A Language to Support Automatic Composition of Newscasts[1]

G. Ahanger and T.D.C. Little

Multimedia Communications Laboratory
Department of Electrical and Computer Engineering
Boston University, 8 Saint Mary's Street
Boston, Massachusetts 02215, USA
(617) 353-8042, (617) 353-6440 fax
{*gulrukh,tdcl*}*@bu.edu*

MCL Technical Report No. 05-28-1998

Abstract–Video production involves the selection, manipulation, and composition of video segments to achieve a refined piece suitable for an intended audience. By associating metadata with each segment it is possible to automate this production process. For example, a mechanism is achievable for the purpose of creating dynamically assembled compositions for information customization applications including news-on-demand.

In this paper we propose a grammar and associated production constraints necessary to facilitate automatic video composition in the news domain. The grammar encompasses composition based on content as well as structure of a newscast. In addition to providing a framework for logical composition of information, the grammar provides constraints for customization of information under bounds on playout duration or content selected by a user. We demonstrate how the language assists automatic information manipulation and composition of a newscast specifically when data are acquired from various sources and delivered under limited resources.

Keywords: News video, metadata, language, information representation, retrieval, composition.

[1]To appear in *Journal of Computer and Information Technology,* Vol. 6, No.3, 1998, pp. 297-310. This work is supported in part by the National Science Foundation under Grant No. IRI-9502702.

1 Introduction

Digital video production involves the process of selection, manipulation, and composition of information for playout. The information required to automate the production process depends on the domain and the content available. The knowledge about the domain resolves the type of information or metadata required for production. Based on the type of metadata established, metadata can be extracted and associated with each segment. These metadata are the foundation for retrieval and composition in automatic video production. Here we focus on collections of news video clips.

The metadata can be stored for each clip independent of the metadata associated with other clips. They can also be used to establish a correlation among clips that have similar concepts using stratification [15]. Therefore, each stratum of a concept (e.g., stratum belonging to concept "Clinton") can be mapped across all clips in the database. If a user issues a query to retrieve all clips relating to concept "Clinton," then the search can be limited to the stratum.

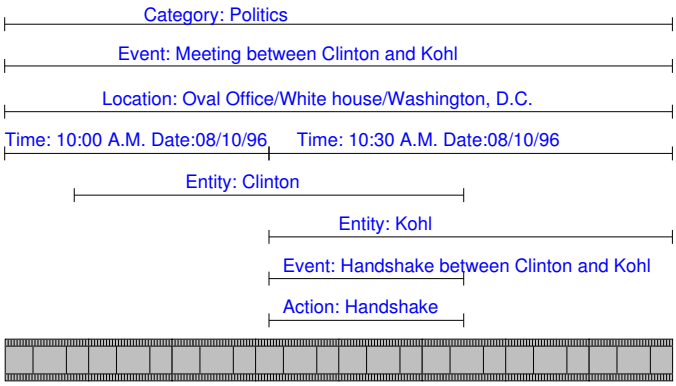

Figure 1: Example Video Clip Annotation

Fig. 1 depicts metadata associated with a single video clip of Clinton and Kohl meeting at the White House. To support search and retrieval on this clip and its components, we use metadata such as category of the clip, event, location, actions, entities, and recording time. Example queries to this metadata include: "Retrieve clips about Clinton's meeting with Kohl at the White House," "retrieve clips showing Clinton and Kohl shaking hands," and "retrieve all clips recorded on 10 August 1996 at 10:30 A.M." Note that parts of or the entire clip can be retrieved as the result of a query.

In our target domain of video-based news, a story is created from a collection of related

clips that relate to an event or *center*. Cohesion in a news story is achieved by ordering the clips to reflect the cause and effect of the center without breaking continuity. Therefore, composition of a cohesive news story from a database of annotated news clips, such as the ones above, requires selection, ordering, and refinement. The ordering, or composition, of clips can be achieved by certain rules, typically applied by a human. For automatic composition, these rules of composition, or their simplifications, must be implementable and yield the desired continuity in storyline. They do not necessarily require that all related clips be incorporated in a final composition. Temporal constraints on playout duration can limit the number and duration of the clips selected in a composition. Therefore, for temporal constraints to be realized, techniques are required to make appropriate selection from a surplus of clips. This leads to inclusion and exclusion rules for composition under temporal constraints. Ozsoyoglu et al. [13] have a similar objective that is achieved by content-based inclusion and exclusion rules.

Related works on automated video composition include ConText and AUTEUR. ConText [8] is a system for automatic temporal composition of a collection of video shots. It allows users to navigate semi-randomly through a collection of documentary scenes associated with a limited range of content metadata describing *character*, *time*, *location*, and *theme*. In this system scenes are delivered to a user sequentially based on a rank determined for all available scenes. This scoring scheme attempts to achieve preferred continuity and progression of detail in the presentation. ConText demonstrates how cognitive annotations of video material can be used to individualize a viewing session by creating an entirely new version through context-driven concatenation. AUTEUR [12] is an application that is used to automatically generate humorous video sequences from arbitrary video material. Composition is facilitated by the available information describing the *characters*, *actions*, *moods*, and *location* and information about camera position (e.g., close-up, medium, long-range shots). Content-based rules are used to compose shots.

The above works use *content-driven* and *user-driven* techniques for composition of video data. We believe that in addition to these techniques we need *format-driven* composition. Format-driven composition is based on the structure of the target video piece (e.g., a news item is composed of introduction, field shots, and comments). Using the format driven technique reduces the amount of knowledge required for composition. For the content-driven composition technique each scenario has a set of rules describing how to compose events while maintaining information continuity. The number of these rules increases with the number of different scenarios acquired. In contrast, the format or structure of a composition does not change and it is possible to use a single set of rules. Format driven composition is guided by

a representation of cinematographic knowledge that attempts to preserve continuity and the use of establishing shots when changing topics in the composition. To achieve this objective, we need to identify which segments make good candidates to start a topic, be part of the body, and end a topic. For this reason we have created a language that allows composition based on the information and relationships, and also on the structure of the composition.

Therefore, in a news video production system we require content and structural information for composition of news items and newscasts. In addition, there can be considerable redundancy in material. This redundancy is common in the news video scenario for the following reasons:

- *Multiple sources:* there are large number of sources from where news data can be acquired. This leads to multiple presentations of similar information in the video database (e.g., a visit of Clinton to South America reported by multiple sources).
- *Identical content:* some of the information contained in segments can be identical (e.g., multiple sources will cover speech of Clinton at a banquet in Brazil).

Therefore, there is a requirement to deal with duplicates and distinct but redundant clips. There is also a requirement to influence the selection of clips based on an individual's preferences. Both of these tasks require analysis of the content of the video to achieve the final composition.

In this paper we present a language for composition of a newscast based on format/structure and content. The language is intended to support the composition of news items based on a story center and the involvement of clips with related information. It is also intended to facilitate the reduction of redundancies and conformance to time constraints. We demonstrate that by using content, user preference, and format information, we not only ensure a cohesive composition of information but also allow easy manipulation of the data. That is, the language possesses sufficient information to appropriately include or exclude video clips in a composition.

The organization of the remainder of this paper is as follows: In Section 2 we overview current techniques used to deliver news video data to a user. In Section 3 we present the language for news video composition. In Section 4 we describe the metadata required to be maintained to support the language. In Section 5 we demonstrate with examples how the proposed language supports composition. In Section 6 we discuss the functionality of the proposed language. Section 7 concludes the paper.

2 Existing News Video Delivery Systems

A number of systems have being proposed for delivery of news video data. Agora [9], an application developed at Bell Labs, uses filtering of multi-channel broadcast news based on a user profile and closed-caption data. The Network Multimedia Information Services [7] system uses start and stop boundaries supplied with individual news items that can be browsed on the Web using an index. Shararay et al. at Bell Labs developed an application that maps the transcripts of the broadcast news with the associated video frames [14]. Transcripts can be browsed via the WWW browser. Brown et al. [6] also have a system that uses closed-caption text to filter information leading to the playout of associated clips.

A common feature of these systems is content filtration. Typically, news items are pre-composed, editing is achieved off-line, and a user has no choice but to see all the content pertaining to a news item. For example, composition of a query, "give me the news item about Waco, Texas with only field scenes" is not possible. The presentation cannot be customized to be comprised of only field scenes. Moreover, time compression of a news item is not possible. For time compression we need information about the relative importance of component clips. In a newscast presentation, news items are accommodated sequentially and in their entirety (i.e., news items are completely played-out prior to subsequent news items). If presentation of a newscast is time-compressed, time cannot simply be trimmed from each news item. Instead, only those news items that can be accommodated in the sequential playout can be completely displayed. Also, because of the constrained playout time, the last news item will be cut-off at an arbitrary point. Fig. 2 depicts a scenario in which a news item is composed from data acquired from multiple sources. Segments are selected in accordance with the query specification, extracted, composed, and rendered.

We envision a customized news video delivery system building on this earlier work to support following functionality:

- Summarizing all news in fixed time period (e.g., provide the latest news in a duration of less than 5 minutes).
- Generating stories based on a story center (e.g., create a news story about the Pope's visit to Cuba).
- Sequencing the topics according to a viewer's preference (e.g., form different tours though the content space).

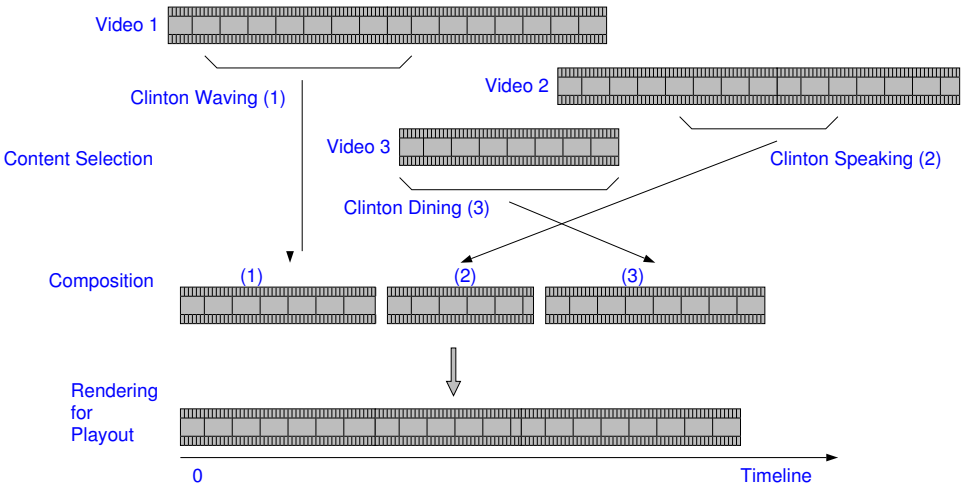

Figure 2: Composition of a News Item from Multiple Sources

- Selecting specific topics in a newscast (e.g., provide news only on sports and politics).
- Customizing content (e.g., filter out violence and nudity).
- Selecting composition of news components (e.g., speech).

To provide above functionalities and to produce a cohesive news item, once the segments are retrieved, we need to manipulate the segments. Manipulation means ordering them in a logical manner or dropping segments when constraint-based (content or playout) queries are issued. Here our ultimate goal is to automate the composition process.

3 Proposed Language

For automatic customization we need to understand how a newscast is composed and the relationships among different concepts contained in the newscast. We use Musburger's [11] and Brabiger's [5] work to identify the information sufficient to compose a coherent newscast. *Content-based* and *structure-based* information can be extracted from raw news video data [2]. The content-based information is used to search for content in a news video archive. The structure-based information is used to extract segments by type (e.g., headline, wild-scene) and are used for composition of a newscast.

To accommodate different news items in a presentation, each news item is allotted a limited duration for presentation. Therefore, the objective of news item composition is

to maximize the presentation of the information related to each event, in spite of the time constraints. The tactic adopted by newscasters is to provide multiple views of an event rather than a single detailed view. For example, multiple views can include field shots, comments, interviews, and re-enactments of an event after its occurrence. A news item is a collage of views and it is possible to rearrange or drop some of the views to convey the same story.

In the proposed language we take advantage of the characteristic of news items that are typically short segments that convey a great deal of information (i.e., sound bites). This is an underlying property on which the proposed language is constructed. In addition, to achieve the aforementioned objectives for the language, the following assumptions are required:

1. A complete news item is considered an event (e.g., Clinton's visit to South America).
2. An event can be composed of sub-events (e.g., interviews, comments from by-standers, and field shots).
3. Thematic continuity is maintained in a news item when sub-events are presented in an arbitrary order provided:
   (a) all sub-events belong to the same event, and
   (b) each sub-event is completely played-out.
4. Within a news item, all types of segments carry the same theme; however, different types of segments depict the theme from different views and are not redundant. Moreover, these segments are not dependent on one another for presentation.

Here we define a presentation possessing thematic continuity as one comprised of segments with related information and ordering to maintain temporal continuity. Assumption (1) ensures that each news item contains information related to a single event and no irrelevant information is presented. Assumption (2) is required to identify types of segments in a body of an event. Assumption (3) is required to maintain a storyline, or theme, and to avoid abrupt discontinuities in a news item by presenting complete information about each segment. In assumption (4) we treat segments as having content independence so that we can include, exclude or arrange them in any order in the segments in the body. A segment is a complete information unit and all dependent content is encompassed in a single segment. For example, if an anchor person introduces a scene, then an introduction is included as part of the scene. Rearrangement of clips is valid only if the segments are from same instance in

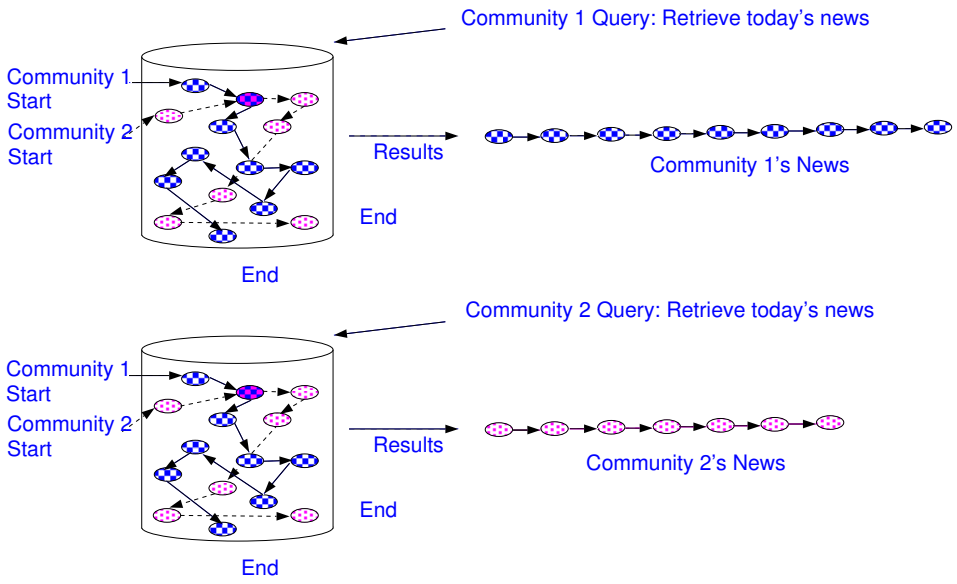

Figure 3: Users Offered Different Tours Through the Content

chronological time. If available segments are from a period, the flexibility to rearrange them is limited.

An EBNF-representation of the language is shown in Table 1. The language is defined as a production grammar ($p \rightarrow \delta$) [4] as shown in Table 1. Each symbol p (e.g., <newsitem>) in a production can be interpreted as a node for holding information. The types of information associated with these nodes are defined by the semantic rules of a production.

Production (1) specifies that a newscast is composed of one or more *tours*. For example, a newscast requested by two users can contain the same content but in a different order (Fig. 3). Production (2) is a recursively defined rule. The syntactic category or a nonterminal <tour> is defined in terms of itself by right-recursion. A tour is a production of a syntactic category <newsitem> or its recursions.

A tour can be represented as a path in a directed graph. Assume that NC is a set of unique news items that can be formed from the available data and $NC = (NI, E, l)$ is represented as a directed graph. Where NI (news items) are the vertices, E edges that connects vertices, and l is a function from E to a set U of users. The function $l(E)$ assigns the users for whom the edge E can be traversed to compose a news item. In Fig. 4, $l(e_2) = (2, 3)$ means that to compose a newscast for users 2 and 3, the edge e_2 is traversed in the direction

Table 1: The Proposed Language in EBNF

1. $<$ newscast $>$ $\rightarrow \{<$ tour $>\}_1^n$
2. $<$ tour $>$ $\rightarrow <$ newsitem $> | <$ newsitem $><$ tour $>$
3. $<$ newsitem $>$ $\rightarrow <$ headline $>$
4. $<$ newsitem $>$ $\rightarrow [<$ headline $>]\{<$ introduction $>\}^1[<$ tmp $>]$
5. $<$ tmp $>$ $\rightarrow <$ b-list $><$ enclose $>|<$ b-list $>$
6. $<$ b-list $>$ $\rightarrow <$ b-list $><$ b-list$_2$ $>|<$ b-list$_2$ $>$
7. $<$ b-list$_2$ $>$ $\rightarrow <$ speech $> | <$ wild-scene $> | <$ interview $> | <$ comment $> | <$ enactment $>$
8. $<$ interview $>$ $\rightarrow <$ question & answer (qa) $> | <$ qa $><$ interview $>$
9. $<$ headline $>$ $\rightarrow <$ shot $>$
10. $<$ headline $>$.entity-list $:= <$ shot $>$.entity-list;
11. $<$ headline $>$.location-list $:= <$ shot $>$.location-list;
12. $<$ headline $>$.category-list $:= <$ shot $>$.category-list;
13. $<$ headline $>$.event-list $:= <$ shot $>$.event-list;
14. $<$ headline $>$.time-list $:= <$ shot $>$.time-list;
15. $<$ headline $>$.action-list $:= <$ shot $>$.action-list;
16. $<$ headline $>$.graphics-list $:= <$ shot $>$.graphics-list;
17. $<$ headline $>$.audio-type-list $:= <$ shot $>$.audio-type-list;
18. $<$ headline $>$.video-type-list $:= <$ shot $>$.video-type-list;
19. $<$ headline $>$ $\rightarrow <$ shot $><$ headline $>$
20. $<$ headline $>$.entity-list $:= \cup(<$ shot $>$.entity-list, $<$ headline $>$.entity-list);
21. $<$ headline $>$.location-list $:= \cup(<$ shot $>$.location-list, $<$ headline $>$.location-list);
22. $<$ headline $>$.category-list $:= \cup(<$ shot $>$.category-list, $<$ headline $>$.category-list);
23. $<$ headline $>$.event-list $:= \cup(<$ shot $>$.event-list, $<$ headline $>$.event-list);
24. $<$ headline $>$.time-list $:= \cup(<$ shot $>$.time-list, $<$ headline $>$.event-list);
25. $<$ headline $>$.action-list $:= \cup(<$ shot $>$.action-list, $<$ headline $>$.action-list);
26. $<$ headline $>$.graphics-list $:= \cup(<$ shot $>$.graphics-list, $<$ headline $>$.graphics-list);
27. $<$ headline $>$.audio-type-list $:= \cup(<$ shot $>$.audio-type-list, $<$ headline $>$.audio-type-list);
28. $<$ headline $>$.video-type-list $:= \cup(<$ shot $>$.video-type-list, $<$ headline $>$.video-type-list);
29. $<$ introduction $>$ $\rightarrow <$ shot $> | <$ shot $><$ introduction $>$
30. $<$ enclose $>$ $\rightarrow <$ shot $> | <$ shot $><$ enclose $>$
31. $<$ qa $>$ $\rightarrow <$ shot $> | <$ shot $><$ qa $>$
32. $<$ speech $>$ $\rightarrow <$ shot $> | <$ shot $><$ speech $>$
33. $<$ wild-scene $>$ $\rightarrow <$ shot $> | <$ shot $><$ wild-scene $>$
34. $<$ comment $>$ $\rightarrow <$ shot $> | <$ shot $><$ comment $>$
35. $<$ enactment $>$ $\rightarrow <$ shot $> | <$ shot $><$ enactment $>$

shown. For clarity, $l(E)$ can be written as $l(NI_i, NI_j)$, where NI_i, NI_j is an ordered set of vertices which are included in a newscast. The path used to compose a newscast for a user in the graph is simple and elementary (i.e., no news item is visited twice). A news item is presented only once in a newscast for a single user. In Fig. 4, path (e_9, e_7, e_6) is traversed to compose a newscast for user 1.

Productions (4) and (5) specify the syntactic category <newsitem> as comprised of <headline>, <introduction>, <b-list>, and <enclose>. A <newsitem> can be composed of only a single headline (see production 3). According to productions (4) and (5) a news item can be produced with a single headline segment, a single introduction segment, a single b-list, and a single enclose segment. An enclose is only present if <b-list> is present. Productions (5), (6), and (7) convey that the syntactic category <b-list> or "body" can be composed of any combination of multiple segments belonging to "speech," "wild-scene," "interview," "comment," and "enactment." As mentioned before this kind of composition is valid only if it is based on chronological time. For example, consider a list of segments of type speech, interview, comment, and wild-scene belonging to "body." We show that it is reduced to production (5) as follows:

speech, interview, comment, comment, wild-scene	
b-list$_2$, interview, comment, comment, wild-scene	(*production 7*)
b-list, interview, comment, comment, wild-scene	(*production 6*)
b-list, b-list$_2$, comment, comment, wild-scene	(*production 7*)
b-list, comment, comment, wild-scene	(*production 6*)
b-list, b-list$_2$, wild-scene	(*production 7*)

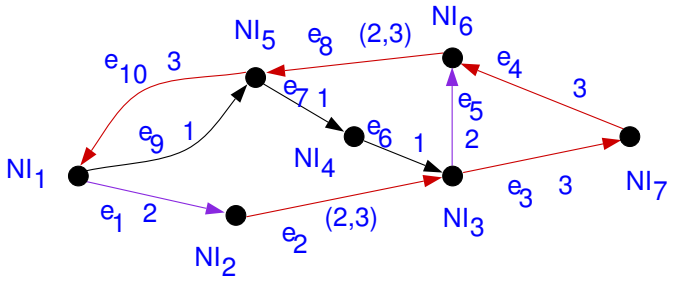

Figure 4: Tour Formation from Retrieved News Items

b-list, wild-scene (*production 6*)
b-list, b-list$_2$ (*production 7*)
b-list (*production 6*)

Production (8) specifies that a syntactic category <interview> is composed of a "question & answer" or its recursions. The syntactic categories <headline>, <introduction>, <enclose>, <qa>, <speech>, <wild-scene>, <comment>, and <enactment> are composed of terminal symbol <shot> or its recursion.

The symbols <headline>, <introduction>, <enclose>, <qa>, <speech>, <wild-scene>, <comment>, <enactment> have *synthesized attributes* [4] associated with them. In Table 1, entity-list, location-list, category-list, event-list, action-list, graphics-list, audio-type-list, and video-type-list are synthesized attributes of <headline>. Not shown are that these attributes are also associated with other symbols like <introduction>, <enclose>, <qa>, <speech>, <wild-scene>, <comment>, and <enactment>.

An entity-list represents all conceptual (any object part of the commentary, e.g., people) and tangible objects (objects part of a video stream). A location-list consists of all locations shown in the video or conceptual locations, i.e., associations with certain places and countries that are discussed but not part of the visuals (e.g., a news item with discussion on Iraq or shots taken in Baghdad). A category-list consists of the classification of the video data (e.g., accidents, political, sports). An event/action-list represents any happening in a news item (e.g., Clinton's controversy, standoff in Iraq, games at Nagano). A time-list contains the historical time or date of an event or when the event actually took place (e.g., 19 February 1878 phonograph invented by Thomas Edison). A graphics-list represents stills or graphics shown in video (e.g., photographs, maps). An audio-type-list represents the type of audio (i.e., lip-sync, when audio requires tight synchronization with the video), wild-dialogue (dialogue that does not sync with a visible speaker), and voice over (when a story uses continuous visuals without showing the speaker). A video-type-list represents the type of video shots (e.g., close-up shot and wild shot).

Each production grammar $p \rightarrow A_1\, A_2 \ldots A_n$ has an associated set of semantic rules of the form $p.s := f(A_1.a_1, A_2.a_2, A_3.a_3, ..., A_4.a_n)$, where s is a synthesized attribute of p, f is a function, and $a_1, a_2, ..., a_n$ are the attributes belonging to the grammar symbols of the production. Consider the nodes <headline> → <shot> and <headline> → <shot><headline> in the parse tree. The value of the attribute <headline>.entity-list at this node is defined by:

Production	Semantic Rule
< headline > → < shot >	< headline > .entity-list := < shot > .entity-list;
< headline > → < shot >< headline >	< headline > .entity-list := ∪(< shot > .entity-list, < headline > .entity-list);

Suppose that a headline segment is composed of three shots. The first shot has three associated entities (*a, b,* and *c*). The second shot has four associated entities (*a, d, e,* and *f*). The last shot has two associated entities (*c* and *g*). Function ∪ performs a union of the two argument lists passed to it. Therefore, the <headline>.entity-list will consist of entities *a, b, c, d, e,* and *f*. Conceptually this semantic rule means that even if an entity is not present in a complete segment it is still assumed to belong to the complete segment.

Before we present example queries to illustrate how the proposed language is used in composition, we summarize how information (metadata) for the video data are represented to support these queries.

4 Information Representation

In addition to the constraints imposed by the language, structural and content information is required for the composition of a newscast. A newscast is composed of news items presented in a temporal order. Each news item is comprised of one or more objects. Furthermore, an object can be composite, being made up of other objects. An object can belong to multiple news items and a news item can belong to multiple newscasts. Segments and the content within these news items are treated as objects (e.g., headline, introduction, entity, location, and action are treated as objects). In Fig. 5 we represent the structural objects and the relationship between these objects as a hierarchy and use this representation to store the information in a database. This format is similar to treating a movie as composed of scenes and the scenes composed of shots [3]. The order of the scenes in a movie is identified by the events in the scenes. However, for a newscast the segments are ordered according to their type under the assumption that all the segments belong to the same event.

As mentioned, the content items within the segments (e.g., wild-scenes) are treated as objects (e.g., "entity," "location," "category," and "graphics"). An object can be composed of other objects, thus forming a hierarchy of object types. An event, which we treat as synonymous to a news item, forms the root of an object hierarchy for the news item. Thus, Figs. 5 and 6 represent the hierarchy of information stored in the metadatabase.

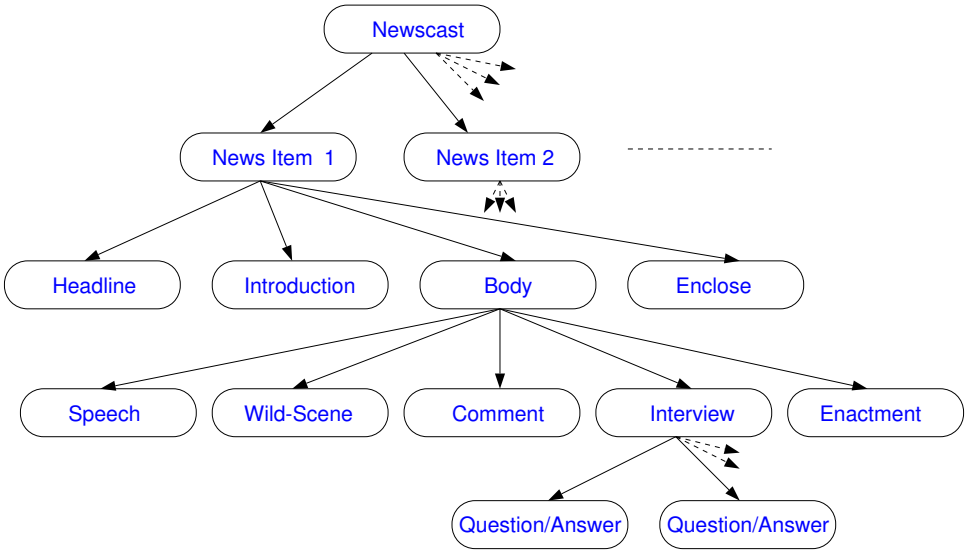

Figure 5: Structural Representation for Newscast Composition

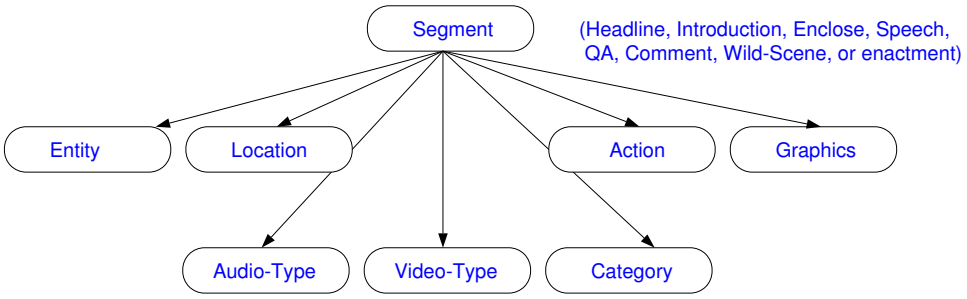

Figure 6: Content Representation in a News item

In this hierarchy, "headline," "introduction," "enclose," "speech," "wild-scene," "qa," "comment," and "enactment" are the *leaves* of the object type tree. Each object is represented by a set of attributes: <object-id, type, name, metatype, medium, popularity, date of creation, time of creation, origin, video-filename, start-frame, end-frame, compression format, playout rate> (Section 5). The cinematographic attributes "compression format" and "playout rate" are maintained for playout as are the attributes of "video-filename," "start-frame," and "end-frame." Metatype qualifies the type (e.g., an entity-type can be a "person" and its metatype can be "president"). Metatypes are stored so that queries like "give me the reaction of the President" can be satisfied. "Headline," "introduction," "wild-scene," "enactment," and "enclose" are the metatypes for "segment." "Speech," "interview," and "comment" are the metatypes for "reaction." "Country," "city," and "place" are the metatypes for "loca-

Table 2: Sample Metadata

Object ID	Type	Metatype	Name	Source	Creation Time	Creation date
O_{01}	Segment	Intro		CNN	13:00:00	26/06/96
O_{02}	Segment	Wild-Scene		CNN	13:00:00	26/06/96
O_{03}	Reaction	Speech		CNN	13:00:00	26/06/96
O_{04}	Reaction	Comment		CNN	13:00:00	26/06/96
O_{05}	Segment	Body		CNN	13:00:00	26/06/96
O_{06}	Event		VTV	CNN	13:00:00	26/06/96
O_{07}	Segment	Wild-Scene		CBS	19:00:00	26/06/96
O_{08}	Segment	Wild-Scene		CBS	19:00:00	26/06/96
O_{09}	Segment	Enclose		CBS	19:00:00	26/06/96
O_{10}	Reaction	Interview		CBS	19:00:00	26/06/96
O_{11}	Reaction	QA		CBS	19:00:00	26/06/96
O_{12}	Reaction	QA		CBS	19:00:00	26/06/96
O_{13}	Reaction	QA		CBS	19:00:00	26/06/96
O_{14}	Segment	Body		CBS	19:00:00	26/06/96
O_{15}	Segment	Intro		CBS	19:00:00	26/06/96
O_{16}	Event		VTV	CBS	19:00:00	26/06/96

tion." The information whether an object is associated with audio, video, or both audio and video is maintained in the "medium" attribute. The creation time and date represent when an event was recorded. The objects and the information about their attributes are stored as metadata in the form of a regular expression to support automatic composition. We consider queries that can be satisfied by the capabilities of the language next.

5 Example Queries Satisfied by the Language

In this section we demonstrate how the language can be used to compose and customize a newscast. We assume the acquisition of the following data from two sources about "Clinton's visit to Venezuela" (abbreviated to VTV to accommodate Table 2).

As previously explained, we store metadata describing the video segments. In Table 2 only the structural metadata are shown. For example, O_{01} is an ID of an object/segment that is of type "Intro" and is acquired from "CNN." The creation time and date of the segment is "13:00:00" and "26/06/96" respectively. The hierarchy of the above objects is shown in Fig. 7. Object O_{06} is an event and is comprised of two segments/objects O_{01} and

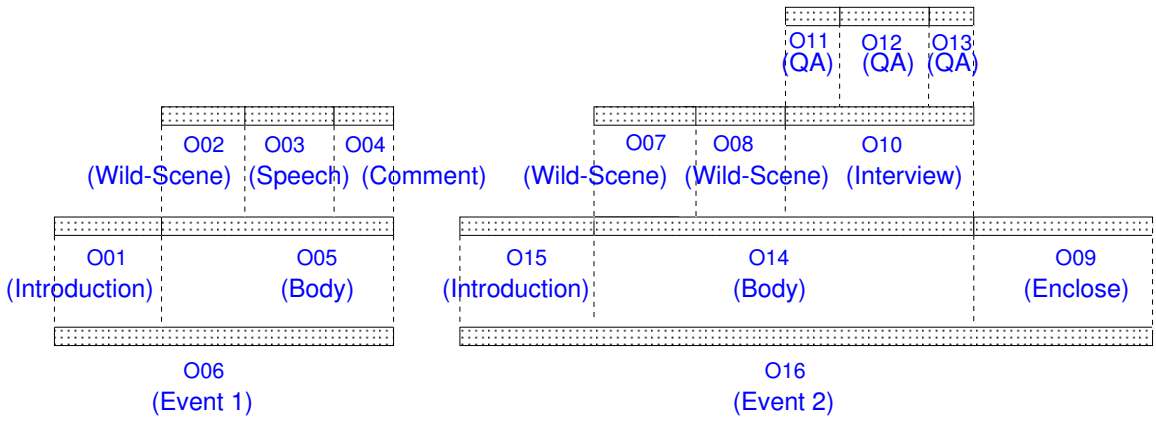

Figure 7: Structural Hierarchy for the Content of the Example

O_{05}. Object O_{05} is comprised of three objects O_{02}, O_{03}, and O_{04}. Object O_{16} is another event that is comprised of objects O_{15}, O_{14}, and O_{09}. Object O_{14} is comprised of objects O_{07}, O_{08}, and O_{10}. Finally, object O_{10} is comprised of objects O_{11}, O_{12}, and O_{13}.

With help of queries that are based on the above metadata, we can demonstrate how to form a coherent news item. We also demonstrate how to merge content from various sources, customize content based on a user's temporal constraints, and customize the selection based on content preferences.

Coherency: A cohesive news item can be formed by using the language.

Query 1: "Compose the news from the most recent material in the system."

In the database the recent objects acquired are from ID O_{07} to O_{16}. After finding the objects that are recent (e.g., news less than one hour old), we try to form a coherent composition of the objects for playout. As seen from the object hierarchy (Fig. 7), objects O_{07}-O_{15} belong to the event (O_{16}) "Clinton's visit to Venezuela." We can compose these objects to form a single news item. This is achieved by constraints imposed by the language as follows:

$O_{15} \rightarrow O_{14} \rightarrow O_{09}$	*production 4*
$O_{15} \rightarrow O_{07} \rightarrow O_{08} \rightarrow O_{10} \rightarrow O_{09}$	*productions 5, 6, and 7*
$O_{15} \rightarrow O_{07} \rightarrow O_{08} \rightarrow O_{11} \rightarrow O_{12} \rightarrow O_{13} \rightarrow O_{09}$	*production 8*

The last row represents the final composition of the news item for playout. It conforms

to production rule (4), i.e., there is no headline in the news item; and it is composed of a single introduction segment (O_{15}), a b-list (O_{14}), and an enclose segment (O_{09}). The b-list consists of segments O_{07}, O_{08}, and O_{10}. Object O_{10} is further decomposed according to production rule (8). According to production rule (6), a b-list can be composed of segments belonging to the body in any combination. Therefore, segments O_{07}, O_{08}, and O_{10} can be sequenced in any order.

Merging: We can combine content from multiple sources belonging to the same event into a single news item.

Query 2: "Retrieve all information on Clinton's visit to Venezuela."

Objects O_{06} and O_{16} are associated with Clinton's trip. All of the sub-objects that comprise these two objects can be merged to form a single news item. To form a coherent news item we require an "introduction," "body," and "enclose." To maintain temporal continuity and chronology, we include the oldest "introduction," and the latest "enclose." Objects belonging to the "body" are also composed in temporal order (most recent objects shown last). In addition to the language, we impose the additional constraint that all objects in the body appear in chronological order. This constraint is imposed to achieve temporal continuity in presentation. The final composition is as follows:

$$O_{01} \rightarrow O_{02} \rightarrow O_{03} \rightarrow O_{04} \rightarrow O_{07} \rightarrow O_{08} \rightarrow O_{11} \rightarrow O_{12} \rightarrow O_{13} \rightarrow O_{09}$$

Objects O_{02}, O_{03}, O_{04}, O_{07}, O_{08}, O_{11}, O_{12}, and O_{13} form the body of the b-list. Object O_{01} is the introduction. Production rule (4) states that an introduction segment is necessary for composition. Object O_{09} is an enclose and is incorporated according to production rule (5). According to our assumptions (Section 3) and the constraints imposed by the language, the above composition results in a coherent news item.

Preferences: Content-based customization, or "preferences," can be achieved by using the production rules of the language.

Query 3: "Retrieve all field shots with information on Clinton's visit to Venezuela."

We gather all the information we have about the event "Clinton's visit to Venezuela" and then apply content-based customization. According to user preferences, only wild-scenes

Table 3: Playout Duration of the Segments

Object ID	Time (s)
O_{07}	30
O_{08}	45
O_{09}	5
O_{11}	120
O_{12}	55
O_{13}	67
O_{15}	15

need to be shown. According to production rule (4) the minimum information to have a coherent news item is an introduction followed by the segments of the type wild-scene. From the table, objects O_{02}, O_{07}, and O_{08} belong to the wild-scenes category. Using production rules (4), (5), (6), and (7) yields the final composition for the playout as follows:

$$O_{01} \rightarrow O_{02} \rightarrow O_{07} \rightarrow O_{08} \rightarrow O_{09}$$

Temporal Constraints: We can achieve time-based customization using the language.

Query 4: "Compose the latest news about Clinton's visit to Venezuela for four minutes of playout."

For this type of a query we need to know the playout duration of each clip to produce a news item within the temporal playout constraint. Assume the timings for the complete playout of objects/segments as shown in Table 3.

Here, in addition to using the production rules, we also use the rules for resolving temporal playout constraints [1]. This can be achieved by presenting information from as many views as possible. If an information presentation is from an instance of chronological time, we cluster different views separately (e.g., wild scenes, comments, interviews). During composition we iterate through the clusters selecting a segment from each cluster (if the playout duration of a segment permits) until the specified duration has been accommodated.

According to the query, we must form a coherent and complete news item within the constraint of $240s$. Event O_{16} and its associated objects have the most recent information; therefore, we initially attempt to compose a news item from these objects with consideration

for the duration of each segment. The following objects can be selected to meet the playout constraint of $240s$:

$$\begin{array}{ll} \text{Iteration 1:} & \mathrm{O}_{15} \rightarrow \mathrm{O}_{11} \rightarrow \mathrm{O}_{07} \\ \text{Iteration 2:} & \mathrm{O}_{15} \rightarrow \mathrm{O}_{11} \rightarrow \mathrm{O}_{12} \rightarrow \mathrm{O}_{07} \rightarrow \mathrm{O}_{09} \end{array}$$

According to the temporal composition criteria, we iterate through the clusters of question and answers and wild-scenes. In each iteration we select an object from a cluster until all time is accommodated. If presentation is from a period of chronological time (e.g., from 15/05/96 to 26/06/96) we divide the timeline into sub-periods. During composition we iterate through the sub-periods and select a single segment from each sub-period in each iteration.

6 Discussion

The proposed language is useful for automatic news composition based on a number of assumptions and does not satisfy all possible functionalities. Here we consider these shortcomings.

The language helps construct a valid news item based on the imposed structural constraints. To achieve time compression, segments can be dropped and then parsed for validity using the grammar. As discussed in Section 2, we adopt the following reasoning to select a segment to drop to achieve time compression:

1. If the duration of a segment is longer than the specified time constraint then it is not used.
2. Based on content customization, if a user does not want to see particular content or particular segments (e.g., violence, comments) then they are dropped.
3. Based on selection of multiple views, if a user does not want to see particular views (e.g., comment) then segments containing these views are dropped.

Even if same segments are dropped, we expect to be left with a viable set of candidate segments. Any segment from this set can be incorporated in the composition. Further refinement is necessary if the playout time does not allow all the segments to be incorporated, we need a selection criterion. We propose the following approaches:

1. Select a segment at random. Using this approach, when the same query is repeated, the probability of the composition being different is high. Therefore, two users issuing the same query will likely get different composition.

2. Associate a parameter of interest with each segment. The value of the parameter will depend on the interest level of the user population for the content (acquired from user feedback). Here we can use concepts such as community filtering to achieve this goal. Using this approach, the composition will be highly influenced by other community members' preferences.

Currently we are investigating whether it is possible to extend the language to make it sensitive to the importance of a segment based on its content.

Another issue is that the language is based on the assumption that structures belonging to the body of a news item can be presented in any order without breaking information continuity. This is true if and only if the retrieval is from an instance of chronological time, i.e., the dependency of the segments on one another, if any, is negligible. If the retrieval is from a specific period then it becomes imperative that we present information in correct chronological order as there can exist temporal dependencies among segments. The facts of a story can change as an event evolves; therefore, chronology of information is important to presentation order. For example, to summarize a basketball game we need to present the clips in correct temporal order to accurately depict the scoring time series. Period-based temporal ordering is incorporated into the language for this purpose.

If the presentation of a newscast is compressed in time, we can distribute available presentation time equally among all news items in the newscast. From our observation of conventionally-produced news broadcasts, time for news items is most often apportioned according to their importance. This property can be supported by the language.

Once the segments belonging to a news item are selected and ordered, the segments are parsed based on the grammar to check validity. The language ensures that a newscast composition is cohesive. In summary, the language is structure and content-sensitive to support various constraints.

7 Summary

In this paper we have presented a language to support automatic composition of news items and newscasts. The language supports composition based on content and applies structural constraints to achieve a coherent composition. The structural constraints not only sequence the video segments in a logical manner depending on their type but also accommodate structural redundancies. For example, if multiple segments of type "introduction" are present in a candidate set of video segments then, only a single segment is selected and the remainder are dropped.

The language also provides structure to a news item that is composed under temporal constraints. In such circumstances the language provides constraints based on the necessity of inclusion for a particular type of segment. For example, at least one segment of type "introduction" should be present in a composition.

The language is a result of a need for automatic cohesive composition of segments containing desired content. Content alone, though important, cannot be used to create a coherent piece of video. Therefore, by incorporating constraints based on both content and structure in the language it is possible to automate the news video production process. Using a variety of examples, we demonstrated that the news video production process assisted by the language results in logical composition of newscasts.

References

[1] G. Ahanger and T.D.C. Little, "Automatic Composition Techniques for Video Production," to appear in *IEEE Transactions on Knowledge and Data Engineering*, Vol. 10, No. 6, 1998.

[2] G. Ahanger and T.D.C. Little, "A System for Customized News Delivery from Video Archives," *Proc. Intl. Conf. on Multimedia Computing and Systems*, Ottawa, Canada, June 1997, pp. 526-533.

[3] G. Ahanger and T.D.C. Little, "A Survey of Technologies for Parsing and Indexing Digital Video," *Journal of Visual Communication and Image Representation*, Vol. 7, No. 1, March 1996, pp. 28-43.

[4] A.V. Aho, R. Sethi, and J.D. Ullman, "Syntax-Directed Translation," *Compilers: Principles, Techniques, and Tools*, Addison-Wesley Publishing Company, Reading, Massachusetts, March 1988, pp. 279-342.

[5] M. Brabiger, *Directing the Documentary*, Focal Press, Boston, 1992.

[6] M.G. Brown, J.T. Foote, G.J.F. Jones, K.S. Jones, and S.J. Young, "Automatic Content-Based Retrieval of Broadcast News," *Proc. ACM Multimedia '95*, San Francisco, November 1995, pp. 35-43.

[7] K. Compton and P. Bosco, "CNN Newsroom on the Internet: A Digital Video News Magazine and Library," *Proc. Intl. Conf. on Multimedia Computing and Systems*, Washington D.C., May 1995, pp. 296-301.

[8] G. Davenport and M. Murtaugh, "ConText Towards the Evolving Documentary," *Proc. ACM Multimedia '95*, San Francisco, November 1995, pp. 377-389.

[9] E. Hyden and C. Sreenan, "Agora – A Personalized Digital Newsfeed," *Proc. 6th Intl. Workshop on Network and Operating Systems Support for Digital Audio and Video*, Zushi, Japan, Short Papers, April 1996.

[10] G. Miller, G. Baber, and M. Gilliland, "News On-Demand for Multimedia Networks," *Proc. ACM Multimedia '93*, Anaheim, August 1993, pp. 383-392.

[11] R.B. Musburger, *Electronic News Gathering*, Focal Press, Boston, 1991.

[12] F. Nack and A. Parkes, "The Application of Video Semantics and Theme Representation in Automated Video Editing," *Multimedia Tools and Applications*, Vol. 4, No. 1, January 1997, pp. 57-83.

[13] G. Ozsoyoglu, V. Hakkoymaz, and J. Kraft, "Automating the Assembly of Presentation for Multimedia Data," *Proc. 12th IEEE Intl. Conf. on Data Engineering*, New Orleans, Louisiana, February 1996, pp. 593-601.

[14] B. Shahrary and D. Gibbon, "Automatic Generation of Pictorial Transcripts of Video Programs," *Proc. Intl. Conf. on Multimedia Computing and Networking, SPIE*, San Jose, February 1995, pp. 512-518.

[15] T.G. Aguierre Smith and G. Davenport, "The Stratification System: A Design Environment for Random Access Video," *Proc. 3rd Intl. Workshop on Network and Operating System Support for Digital Audio and Video*, La Jolla, CA, November 1992.