Andrew Salway and Khurshid Ahmad

“Talking Pictures: Indexing and Representing Video with Collateral Texts”

Procs. 14th Twente Workshop on Language Technology – Language Technology for Multimedia Information Retrieval, pp. 85-94. ISSN / ISBN: 0929-0672.

Talking Pictures: Indexing and Representing Video with Collateral Texts

Andrew Salway and Khurshid Ahmad
Department of Computing, University of Surrey
Guildford, GU2 5XH United Kingdom
a.salway@surrey.ac.uk, k.ahmad@surrey.ac.uk

ABSTRACT

The relevance of collateral texts for building knowledge-based visual information systems is discussed, with reference to moving images of dance. The knowledge acquisition technique Protocol Analysis is applied to elicit verbal reports from experts watching moving images with complex contents and interwoven meanings. These reports are analysed at lexical, clausal and discourse levels, using text analysis methods. The results show a potential for using these reports to index and represent moving images, through the creation of lexical resources and knowledge-bases. The KAB system integrates text analysis modules with a video object database to process collateral texts.

Keywords: Knowledge-based Visual Information Systems, Video Indexing, Collateral Text, Protocol Analysis

1 INTRODUCTION

There is a need for technologies that can assist in the retrieval and presentation of visual information. These technologies must provide ways of attaching useful indices to both still and moving images so that they can be matched against user queries. Furthermore, representations of image content in knowledge-based visual information systems should address the fact that an image can mean many things to many people.

The inclusion of human knowledge may be crucial in domains where the perception and understanding of images is an expert task. The knowledge of experts is realised in the language of their written and spoken discourse. The texts produced by experts can be used as high-level expressions of image contents - from which indices and representations for computer systems could be derived.

Typically, image and video retrieval research has sought to compute indices from raw image data - attaching colour, texture and shape features to still images, and segmenting and selecting key-frames for moving images. In the field of computer vision, researchers have developed picture grammars which build up image descriptions in terms of primitives such as edges, corners and surfaces. These approaches work well in many cases when matching on *perceptual similarity* is required.

However, much visual information is complex, comprising interwoven strands of meaning that may be confounded in the image. In these cases, indices and representations of images need to be high-level and symbolic - to refer beyond the physical image contents.

Certain visual information may be best understood, and hence explicated, by the experts of a particular domain: consider, a surgeon examining an X-ray image; a meteorologist making predictions from moving images of weather systems; a scene-of-crime officer recording photographed evidence; and art critics and dance scholars who can elucidate meanings in complex images which would not be apparent to a lay person.

In these cases the words spoken and written by the experts about the visual artefacts will be high-level expressions of their contents. In order to exploit this fact for building knowledge-based visual information systems it is important to understand how experts articulate their knowledge, and how their articulations relate to visual information.

Researchers have already exploited textual information that co-occurs with everyday visual information. For example, Srihari reported how newspaper photograph captions were processed to constrain subsequent image analysis algorithms that detected the faces of people described in the caption [1]. She used the phrase ‘collateral text’ to denote textual information which related in some way to visual information. Recent research has used the textual component of video email and of news video for indexing purposes, see [2] for examples.

This paper considers how the written and spoken words of experts can be utilised as collateral texts to build knowledge-based visual information systems for specialised images. This question is explored in the exemplar domain of dance - chosen for its stylized moving images, diverse multimedia information and consolidated expert knowledge.

The work of dance scholars is first discussed to highlight the ways in which a moving image can be analysed, and to suggest how texts written by dance experts could be collateral to dance images (Section 2). An investigation is then reported in which spoken texts were elicited from dance experts as they watched dance videos; following the knowledge acquisition technique of Protocol Analysis. Their verbal reports were analysed for their lexical, clausal and discourse-specific behaviour. This analysis sought to explicate the ways in which the experts articulate their knowledge about moving images (Section 3). The findings of these investigations have been used to refine the specification of the KAB system. KAB (Knowledge-rich Annotation and Browsing) integrates digital video and collateral texts to give knowledge-rich representations of moving images (Section 4). The paper closes with some remarks about how collateral texts might be used in knowledge-based visual information systems with integrated video analysis functionality (Section 5).

2 DANCE EXPERTS' DISCOURSE ON MOVING IMAGES

Dance comprises stylized movements which are rhythmic and usually set to music. Recorded dance images provide a source of diverse, but interrelated, multimedia information about human movements, music, costume and the stage set. The dancing body may be described in terms of its parts, or itself may be described as a part of a greater whole – when groups of dancers form patterns. Dance can convey emotions, tell stories and make social comments and cultural statements. In dance, whether recorded images or live performance, there are interwoven strands of meaning.

2.1 DANCE ANALYSIS

There are specialists who study dance as an academic, and as an applied subject. They deal with the perceptual and cognitive aspects of dance and discuss dance in historical, cultural and political contexts.

Dance scholars have evolved notation systems for recording muscular movement (with similarities to musical notation). The emphasis here is on recording the muscular-skeletal movements of the dancers and their positions in space. Movement notation systems are expected to 'provide the key to relatively unambiguous communication through the creation of an agreed symbol system' [3]. An example of a prominent system, *Labanotation*, is shown in Fig. 1. The notation is read from bottom to top, 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 the left hand side of the body: the foot, leg, torso, arm, hand and fingers - in that order. The symbols' points, shadings and size capture the movement dynamics of direction, level of extension and duration.

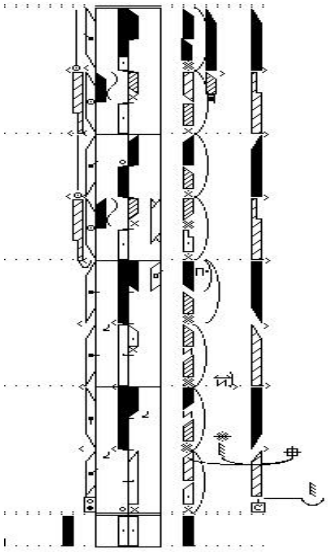

Fig. 1: An example of *Labanotation*

Movements are also described with the terminology of established dance genres, like classical ballet's *plié*, *relevé*, *attitude* and so on. The dance analyst then discerns how individual actions and gestures are combined in spatial and temporal forms, e.g. to show the interaction between dancers, or the recurrence of a *motif*.

It is problematic to separate the *objective* description of dance (movement durations, directions, accelerations) from its *subjective* facet (meanings and emotions). For, whilst a *slow movement* can be assessed in biomechanical terms, it might also be, depending upon contextual factors, a *tender movement* or indicative of *wandering*.

These movement qualities guide the dance scholar in explicating the narrative and intention of a dance.

Individual dances and choreographers are considered and a view of genres and styles of dance is developed on a case-by-case basis. Thus, Judith Mackrell has written that classical ballet tries to create 'the illusion of flight', and that some classical Indian dances 'are grounded on earth' [4]. 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.]. When making such interpretations, it has been argued, it is important to recognise that 'the conventions and traditions of the context, genres and styles presuppose and, therefore, prescribe specific ranges of *subject matter and the manner of treatment*' (emphasis original) [3].

The dance theorist draws on theoretical perspectives in order to place the dance in an historical, cultural and theoretical framework. Dance genres and styles may be used to classify a dance, and relate it to its forebears. Other political and cultural theories may be adapted, for example, Marxism, feminism and psychoanalysis. In the course of evaluation, reference is made to other dances as well as sources including books, films, historical events and cultural phenomena.

Janet Adshead-Lansdale has outlined a four stage method for analysing dance [3]. The method involves describing movements, discerning spatial and temporal forms, interpreting meanings and evaluating the aesthetic merits of a dance. This framework was used in the research reported here for focussing experts on particular aspects of moving images during knowledge acquisition. Similar frameworks have been used in other research concerning visual information: Erwin Panofsky's three levels of meaning in fine art [5] have been adapted by information scientists for picture classification [6]; and, Christian Metz's five levels of cinematic analysis [7] motivated a semantic data model of video [8].

2.2 DANCE TEXTS AND DANCE KNOWLEDGE FOR IMAGE RETRIEVAL

The knowledge used by dance experts to analyse and understand dance at perceptual, cognitive and aesthetic levels is relevant to the ways in which images of dance can be stored and retrieved. The content of a dance, from a retrieval perspective, may include isolated muscular movements, patterns formed by dancers, metaphoric references, e.g. to 'pull of gravity' or 'information overload', as well as historical, cultural and possibly political allusions.

For us, a verbal expression about a dance can be collateral to still and moving images of that dance. This collaterality can be found in the program of a dance performance, in a dance critic's newspaper review, in popular books explicating the works of choreographers and in learned journal articles.

The question is how to use these different types of texts for the task of labelling and retrieving dance sequences. A dance program could be used to label long stretches of movement, like acts and scenes (circa 15 - 20 minutes). A dance critic's output might highlight noteworthy aspects of a dance to be labelled, e.g. certain sequences of movement, the performance of certain dancers or the use of costume and sets. More learned writing includes the dissection of short movement sequences (*motifs*) as well as the principled grouping (*classification*) of dances or choreographers into genres and styles.

The information retrieval literature would suggest the building of a thesaurus, comprising terms related to dance movements, key historical events and even philosophical and political trends: these would then be used to label all or parts of a dance video - much as is practised in the keywording of journal articles and textbooks.

The different types of dance text discussed above are produced for information dissemination to a well-defined readership. The intention of the authors is not to index a dance for retrieval purposes. These texts can merely serve as mnemonics, placing the burden of interpretation of collateral words, phrases, clauses and whole texts, on the reader. Nevertheless, it is these texts that help their readers to see beyond the physical image: beyond the surfaces, edges and contrasts of light and shade. These texts have the potential for helping end-users of an image retrieval system by expanding and refining queries. For instance, these texts might be used to produce a specialist lexicon of dance, or of a particular genre of dance with contextual information about most of its lemma entries. Key-word-in-context analysis may help to build a lexical semantic network linking lemmas with arcs labelled by semantic relations like synonymy and hyponymy.

Though the extant texts of the dance domain are collateral to moving images of dance, the fact that they are written means they do not share a straightforward temporal relationship with their subject matter. Authors bring together examples of movement from different parts of a dance, or from different dances. The collaterality is further obscured by the changing

focus of the text - in one paragraph the discussion might be of muscular movement, in the next it might turn to the historical building in which the dance was staged. Thus, whilst these extant texts may be collateral to dance images, they would pose problems if used in the initial development of a knowledge-based visual information system.

In contrast, a recorded running commentary on a dance sequence will preserve collaterality, whilst still being a linguistic artefact – amenable to subsequent analysis. To ensure that the commentary is knowledge-rich, i.e. it contains so-called domain objects and heuristics, it can be elicited from a domain expert. Finally, the expert should be instructed to focus on a particular aspect of the dance in a single recording. Protocol Analysis is a technique for prompting experts to articulate aspects of their knowledge in a *verbal report*.

3 VERBAL REPORTS OF DANCE SEQUENCES

Researchers in various disciplines have been concerned with how subjects talk and write about images. Firschein and Fischler elicited descriptions of aerial city photographs from subjects with a variety of tasks; their investigation concerned the descriptive representation of pictorial data for computer vision systems [9]. In information science, researchers have studied both how indexers attach keywords to images [10], and how searchers of images phrase their queries [11]. Linguists have recorded narratives from subjects about a film they watched; the transcripts became the basis for a cross-cultural study of discourse and the relationship between conscious experience and the spoken word [12].

The artificial intelligence and cognitive psychology literatures discuss knowledge acquisition techniques which can be used to elicit, analyse and represent aspects of human knowledge for use in computer systems. Knowledge engineers have access to two sources of knowledge: human experts and the texts the experts have produced. Knowledge can be elicited from experts through brainstorming, interviews, questionnaires and the like. Techniques to extract knowledge from text may be applied to the transcripts of such knowledge acquisition sessions, and also to texts extant in the domain. Text analysis can assist in knowledge engineering, with the elaboration of domain terminology leading to the modelling of concepts and then propositions and rules [13].

Protocol Analysis is a knowledge acquisition technique, in which an expert is asked to 'think aloud' as they perform a task: the resultant verbalization is taken to reflect their cognitive processes - and hence their expertise [14]. Protocol Analysis is used to access the steps that an expert takes in performing a task, i.e. to understand 'how' the expert does it. The expert's verbalization is recorded and becomes the object of investigation - hence the claim that verbal reports provide us with data about the subject's cognitive processes. There is a similarity with the research method of Content Analysis, which 'procedures to make valid inferences from text' [15].

In the current research, Protocol Analysis was applied to access an expert's perception and understanding of complex visual information in real time. The goal was to elicit verbal reports that (i) would help to understand the expertise used in analysing moving images; and, (ii) would serve as collateral texts for indexing and representing them.

Experts were prompted to talk about a moving image as they watched it: the instructions they were given before talking were used to focus their verbalization on particular aspects of the image contents.

3.1 METHOD

Five dance experts[1] were twice recorded speaking as they watched a video compilation of dance excerpts lasting 20 minutes. Four of the excerpts were duets, a fifth featured 12 dancers. The types of dance included neo-classical ballet and modern dance.

Before the first recording the expert was prompted with an instruction to 'Describe'[2] the dances, speaking as they watched. For the second recording the instruction was to 'Interpret'[3]. These instructions were worded by the co-ordinating expert, Prof. Janet Adshead-Lansdale, to reflect tasks familiar to the other experts.

The experts' verbalizations were recorded onto one of the sound tracks of the video cassette that they were watching; this maintained the temporal relationship between word and image. They were transcribed by

[1] A university lecturer and four post-graduate students from the same Dance Studies MA course at Surrey.

[2] The written instruction advised *'by describe we mean, focus particularly on the detail of the movement, its use of space and its dynamic emphasis'*.

[3] This time the elaboration was *'by interpret we mean, outline one or many kinds of significance you might attribute to the interaction in this section'*.

one of the authors. The flow of spoken language was broken into speech fragments on the basis of the speaker's pauses and the perceived completeness of each speech fragment. A start-time (m:s) was manually inserted before each transcribed speech fragment. For the current research there was no need to transcribe intonation information and lengths of pauses, but this information might be valuable in future studies.

The verbal reports for the 'Describe' task, henceforth D-texts (for the sake of clarity in this paper), totaled 11,300 words, and 1600 speech fragments; from 100 minutes of speech. The most fluent speaker averaged a rate of about 150 words per minute, the least about 75: there was no significant variation in the rate of speech between different dance excerpts. Half the transcripts from the 'Interpret' task (I-texts) have been analysed, these total 6289 words.

3.2 Analysis of Verbal Reports

Initial analysis of the verbal reports manually examined the extent (i) to which the linear order of their contents matched those of the video; and, (ii) to which subjects agreed on what to speak about and how. Observation of aligned verbal reports suggested a reasonable correlation between their contents and the videos' contents. Table 1 shows samples from two D-texts (from different experts) and one I-text, corresponding to 40 seconds of an excerpt from Matthew Bourne's *Swan Lake*.

It is clear that in the D-texts the experts have chosen to focus on similar aspects of the video, however they describe them at different levels of detail, and using different words. For example, in describing the man walking, one expert noted *where* - 'across the stage area' and another noted *how* - 'at a very slow pace'; and the arms of the swans are 'extending .. outwards and behind their backs' (D-text 2) or 'locked behind their backs' (D-text 2).

The I-text in Table 1 says less about the individual actions, and more about the mood of the characters - 'a sense of loneliness' and 'looking longingly'; and, about their relationship - the man is 'separated from the rest of the characters'.

Table 1: Excerpts from two corresponding D-texts and one I-text

D-text 1	D-text 2	I-text
[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 back 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

Generally the experts' expressions were observed to range from single words - naming individual actions and gestures almost as soon as they recognised them, to longer speech fragments – detailing the time, location and quality of the movements; and, in the case of the I-texts, making cases for their interpretations. The experts generally kept up with the moving image – their words tended to lag no more than a few seconds behind the subject matter. However, there were examples of experts referring back to earlier sequences, e.g. 'back to the pas de deux type lifting movement'; and of experts referring to longer sequences in their interpretations, e.g. 'their relationship to each other alters during the course of the piece.'

The verbal reports were analysed with regard to the distribution of their lexical items; the information content of their clauses; and cohesion phenomena.

The analysis was performed with Surrey's text analysis package *System Quirk ©.* Results show how experts' knowledge for analysing moving dance images can be articulated to: name movements and their qualities; elaborate on gestures, actions and poses; highlight important sequences; and make interpretations.

3.2.1 Terminology for describing dance

A statistic which divides the relative frequency of a word in a collection of specialist language texts (SL) by its relative frequency in a general language sample (GL) gives a word list in which words peculiar to the texts rise to the top. Table 2 shows some words which appeared in the D-texts 100+ times relatively more often than they did in a general language sample (10 million words of text from the Longman Corpus). The 25 words are from a specialist dance terminology; they constitute just under 20% of the D-text words with SL / GL > 100. Apart from the generally used terms *bodyshape, choreography, duet(s), footwork* and *motif*, Table 2 contains terms from a balletic vocabulary – perhaps reflecting the background of the experts, and the kinds of dance they spoke about.

Table 2: Dance Terminology in the D-texts

adage, adagio, arabesque(s), balletic, battement, batterie, bodyshape, choreography, développé(s), duet(s), footwork, jeté, motif, passé, penché, pirouette(s), planche, plié(ing), promenaded

A second subset of words with SL / GL > 100, comprising a further 43 % of the total, is listed in Table 3. These are words which would be familiar to non-experts, but that have been appropriated, sometimes with a shift in meaning, by the dance experts; e.g. *pedestrian* which refers to everyday movement in this context. They are split into two groups: (i) nouns and verbs which denote movement and actions; and (ii) adjectives and adverbs which denote the quality of the movement and actions. The descriptions of quality at times cross the boundary into interpretation, e.g. *animalistic* and *robotic*.

Table 3: Preponderant general language lexical items in the D-texts

Movement & Action	arching, balances, balancing, caresses, clasping, flicks, hops, interlocking, jumping, jumps, kneeling, leans, leaps, lunge, lunges, manipulates, manipulating, pivoting, pivots, pushes, quivers, rotates, slicing, spins, spiralling, spirals, splaying, stroking, sways, tilts, tottering, totters, twisting, twitching, undulations, weaved, wiggles, wraps
Quality	animalistic, dynamic, flexed, gestural, jerky, lyrical, pedestrian, rhythmical, robotic, stuttered, swirly, synchronised, undulating, unison, unisonal, unisonally, virtuostic, wiggly

The mid-ranges of the SL / GL list were filled with more generally familiar words which describe movement and actions, e.g. *bend, come, hold, roll, kneel, walk*. Words that locate movements and actions in space and time also appear, e.g. *diagonal, forward, left, right, across* and *continuing, occasional, while, sporadic*.

The absolute frequency of words denoting body parts perhaps says something about how the dance analyst attends to movement - it may also say something about the genres of dance being described. The D-texts gave the following result, in descending order of absolute frequency: *arm(s), leg(s), hand(s), head(s), foot / feet, body/ies, back(s), shoulder(s), chest, neck, torso, waist, face(s), elbow(s), hair, knees, palms, spine*.

Compound terms were identified using a method that extracts strings occurring between lexical items given in boundary lists, giving, e.g. *corps de ballet, rond de jambe* and *pas de deux*.

Distinct collocation patterns involving the words *position, gesture* and *action* were noted - in each case a particular type of preceding word was seen to collocate about 50% of the time, Table 4. These examples might be considered to be 'semi-fossilized phrases' [16] - in which one word predicts a limited number of collocating words.

Table 4: Semi-fossilized phrases in the D-texts

Nucleate	Total Freq.	Typical preceding word	Example Phrases
position	68	first – fifth (44%)	first position ... fifth position
gesture	19	'BODY_PART' (53%)	leg gesture head gesture
action	18	'METAPHOR' (56%)	pendulum action sawing action

3.2.2 Clauses bearing information about *gesture, action* and *pose*

Classification of the speech fragments was made in terms of the information content of their clauses. Observation of the linguistic data suggested three categories of clause for this purpose: these were validated by the co-ordinating dance expert. *Gesture* clauses describe a spatial reconfiguration of body parts in relation to one another; *Action* clauses describe a spatial relocation of the whole body along spatial pathways, this includes locomotion; *Pose* clauses describe dancers' locations on stage, held positions and gazes.

A manual analysis of one D-text (80 speech fragments describing 333 seconds of dance) gave the distribution of clauses shown in Table 5 (a speech fragment may contain more than one clause). Twelve speech fragments did not fit the scheme; they referred to costumes or camera actions.

Table 5: Distribution of clauses by information content in one D-text

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

The even spread of clauses suggests that this is a useful classification for further analysis to be based on. Each clause can refer to 1, 2 or 3+ dancers - dancing in unison or taking different roles in an interaction. The contents of a clause can be modified by adjectives and adverbs to describe quality, and by prepositional phrases to situate them in space and time.

3.2.3 Lexical and syntactic cohesion

Halliday has described linguistic phenomena relating to cohesion and the creation of 'texture' [17]. Two textural characteristics of the D-texts are noted here. Some passages are marked by repetition of semantically-related words, exemplifying lexical cohesion, Table 6.

Table 6: Cohesion through lexical repetition in D-texts

...
[2:44] there are a lot of **leaps**
[2:45] **hops**
[2:46] **jumps**
[2:48] with legs usually extended
[2:51] there are different qualities of **aerial steps**
...
[5:11] they are now face-to-face in an **embrace**
[5:15] the swan character **clasps** the male character in an almost foetal position
[5:20] he is **clinging** his arms around the back of the swan character's neck
...

Other passages focus on a particular dancer, or group of dancers. Here cohesion is maintained by reference, Table 7.

Table 7: Cohesion through reference in D-texts

[1:00] **the** central character is kneeling

[1:05] there is now **another** character entered
[1:07] **who** turns, spirals and goes into an arabesque position
[1:14] **_** curving
[1:15] **_** twisting
[1:16] **his** arms are undulating
[1:18] sometimes **he** is stretching upwards
[1:20] sometimes **he** is curving inwards

[1:22] **the** central character remains kneeling, glancing upwards

In the above passage, a mention of a previously unseen dancer is cued by an indefinite article, *another* in this case. Subsequent mentions are referential - *who, his, he* - or elliptical. The return of a previous dancer is cued by the definite article *the*.

3.2.4 From description to interpretation

The fuzzy boundary between description and interpretation marks the passage from the literal to the metaphorical. Table 8 lists words with SL / GL > 100 in the I-texts (cf. Tables 1 & 2 for D-texts). The contrast between Tables 1 & 2 and Table 8 suggests that the experts successfully differentiated the 'Describe' and 'Interpret' tasks. However, words underlined in Table 8 also appear in Tables 1 & 2. Some of these are such general words they might be expected in any kind of dance text, e.g. *duet* and *footwork*; others are indicative of the problem in keeping movement descriptions objective, e.g. *animalistic*. What properly distinguishes the I-texts is

the unusual occurrence in Table 8 of abstract nouns like *ethereality*, *mentalities* and *recalcitrance*.

Table 8: Words with SL / GL > 100 in the I-texts

allures, animalistic, balletic, constriction, counterbalances, duet, dynamic, ecstasy, endlessness, ethereality, figment, floorspace, flurries, footwork, hinging, homoeroticism, impacting, instigating, interlinked, layerings, magnets, manipulative, mentalities, palpability, quirkiness, recalcitrance, repelling, seducing, skyscrapers, soulful, swan, swans, thematically, togetherness, torsos, transversed, twittering, unisonal

Some interpretative statements take a metaphoric form so that a phrase with relatively objective content is linked with one that is more imaginative: in this way the signification of muscular movements is expounded. The phrases are linked by one of six phrases, Table 9.

Table 9: Linking of phrases to form interpretations in the I-texts

Linking phrase	Freq.	Example
seem*	32	although her involvement in this seems perhaps more vital
sense (of)	19	holding the wrists, a sense of being bound
suggest*	17	aerial steps, which could also suggest flight
as if	16	it is in blue, sort of dark, as if he is dreaming
like	16	the stretching of the neck, like a swan
appear to be	4	the constriction also appears to be a support

4 PROCESSING VERBAL REPORTS AS COLLATERAL TEXTS

The textual component of video has been exploited for indexing purposes by using combinations of established language technologies like speech recognition, information retrieval and information extraction, for examples see [2]. The analysis of the verbal reports suggests that they could be analysed by IR and IE techniques to index moving dance images: they are rich in specialist terminology and other keywords, linked in time to the moving image; the semi-formal characterisation of clausal information content might help to extract representations of image contents; and, the observed cohesion phenomena might support video segmentation.

As the first stage in building a knowledge-base, the elicitation of verbal reports on moving images gives the knowledge engineer a rich source of domain concepts and a set of cases. The cases can form the basis of further sessions in which the experts might elaborate the reasoning behind their descriptions and interpretations.

The collaterality between verbal reports suggests the possibility for extracting terminologies rich in lexical relations, following an approach like that used to process aligned texts in multilingual systems. Furthermore, the link to the moving image makes ostensive definitions available.

The KAB system (Knowledge-rich Annotation and Browsing) uses collateral texts to index and represent digital video. An early version facilitated the browsing of extant domain texts alongside video [18]. This was implemented in Macromedia® Inc.'s *Director* - a commercial multimedia authoring system. Following the study of verbal reports, the specification of KAB has been extended to implement a *video object* database, alongside text analysis modules. Collateral text is processed in order to semi-automate video annotation by generating video objects with reference to a knowledge-base, Fig. 2.

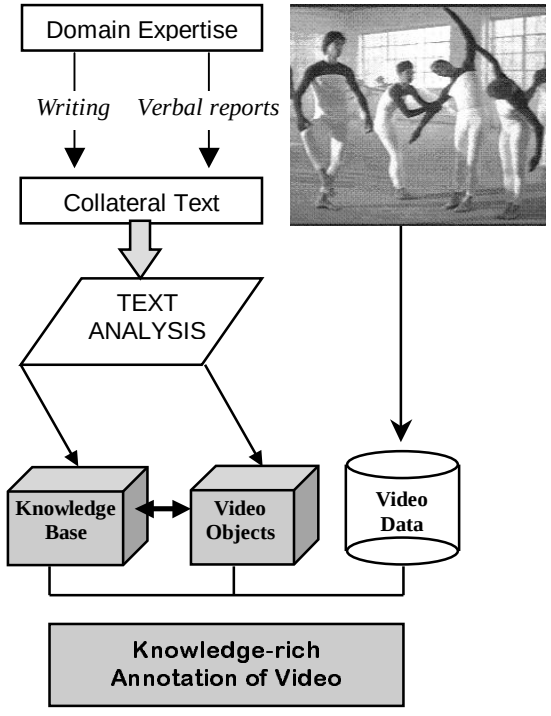

Fig. 2: The KAB Video Annotation Overview

Following Oomoto and Tanaka, a video object is a data structure which refers to a (possibly discontinuous) sequence of video frames, marked by start and end times, with an expression which represents the image contents [19]. This model is chosen for KAB because it can be easily adapted to handle different forms of representation - including: keyterms, attribute-value pairs, predicated expressions and links to other multimedia documents. With current video coding standards it is sufficient for video objects to refer only to sequences of video frames: this situation may change with the advent of object-based video coding - the nascent MPEG-4 standard[4] will give the objects portrayed in moving images their own identities.

The KAB system is being implemented in the object-oriented Java programming language, extended by the Java Media Framework[5] (JMF). This combination offers high-level operations for both text analysis and video presentation. The JMF abstracts from the physical layer of video so that programs can be written as platform and *codec* independent. The JMF API (Application Programming Interface) provides high-level commands for controlling video, e.g. 'stop', 'start' and 'go to time X'. The KAB system incorporates the text analysis software used previously in the analysis of the verbal reports.

The KAB prototype, Fig. 3, lets the user build collections of linked videos and collateral texts. Annotations can be attached to the video in the form of video objects through a series of dynamic menus which show a selection of available representations (updated through the 'Add Lexical Knowledge' option). Searching is achieved by making a selection from similar menus, which returns a set of matching video objects. Current work is implementing the 'Process Texts' function so that collateral texts are analysed to automatically suggest video objects – grounded in lexical resources and knowledge-bases. As well as being used to match queries for retrieval purposes, the expressions attached to video objects can also be used to explain the video contents to the viewer when browsing, e.g. by showing an expert's commentary on a sequence or offering a link to related media. For further information about the development of KAB see [20].

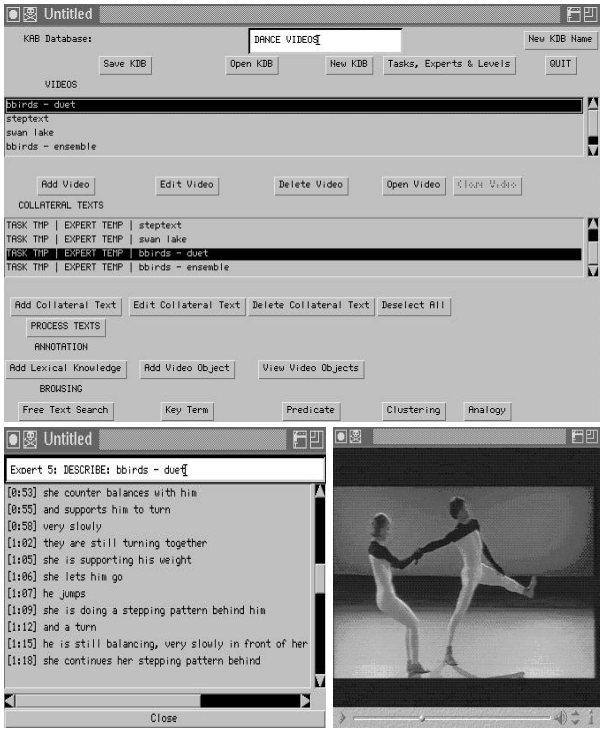

Fig. 3: The KAB Prototype main menu and example video with collateral text

5 CLOSING REMARKS

The use of keywords and other linguistic expressions as pointers to image contents is contentious philosophically in that it posits a primacy for language over other modes of communication. However, the words of experts elaborating upon the visual artefacts of their domain can help in understanding the artefact, especially for pedagogic purposes, and by extension for information retrieval tasks. The collaterality of texts with moving images was discussed and a suggestion made for how spoken collateral texts can be elicited from experts. It was shown how experts analyse an image sequence, both literally and metaphorically, and as it were take us beyond the image. Verbal reports generated by experts serve as content-rich running commentaries of moving images, as in the KAB system above.

The use of verbal reports for indexing images appears to be labour intensive and this fact should perhaps discount the approach for routine or large-scale image indexing tasks. Dance images, though,

[4] http://drogo.cselt.stet.it/mpeg/standards/mpeg-4/mpeg-4.htm
[5] http://www.javasoft.com/products/java-media/jmf/index.html

may be a special case in that they comprise themed sequences, realised by idiosyncratic motion patterns. Thus, particular genres, dancers, choreographers and narratives, for example, may be linked with 'signature sequences'. The verbal reports help to identify where examples of these idiosyncratic motion sequences occur. A system like KAB can then elaborate indexes and annotations of relevant image sequences for a domain.

Research in video analysis, particularly of human movement [21] [22], will perhaps lead to systems for recognising these idiosyncratic sequences in previously unseen video data. Such a system could then interact with a system like KAB to attach linguistically-based expressions to the new sequence, based on previous examples already aligned with collateral texts. This scenario would exploit automatic image analysis techniques for low-cost indexing, and would give knowledge-rich representations from the results of an initial, 'one-off', knowledge acquisition stage.

REFERENCES

[1] R. Srihari. Use of Captions and Other Collateral Text in Understanding Photographs. In *Artificial Intelligence Review* 8 (5-6), pages 409-430, 1995.

[2] M. T. Maybury, editor. *Intelligent Multimedia Information Retrieval*. Menlo Park CA: AAAI Press / MIT Press, 1997.

[3] J. Adshead, editor. *Dance Analysis: Theory and Practice*. London: Dance Books, 1988.

[4] J. Mackrell. *Reading Dance*. London: Michael Joseph, 1997.

[5] E. Panofsky. *Meaning in the Visual Arts*. Harmondsworth: Penguin, 1970.

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

[7] C. Metz. *Film Language: a semiotics of the cinema*. New York: Oxford University Press, 1974.

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

[9] O. Firschein and M. A. Fischler. A Study in Descriptive Representation of Pictorial Data. In *Pattern Recognition* 4, pages 361-377, 1972.

[10] P. G. B. Enser. Pictorial Information Retrieval. In *J. of Documentation* 51 (2), pages 126-170, 1995.

[11] L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. In *J. of Information Science* 23 (4), pages 287-299, 1997.

[12] W. L. Chafe, editor. *The Pear Stories: cognitive, cultural and linguistic aspects of narrative production*. Norwood NJ: Ablex Pub. Corp., 1980.

[13] J. Boose. Knowledge Acquisition. In S. C. Shapiro, editor, *The Encyclopedia of Artificial Intelligence, Vol. I*, 1992.

[14] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. 2nd Edition, Cambridge MA and London: The MIT Press, 1993.

[15] R. P. Weber. *Basic Content Analysis*. 2nd Edition, London: Sage Pubns., 1990.

[16] G. Kjellmer. A Mint of Phrases. In K. Aijmer and B. Altenberg, editors, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, pages 111-127, 1991.

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

[18] K. Ahmad, A. Salway and J. Adshead-Lansdale. (An)notating Dance: Multimedia Storage and Retrieval. In H. Selvaraj and B. Verma, editors, *Procs. ICCIMA '98*, pages 788-793, 1998.

[19] E. Oomoto and K. Tanaka. Video Database Systems - Recent Trends in Research and Development Activities. In W. I. Grosky, R. Jain and R. Mehrotra, *The Handbook of Multimedia Information Management*, pages 405-448, 1997.

[20] A. Salway. *Forthcoming Ph.D. dissertation*, Dept. of Computing, University of Surrey.

[21] C.-C. Lien and C.-L. Huang. Model-based articulated hand motion tracking for gesture recognition. In *Image and Vision Computing* 16, pages 121-134, 1998.

[22] T. Ahmad et al.. Tracking and recognising hand gestures, using statistical shape models. *Image and Vision Computing* 15, pages 345-352, 1997.