Bound Together: Signs and Features in Multimedia Content Representation

Edward Hartley
Computing Department
Lancaster University
United Kingdom
hartley@comp.lancs.ac.uk

ABSTRACT

This paper introduces new extensions to the semiotic model that allow the model to account for image features that characterize an audio, visual or audio-visual object. The treatment in this paper emphasizes visual content description. The framework and the associated construct of *image features* characterizing the *visual object* "binding" to *conceptual terms* used to describe the *visual object* is described and illustrated in terms of UML diagrams. Visual object identification, location and media temporal segmentation approaches are outlined. The setting construct developed by Parkes [17] is then replaced with the introduction of the existent context as the minimal unit for temporal decomposition in content description. This construct forms the basis of the content modeling for the upper level representation scheme adopted in the Automating Video Annotation (AVA) moving picture content annotation prototype tool. The analysis is supported through a worked example developed on a content sequence from an episode of children's television.

Keywords

Computational Semiotics, Multimedia Annotation, Video Annotation, Content Description, Content Representation, Visual Feature Extraction.

1. INTRODUCTION

Multimedia content representation is concerned with constructing models of that are used to facilitate content description or annotation. In general these models have adopted a bottom-up or top-down approach. The bottom-up approach is based on the analysis of still and moving picture features typified by the descriptors in MPEG-7 Visual part [10]. These features typically characterize a visual object in terms of color, texture or both using methods similar to those described in section 3. The top-down approach is typified by the content modeling portions of the MPEG-7 MDS part [12]. In contrast the approach adopted in the Automating Video Annotation (AVA) project at Lancaster is bi-directional in that it combines elements of both approaches. The integration of both approaches has needed new theoretical developments that extend the semiotic model. These new theoretical developments are based on previous revisions and extensions to the semiotic model developed by Hartley *et al.* [8].

First published at COSIGN-2004,
14 – 16 September 2004, University of Split (Croatia)

An outline of the Revised and Extended semiotic models is given in section 2 these introduce the data and description planes into the model. This is followed by a description of the new revisions to the model. These revisions enable the model to account for image features together with denotative and connotative descriptions. To simplify the analysis the emphasis in this paper is on visual media and objects, however the approach can be readily extended to the audio and multimedia domains.

Section 3 outlines an example of image processing techniques for visual object identification and location and an overview of approaches to temporal segmentation. Known limitations of these techniques are then discussed. On the basis of this analysis assumptions are made about the capabilities of these algorithms that allow an extension of the semiotic modeling from section 2 into the time domain.

In section 4 the adoption of the term existent by Chatman [4] in his semiotic analysis of narrative structure in fiction and film is described in relation to the modeling from section 2. This results in the definition of the existent context as the minimal unit of spatio-temporal segmentation under constraints determined by the capabilities of the chosen segmentation algorithms from section 3. It is then shown that the existent context set can replace the setting construct developed by Parkes [17] as the minimal unit of temporal segmentation for moving picture content description.

A qualitative worked example of the application of the model is given in section 5, which is followed by conclusions about the adoption of the approach in section 6.

2. Revising the Semiotic Model

It is well known that the semiotic analysis [5] of textual media distinguishes between the content plane and the expression plane. The content plane contains the meaning carried by the words on the page and invoked in the readers mind. Whereas the expression plane contains the words seen on the page through expressed through the print medium. The analysis then goes on to distinguish between the idea of the words denoted by the text i.e. the base meaning of the word in a given context and the ideas that are conotated by the word from associations in the readers mind. Nack [16] adopted a semiotic approach to multimedia content modeling and description in his work on the Auteur, automatic editing prototype. It was found subsequently by Hartley *et al.* [8] that whilst the semiotic approach to multimedia content modeling was valuable the distinction between content and expression could not fully account for digital multimedia content and its description. This was not due to inadequacies in the semiotic approach to content modeling but is due to limitations in the

extent of the model. Revisions and extensions to the model were undertaken which this paper continues to allow features characterizing objects to be included.

2.1 Revised Semiotic Model

The revised semiotic model (see Figure 1) was developed [8] to extend the application of the semiotic approach to multimedia content. The data plane was added to the basic semiotic model to separate the information being processed for rendering or sonifying from its content and the computer based expression mechanism. The description plane was also introduced to distinguish content description from the content itself. The description and data planes are needed because content and its expression are not linked in digital multimedia in the way they are in printed textual media. The traditional semiotic premise is that the text content is the ideas invoked in the viewers mind. This premise is maintained by placing the observer adjacent to the content plane quadrant in the model.

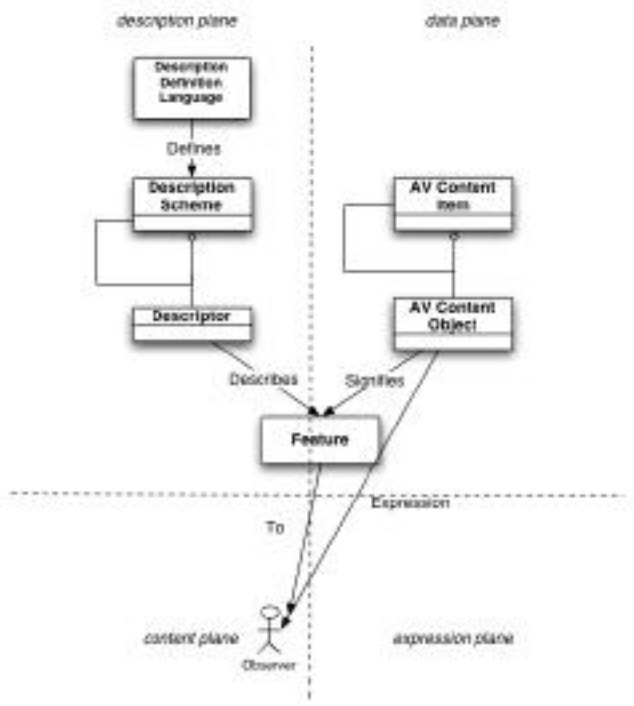

Figure 1.Revised semiotic model.

The model is useful because it provides a conceptual bridge between software modeling and a semiotic analysis of content. The model emphasizes that there is no immediate one to one relationship between the ideas invoked in the observers mind and their textual description. It can be seen that this contrasts strongly with the case of the ideas denoted and expressed by a piece of text. The model was originally developed to clarify terminology being proposed for adoption during the analysis of MPEG-7 [11] requirements. This clarification centered on exposing the terminological deficiencies in the shot and scene break terminology then being considered as the basis for temporal decomposition in MPEG-7. This points to the need to introduce mechanisms to account for the time based nature of much of multimedia content, which is undertaken in section 4.

2.2 Extended Semiotic Model

The extended semiotic model (see Figure 2) introduces the concepts of compressed data and compressed descriptions into the model. This was seen as essential in the context of multimedia content as much of the content of interest is compressed. The assumption that binary descriptions would need to taken into account in the modeling has been vindicated by the development of the Binary Format for MPEG-7 the BiM. This is MPEG-7's XML schema-based compression standard [9]. Hartley *et al.* [8] show the application of the extended model to a short segment of video content.

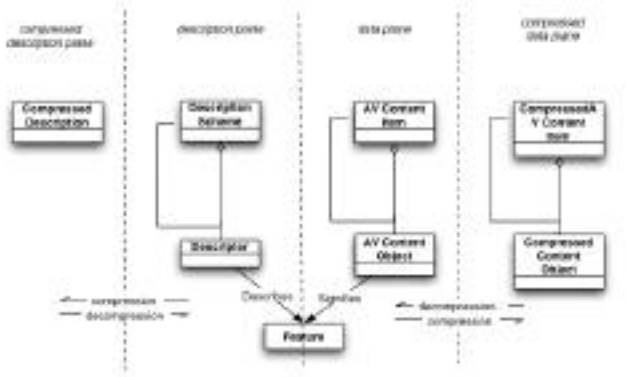

Figure 2. Extended semiotic model.

2.3 Further Revisions to the Semiotic Model

Consideration of Figures 1 and 2 leads to the conclusion that there are still deficiencies in modeling the relationships between the visual feature, the visual content object, the object description and the observer. Further revisions of the semiotic model have now been introduced to accommodate the need to distinguish: the feature sets used to *characterize* the object perceived by the viewer, the visual object itself, the concept invoked in the viewers mind and the schemata describing the visual object. In MPEG-7 [15] terminology the descriptors and description schemes describing the visual object include the features characterizing the object and the descriptive terms. The features *characterize* the visual content object at some level of decomposition of the audio-visual content item as opposed to describing the visual content object in any meaningful sense. The features characterizing an object are typically color or texture metrics. The introduction of image features into the model provides an improved framework for the description of still images. The framework and the associated construct of the image features characterizing the visual object "binding" to the conceptual terms used to describe the visual object is described and illustrated in diagrammatic form in terms of UML (see Figure 3). The model is also consistent with the earlier model. It becomes apparent that to provide an effective meaningful description of visual objects both visual features *characterizing* the visual object and meaningful terms *describing* the visual object are needed. These feature sets and other descriptive components are then combined or "bound" together either by a content describer of by the system. This introduces an important separation of concerns since it allows content modeling to be carried out

independently of feature extraction. It also allows meaningful descriptions and feature sets to be instantiated independently without the binding operation being carried out. In implementation the descriptive terms corresponding to the concept denoted by the visual object, the ideas connoted to the describer and the features characterizing the object are all separate class or attribute instances.

The ability to instantiate object description and feature sets without completing the binding operation allows greater flexibility. This is one of the factors that distinguish the approach adopted in the Automating Video Annotation (AVA) project and the associated AVA demonstrator developments at Lancaster from other approaches in that the binding operation is made overt rather being implicit.

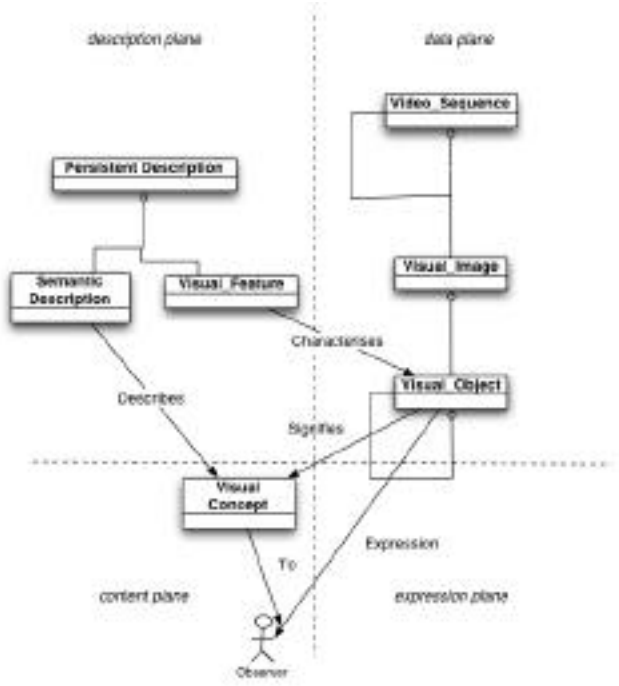

Figure 3. R2 semiotic model.

2.4 Knowledge Representation or Feature Extraction

It will be readily appreciated that there is minimal correspondence between the color and texture metrics referred to above and any meaningful classification in the knowledge engineering sense of the visual objects under consideration. This is supported by noting that only the correspondence between color terms such as red green etc. and physical colors has been investigated in any depth by Berlin [2] and that few descriptive terms exist for texture.

However the use of color and texture is now common and other metrics for visual object recognition and location in content-based retrieval applications have no common usage denotative basis. So typically these applications rely on using a query image and return images similar to the query. This highlights the arbitrary nature of the identifiers for visual objects and their association with descriptive terms when viewed from the feature space viewpoint.

2.5 Modeling Time

The modeling so far has been related to visual objects in images, which whilst allowing for decomposition of time based media into media objects has been primarily static. This is acceptable for still images however in multimedia content representation time dependent media must be modeled as well. To facilitate this modeling the distinction between separate time lines relating to content will be introduced namely; •*capture time,* •*expression time and* •*represented time.*

Content capture time is the time at which the content is recorded or captured. Content expression time is the time at which content is expressed to an observer and content represented time is the time portrayed in the content. These separate time lines are needed to allow the model to account for time-based media content. In turn the temporal relationships within each time line can be represented using interval graphs. Separating these time lines allows the entities visible to the observed at content expression time to be modeled in a way that supports Chatman's [4] distinction between process and stasis statements. These are described in section 4 after an outline of data-plane spatio-temporal decomposition techniques is given in section 3.

3. Spatio-temporal Segmentation

The bi-directional approach adopted in the AVA project combines both low-level feature extraction and high-level knowledge representation techniques. The low-level techniques that are used to provide both a temporal and spatial decomposition of moving picture data into temporal segments and visual objects will now be described. Implicit in this account is that some manual segmentation has to be carried out at some point to provide a basis for retrieval. The spatial techniques for object recognition and location are followed by an outline of the temporal techniques.

3.1 Visual Object Identification and Location

There are effectively two tasks that were identified by Swain and Ballard [19] that must be addressed that relate to segmentation of still images namely; visual object identification and location.

3.1.1 Histogram Intersection

Swain and Brown [20] described image histogram intersection as a method for visual object retrieval. More recently the technique has been investigated extensively by Smith [19] and Schiele [18] who both compare the effectiveness of extensions to the basic approach described and describe different combined color and texture approaches. An image histogram is defined (see equation 1) as an n-dimensional vector:

$$H_i(j), j = 1,...,n, \qquad (1)$$

Where n is the number of bins representing the number of grey levels or colors and *Hj* is the number of pixels in the image with the color *j*. Normalized histogram intersection provides a confidence value estimation method and is defined on a pair of

histograms designated the image I and the model M each containing n bins (equation 2).

$$H(I_j, M_j) = \frac{\sum_{j=1}^{n} \min(I_j, M_j)}{\sum_{j=1}^{n} M_j} \qquad (2)$$

This method provides an approach that has been used successfully for identifying the the presence of a visual object in a given image.

3.1.2 Histogram Back Projection

Visual object location in a still image can be achieved through histogram back projection. This algorithm first computes a ratio histogram from the model histogram and the image histogram (equation 3).

$$R_i = \min\left(\frac{M_i}{I_i}, 1\right) \qquad (3)$$

This operation results in a look up table in the color space representing how much of the searched object color is present in the image. Then in *back-projection* each pixel (x-y) of color *i* in the original image is replaced by R_i, and resulting peaks in the distribution of values will represent the expected locations of the object in the image.

3.1.3 Limitations in Histogram Intersection and Back Projection

Schiele [18] identifies that the histogram intersection technique has good immunity to changes in scale and rotation and limited occlusion. His findings indicate that the algorithm can support changes of scale of around 4:1. Ennesser and Medioni [6] have successfully applied a modified back projection algorithm that uses weighted histogram intersection to object location in cluttered scenes. This gives improved results to those obtained using the back projection algorithm. However even in relatively simple content such as the children's TV example considered here there are changes of scale and levels of occlusion are likely to defeat the capability of these algorithms. This is the problem that the binding concept is introduced to overcome.

3.2 Temporal Segmentation

An extensive body of literature exists on shot and other film production effect detection, which is reviewed by Aas et al. [1], Brunelli et al. [3], Koprinska and Carrato [13] and Lienhart [14]. This describes a wide variety of temporal segmentation approaches that map moving picture production effects such as cuts, dissolves and wipes etc. onto metrics derived from the expression plane time distributions of the one or more image features. These features can be extracted from the data or compressed, data planes. The majority of these techniques have difficulty discriminating slow moving objects from the effects that they seek to recognize. From the point of view of automating content description this limits the usefulness of the algorithms. At present in the AVA project a simple color difference threshold based approach has been adopted.

3.3 Open Issues and Assumptions

In the remainder of this analysis the following assumptions will be made about the capabilities of spatial and temporal segmentation techniques.

3.3.1 Spatial Segmentation Assumptions

It will be assumed in rest of this paper that a pair of algorithms can be used to identify and locate a given visual object over a range of scales up to 4:1 at a high level of confidence.

3.3.2 Temporal Segmentation Assumptions

Annotation obtained from production data or one of the techniques outlined in section 3.2 above can be used to provide a *expression* time temporal segmentation. However the level at which this segmentation will correspond to a temporal segmentation of *represented time* dependent on which segments are stasis statements and which process statements in the content. The distinction between stasis and process statements is described in section 4. This is regarded as an issue for future study and it will be assumed that this is a concern for higher level modeling than the instantiation of existents and existent contexts. So the assumption is made that expression time segmentation can be achieved that has some level of correspondence with production of effects.

4. Modeling and Describing Media Objects in Time

In the subsequent discussion some terminology will be adopted from narrative theory applied to moving pictures by Chatman [4] that characterizes all the actors and objects in a narrative as existents. The introduction of this terminology allows the modeling to be extended to include time whilst still maintaining the semiotic basis of the analysis.

4.1 Describing Visual Objects

Chatman's approach is to characterize the actors, objects and scene elements in a narrative as different classes as existents. This terminology is adopted here because it provides a convenient method to identify a base class for all objects visible in an image. He also distinguishes between stasis statements in narrative process statements. Stasis statements are those that do not move the story along but state the existence of something. In contrast process statements are those that denote a change of state. In moving picture sequences both types of statement result in elapsed expression time but not necessarily elapsed narrative or *represented time* in the terminology from section 2. This distinction confirms that there are limitations in the interpretation that the behavior of the temporal segmentation algorithms and supports the assumptions outlined above. In moving picture content where the process statements are not denoted by a voice track a process statement will typically involve motion which again will be considered as an issue for higher level modeling.

4.1.1 Describing Visual Objects in Still Images

Considering only the visual object description portion of figure 3 an existent description is obtained by combining the attributes needed to completely describe a visual object in a

still image. The term existent description is adopted to denote such a description.

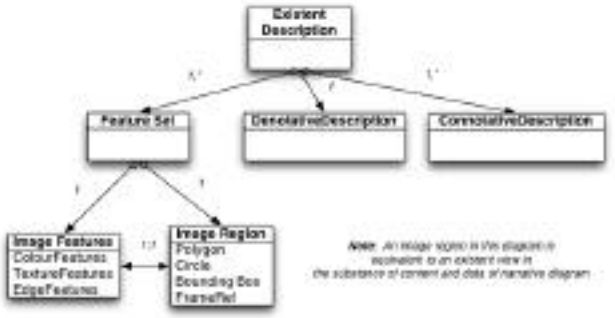

Figure 4. Existent Description.

The attributes that are needed are; (i) a feature set defining the visual features needed to identify and locate the visual object and their boundary, (ii) the objective or denotative description of the object and (iii) any additional connotative terms the describer may choose.

4.1.2 Describing Visual Objects in Time

The existent context is now introduced as the minimal unit of spatio-temporal decomposition for moving picture sequences. The existent context is defined as the set of images from a moving picture sequence that contain a visual object that can be recognized using the same set of image features for a *specific* pair of recognition and location algorithms. Clearly the temporal extent of the existent context is dependent on the capabilities of the recognition and location algorithms in question but the definition is independent of these capabilities. In fact there is scope for competitive evaluation of algorithms for existent context instantiation. So the existent description is extended to include temporal references or frame numbers (see Figure 5). The existent context provides a bridging mechanism between image features characterizing the data plane entities that correspond to a visual object perceived by an observer and the denotative and connotative description of the visual object. It does this in a way that is objectively verifiable whilst being limited by visual object recognition and location algorithm performance. For this reason it is preferred to the setting construct developed by Parkes [17], which is outlined below for comparison.

Visual objects in moving picture content are not just visible for single sequences of frames but may appear disappear or be occluded many times from the observer's viewpoint in any meaningful sequence. The existent context is therefore allowed to reference any number of different feature sets that are needed to support the objects recognition and location throughout the entire sequence. If an object is present in a sequence at a scale or level of occlusion different from that at which the bound existent context was initially instantiated another binding between the description and feature set occurs. This gives a failure driven mechanism for binding the feature sets to the descriptive elements.

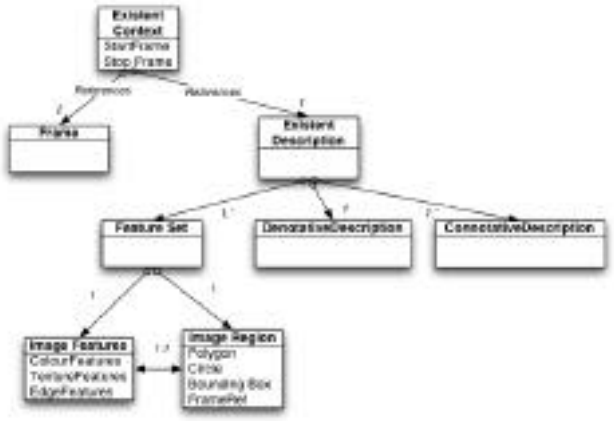

Figure 5. Existent Context.

4.1.3 The Setting

Parkes [17] in his approach to videodisc content description introduced the setting construct. He used the construct to provide a conceptual bridge between entities that were objectively visible in a still image and their objective description. Parkes defined the setting to be a set of images in a moving picture sequence sharing the same objectively visible state. The objectivity was dependent on the describer. The associated setting description was an objective description of that visible state. Underpinning the setting concept was the assumption that a still image of a motion image sequence could only have motion inferred from it rather than the still image being able to denote motion. This is assumption is invalidated by functional nuclear magnetic resonance imaging results [7] showing the same brain areas are excited when still images with implied motion are viewed as are excited when moving image sequences with comparable movement are viewed. The existent context in contrast provides a definition that is verifiable against feature extraction algorithm performance. It is also not dependent upon assumptions about the depiction of motion.

5. Applying the Model

A descriptive account of how the model has been applied to an episode of the UK children's television program Teletubbies[1] will now be given; a comparable numerical analysis is currently in progress. A small number of frames from the 'Favourite Things" episode of this popular program are reproduced. This account serves to illustrate the role of the constructs described in the earlier sections. This content was chosen because it combines simplicity of character, location and plot to be tractable enough to illustrate the concepts under discussion. Yet it is complex enough to be interesting and expose many of the issues present in more complex material. The "Favourite Things" episode in particular was chosen because the quest theme that it realizes is repeated four times involving a different teletubbie in turn.

[1] Images reproduced from Teletubbies Favourite Things are copyright BBC/Ragdoll Productions Ltd 1996 and are reproduced with permission.

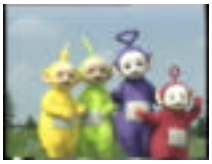

Figure 6 (1:50).

Figure 7 (3:40).

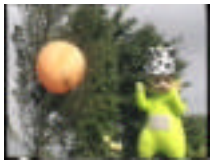

Figure 12 (05:17).

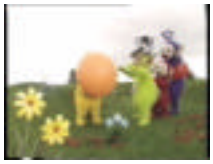

Figure 13 (05:38).

Many episodes of teletubbies begin with a title sequence where each of the characters is introduced in turn followed by a group shot (see Figure 6). The numbers correspond to approximate time codes for the frames. During the introductory sequence the describer would bind the name of each teletubby to the feature set characterizing each teletubby. Each of the teletubbies is then shown with their corresponding favorite thing. So additional feature sets would have to be instantiated and bound to the favorite things existent descriptions corresponding to the visual objects now visible (see Figure 7). The narrator informs us that LaLa's ball is missing, which is clearly visible in the shot shown in figure 7.

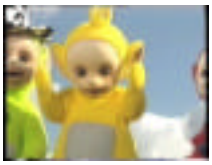

Figure 8 (4:05).

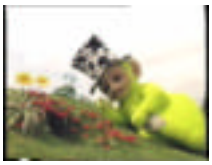

Figure 9 (04:21).

LaLa then expresses distress about the lost ball and there is a significant change of scale (see Figure 8). This results in one teletubby being obscured completely and the other two becoming severely occluded. So the existent context for Tinky Winky would end and additional feature sets instantiated and bound for Dipsy and Po. The search for the missing ball then begins. A number of different locations are visited and a cut occurs between each. The scale/occlusion level in these sequences is consistent with those already instantiated in the earlier sequences (see Figure 9).

Figure 10 (04:31).

Figure 11 (05:00).

In Figures 10 and 11 more locations are shown being visited by the teletubbies. Figure 11 in particular shows a significant level of scale change affecting two of the teletubbies. So a new feature set would need to be bound to the existent descriptions concerned. Several cuts also occur and there is considerable object motion in these sequences so the frames shown are representative only. It has already been noted that motion modeling is not considered in this analysis.

LaLa's missing ball is then shown in a new location requiring that a new existent description and existent context must be instantiated. The modeling of relationships between existents and existent contexts will be considered in further work.

Dipsy finds the ball (see Figure 12). The scales of both the ball and Dipsy are consistent with an existent context to feature set binding for both to objects. The last image shown in this analysis is that of the ball being returned to LaLa by all the teletubbies (see Figure 13). This image is particularly interesting because LaLa is severely occluded and the other teletubbies reenter the picture at different levels of occlusion. In subsequent sequences each of the other teletubbies favorite things is lost and subsequently found. In these stories many sequences already described are repeated. So the previous existent context bindings can be reused in the subsequent realizations of the search story except where the story differs from the one described here.

6. Conclusions

The overt binding of feature sets derived from image processing to denotative descriptions has been introduced. Both feature sets and existent descriptions can be instantiated without being bound to each other. This allows analysis to be carried out without modeling and vice-versa. The existent description and existent context constructs have been introduced. The existent description combines the feature sets with denotative descriptions through the binding mechanism. The existent context has been shown to be a preferable minimal unit of temporal decomposition to the setting. These components are used in the AVA approach to content description at Lancaster. Further extensions of the modeling to accommodate higher-level concepts, object relationships and object motion from a semiotic perspective will be the subject of a subsequent paper.

7. ACKNOWLEDGMENTS

UK EPSRC provided financial support for work contributing to this paper under grants GR L81/604 (MIM) and GR M81/755 (AVA). The author's participation in MPEG-7 standardization work was supported by grants GR S421871/01 and GR S421871/01

Hartley family resources have also provided financial support for this work.

8. REFERENCES

[1] Aas K. and Line L., A survey on: Content-based access to image and video databases. Technical Report Report Number 915, Norsk Regensentral (Norwegian Computing Center), Gaustadalleen 23, Post Office Box 114 Blindern, N-0314 Oslo, Norway, 1997.

[2] Berlin B. and Paul K., Basic Color terms: their universality and evolution. Berkeley University of California Press, 1969.

[3] Brunelli R., Mich O., and Modena C.M., A survey of the automatic indexing of video data. Journal of Computer Vision and Image Representation, 10:78–112, 1999.

[4] Chatman S. Story and Discourse: Narrative Structure in Fiction and in Film. ISBN 0-8014-9186-X. Cornell University, 1978, 1993.

[5] Eco U. A Theory of Semiotics. ISBN 0-253-35955. Indiana University Press, Bloomington, London, 1976.

[6] Ennesser F. and Gerard Medioni G., Finding Waldo or focus of attention using local color information. IEE Transactions on Pattern Analysis and Machine Intelligence, 17(8): 805–809, August 1995.

[7] Goebel G., Khorram-Sefat D., Muckli L., Hacker H., and Singer W., The constructive nature of vision: direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. European Journal of Neuroscience, 10:1563–1573, 1998.

[8] Hartley E., Parkes A.P., Hutchison D. A Conceptual Framework to Support Content-Based Multimedia Applications, pages 297–315. ISSN 0302-9743 LNCS 1629. Springer, 1999.

[9] ISO/IEC 15938 Multimedia content description interface Part 1 Systems ISO IEC, 2000.

[10] ISO IEC SC29 WG11. ISO/IEC 15398 Multimedia content description interface Part 5 Visual. ISO IEC, 2000.

[11] ISO MPEG-7 Requirements

[12] ISO IEC SC29 WG11. ISO/IEC 15398 Multimedia content description interface Part 5 Multimedia Description Schemes. ISO IEC, 2000.

[13] Koprinska I. and Carrato S., Detecting and classsifying video shot boundaries in mpeg compressed sequences. In Proceedings of the IX European Signal Processing Conference (EUSIPCO), pages 1729–1732, Rhodes, 1998.

[14] Lienhart R., Comparison of automatic shot boundary detection algorithms. Image Video Processing, VII, 1999

[15] Manjunath B., Salembier P., and Sikora T., editors. Introduciton to MPEG-7 Multimedia Content Description Interface. ISBN 0 471 48678 7. Wiley, 2002.

[16] Nack F. and Parkes A. P., Towards the automated editing of theme orientated video sequences. Applied Artificial Intelligence, 11(4):331–366, 1997.

[17] Parkes A.P., Settings and the settings structure: The description and automated propagation of networks for perusing videodisc image states. Proceedings of the Twelfth Annual International ACMSIGIR on Research and Development in Information Retrieval, 229–238, June 1989.

[18] Schiele B. and Crowley J.L. Recognition without Correspondence using Multidimensional Receptive Field Histograms, 36 (1):31-50, 2000

[19] J. R. Smith J.R. and Chang S.-F., VisualSeek: a fully automated content-based image query system. ACM Multimedia '96, November, 1996

[20] Swain M.T. and Ballard D.H., Color indexing. International Journal of Computer Vision (IJCV), 7(1): 11–32, 1991