

Video Retrieval using an MPEG-7 Based Inference Network

Andrew Graves
Department of Computer Science
Queen Mary, University of London
London, England, E1 4NS
andrew@dcs.qmul.ac.uk

Mounia Lalmas
Department of Computer Science
Queen Mary, University of London
London, England, E1 4NS
mounia@dcs.qmul.ac.uk

ABSTRACT

This work proposes a model for video retrieval based upon the inference network model. The document network is constructed using video metadata encoded using MPEG-7 and captures information pertaining to the structural aspects (video breakdown into shots and scenes), conceptual aspects (video, scene and shot content) and contextual aspects (context information about the position of conceptual content within the document). The retrieval process a) exploits the distribution of evidence among the shots to perform ranking of different levels of granularity, b) addresses the idea that evidence may be inherited during evaluation, and c) exploits the contextual information to perform constrained queries.

Keywords

Structured Video Retrieval, MPEG-7, Inference network, Combination of evidence

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Retrieval models*; H.3.7 [Information Storage and Retrieval]: Digital Libraries—*Standards*

General Terms

Theory, Management

1. INTRODUCTION

Recent years have witnessed a large increase in the usage and generation of digital video information. As a consequence, there is a need for advanced video storage, transmission, indexing and retrieval techniques. The goal of video retrieval systems is to find video data upon demand, but not necessarily to understand it. An initial step in video processing systems is to ascertain the shots and scenes in the video by performing shot-boundary-detection and scene-change-detection. This is an active research area whereby a video is decomposed into its constituent shots (syntactic boundaries) and scenes (semantic boundaries) [17].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '02, August 11-15, 2002, Tampere, Finland.
Copyright 2002 ACM 1-58113-561-0/02/0008 ...\$5.00.

The common video retrieval method has been to adopt a query-by-example content-based approach, whereby the video data is quantified according to the detectable features (such as colour, texture or motion) and compared against a submitted query clip (or keyframe) to form a similarity metric, which is then used to perform ranking. The effectiveness of this approach is limited owing to the semantic gap [5] that exists between the user and the system. Whereas the user has an inherent information need expressed in semantics, or high-level concepts, the system operates according to the low-level features. Either the user has to make the semantic-content translation or has to find a suitable video clip (or keyframe) to represent the query. Adopting this approach also makes the assumption that semantically similar video-clips have a small distance between them in feature space which is not always the case.

Clearly there is the need to adopt a semantic based approach, whereby the content-semantic translation is done by the system and the query is expressed in semantic keywords. Such a system would involve the automatic extraction of semantics (or automatic generation of video annotation) [13], the storage of these semantics in some suitable arbitrary metadata format, and finally, performing keyword based retrieval based upon the metadata [9].

MPEG-7, formally called “*Multimedia Content Description Interface*”, is a new standard that aims at describing the content of multimedia data by attaching metadata to multimedia content. MPEG-7 specifies a standard set of description tools, which can be used to describe various types of multimedia information. In this paper, we are concerned with the development of access methods for searching video data using MPEG-7 annotation. Our aim is not to derive the MPEG-7 annotation that shall be associated with a video stream, but to develop a model that exploits the characteristics of MPEG-7 for the effective retrieval of video data.

We use the inference network as a basis for the model because the inference network is able to represent the distributed multitude of evidence abundant in MPEG-7, in particular structural, conceptual, and contextual aspects. The video structure (ie: video breakdown into scenes and shots) is extensively dealt with in MPEG-7 and allows us to produce a ranking for each structural element. The concepts within the video are extracted from certain MPEG-7 elements known to contain concepts (eg: using the terms within the video <Abstract> element). The information regarding the document-content relationship in MPEG-7 (eg: the position of the concept within the document metadata) is considered to provide context. Such contextual information can be exploited to perform constrained queries (eg: retrieve documents with a certain concept within a certain context).

The rest of the paper is structured as follows: Section 2 introduces the background material of MPEG-7 and inference network; Section 3 introduces our model for MPEG-7 based retrieval; Sec-

tion 4 presents an example network showing how the probabilities are estimated; Section 5 presents an illustration of video retrieval; and Section 6 presents some concluding remarks.

2. BACKGROUND

We consider video indexing and retrieval systems to consist of two modules: the video analysis module (or *indexer*); and the video retrieval module. The video analysis module processes the video to extract low-level and high-level information, both of which can then be stored in an MPEG-7 file. In this work we are concerned with exploiting high-level MPEG-7 annotation, as well as the video structure, for performing retrieval using the inference network model.

2.1 Structured Document Retrieval

We consider a video document to be a structured document as illustrated in figure 1. The video consists of scenes which themselves consist of shots. This structure can be ascertained using video processing techniques such as shot-boundary-detection and scene-boundary-detection.

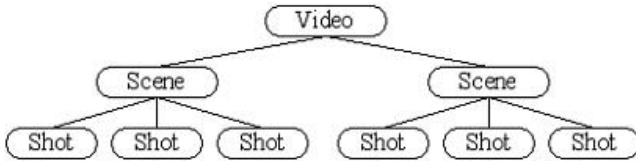


Figure 1: The structural nature of video

We aim to build a retrieval system that is capable of ranking different levels of granularity [7] (ie: video, scene and shot in the result ranking), as the individual scenes and shots may be more relevant to the query than the entire video. We consider the evidence pertaining to a video object as being contained within the scenes, and a similar dissemination of scene evidence occurs among the shots. To perform retrieval the question of evidence aggregation must be answered.

In previous work [16], we aimed to provide the best entry points into a hierarchical structure, given the aggregated evidence in the individual objects. In the model presented in this paper, we perform an estimation of document relevance for each retrievable element in turn where the distributed evidence is considered. The result is a ranked list consisting of the various document component types

2.2 MPEG-7

MPEG-7 is a standard derived by the Moving Pictures Expert Group [11]. It has the ability to describe the low-level features, high-level semantics and structural aspects of any multimedia file. The standard is both extensive and extendible and is becoming increasingly popular [6, 15, 14] as the metadata format of choice for building practical multimedia information systems.

MPEG-7 consists of the following three main components, the relationship between which is shown in figure 2.

- Description Definition Language (DDL), based on the XML Scheme language.
- Descriptors (D), representing the individual items of information.
- Description Schemes (DS), modelling the organisation of the Descriptors.

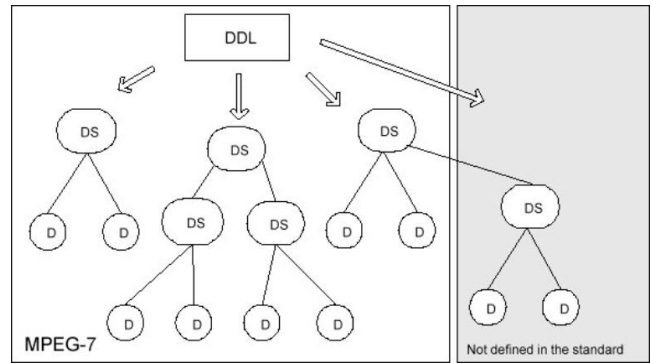


Figure 2: The main MPEG-7 components

We are concerned with the representation of the structural aspects (ie: video, scene and shots) and the high-level semantics (or concepts). The structural breakdown of a video can be described using the `<Segment>` and `<SegmentDecomposition>` DSs. The concepts associated to the video, scene and shot elements are then described in various Descriptors including the `<TextAnnotation>` Descriptor, an example of which is given below.

```

<TextAnnotation>
  <FreeTextAnnotation>
    Basil attempts to mend the
    car without success
  </FreeTextAnnotation>
  <StructuredAnnotation>
    <Who>Basil</Who>
    <WhatObject>Car</WhatObject>
    <WhatAction>Mend</WhatAction>
    <Where>Carpark</Where>
  </StructuredAnnotation>
</TextAnnotation>
  
```

The above excerpt contains high-level semantics that could be associated with either the video or one of the scenes or shots. It is important to note the unusual behaviour of evidence within this structure: if this information were associated with the video, then it also applies to the scenes and shots (a form of *evidence inheritance*); if this information were associated with a scene or shot, then it inherently applies to the video as the video contains the scene or shot.

2.3 Inference Network Model

To model the evidence contained within the MPEG-7 documents, and specifically, the structural and unusual conceptual characteristic of evidence inheritance, we adopt the Inference Network model for IR as developed by Turtle [18]. The Inference Network (IN) model has ability to perform a ranking given many sources of evidence by performing a combination of evidence. The IN model is basically a Bayesian Network used to model documents, the document contents, and the query. The IN consists of two sub-networks: the Document Network (DN) produced during indexing and then static during retrieval; the Query Network (QN) produced from the query text during retrieval.

The DN represents the document collection and consists of nodes for each document (called document nodes) and nodes for each concept with the collection (document concept nodes). The document nodes represent the retrievable units within the network, that is, those items we wish to see in the resultant ranking. A causal link (represented as \rightarrow) between document node and the document concept node indicates that the document content is represented by the

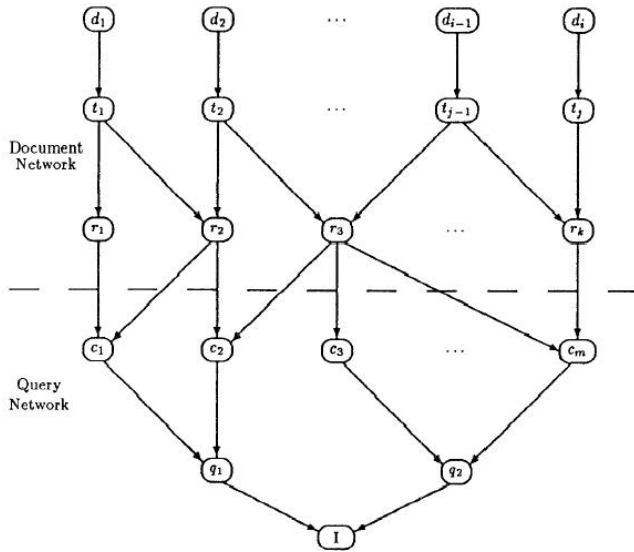


Figure 3: An Inference Network

concept. Each link contains a conditional probability, or weight, to indicate the strength of the relationship. The evaluation of a node is done using the value of the parent nodes and the conditional probabilities.

The QN represents the submitted query and consists of a framework of nodes that represent the required concepts (query concept nodes) and the operators (query operator nodes), connected in an inverted tree structure. The QN is constructed with a final leaf node I that represents the user Information Need. The framework permits statistical operators and statistical approximations of the Boolean operators, a number of which are given in Table 1 (as in the INQUERY implementation [2]).

#and	AND the terms
#or	OR the terms
#not	Negate the term (incoming belief)
#sum	Sum of the incoming beliefs
#wsum	Weighted sum of the incoming beliefs
#max	Maximum of the incoming beliefs

Table 1: Operators supported by the IN model

Two further processes are done to perform retrieval: the attachment process, where by the QN is attached to the DN to form the complete IN and is done where concepts in both networks are the same; the evaluation process, whereby the complete IN is evaluated for each document node to form the probability of the relevance to the query. The evaluation is initialised by setting the output of one document node to 1 and all the other document nodes to 0. This is done for each document node in turn and the network is evaluated (see [19] for exact detail and examples on how nodes are evaluated). The probability of document relevance is taken from the final node I and is used to produce the ranking.

3. MODEL

The basis for our model is the Inference Network (IN) model as it has the ability to model the structural, conceptual and contextual aspects available in MPEG-7. This framework has been successfully

used before [12] to perform retrieval of SGML documents by modelling the structure and contents. This section discusses our Document and Query sub-networks; the modifications to the attachment and evaluation processes; and how the probabilities within the network are estimated.

3.1 Document Network

The DN captures the structural elements of the documents, the concepts found in the collection, and the relationships (and strengths) between each document and concept.

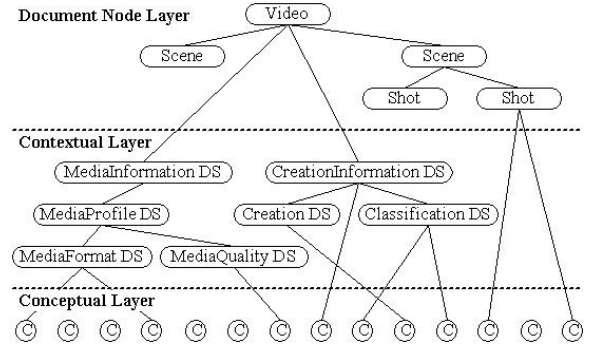


Figure 4: The three-layered Document Network

It consists of three layers as follows:

- The document node layer, which contains nodes that represent the retrievable units in the collection.
- The contextual layer, which contains nodes that represent the contextual information about the document→concept links.
- The conceptual layer, which contains nodes that represent all the concepts in the collection.

In the document node layer we have a hierarchical structure of document nodes to represent the structure of the video as represented in the MPEG-7 file. Each video is thus represented by a Video node, each of which can contain Scene nodes to represent scenes, each of which can contain Shot nodes to represent the shots. This structure can be extracted from the MPEG-7 structural components [11]. The Video→Scene and Scene→Shot links represents the decomposition of the Video into smaller parts, and the conditional probabilities on these links indicate the importance of the child element with respect to the parent (ie: how much a particular Scene contributes to the content of the Video).

The conceptual layer contains concept nodes that represent the identified concepts within the document collection. Concepts are identified during an extraction process where each document is parsed and concepts recognised (as in INQUERY). The presence of a concept within a document is represented as a relationship (direct or indirect) between the document node and the concept node. In reality, such a link exists for every document→concept relationship and the non-presence of a concept is indicated with a conditional probability of 0.5 (as thus, during evaluation, the value of the parent will not influence the value of the child). Our methods for estimating these conditional probabilities are discussed in section 3.5.

The contextual layer contains context nodes. The layer attempts to exploit the rich structure that is available within the MPEG-7 documents, namely the location of the concept within DS hierarchy.

This structural information can be used for both:

- Enabling a more precise estimation of how much a concept represents a document considering the context (eg: contexts may be of variable quality as some may be a better indicator of content than others). Concept nodes are included in the DN for all of the MPEG-7 elements we wish to monitor, resulting in Document→Context(s)→Concept relationships;
- Enable the identification of concepts that occur only within a particular context (eg: the query concerns the “bbc” concept that occurs only within the “Creation DS” context). This is exploited in our query network using a constrained concept.

3.2 Query Network

Our QN consists of the standard query concept nodes and query operator nodes as previously described. The operators supported are those in Table 1, which operate in the same manner as in INQUERY. Our QN also includes query context nodes, which are associated to a query concept node to form a constraint. The constraints specify the desired context of the concept within the expected matching documents (eg: we wish the “bbc” concept to occur under the “Creation DS” context). Our constraints are of two types, normal or complex.

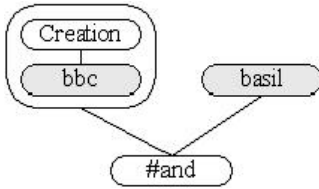


Figure 5: A normal constraint

The QN in figure 5 contains two query concept nodes (these are grey), a single query operator node (begins with #) and a query constraint node. The query constraint node is attached to the “bbc” concept node to form a normal constraint. This is denoted diagrammatically by enclosing the query concept node with the query constraint node. This query structure specifies that the query concept node “bbc” should only be attached to document concept nodes “bbc” which satisfy the constraint, that is, those nodes which occur under the context “Creation”.

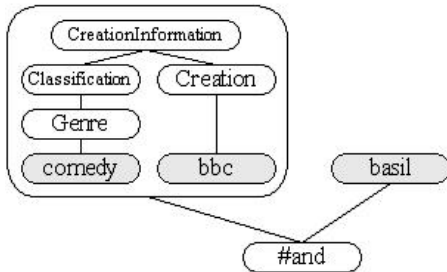


Figure 6: A complex constraint

The QN in figure 6 shows an example of a complex constraint. In this case, a structure is specified in the query that we wish to occur in the matching documents. The structure shown consists of two concepts and their relationship via a number of contexts. The query therefore demands those documents that match this sub-query structure. The constraints are considered in the attachment process described in section 3.3.

In addition therefore to the standard INQUERY query text, we propose two new keywords that allow the construction of normal and complex constraints: #constraint for placing a normal constraint upon a concept; and #tree for creating a complex constraint. Two examples are given below which represent the QNs shown in figures 5 and 6.

```
#and(#constraint(Creation, "BBC") "Basil")

#and(#tree(CreationInformation,
  #constraint(Classification/Genre, "Comedy")
  #constraint(Creation, "BBC") "Basil")
```

3.3 Attachment

The attachment process is performed during retrieval once the QN has been built and requires attaching to the DN. The result of the process is a number of DN→QN links where the concepts are considered to be the same, and where any specified constraints are matched. This link contains a weight, or conditional probability, that we can use to indicate the strength of the attachment (ie: how close the concepts and constraints matched).

The attachment process performs the following tasks:

1. Create candidate attachments (DN→QN links) by comparing the concepts in the QN against the concepts in the DN. In our implementation we compare the text of the QN concept nodes against the text of the DN concept nodes and create a link when it is the same. At this point we could consider miss-spellings, localisation, thesauri and stemming [1] in order to intelligently group similar concepts. We call this *conceptual fusion*.
2. Create attachments according to the constraints.
 - (a) For unconstrained candidate attachments we create a firm attachment (the same for every document node).
 - (b) For constrained candidate attachments we analyse the constraints. This is performed individually for every document node as only nodes that occur under the document node are considered to be part of the candidate DN structure. We calculate the *Edit Distance* (ED) [20] between the candidate DN structure and the QN concept constraint by counting the number of insertions, deletions and amendments that would be necessary to transform the former into the latter. This results in an ED metric for each constrained query (normal or complex) for each document node. This metric measures the closeness of fit.
3. For each document node we then create document specific attachments from the candidate attachments either using a threshold or a weighted link method. In the former we create the attachment using a weight of 1 (the conditional probability on the DN→QN link) if the ED is below a specified threshold. In the latter we create the attachment and use the ED to form the weight using the formula:

$$P(\text{QN node} \mid \text{DN node}) = 1 / (\text{ED} + 1)$$

The conditional probability represents the strength in the belief in the attachment considering how closely the structures fit (as ED tends to 0 the weight tends towards 1). At this point, we can also use the information about the conceptual fusion process (use of the thesauri, localisation, etc) in calculating the conditional probability on the DN→QN link. One advantage of using the weighted link method is that it produces a larger results list.

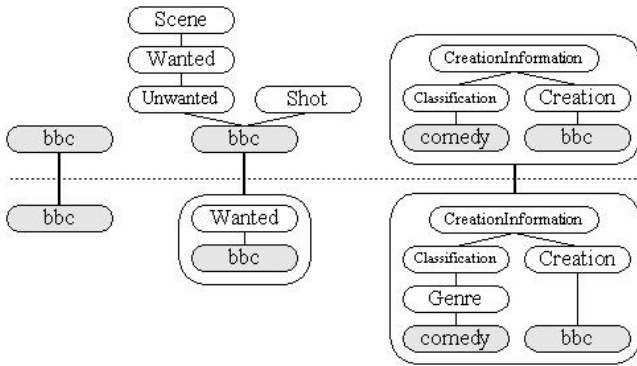


Figure 7: (a)(b)(c) Three examples of attachment

The result of the attachment process is a set of firm attachments and a set document specific attachments for each document node. This is known as the *complete network*. Figure 7 demonstrates three examples of attachment where the structures above the dotted line are in the DN and those below are in the QN. These examples are described below.

a) Shows an example of an unconstrained query concept and a matching document concept node, so a firm attachment is created with $P(QN|DN) = 1$.

b) Shows an example of a normally constrained query concept. The “Creation” document node satisfies the “Wanted” constraint with an $ED = 1$ if we delete the “Unwanted” node. We then create a document specific attachment. Using the threshold method we create the attachment with $P(QN|DN) = 0$ or 1 according to whether the ED is below the threshold. Using the weighted link method we create the attachment with $P(QN|DN) = 1 / (1+1) = 0.5$.

c) Shows an example of a complex constraint. The DN candidate structure shown satisfies the constraint as the root context node “CreationInformation” and the two concept nodes “comedy” and “bbc” all occur.

3.4 Evaluation

The evaluation process is performed for each document node using the complete network. The network is evaluated for each document node using the document specific attachments for that document. The result of the network evaluation is taken from the information need node in the QN. In addition to the standard method of complete IN evaluation [18, 19, 2], we consider two methods that exploit the structure within the DN:

- Link Inheritance (LI), is the situation when a child node can inherit context nodes from the parent nodes.
- Path Cropping (PC), is the situation where a parental contribution to a DN concept node is cropped (removed) as it does not satisfy the constraint on the attached QN query concept node.

LI is illustrated in figure 8. The Video node has two child document nodes that indicate the structure, the Scene and Shot nodes. The Video node also contains a context “Creation” (which can contain links to concept nodes). The idea of link inheritance is that this context also applies to the children of the Video node, the Scene and Shot nodes. For example, the context may contain concept nodes that describe when the video was created, information that is applicable to the scenes and shots within the video. The child document nodes therefore inherit the Video→Creation link from their parent during evaluation.

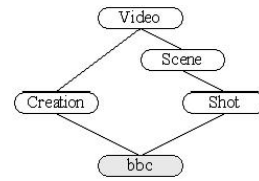


Figure 8: Link Inheritance

Although the inherited context is applicable to the child nodes, it becomes less influential as the number of generations between the child and the context increases. In our example, the “Creation” context is mostly applicable to the Video node, then the Scene node, and least applicable to the Shot node. To capture this fact, we degrade the conditional probability on the inherited Video→Creation link. We refer to this as Link Inheritance with Degradation (LID). The size of degradation is calculated according to size of the generation gap (the node distance between the real parent of the context and the node inheriting the context). The degradation could also consider the frame duration ratio.

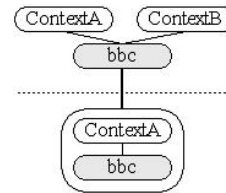


Figure 9: Path Cropping

PC is illustrated in figure 9. The figure shows a document concept node “bbc” with two parental contexts, attached successfully to a normally constrained query concept node. During the evaluation of the document concept node, we can choose to ignore the parental influence “ContextB” as this particular concept is not specified within the constraint. That is, the influence of this parent is not of interest to the query and hence is cropped; the document context node in the figure is evaluated as though it has only one parent.

3.5 Probability Estimation

The final part of our model is concerned with specifying how the conditional probabilities (or weights) between the nodes are estimated. A high conditional probability between two nodes would indicate that they are closely coupled and the child value should closely reflect the value of the parent. The weights need only be estimated for the DN as the QN uses a fixed behaviour according to the operators.

Two types of conditional probability need to be estimated:

- Context→Context. Between a parental context and a child context. Two subtypes:
 - Structural (eg: Video→Scene)
 - Contextual (eg: Video→CreationInformation)
- Context→Concept. Between a context and the concepts that are associated with it.

The structural probabilities are estimated using *duration ratio* information (duration-of-parent/duration-of-child), as each document node has a frame duration encoded in MPEG-7. The contextual probabilities are estimated using *context sibling information*

using $(1/\text{total-number-of-siblings})$. Alternatively, we could consider the context size, frequency, or perceived quality of the child context in this estimation [8].

We estimate the Context→Concept probability according to two factors: the statistical properties of the concept (eg: term frequency, inverse document frequency [1]); and information about how the concept was extracted.

$$\text{Weight} = \text{tf}(t,d) * \text{idf}(t)$$

$$\text{Pweight} = 0.5 + (0.5 * \text{weight})$$

The statistical properties of the concept provide information about the quality of the term as a discriminating factor and information about how much a document (or context in our case) is represented by the term. We use the above Weight formula to calculate the conditional probability from the term statistics. We use the above Pweight formula [19] throughout the network in order to ensure that the weight is above 0.5 and therefore a positive influence on the result.

```
<StructuredAnnotation>
  <Who>Basil,Sybil,Andre</Who>
</StructuredAnnotation>
```

The MPEG-7 file itself can provide clues about the quality of a term which we can analyse during the extraction process. In the above example MPEG-7 excerpt, three concepts (Basil, Sybil and Andre) are contained within a single MPEG-7 Descriptor using a comma delimiter. Each concept is therefore not as representative as a single concept as it has two descriptor siblings. We use this information to refine the probabilities using $(1/\text{total-number-of-siblings})$.

4. EXAMPLE NETWORK

This section describes an example of how the probabilities within the document network are estimated. The resultant example network is used later in section 5.

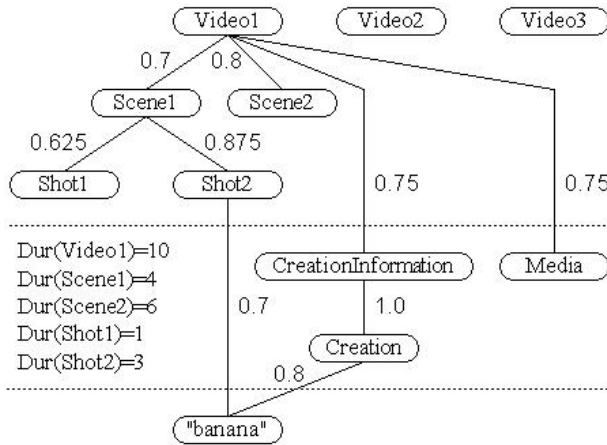


Figure 10: An Example DN

The example DN shown in figure 10 contains 3 Videos, 2 Scenes and 2 Shots, and also two occurrences of the concept “banana” within two different contexts:

```
Video1->CreationInformation->Creation->Concept
Video1->Scene1->Shot2->Concept.
```

Given the structure in figure 10, where Dur(Element) states the duration of the document node, the structural Context→Context probabilities are estimated as follows: (using the duration ratio and Pweight formula)

$$\begin{aligned} P(\text{Scene1}|\text{Video1}) \text{ Duration ratio} &= 4/10 = 0.4 \\ P\text{Weight} &= 0.5 + (0.5 * 0.4) = 0.7 \\ P(\text{Scene2}|\text{Video1}) \text{ Duration ratio} &= 6/10 = 0.6 \\ P\text{Weight} &= 0.5 + (0.5 * 0.6) = 0.8 \\ P(\text{Shot1}|\text{Scene1}) \text{ Duration ratio} &= 1/4 = 0.25 \\ P\text{Weight} &= 0.5 + (0.5 * 0.25) = 0.625 \\ P(\text{Shot2}|\text{Scene1}) \text{ Duration ratio} &= 3/4 = 0.75 \\ P\text{Weight} &= 0.5 + (0.5 * 0.75) = 0.875 \end{aligned}$$

The contextual Context→Context probabilities are estimated as follows: (using the total-number-of-siblings and Pweight formula)

$$\begin{aligned} P(\text{CreationInformation}|\text{Video1}) \text{ and } P(\text{Media}|\text{Video1}) \\ \text{Total number of siblings} &= 2 \\ \text{Sibling weight} &= 1 / 2 = 0.5 \\ P\text{Weight} &= 0.5 + (0.5 * 0.5) = 0.75 \\ P(\text{Creation}|\text{CreationInformation}) \\ \text{Total number of siblings} &= 1 \\ \text{Sibling weight} &= 1 / 1 = 1.0 \\ P\text{Weight} &= 0.5 + (0.5 * 1.0) = 1.0 \end{aligned}$$

In our model the Context→Concept probabilities are estimated using term statistics. However, in our example DN we have too few terms to perform this calculation so we have assigned the values:

$$\begin{aligned} P(\text{banana}|\text{Creation}) &= 0.8 \\ P(\text{banana}|\text{Shot2}) &= 0.7 \end{aligned}$$

The probability estimation is conducted during the indexing phase resulting in the DN shown in figure 10. The DN consists of a number of nodes and probabilistic links between them. All of conditional probabilities are positive (ie: >0.5, owing to the Pweight formula), meaning that those units that influence the parents of the “banana” concept (ie: Creation and Shot2) will be ranked higher.

5. ILLUSTRATION

To illustrate our approach, we implemented our model with two major modules: the extraction module; the evaluation module. The extraction module builds the DN from the MPEG-7 files, using the structural aspects when building the document node layer and other select MPEG-7 DSs and Descriptors for building the contextual and conceptual layers. At this point the conditional probabilities are also estimated. The evaluation module performs the query processing and evaluation according to the system parameters.

We conducted two experiments on the system which illustrate the working of the model. In particular we wish to examine whether the model can produce a good ranking; whether different levels of granularity can be retrieved; and whether the novel elements (constrained queries, link inheritance, path cropping) influence the results.

```
#constraint(CreationInformation, "banana")
```

The first experiment uses the example DN described in section 4 and the above query. The query states that we wish to retrieve elements (document nodes) containing the concept “banana” that adhere to the constraint “CreationInformation”. Note that the constraint will only be enforced on tests where the functionality is enabled.

0.3640 Video1	0.4550 Shot2	0.3850 Shot2
0.2450 Shot2	0.4225 Scene1	0.3738 Scene1
0.2275 Scene1	0.3640 Video1	0.3640 Video1
0.1050 Shot1	0.1950 Shot1	0.1725 Scene2
0.1050 Scene2	0.1950 Scene2	0.1650 Shot1
0.1050 Video2	0.1050 Video2	0.1050 Video2
0.1050 Video3	0.1050 Video3	0.1050 Video3

Table 2: The results of tests 1, 2 and 3

Test 1 was performed with no parameters (no link inheritance, no constraints, no path cropping). Test 2 was performed with LI only. Test 3 was performed with LID only. The results shown in table 2 are thought to show promise.

With no parameters, Video1 is ranked 1st as this has influence over both parental influences of the concept (the first path is via Video1→CreationInformation→Creation→Concept, second path is via Shot2→Concept). Shot2 is ranked 2nd as this is the closest to the concept on the second path. The other elements are ranked according to the second path influence as only Video1 influences the first path. Note that if Shot2=0 and Creation=0 then no evidence exists for the concept, however, due to the uncertain probabilistic nature of the network the result is 0.1050.

With the introduction of LI in Test 2, the child nodes of Video1 now inherit first path and thus are ranked higher. Shot2, Scene1 and Video1 are now ranked 1st, 2nd and 3rd, according to the amount of influence they exert on the second path. Shot1 and Scene2 are ranked lower as these have no influence on the second path, but are ranked above Video2 and Video3 as these have no influence on either path. This in particular was thought to be a good result.

In Test 3 LID produces a similar result but is thought to be more satisfactory. Scene2 is now ranked above Shot1 because the degradation is larger according to the number of generations. Shot1 is two generations from Video1 whereas Scene2 is only one generation.

We then performed tests with the constraints enabled. The Video1 node is the only document node that satisfies the constraint with an ED of 1 (as there is one extra node “Creation”). Using the threshold method with threshold = 3 the Video1 obtains the same probability as in Test 1 as the document specific attachment is used. All the other document nodes score 0. With threshold = 0 the score is 0 for the Video1. These results proved that the constraint was being enforced. Using the weighted link method we found that the score was slightly lower. When we introduced PC we found that the score was slightly higher as the concept node is only calculated with a single parent and the uncertainty introduced from the second parent is removed.

The second experiment uses a DN generated from a small MPEG-7 collection consisting of: two comedy clips [3, 4] and one drama clip [10]; each approximately 10 minutes long. We performed shot detection of 329 shots. Scenes were then created manually by grouping the semantically similar shots together. Finally annotation was manually added:

- <Abstract> using the official video box abstract.
- <StructuredAnnotation> for each scene specifying exactly the characters, the location and additional facts.
- <FreeTextAnnotation> for each video, scene and shot, written to describe the action.
- <FreeTextAnnotation> for each shot containing all of the speech that occurs.



Figure 11: Example Keyframes

Although automatic generation of video metadata is desirable for many applications, manual annotations are still very much in use. The model described in this paper can be applied to MPEG-7 metadata whether created automatically or manually. It must be noted that the quality of the retrieval is dependent upon the quality of the annotations.

The MPEG-7 files in the collection were then parsed to form the DN. The DN is stored as XML an example of which is given below:

```
<Video id="GourmetNight"\>
  <Scene id="TheMajorComplains"\>
    <Concept weight="0.5106"\>mushroom</Concept>
    <Concept weight="0.5211">mushrooms</Concept>
    <Shot id="Shot_68"/>
    <Shot id="Shot_69">
      <Concept weight="0.5211">mushrooms
    </Concept>
    </Shot>
    <Shot id="Shot_70"/>
  </Scene>
</Video>
```

We then conducted four tests using the queries:

```
Q1) #or("mushroom" "mushrooms")
Q2) #or("chips" #and("salad" "cream"))
Q3) #constraint(Classification "comedy")
Q4) #or("bedroom" "room" "rooms")
```

We enabled LID, constraints and PC. The results for the four queries were as expected. Q1 produced a ranking of Scene, Video, Shot which replicates the findings from the first experiment. Q2 also ranked highly the scenes and shots that contained the evidence. Q3 produced a ranking for the two comedy videos but not the drama video which was the expected result considering the constraint. Q4 was used for generating recall/precision metrics using a subjective assessment of the most relevant documents. The average was 69.25 however the results were not significant owing to a) the small size of the test collection, and b) the lack of independent queries and relevance assessments. However, these initial results demonstrate the potential of the model working with an MPEG-7 collection.

6. CONCLUDING REMARKS

In this work we have adopted a metadata-based approach to video retrieval. We propose the use of MPEG-7 files to encode the content and to hold the video structure. MPEG-7 provides a rich set of tools for this purpose. We consider the MPEG-7 Description Scheme structure to contain additional knowledge about the document-content relationship (we call this context).

We presented a model based upon the Inference Network model that preserves and uses the structural, conceptual and contextual aspects of MPEG-7. We use the structural aspects to allow the retrieval of different levels of granularity where the evidence for a higher-level unit is contained within its children (eg: the evidence for a scene is contained in the shots). We use the duration ratio between structural units to estimate the conditional probabilities

between them. We extract document concepts from the conceptual aspects of MPEG-7 and use term statistics and extraction information to estimate the probabilities. We use the contextual aspects to decompose the documents into smaller contexts and to allow constrained queries. In addition, we perform evidence inheritance during the evaluation process.

The quality of the results when performing a metadata-based approach is based upon the quality of the metadata itself. Presently this is low and thus the experimental results presented in this paper are not conclusive. However, the initial results presented indicate that the metadata-based approach is well founded. Our next steps are to establish a larger more consistent MPEG-7 collection and to consider further how the inheritance of evidence should effect the retrieval process. This can be done in conjunction with TREC when an MPEG-7 collection becomes available.

7. ACKNOWLEDGEMENTS

We wish to acknowledge the MPEG-7 generation tool provided by Laboratoires d'Electronique Philips. The first author was funded by the EPSRC.

8. REFERENCES

- [1] R. Baeza-Yates and N. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, Essex, England, 1999.
- [2] J.P. Callen, W.B. Croft, and S.M. Harding. The inquiry retrieval system. In *3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.
- [3] J. Cleese and C. Booth. *Fawlty towers: Gourmet night*. BBC ©, 1975.
- [4] J. Cleese and C. Booth. *Fawlty towers: Communication problems*. BBC ©, 1979.
- [5] C. Colombo, A. Del Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*, 6(3):38–53, July-September 1999.
- [6] N. Fatemi and O. Abou Khaled. Indexing and retrieval of tv news programs based on mpeg-7. In *IEEE International Conference on Consumer Electronics*, Los Angeles, California, USA, June 2001.
- [7] N. Fuhr and K. Großjohann. Xirql: A query language for information retrieval in xml documents. In *24th ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 172–180, New Orleans, Louisiana, USA, 2001.
- [8] A. Graves. Video indexing and retrieval using an mpeg-7 based inference network. Master's thesis, Queen Mary, University of London, 2001.
- [9] J. Hunter and R. Iannella. The application of metadata standards to video indexing. In *Second European Conference on Research and Advanced Technology for Digital Libraries*, Crete, Greece, September 1998.
- [10] J. Ivory. *A room with a view*. Merchant Ivory Productions ©, 1996.
- [11] ISO MPEG-7. Text of iso/iec cd 15938-2 information technology - multimedia content description interface - part 5 multimedia description schemes, iso/iec jtc 1/sc 29/wg, March 2001.
- [12] S. Myaeng, D.H. Jang, M.S. Kim, and Z.C. Zhoo. A flexible model for retrieval of sgml documents. In *21st ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 138–145, Melbourne, Australia, 1998.
- [13] M.R. Naphade and T.S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Multimedia*, 3(1):141–151, March 2001.
- [14] A. Pearmain, M. Lalmas, E. Moutogianni, D. Papworth, P. Healey, and T. Rölleke. Using mpeg7 at the consumer terminal in broadcasting. In *EURASIP (European Association for Signal, Speech and Image Processing) Journal on Applied Signal Processing*, volume 4, pages 354–361, April 2002.
- [15] W. Putz. The use of mpeg-7 metadata in a broadcast application. In *Media Future*, Florence, Italy, May 2001.
- [16] T. Rölleke, M. Lalmas, G. Kazai, I. Ruthven, and S. Quicker. The accessibility dimension for structured document retrieval. In *24th European Colloquium on Information Retrieval Research*, Glasgow, Scotland, March 2002.
- [17] A. Smeaton. Indexing, browsing and searching of digital video and digital audio information. In *European Summer School in Information Retrieval*, pages 93–110, Varenna, Lago di Como, Italy, 2000.
- [18] H.R. Turtle. *Inference Networks for Document Retrieval*. PhD thesis, University of Massachusetts, 1990.
- [19] H.R. Turtle and W.B. Croft. Efficient probabilistic inference for text retrieval. In *RIAO 3*, pages 644–661, 1991.
- [20] J.T.L. Wang, K. Zhang, and C. Chang. Identifying approximately common substructures in trees based on a restricted edit distance. *Information Sciences*, 121(3-4):367–386, December 1999.