motion properties can be automatically generated from video data. The capture of 'semantic' features requires language-based annotations that may be automatically generated, from video data in limited cases, and more generally from collateral text.

A look at the state-of-the-art in computer vision technology suggested that systems generating language-based annotations may need to use collateral texts as a source, complementary to the visual component of video data. The use of so-called integral text, i.e. the spoken words of presenters in video sequences, has been widely explored, but this text often has an ad-hoc relationship with the content of the moving images. We have concentrated on collateral text that was produced specifically to elucidate moving images. The developers of two previous systems have used such 'external' text but discussed only short text fragments, and said little about the language used to produce the text. For us, a system that processes collateral text should be grounded in an appreciation of the link between vision and language.

What is of interest here is this: how do people understand moving images, and in particular how do they articulate their thoughts about them? It was perhaps beneficial to have constrained these questions by focusing on what we called specialist moving images. Such moving images are restricted in their intent, content, production and usage so they can be analysed systematically, at least by experts. We have argued that the systematic analysis of a set of restricted images leads to well organised knowledge and that this knowledge is reflected in a special language used to articulate analyses.

So what can we say about the knowledge of experts who analyse moving images? The aesthetic frameworks that experts follow suggest a degree of shared *procedural* knowledge, and the apparent proliferation of interrelated terminology suggests a conceptual framework for developing *declarative* knowledge. Both procedural and declarative knowledge may support the systematic description and interpretation of images. In their descriptions experts pick out salient entities and actions from a mass of visual information and sometimes group them in order to simplify their analysis. For their interpretations, experts draw on information which is *beyond the image*, in order to explain its meaning.

Mentions of 'experts', 'organised knowledge' and 'special language' resonate with parts of the knowledge acquisition literature; particularly the literature about knowledge acquisition from text. Perhaps then it is appropriate to view the task of attaching surrogates to video data (at least for specialist moving images) as involving the transfer of experts' knowledge to a machine. It was this viewpoint that motivated the investigation into a 'language of dance' and the application of verbal reporting techniques.

The examination of a corpus of dance texts found a range of evidence for the hypothesised 'language of dance'. An automated analysis highlighted statistically-based contrasts between the dance corpus and a collection of general language texts at lexical and collocational levels. Then a manual analysis of different text types showed how the intention of the text's author may determine the selection and organisation of information about a dance. Signs of an underlying conceptual structure for the domain were especially apparent in dictionary definitions of specialist movement terms. All these indications of order are encouraging for the enterprise of transferring dance experts' knowledge to a machine.

Some of a dance expert's thoughts about a moving image are in a sense 'ephemeral' because they are not usually articulated. When an expert watches a moving image they may be able to describe and interpret much more than is usually required for an analysis. The application of verbal reporting, due originally to Ericsson and Simon, showed that it was possible to elicit experts' descriptions and interpretations of image sequences in a systematic fashion. Contrasts were observed between these two kinds of verbal reports at lexical, syntactic and textual levels. The consistent information content and strong temporal

alignment with the moving image suggest verbal reports as especially good collateral texts for video annotation purposes.

The processing of collateral texts for video annotation requires the organisation of video data and texts, and the realisation of the link between them, in a computational environment. The KAB system was designed so that collections of moving images and texts could be built and so that the texts could be exploited for accessing video data. In the current implementation, keywords are identified in time-coded text fragments and are used to label intervals of video data: relevant texts are also made available to enable the user to browse between a text and a moving image. The system was evaluated with the concurrent verbal reports elicited from dance experts and users seemed to appreciate the video-text link for the annotation, retrieval and browsing of video sequences.

Annotation, retrieval and browsing in the KAB system depend upon lists of terms and other lexical items. The compilation of these lists benefited from the investigations of special language and of verbal reporting. The computation of linguistic variance at a lexical level in the dance corpus extracted some terminology which was used for labelling video sequences: and the analysis of verbal reports pointed to 'interpretation cues' which were highlighted in texts when a user browsed them. In turn, the investigation of communicative and cognitive processes through corpus linguistics and verbal reporting requires the organisation and systematic analysis of text collections alongside moving images. These examples are perhaps indications of the synergy between the theoretical and practical aspects of this research.

7.2 Future Work

This thesis may be seen as a critique of video annotation which has considered how images are described and interpreted, the structure of the language in which such analyses are articulated, and methods by which they can be elicited through verbal reporting. The approach to video annotation which has been advocated here is different to other approaches in the field in that ideas relating to special language and specialist images were explored through an object-oriented system. The development of this system to date suggests to us a functional specification of a knowledge-rich video annotation system:

- **Text Gathering.** The system should *collect extant texts* from the WWW and other sources and facilitate the *elicitation of verbal reports* with speech recognition technology.
- **Text and Video Organisation.** The system should include a *text database* and a *video database* to allow the creation of virtual hierarchies of texts and videos according to user-specified attributes.
- **Text Analysis / Natural Language Processing.** The system should exploit *text analysis* techniques for terminology extraction; definition analysis; proper noun elaboration; and summarisation. More generally it should integrate established technologies for *information retrieval*, *information extraction* and *natural language understanding*.
- **Machine-executable Surrogates.** The system should attach a *variety of surrogates* to image sequences, including: visual features; individual key terms; key term vectors; and structured surrogates that capture, for example, the relationships between entities and actions in space and time, causal relationships, and the structure of interpretations. (Cf. *knowledge representation*).
- **Terminology/Knowledge-base.** The system should have access to *knowledge about domain artefacts* including specialist movement terms, theoretical constructs and the people involved in producing images. This knowledge should be used to make inferences, e.g. for *query expansion*.
- **Video Processing.** The system should incorporate algorithms for the computation of visual features to allow retrieval by *visual similarity*, and to encourage 'fusion' with semantic features.
- **Querying.** The system should allow the user to make language-based queries and visual queries for retrieving video sequences, and to refine their queries in the light of retrieved sequences.
- **Browsing.** The system should present video data and collateral texts in a co-ordinated fashion so that the user can 'jump' between moving images and texts. (Cf. *hypertext/hypermedia* / MHEG standard).

An architecture for the proposed system is shown below, Figure 7.1.

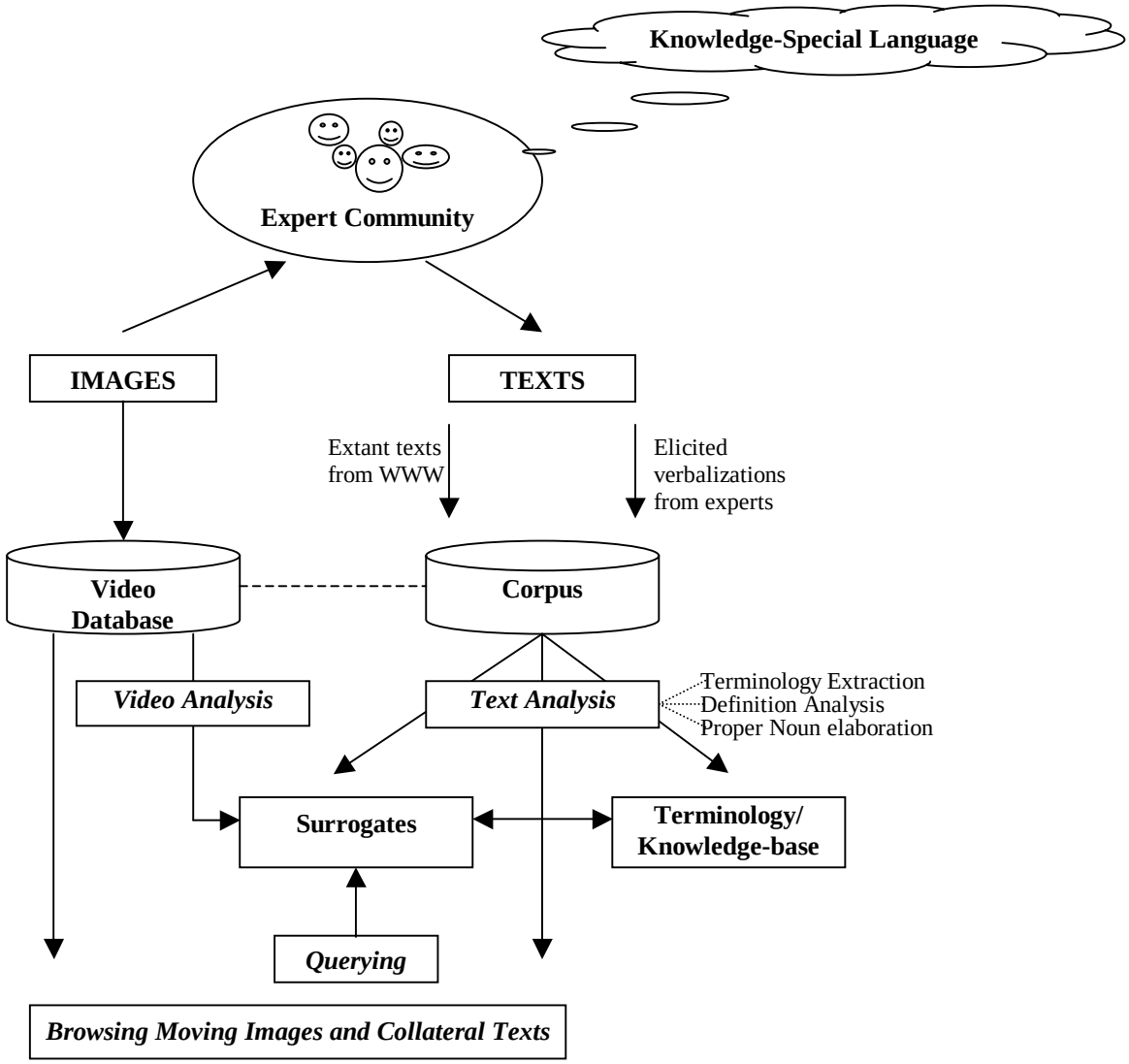


Figure 7.1: The architecture for a proposed knowledge-rich video annotation system. Texts are gathered and organised in a corpus where they are associated with video sequences. The texts are analysed on the one hand, to generate machine-executable surrogates for video retrieval, and on the other hand, to develop a terminology/knowledge-base. The terminology/knowledge-base supports the retrieval process by providing inferencing capabilities. Video surrogates can also be generated by analysing video data directly. The user accesses the video data by making queries that are matched against surrogates, and by browsing through moving images and texts that are presented in a co-ordinated fashion.

There are a number of domains in which knowledge-rich image and video systems could be developed, as well as dance: the variety of these domains provides interesting challenges for the endeavour, and all appear to have a need for systems to access image and video data. Consider, remaining in the aesthetic realm, images of fine art and art-house films; and, moving to the scientific realm, consider still and moving images from the microscopes of cell biologists, medical X-rays and scans, images taken at the scene of a crime, and potentially

vast collections of archaeological and historical artefacts. In all these examples, specialist images are analysed by experts who articulate descriptions and interpretations.

There is one other example to be mentioned here which involves specialists, although the challenges in this case are less constrained because the specialists analyse everyday images. Recent advances in digital technology mean that television broadcasts can include extra soundtracks, including one to transmit an *audio description* of the moving image for the visually impaired. These audio descriptions are scripted by trained individuals who follow guidelines that advise on how to describe (and at times interpret) a range of programmes, including soap operas, documentaries and films. In the first instance such text could be used for accessing video data; further on, it may be that investigations into the link between moving images and collateral texts can ease the burden of producing the audio descriptions.

The development of systems that are based on the link between images and collateral texts may benefit from, and contribute to, theoretical investigations of vision and language. There are many questions that can be asked about the link between the two. Leaving aside, for now, interesting debates from aesthetics and semiotics about the differences between image and text, we consider questions that might be asked based on collections of collateral texts.

Scholars of special language who examine text corpora have contributed to debates about the development of knowledge in specialisms. Maybe then the examination of a text corpus associated with moving images, especially a diachronic corpus, will shed light on the development of knowledge about images and about how they can be analysed. Whilst special language corpora are often studied in terms of a register variance, i.e. field of discourse, they can also be studied on other planes. A collection of texts which described images in a variety of domains would maybe reflect a 'language of description' with common linguistic features used by experts in different specialisms. Some of these features might relate to the need to convey a sense of movement when describing moving images, and so we might expect texts to follow moving images in their temporal order, and to have higher than normal numbers of verbs and temporal and spatial prepositions. As for an analogous 'language of interpretation', maybe it would reflect the need to go beyond the literal contents of an image with a high number of unconventional lexical choices, interpretation cues and metaphors.

Whereas corpus-based studies normally deal with extant texts that are sophisticated in their mixture of information about a topic, verbal reporting can be used to isolate aspects of an expert's analysis. From the point of view of cognitive psychology, verbal reports produced as an expert watches a moving image might give insight into the transformation of visual information into conceptual information: issues here would include attention, recognition of objects and actions, and the organisation of conceptual information. From the point of view of language production, verbal reports can be examined to characterise the selection of information, and the lexical and syntactic choices made by the speaker.

Reiterating the point, we note that the enhanced understanding of how people put moving images into words can contribute, and benefit from, the development of systems in which video data and collateral texts are organised and processed alongside one another. And then, why stop at images and collateral texts, vision and language – what about other combinations of media and modes? As noted at the start of the chapter, textual information appears to be collateral with sounds, smells and tactile sensations, as well as images, and often more than one of these at a time. Although visual, auditory, olfactory and tactile sensations can be reduced to digital streams, there remains an important challenge in "providing an integration of representations – a media interlingua – or at least access to heterogeneous media representations" (Maybury 1997:xxi). So perhaps there is a digital future of 'collateral text with everything', or 'collateral media' if we do not want to privilege language.

Bibliography

Adshead (1988). Janet Adshead (ed.), *Dance Analysis: Theory and practice*. London: Dance Books.

Adshead-Lansdale (1999). Janet Adshead-Lansdale (ed.), *Dancing Texts: Intertextuality in Interpretation*. London: Dance Books.

Ahanger (1999). Gulrukh Ahanger, *Techniques for Automatic Digital Video Composition*. PhD Thesis, Boston University.

Ahmad (1999). Khurshid Ahmad, 'Linguistic Variance in English Special Language Texts.' Submitted to *Terminology*.

Ahmad et al. (1999). Khurshid Ahmad, Tracey Bale, Darren Burford, W. M. van den Bergh and M. H. van Enschot, 'Choosing Codebooks for Self-Organising Maps.' Accepted for publication in *Neural Computing and Applications*.

Ahmad, Salway and Adshead-Lansdale (1998). Khurshid Ahmad, Andrew Salway and Janet Adshead-Lansdale, '(An)notating Dance: multimedia storage and retrieval.' In: Henry Selvaraj and Brijesh Verma (eds.), *ICCIMA '98* - Proceedings of the International Conference on Computational Intelligence and Multimedia Applications, Victoria, Australia, 9-11 February 1998, pp. 788-793. Singapore: World Scientific.

Ahmad, Salway and Adshead-Lansdale (forthcoming). Khurshid Ahmad, Andrew Salway and Janet Adshead-Lansdale. 'The Moving (Human) Image and its Annotation', submitted by invitation to *Image and Vision Computing*.

Aigrain, Zhang and Petkovic (1996). Philippe Aigrain, HongJiang Zhang and Dragutin Petkovic, 'Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review.' *Multimedia Tools and Applications* 3, pp. 179-202.

Al-Jabir (1999). Shaikha Al-Jabir, *Terminology-Based Knowledge Acquisition*. PhD thesis, University of Surrey.

Allen (1983). J. F. Allen, 'Maintaining Knowledge About Temporal Intervals.' *Communications of the ACM* 26 (11), pp. 832-843.

Badler and Smoliar (1979). Norman I. Badler and Stephen W. Smoliar, 'Digital Representations of Human Movement.' *Computing Surveys* 11 (1), pp. 19-38.

Baecker et al. (1995). Ronald M. Baecker, Jonathan Grudin, William A. S. Buxton and Saul Greenberg (eds.), *Readings in Human-Computer Interaction: Toward the Year 2000*. 2nd Edition, San Francisco CA: Morgan Kaufmann.

Barnbrook and Sinclair (1995). Geoff Barnbrook and John Sinclair, 'Parsing Cobuild Entries.' In: John Sinclair, Martin Hoelter and Carol Peters (eds.), *The Language of Definition: The Formalisation of Dictionary Definitions for Natural Language Processing*, pp. 13-58. Luxembourg: European Commission.

Beaumont (1949). Cyril W. Beaumont, *Dancers Under my Lens: essays in ballet criticism*. London: Beaumont.

Beaumont (1952/1982). Cyril W. Beaumont, *The Ballet Called Swan Lake*. New York: Dance Horizons.

Biber, Conrad and Reppen (1998). Douglas Biber, Susan Conrad and Randi Reppen, *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bobick (1997). Aaron F. Bobick, 'Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion.' *Philosophical Transactions of the Royal Society of London* Series B – Biological Sciences 352 (1358), pp. 1257-1265.

Bordwell and Thompson (1997). David Bordwell and Kristin Thompson, *Film Art: an introduction*. 5th Edition, New York: McGraw-Hill.

Bove (1996). V. M. Bove, 'Multimedia based on object models: Some whys and hows.' *IBM Systems Journal* 35 (3&4), pp. 337-348.

Calvert (1986). Thomas W. Calvert, 'Toward a Language for Human Movement.' *Computers and the Humanities* 20, pp. 35-43.

Campbell and Bobick (1995). Lee Campbell and Aaron F. Bobick, 'Recognition of Human Body Motion Using Phase Space Constraints.' M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 309. (Abbreviated version appears in ICCV '95).

Carter (1998). Alexandra Carter (ed.), *The Routledge Dance Studies Reader*. London and New York: Routledge.

Cédras and Shah (1995). Claudette Cédras and Mubarak Shah, 'Motion-based recognition: a survey.' *Image and Vision Computing* 13 (2), pp.129-155.

Chafe (1980). Wallace L. Chafe (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood NJ: Ablex.

Chang et al. (1998). Shih-Fu Chang, William Chen, Horace J. Meng, Hari Sundaram and Di Zhong, 'A Fully Automated Content-Based Video Search Engine Supporting Spatiotemporal Queries.' *IEEE Transactions on Circuits and Systems for Video Technology* 8 (5), pp. 602-615.

Chang et al. (1999). Shih-Fu Chang, Qian Huang, Thomas Huang, Atul Puri and Behzad Shahraray, 'Multimedia Search and Retrieval.' In: A. Puri and T. Chen (eds.), *Advances in Multimedia: Systems, Standards and Networks*. New York: Marcel Dekker.

Chang, Chen and Sundaram (1998). Shih-Fu Chang, William Chen and Hari Sundaram, 'Semantic Visual Templates: Linking Visual Features to Semantics.' In: *Proceedings of IEEE International Conference on Image Processing*, Chicago, October 1998.

Christel et al. (1996). Michael Christel, Scott Stevens, Takeo Kanade, Michael Maudlin, Raj Reddy and Howard Wactlar, 'Techniques for the Creation and Exploration of Digital Video Libraries.' In: Furht, pp. 283-327.

Cohen and Feigenbaum (1982). Paul R. Cohen and Edward A. Feigenbaum, *The Handbook of Artificial Intelligence*. Reading MA: Addison-Wesley.

Corridoni et al. (1996). Jacopo M. Corridoni, Alberto Del Bimbo, Dario Lucarella and He Wenxue, 'Multi-perspective Navigation of Movies.' *Journal of Visual Languages and Computing* 7, pp. 445-466.

Davenport, Aguierre Smith and Pincever (1991). Glorianna Davenport, Thomas Aguierre Smith and Natalio Pincever, 'Cinematic Primitives for Multimedia.' *IEEE Computer Graphics and Applications* July 1991, pp. 67-74.

Davis (1995). Marc Eliot Davis, 'Media Streams: An Iconic Visual Language for Video Representation.' In: Baecker et al., pp. 854-866. (For more details, see *Media Streams: Representing Video for Retrieval and Repurposing*. PhD Thesis, Massachusetts Institute of Technology).

de Jong (1998). Franciska de Jong, 'Human Language as ELSNET Interlingua: Intelligent Multimedia Indexing.' *Proceedings ELSNET in Wonderland*, pp. 51-57. Utrecht: ELSNET.

de Marinis (1993). Marco de Marinis, *The Semiotics of Performance*. Bloomington and Indianapolis: Indiana University Press.

Del Bimbo, Vicario and Zingoni (1995). Alberto Del Bimbo, Enrico Vicario and Daniele Zingoni, 'Symbolic Description and Visual Querying of Image Sequences Using Spatio-Temporal Logic.' *IEEE Transactions on Knowledge and Data Engineering* 7 (4), pp. 609-621.

Downing (1980). Pamela Downing, 'Factors Influencing Lexical Choice in Narrative.' In: Chafe (1980), pp. 89-126.

Eco (1979). Umberto Eco, *The Role of the Reader*. Bloomington: Indiana University Press.

Eco (1995). Umberto Eco, *The Search for the Perfect Language*. Oxford and Cambridge MA: Blackwell.

Enser (1995). P. G. B. Enser, 'Pictorial Information Retrieval.' *Journal of Documentation* 51 (2), pp. 126-170.

Ericsson and Simon (1993). K. Anders Ericsson and Herbert A. Simon, *Protocol Analysis: Verbal Reports as Data*. Revised Edition, Cambridge MA and London: The MIT Press.

Eriksson and Penker (1998). Hans-Erik Eriksson and Magnus Penker, *UML Toolkit*. New York: John Wiley & Sons.

Fellbaum (1998). Christiane Fellbaum (ed.), *WordNet: an Electronic Lexical Database*. Cambridge MA and London: The MIT Press.

Fischler and Firschein (1987). Martin A. Fischler and Oscar Firschein, *Intelligence: The Eye, the Brain, and the Computer*. Reading MA: Addison-Wesley.

Flickner et al. (1997). Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele and Peter Yanker, 'Query by Image and Video Content: The QBIC System.' In: Maybury, pp. 7-22.

Foster (1986). Susan Leigh Foster, *Reading Dancing*. Berkeley CA: University of California Press.

Furht (1996). Borko Furht (ed.), *Multimedia Tools and Applications*. Boston MA: Kluwer Academic.

Furht (1999). Borko Furht (ed.), *Handbook of Multimedia Computing*. Florida: CRC Press.

Furht, Smoliar and Zhang (1995). Borko Furht, HongJiang Zhang and Stephen W. Smoliar, *Video and Image Processing in Multimedia Systems*. Boston MA: Kluwer Academic.

Gavrila (1999). D. M. Gavrila, 'The Visual Analysis of Human Movement: A Survey.' *Computer Vision and Image Understanding* 73 (1), pp. 82-98.

Gläser (1993). Rosemarie Gläser, 'A Multi-level Model for a Typology of LSP Genres.' *Fachsprache: the International Journal of LSP* 1993 (1-2), pp. 18-26.

Golshani and Dimitrova (1998). Forouzan Golshani and Nevenka Dimitrova, 'A Language for Content-Based Video Retrieval.' *Multimedia Tools and Applications* 6, pp. 289-312.

Gong et al. (1996). Yihong Gong, Chua Hock Chuan, Zhu Yongwei and Masao Sakauchi, 'A Generic Video Parsing System with a Scene Description Language (SDL).' *Real-Time Imaging* 2, pp. 45-59.

Haase (1996). Kenneth Haase, 'FramerD: Representing knowledge in the large.' *IBM Systems Journal* 35 (3&4), pp. 381-397.

Halliday (1994). M. A. K. Halliday, *An Introduction to Functional Grammar*. 2nd Edition, London: Edward Arnold.

Halliday and Martin (1993). M. A. K. Halliday and J. R. Martin, *Writing Science: Literacy and Discursive Power*. London: The Falmer Press.

Hampapur and Jain (1998). Arun Hampapur and Ramesh Jain, 'Video Data Management Systems: Metadata and Architecture.' In: Amit Sheth and Wolfgang Klas (eds.), *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*, pp. 245-286. New York: McGraw-Hill

Hampapur, Jain and Weymouth (1996). Arun Hampapur, Ramesh Jain and Terry E. Weymouth, 'Production Model Based Digital Video Segmentation.' In: Furht, pp. 111-153).

Hanna (1994). Judith Hanna, 'Dance.' In: Thomas A. Sebeok (ed.), *Encyclopedic Dictionary of Semiotics*. 2nd Edition, Berlin and New York: Mouton de Gruyter.

Haykin (1994). Simon Haykin, *Neural Networks: a comprehensive foundation*. New York: Macmillan.

Hiemstra, de Jong and Netter (1998). Djoerd Hiemstra, Franciska de Jong and Klaus Netter (eds.), *Proceedings of the 14th Twente Workshop on Language Technology – Language Technology for Multimedia Information Retrieval*. Enschede: University of Twente.

Hoffman (1984). Lothar Hoffman, 'Seven Roads to LSP.' *Fachsprache: the International Journal of LSP* 1984 (1-2), pp. 28-38.

Hopper, Owens and Croll (1999). Richard Hopper, Carol Owens and Mike Croll, 'Achieving Full Media Interoperability Using Information Systems and Indexing Schemes.' In: Proceedings of IEE Colloquium on Multimedia Databases and MPEG-7, London, 29 January 1999, pp. 8/1-8/7.

Hutchinson Guest (1984). Ann Hutchinson Guest, *Dance Notation: the process of recording movement on paper*. London: Dance Books.

Jones et al. (1997). Gareth Jones, Jonathan Foote, Karen Sparck Jones and Steve J. Young, 'The Video Mail Retrieval Project: Experiences in Retrieving Spoken Documents.' In: Maybury, pp. 191-214.

Khoshafian and Abnous (1995). Setrag Khoshafian and Razmik Abnous, *Object Orientation*. 2nd Edition, New York: John Wiley & Sons.

Kim and Shibata (1996). Yeun-Bae Kim and Masahiro Shibata, 'Content-Based Video Indexing and Retrieval – A Natural Language Approach.' *IEICE Transactions on Information and Systems* E79-D (6), pp. 695-705.

Krueger (1991). Myron, W. Krueger, *Artificial Reality II*. 2nd Edition, Reading MA: Addison-Wesley.

Landau (1989). Sidney I. Landau, *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Lindley and Srinivasan (1998). C. A. Lindley and U. Srinivasan, 'Query Semantics for Content-Based Retrieval of Video Data: an Empirical Investigation.' To appear in: *Proceedings of Storage and Retrieval Issues in Image and Multimedia Databases, DEXA '98*.

Lodge (1988). David Lodge (ed.), *Modern Criticism and Theory*. London and New York: Longman.

Mackrell (1997). Judith Mackrell, *Reading Dance*. London: Michael Joseph.

Mandal, Idris and Panchanathan (1999). M. K. Mandal, F. Idris and S. Panchanathan, 'A critical evaluation of image and video indexing techniques in the compressed domain.' *Image and Vision Computing* 17, pp. 513-529.

Mani et al. (1997). Inderjeet Mani, David House, Mark T. Maybury and Morgan Green, 'Towards Content-Based Browsing of Broadcast News Video.' In: Maybury, pp. 241-258.

Margolis (1980). Joseph Margolis, *Art and Philosophy*. Brighton: Harvester Press.

Marr (1982). David Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman.

Maybury (1997). Mark T. Maybury, *Intelligent Multimedia Information Retrieval*. Menlo Park CA / Cambridge MA: AAAI Press / The MIT Press.

McKeown et al. (1998). Kathleen R. McKeown, Steven K. Feiner, Mukesh Dalal and Shih-Fu Chang, 'Generating multimedia briefings: coordinating language and illustration.' *Artificial Intelligence* 103, pp. 95-116.

Metz (1974). Christian Metz, *Film Language: a semiotics of the cinema*. New York: Oxford University Press.

Minka and Picard (1997). T. P. Minka and R. W. Picard, 'Interactive Learning with a "Society of Models".' *Pattern Recognition* 30 (4), pp. 565-581.

Minsky (1975). Marvin Minsky, 'A Framework for Representing Knowledge.' In: Patrick Winston (ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill.

Nack and Parkes (1997). Frank Nack and Alan Parkes, 'Toward the Automated Editing of Theme-Oriented Video Sequences.' *Applied Artificial Intelligence* 11, pp. 331-366.

Nagel (1988). Hans-Hellmut Nagel, 'From image sequences towards conceptual descriptions.' *Image and Vision Computing* 6 (2), pp. 59-74.

Nagel (1994). Hans-Hellmut Nagel, 'A Vision of "Vision and Language" Comprises Action: An Example from Road Traffic.' *Artificial Intelligence Review* 8, pp. 189-214.

Nakamura and Kanade (1997). Yuichi Nakamura and Takeo Kanade, 'Semantic Analysis for Video Contents Extraction – Spotting by Association in News Video.' *ACM Multimedia 97* – Electronic Proceedings, http://www.uni-mannheim.de/acm97/papers/nakamura/main.html .

Netter (1998). Klaus Netter, 'POP-EYE and OLIVE – Human Language as the Medium for Cross-lingual Multimedia Information Retrieval.' *The ELRA Newsletter (European Language Resources Association)* November 1998, pp. 5-6.

Oomoto and Tanaka (1993). Eitetsu Oomoto and Katsumi Tanaka, 'OVID: Design and Implementation of a Video-Object Database System.' *IEEE Transactions on Knowledge and Data Engineering* 5 (4), pp. 629-643.

Owen and Makedon (1999). Charles B. Owen and Fillia Makedon, 'Cross-Modal Information Retrieval.' In: Furht, pp. 403-423.

Panofsky (1939/1970). Erwin Panofsky, *Meaning in the Visual Arts*. Harmondsworth: Penguin.

Parkes (1989). Alan P. Parkes, 'The Prototype CLORIS System: Describing, Retrieving and Discussing Videodisc Stills and Sequences.' *Information Processing and Management* 25 (2), pp. 171-186.

Pentland (1997). Alex Pentland, 'Content-based indexing of images and video.' *Philosophical Transactions of the Royal Society of London* Series B – Biological Sciences 352 (1358), pp. 1283-1290.

Plant and Smith (1997). Darrel Plant and Doug Smith, *The Lingo Programmer's Reference*. Ventana Communications.

Quirk et al. (1985). Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik, *A Comprehensive Grammar of the English Language*. London and New York: Longman.

Robin (1992). Harry Robin, *The Scientific Image: From Cave to Computer*. New York: W. H. Freeman.

Rorvig (1990). Mark E. Rorvig, 'Introduction to a Special Issues on Intellectual Access to Graphic Information.' *Library Trends* 38 (4), pp. 639-643.

Roth (1999). Volker Roth, 'Content-based retrieval from digital video.' *Image and Vision Computing* 17, pp. 531-540.

Rowe, Boreczky and Eads (1994). Lawrence A. Rowe, John S. Boreczky and Charles A. Eads, 'Indexes for User Access to Large Video Databases.' *SPIE* (International Society for Optical Engineering) Vol. 2185, pp. 150-161.

Sager, Dungworth and McDonald (1980). Juan C. Sager, David Dungworth and Peter F. McDonald, *English Special Languages: Principles and practice in science and technology*. Wiesbaden: Brandstetter.

Satoh, Nakamura and Kanade (1999). S. Satoh, Y. Nakamura and T. Kanade, 'Name-it: Naming and detecting faces in news videos.' *IEEE Multimedia* 6 (1), pp. 22-35.

Sawhney, Balcom and Smith (1997). Nitin Sawhney, David Balcom and Ian Smith, 'Authoring and Navigating Video in Space and Time.' *IEEE Multimedia* (October-December), pp. 30-39.

Schachlbauer and Weiss (1998). H. Schachlbauer and S. M. Weiss, 'EBU / SMPTE Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams – Final Report: Analysis and Results.' *SMPTE Journal* 107 (9), pp. 605-815.

Schank (1975). Roger C. Schank, 'Identification of Conceptualizations Underlying Natural Language.' In: Schank and Colby (eds.), *Computer Models of Thought and Language*, pp. 187-247. San Francisco: W. H. Freeman and Company.

Shahraray (1999). Behzad Shahraray, 'Multimedia Information Retrieval Using Pictorial Transcripts.' In: Furht, pp. 345-359.

Shatford (1986). S. Shatford, 'Analyzing the Subject of a Picture: a Theoretical Approach.' *Cataloging and Classification Quarterly* 6 (3), pp. 39-62.

Smadja (1994). Frank Smadja, 'Retrieving Collocations from Text.' In: Susan Armstrong-Warwick (ed.), *Using Large Corpora*, pp. 143-177. Cambridge MA and London: The MIT Press.

Small (1996). Peter Small, *Lingo Sorcery: the magic of lists, objects and intelligent agents*. Chicester: John Wiley & Sons.

Smith and Chang (1997). John R. Smith and Shih-Fu Chang, 'Visually Searching the Web for Content.' *IEEE Multimedia* July-September, pp. 12-20.

Smith and Smith (1977). R. A. Smith and C. M. Smith, 'The Artworld and Aesthetic Skills: A Context for Research and Development.' *Journal of Aesthetic Education* 11 (2), pp. 117-132.

Sonka, Hlavac and Boyle (1993). Milan Sonka, Vaclav Hlavac and Roger Boyle, *Image Processing, Analysis and Machine Vision*. London: Chapman and Hall.

Sparck Jones and Willett (1997). Karen Sparck Jones and Peter Willett (eds.), *Readings in Information Retrieval*. San Francisco CA: Morgan Kaufmann.

Srihari (1995). Rohini K. Srihari, 'Computational Models for Integrating Linguistic and Visual Information: A Survey.' *Artificial Intelligence Review* 8 (5-6), pp. 349-369.

Still and Worton (1990). Judith Still and Michael Worton, 'Introduction.' In: Worton and Still (eds.) *Intertextuality: Theories and Practices*, pp. 1-44. Manchester and New York: Manchester University Press.

Stolnitz (1960). J. Stolnitz, *Aesthetics and the Philosophy of Art Criticism*. Boston MA: Houghton Miffin.

Subrahmanian (1998). V. S. Subrahmanian, *Principles of Multimedia Database Systems*. San Francisco CA: Morgan Kaufmann.

Svartvik (1992). Jan Svartvik (ed.), *Directions in Corpus Linguistics* (Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991). Berlin and New York: Mouton de Gruyter.

Takeshita, Inoue and Tanaka (1997). Atsushi Takeshita, Takafumi Inoue and Kazuo Tanaka, 'Topic-based Multimedia Structuring.' In: Maybury, pp. 259-277.

Tanaka, Ariki and Uehara (1999). Katsumi Tanaka, Yasuo Ariki and Kuniaki Uehara, 'Organization and Retrieval of Video Data.' *IEICE Transactions on Information and Systems* E82-D (1), pp. 34-44.

Tannen (1980). Deborah Tannen, 'A Comparative Analysis of Oral Narrative Strategies: Athenian Greek and American English.' In: Chafe, pp. 51-87.

Torres and Kunt (1996). Luis Torres and Murat Kunt (eds.), *Video Coding: The Second Generation Approach*. Boston MA: Kluwer Academic.

Wactlar et al. (1999). Howard D. Wactlar, Michael G. Christel, Yihong Gong and Alexander G. Hauptmann, 'Lessons Learned from Building a Terabyte Digital Video Library.' *Computer* February 1999, pp. 66-73.

Weise (1993). Günter Weise, 'Criteria for the Classification of Texts.' *Fachsprache: the International Journal of LSP* 1993 (1-2), pp. 26-31.

Weiss, Duda and Gifford (1995). Ron Weiss, Andrzej Duda and David K. Gifford, 'Composition and Search with a Video Algebra.' *IEEE Multimedia* Spring 1995, pp. 12-25.

Wilks, Slator and Guthrie (1996). Yorick Wilks, Brian Slator and Louise Guthrie, *Electric Words: dictionaries, computers, and meanings*. Cambridge MA and London: The MIT Press.

Winston (1992). Patrick Henry Winston, *Artificial Intelligence*, 3rd Edition. Reading MA: Addison-Wesley.

Yacoob and Black (1999). Yaser Yacoob and Michael J. Black, 'Parameterized Modeling and Recognition of Activities.' *Computer Vision and Image Understanding* 73 (2), pp. 232-247.

Zhang et al. (1997). HongJiang Zhang, Jianhua Wu, Di Zhong and Stephen W. Smoliar, 'An Integrated System for Content-Based Video Retrieval and Browsing.' *Pattern Recognition* 30 (4), pp. 643-658.

Appendices

Appendix A: Clustering and Segmenting Verbal Reports

This appendix reports a preliminary investigation into the use of neural networks for clustering and segmenting moving images on the basis of keyword vectors generated from collateral texts. There are signs that, at least for the concurrent verbal reports spoken by dance experts, this approach may be useful for clustering whole moving images, and for delimiting important sequences within them. Since, like any other learning technique, neural networks are only as good as the data with which they are provided, the signs of success here may be taken as support for the claim that verbal reports provide reliable sources of information about moving images.

A.1 Clustering Verbal Reports on a Self-Organising Map

Texts that describe and interpret moving images comprise words that refer to entities and actions that are depicted by the moving image, and to their significance and meanings. The occurrence of certain words may then be indicative of particular kinds of moving images: as such, frequency counts of important words may convey the content of moving images as a feature vector. These feature vectors can then be processed by statistical pattern recognition techniques, including some artificial neural networks, in order to ascertain degrees of similarity between vectors, and hence between moving images. The use of text feature vectors for indexing still and moving images has been explored by McKeown et al. (1998) and Ahanger (1999).

Kohonen's self-organising feature map is an artificial neural network that reduces the dimensionality of a data set by processing feature vectors from an input layer of nodes onto a 1D or 2D lattice of output nodes. A trained Kohonen map is useful for visualising clusters in multi-dimensional data, and for classifying further feature vectors (Haykin 1994: 397-424). The input nodes are fully-connected to the output nodes: furthermore, there are lateral connections between neighbouring nodes in the output layer. The training of the network

proceeds in an unsupervised fashion, whereby the weights of the output node responding most strongly to the current input are adjusted so that it will respond even more strongly to that particular input. The weights of nodes neighbouring the winner are also updated in order to encourage clustering: the weights of the closest neighbours on the lattice are adjusted so that they respond more strongly toward the current input; the weights of those more distant are adjusted so they respond more weakly.

Once the network has been trained each feature vector in the training set can be seen to have its own winning node in the output layer of the map. The 'principle of topographic map formation', which characterises this network, means that vectors with proximate output nodes are those that have features which make them similar in some respects (ibid.:400-1). Thus, it is possible to 'read off' clusters of vectors which may pertain to classifications of the input data.

The importance of carefully selecting words to form feature vectors for a set of texts, and a method for doing this were discussed by Ahmad et al. (1999). This method, which is based on the idea of 'weird' words, was used here to automatically generate feature vectors for the concurrent descriptions and interpretations of moving images elicited from dance experts. As discussed in Chapter 4, 'weirdness' is a function of the relative frequency of a word in a set of texts divided by its relative frequency in a general language sample. The appropriateness of these words for feature vectors was demonstrated when trained Kohonen maps were shown to cluster texts relating to the same sequences of moving images, and also to cluster texts according to their speaker.

Method for Generating Text Feature Vectors and Training SOM

Feature vectors were automatically generated, first for the 25 descriptions of moving images elicited from five experts watching five dance sequences, and then for their subsequent interpretations. The feature vectors were formed from the 100 'weirdest' words in the set of texts, with a frequency greater than 2. The weirdness criterion ensures the selection of words that are particular to the set of texts, and should thus be able to discriminate between them

more meaningfully; and, the frequency threshold reduces 'noise' and makes it more likely that a selected word occurs in more than one text.

The words automatically selected to form feature vectors for the descriptions and the interpretations are shown respectively in Tables A.1a and A.1b. For the descriptions these words mainly refer to body parts and to movements and movement qualities, using some specialist terminology: the words for the interpretations also include words that are less literal in this context, such as *wings*, *underwater* and *ritual*.

aerial	angular	arabesque	arabesques	arched	arching	armpits
arms	backwards	balance	balances	balancing	balletic	beak
character	clasping	counter	couples	curves	curving	dancer
dancers	demi	deux	diagonal	duet	duo	elbows
extending	flexed	footwork	foreground	forwards	fourth	gestural
gesture	gestures	groupings	hops	hug	hyper	interlocking
isolations	jerky	jeté	jumping	jumps	kicks	kneeling
leans	leaps	leg	lifting	lifts	lots	lunge
lyrical	maintains	male	manipulating	motif	movements	moving
nifty	partner	partnering	pas	pedestrian	pelvis	pirouettes
plié	pushes	quivering	slicing	slow	solo	spins
sporadic	stepping	stroking	supporting	supports	sustained	swan
swans	tilts	torso	torsos	touching	trio	turns
twisting	twitching	underneath	undulating	unison	unisonal	upstage
upwards	wrists					

Table A.1a: The hundred 'weirdest' words with frequencies greater than 2, automatically selected to form feature vectors for texts describing dance sequences.

aerial	aggressive	arms	backwards	balancing	ballet	beak
birds	bodies	camera	caring	couples	curve	curved
curving	dance	dancer	dancers	deux	distorted	duet
echoing	empathy	enjoying	erotic	fantasy	female	flexed
footwork	foreground	forwards	gender	geometric	gestural	gesture
gestures	holding	hugging	imagery	images	independently	intimate
jerky	jumping	jumps	kneeling	leans	leaps	leg
lifting	lifts	longing	lots	male	males	manipulating
manipulative	mating	movement	movements	moves	moving	ok
pas	penguin	picking	possibly	potentially	prince	pulling
pushing	references	relationship	repeating	repetition	ritual	sensual
sinister	solo	static	stillness	stretching	suggesting	supporting
supports	swan	swans	touches	touching	twisting	twitching
twittering	underwater	undulating	unison	upwards	walks	whereby
wings	yearning					

Table A.1b: The hundred 'weirdest' words with frequencies greater than 2, automatically selected to form feature vectors for texts interpreting dance sequences.

The feature vector for each text thus comprised 100 elements, each computed from a word frequency. The absolute frequency, *f*, was normalised to account for the length of the moving image that the text described or interpreted (1); and, a further logistic function was used to ensure that the value fell between 0 and 1 (2). The second function was used because most frequencies were less than 4: a logistic function can differentiate these low frequencies well at the expense of the less important differences between high value frequencies.

$$(1) \quad f_1 = f * \left(\frac{length\_of\_longest\_moving\_image\_in\_set}{length\_of\_current\_moving\_image} \right)$$

$$(2) \quad f_2 = \left(\left(\frac{1}{(1+\exp(-f_1))} \right) - 0.5 \right) * 2$$

The Stuttgart Neural Network Simulator (SNNS)[8] was used to train self-organising feature maps with 10x10 output layers. For training maps with the descriptions, the important parameters were set as: initial learning rate = 0.2; initial neighbourhood size = 6; and, rate of decrease for these = 0.999 – that is to say, the values were multiplied by 0.999 after every pattern was presented. In order to get good results for the interpretations, it was necessary to use a smaller neighbourhood size of 4, and to have the learning rate decreasing more slowly, at 0.99, and the neighbourhood size more quickly, at 0.9999. These parameters helped the networks to avoid the edges of the map and gave them better chances for good clustering.

Results of Training SOM: clusters by dance and by author

The maps that were trained with the descriptions showed good clusters after 100 cycles; for the interpretations, it was necessary to do three lots of 100 cycles, resetting the learning rate and the neighbourhood size at the beginning of each lot. For all the texts, it was found that the location of feature vectors on the maps reflected both the dance that the text was collateral to, and the expert who spoke the commentary.

It can be seen from exemplar maps that the descriptions cluster well according to both dance and speaker, Figures A.1a and A.1b. Here, only three texts missed being clustered by their dance (nos. 5, 10 and 23), and only five by their expert (nos. 7, 11, 12, 13 and 17). It may be noted that texts 5 and 10 are strongly influenced by their expert, and texts 11-13 are strongly influenced by their dance. The clustering of the interpretations is perhaps less clear because of the competing factors of their subject matter and their speaker, so that whilst some texts cluster by dance, others are too strongly influenced by their producer, Figures A.2a and

[8] Available from, ftp.informatik.uni-stuttgart.de/pub/SNNS

A.2b. This would suggest that experts are more individualistic when they are interpreting, perhaps because there is a greater range of subject matter on which to comment. Two of the sequences (texts nos. 1-5 and 6-10), were from the same dance. These ten texts, relating to the one dance, tended to cluster, especially in the descriptions. This suggests that the texts are referring to qualities of the dance that run throughout it. Of course collateral texts for many more dances would be required to properly evaluate this technique, however the results here do at least suggest that verbal reports can be a reliable source of information about moving images.

Statistical pattern recognition techniques need to be run many times in order to be properly evaluated, but the preliminary results presented here are encouraging: in particular they support the method for automatically generating text feature vectors, which in this scenario might also be seen as representative of moving images. Further investigations might study the properties of particular features in the input vectors and look for correlations between particular features: this would help to build up a picture of how sets of words, and their related concepts, refer to moving images.

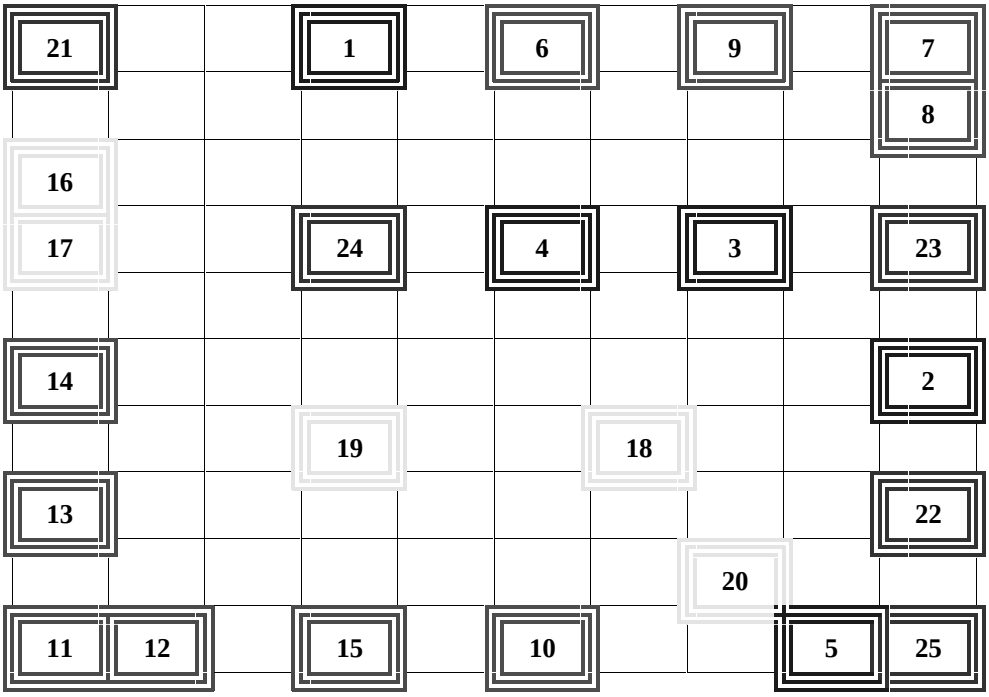


Figure A.1a: A 10x10 self-organising map, trained on 25 feature vectors for texts describing dance sequences. The texts are numbered 1-25, and coloured according to the dance sequence that they are about. Note that the texts can be seen to cluster according to the dance they are about from top-left to bottom-right.

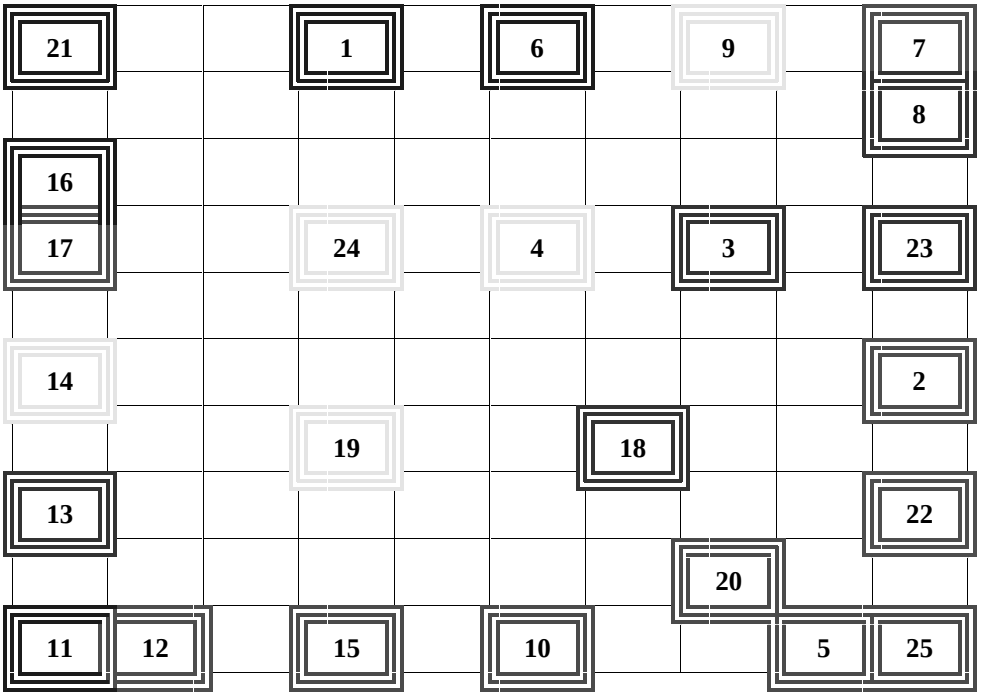


Figure A.1b: The same self-organising map, now coloured according to the speakers of the texts. Note that weaker clusters can now be seen diagonally from bottom-left to top-right.

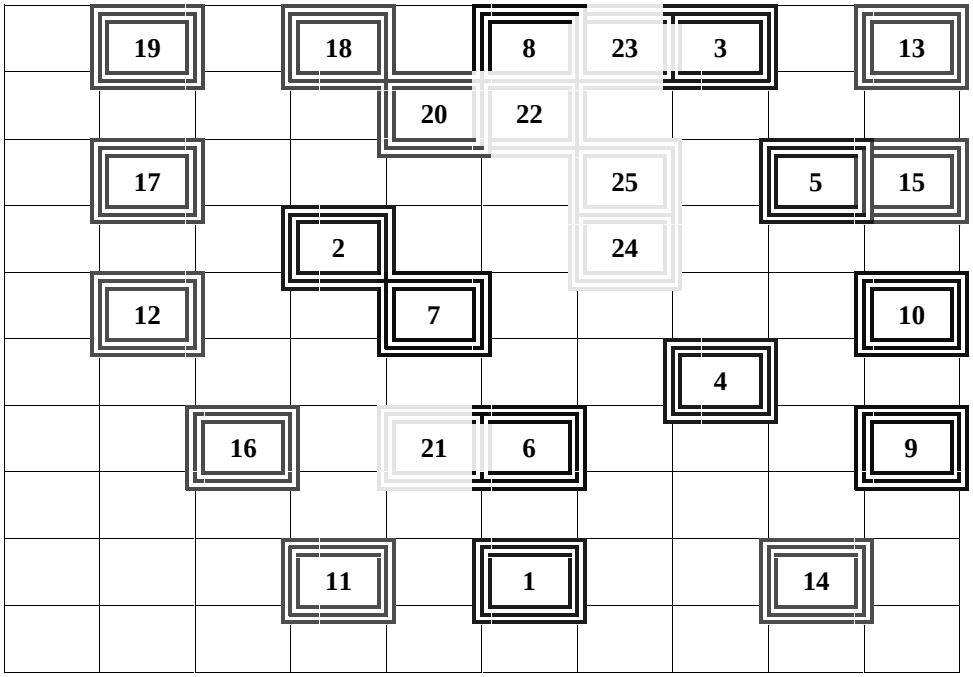


Figure A.2a: A 10x10 self-organising map, trained on 25 feature vectors for texts interpreting dance sequences. The texts are numbered 1-25 and coloured according to the dance sequence that they are about. There are some signs of clusters here, though they are weaker than for the corresponding descriptions. Note that texts 1-5 and 6-10 relate to sequences from the same dance and some of them appear to have clustered.

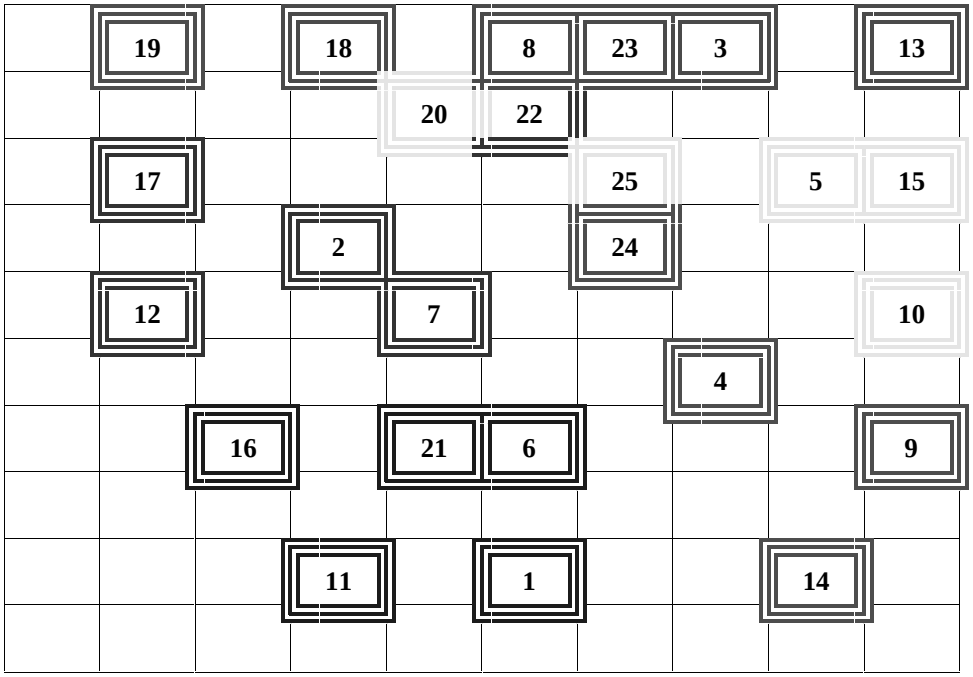


Figure A.2b: The same map, now coloured according to the speakers of the texts. Note that, compared with the descriptions, these texts seem to cluster more strongly according to their speaker.

A.2 Segmenting Moving Images by the Changing Content of Verbal Reports

When text is temporally-aligned with a moving image it is possible that the structure of the text reflects a structure in the moving image: thus, by segmenting the text it is possible to identify significant sequences in the moving image. Researchers have used a method of text segmentation in order to delimit sequences of news video – this method relied on the use of a thesaurus to establish lexical links between portions of the spoken component of a news video (Mani et al 1997).

A method for segmenting text related to a moving image is presented here which does not require any prior linguistic knowledge. This method is based on text feature vectors, generated for consecutive, but overlapping, text fragments. These vectors are used to train a self-organising map, on which clusters of vectors then indicate sections of the text, and hence of the moving image.

Five time-coded descriptions of the same moving image sequence were merged, and then a 20-second long 'window' was passed over the combined text. This window was moved along at 10-second intervals to give 32 text fragments for a dance sequence lasting 334 seconds. Each fragment overlapped by 10 seconds with the preceding one, and hence with the following one. Text feature vectors were generated for each fragment by selecting the 75 'weirdest' words in the merged texts with frequencies greater than 3. The frequencies for these words in each fragment were then normalised with a logistic function.

A 10x10 Kohonen self-organising map was trained on the 32 vectors, with an initial learning rate of 0.2, neighbourhood size of 6, and a rate of decrease for both these of 0.999. Vectors for consecutive text fragments were generally seen to cluster on the map, but in some cases there were relatively large distances between consecutive fragments. Distances of four or more horizontal and/or vertical nodes were taken to delimit clusters, Figure A.3. The 'principle of topographic map formation' suggests that these gaps are due to significant differences between the consecutive text vectors at these points: as such they may be taken as cues for segmenting the text, and thus for segmenting the moving image.

26		**27**		**13**	**12**		**11**		**10**
							7		
	28								**9**
29				**6**		**5**	**4**		
30									**1**
25		**24**		**31**		**32**		**3**	**2**
	23								
22									**16**
21	**20**		**19**	**18**		**14**	**15**		**17**

Figure A.3: A 10x10 self-organising map trained on 32 text feature vectors for fragments of a time-coded text relating to a moving image sequence. The texts are numbered in temporal order. Clusters of texts have been delimited by the criterion that consecutive vectors should be no more than four horizontal and / or vertical nodes apart; for example the distance between texts 6 and 7 means that a new cluster is started. The resulting clusters may thus reflect significant sections in the moving image.

In order to see if such a map could be used for segmenting moving images, 10 maps were trained, each with random starting weights. The clusters read off from each map are listed in Table A.2a. These results are amalgamated in Table A.2b, which shows the number of times that each text fell at the end of a cluster. Those texts which fell at the end of a cluster in four or more of the maps are marked with an asterisk: six texts are thus suggested as candidate segmentation points.

Map no.	Clustered Texts, i.e. consecutive texts are no more than 4 horizontal and vertical square moves away.
1 (See Figure 14)	1-6, 7-13, 14-17, 18-25, 26-30, 31-32
2	1-4, 5-7, 10-11, 12-13, 14-25, 26-32
3	1-6, 7-17, 18-19, 22-32
4	1-17, 18-25, 26-32
5	1-4, 5-13, 14-24, 25-30, 31-32
6	1-4, 5-13, 14-24, 25-30, 31-32
7	1-5, 6-9, 10-17, 18-25, 26-30, 31-32
8	1-18, 19-25, 26-32
9	1-4, 5-13, 14-30, 31-32
10	1-5, 6-14, 15-17, 18-24, 25-30, 31-32

Table A.2a: The clusters observed in 10 self-organising maps, trained on 32 text feature vectors for fragments from time-coded text describing a moving image. In effect each vector, numbered in temporal order, represents 20 seconds of moving image. The clusters were delimited by the criterion that consecutive vectors should be no more than four nodes apart on the map.

Text no. : 1 = 0-20 s; 2 = 10-30 s …; 32 = 310-330s																													
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
			4 *	2	2	1		1		1		5 *	1			5 *	1		1				3	5 *					6 *		10 *

Table A.2b: The number of times, in 10 maps, that a text falls at the end of a cluster. Those occurring four times or more (marked with an asterisk) are considered as candidate segmentation points.

Using these six texts (nos. 4, 13, 17, 25, 30 and 32) as segmentation points gives the six sections of the dance outlined in Table A.3. These sections, generated automatically from time-coded text describing the dance, correspond to significant sequences in the dance. These sequences are characterised by changes in the predominant dancer, and by changes in the way that the dancers are dancing together.

Texts	Equivalent Interval in the Moving Image	Characteristics of Interval and Key Changes (*)
1-4	0-50s	The first dancer is walking around the stage in front of the corps de ballet. *At 45s the dancer kneels down and a second dancer enters.
5-13	40-140s	The second dancer is dancing as the first kneels. *At 131s the second dancer touches the first, and at 135s the first starts to dance with the second.
14-17	130-180s	Both dancers are dancing around the stage. *At 178s the dancers start contact work, i.e. balancing and supporting one another.
18-25	170-260s	The dancers are doing contact work. *At 253s they break from their contact.
26-30	250-310s	The dancers are dancing around the stage. *At 310s they come into close contact again.
31-32	300-334s	The dancers stay in close contact until they exit.

Table A.3: Six sections of a dance automatically delimited by the analysis of a time-coded text describing the dance. These sections can be seen to correspond with significant events in the dance.

It seems unlikely that these sequences would have been detected by an algorithm that analysed only raw video data, since the sequence boundaries do not correspond with sudden changes of colour, texture or optical flow: it is these changes that normally allow such algorithms to detect shot and scene changes. The segmentation technique presented here used arbitrary values for some important parameters: by experimenting with the parameters of the window size and interval, and with the criteria for reading off clusters from the trained maps, it may be possible to segment moving images at different temporal granularities.

Appendix B: KAB Evaluation Study Details

Questionnaire: Closed questions and users' (1-6) responses

KAB Evaluation Study

Many thanks for helping with our study.
Please note that your identity will remain confidential.

For each task tick one rating.

When 'Browsing' videos and texts with KAB, how easy or difficult was it to:

Select, open and view videos and collateral texts

Very Easy	Easy	OK	Difficult	Very Difficult
3, 4, 5, 6	1	2		

Navigate through a video from a collateral text

Very Difficult	Difficult	OK	Easy	Very Easy
		2, 3	1, 4	5, 6

Highlight and Filter keywords in a collateral text

Very Easy	Easy	OK	Difficult	Very Difficult
5	1, 4, 6	2, 3		

Search for video sequences with keywords

Very Easy	Easy	OK	Difficult	Very Difficult
5	4, 6	2, 3	1	

When 'Annotating' videos with KAB, how easy or difficult was it to:

Add a video and a collateral text to the system

Very Difficult	Difficult	OK	Easy	Very Easy
		2, 3, 4, 5	1, 6	

Add terms to the system

Very Easy	Easy	OK	Difficult	Very Difficult
1, 5	3, 6	2, 4		

Manually annotate a video sequence with a keyword

Very Easy	Easy	OK	Difficult	Very Difficult
	4, 5	1, 2, 3, 6		

Semi-automatically annotate video sequences with KAB generating candidate video objects

Very Difficult	Difficult	OK	Easy	Very Easy
		1, 2, 4	3, 5, 6	

Questionnaire – Open-ended questions and users' (1-6) Responses

The Potential Use and Development of KAB

Remembering that KAB is intended to help people access and analyse video sequences…

What are KAB's best features?

(1) The linking of text to video; facility for close analysis; opportunities for the user to create something new; support provided through existing examples.
(2) Context and content-dependent searching of video databases. Ability to annotate video material with different expert commentary.
(3) An interactive resource that could potentially act as a rigorous knowledge base for students / scholars.
(4) Easy searching of videos.
(5) Helps demystify the video annotation process. Customisable.
(6) 'Jump to' the sequence selected in the text. Being able to loop a sequence. Comprehensive searching of texts in several formats.

What are KAB's worst features?

(1) The delay between movement and the text which relates to it – makes locating specific movements tricky.
(2) Difficult user interface. Problems with synchronization / anticipation of commented events.
(3) It is cumbersome to look at and poor quality image while also trying to scan the written text. There seems to be no reference to the role that filming techniques and editing has in the construction of interpreting the dance – this clearly matters with work made / re-worked for the screen. I would also question how relevant all of the descriptive text is.
(4) No automatic update of lists. Creating a new list of keywords is not a clear option. OK to edit existing list.
(5) Some inconsistencies in UI, not really an issue. While not a major issue, performance with the UI stifles the flow. More attention is paid to the UI and not the task at hand.
(6) New windows open over existing windows.

How could you imagine using KAB?

(1) In teaching at U/G and P/G levels in analysis of dance (potentially theatre too). As part of an online course tutorial. To pursue further analysis. For 'publication' in a new form.
(2) Possibly to annotate and describe sound recordings.
(3) This could be a useful resource for students / teachers – I worry, however, that the programme would be doing a lot of work that students should be able / or learning how to do for themselves – I also worry that it would date quickly (both in terms of changing modes of 'dance analysis' and keeping up to date with reviews of works / artists).
(4) [See 'Potential Interest in a System like KAB'].
(5) No limit.
(6) Annotating police video evidence. Analysing video taken of live tests such as crash tests. In teaching – displaying a teaching video alongside a lecture.

How could you imagine changing KAB?

(1) Addition of 'expert' texts to enrich descriptive and interpretive material. As above re timing. By adding more and more…
(2) Improve user interface and ease of use. Speech to text conversion to annotate recordings as they play.
(3) As a system it is reasonably clear in terms of how to interact with it – where it would need changing / refining is if it was to become subject specific and in consideration of its educational remit.
(4) Extension to still images, and sound. Add other ways of characterising images, how about "jesture" input – 'find something like this' [i.e. movement queries 'performed' by user].
(5) [No answer].
(6) Have windows file.

Questionnaire – User profile and further questions, and users' (1-6) responses

Subject Profile (Optional)

Occupation:
(1) University Professor of Dance Studies
(2) Senior Lecturer [in sound recording]
(3) Lecturer in Dance Studies
(4) University Senior Lecturer
(5) Systems Engineer
(6) Student (undergraduate)

Education / Training:
(1) degrees → PhD
(2) Bmus Tonmeister, PhD
(3) BA Creative Arts / MA Dance Studies / PhD – Dance on Screen
(4) PhD in Physics
(5) Business / Computer Science / training / Consultancy
(6) 'O' levels & 'A' levels + City & Guilds

Computing Experience:
(1) word processing, spreadsheets
(2) Considerable operational, some programming
(3) Basic WP
(4) Extensive. Research in Computer Assisted Learning and in Computer Modelling & Simulation
(5) 20 yrs
(6) Final year honours degree Computing student. 1 years industrial experience in an IT department.

Dance Experience:
(1) plenty – over many years, as performer and spectator
(2) none
(3) See above
(4) none
(5) none
(6) none

Potential interest in a system like KAB
(1) From dance researchers and higher education lecturers across Europe and N. America
(2) As a means of annotating and searching sound material
(3) [see 'Further Comments']
(4) Use of KAB-like system to monitor output of a simulation to recognise 'special events'.
(5) Has promise of commercial use with right design and target audience
(6) Would like to take the work forward with a project of my own.

ANY FURTHER COMMENTS:
(1) Much progress has been made since I first saw the early stages. Congratulations.
(2) [No answer].
(3) The basic idea is interesting but I have some serious concerns about its educational value and some bigger questions about frameworks of dance knowledge and how these may be problematised with a system like KAB.
(4) [No answer].
(5) [No answer].
(6) Good luck.