Generation of Video Documentaries from Discourse Structures

Cesare Rocchi
ITC Irst
Trento Italy
rocchi@itc.it

Massimo Zancanaro
ITC Irst
Trento Italy
zancana@itc.it

Abstract

Recent interests in the use of multimedia presentations and multimodal interfaces have raised the need for the automatic generation of graphics and especially temporal media. This paper presents an engine to build video documentaries from annotated audio commentaries. The engine, taking into consideration the discourse structure of the commentary, plans the segmentation in shots as well as the camera movements and decides the transition effects among shots. The output is a complete script of a ”video presentation”, with instructions for synchronizing images and movements with the playing of the audio commentary. The language of cinematography and a set of strategies similar to those used in documentaries are the basic resources to plan the animation. Strategies encompass constraints and conventions normally used in building video documentaries.

1 Introduction

In the last decade there has been an increasing interest in the generation of multimedia presentations and a growing tendency towards the use of multi-modal interfaces (Wahlster et al., 1993; Maybury, 1993). These interests have raised the need for automatic generation not only of natural language, but also graphics and especially temporal media (André, 2000).

In this paper, an engine to build video sequences of images starting from an audio commentary is described. The input for the engine is a representation of (possibly automatically) generated verbal commentary. The engine, taking into consideration the discourse structure of the commentary, retrieves the most appropriate set of images from an annotated database, plans the segmentation in shots as well as the camera movements and finally decides the transition effects among shots. The output of the engine is a complete script of a ”video presentation”, with instructions for synchronizing images and movements with the playing of the audio commentary.

The language of cinematography (Metz, 1974), including shot segmentation, camera movements and transition effects, is the basic resource to plan the animation and to synchronize the visual and the verbal parts of the presentation. In generating animations, a set of strategies similar to those used in documentaries are employed. Two broad classes of strategies have been identified. The first class encompasses constraints imposed by the grammar of cinematography, while the second deals with conventions normally used in guiding camera movements in the production of documentaries.

After a short discussion on related works, relevant concepts and terminology of cinematography are introduced in section 3. Section 4 briefly summarizes the Rhetorical Structure Theory (RST) for the analysis of discourse structure. In section 5 we present some of the heuristics that we have borrowed from the field of cinematography. In section 6 we illustrate the architecture of the engine and its parts. In section 7 we give some examples of how the engine works. Finally, in section 8, we outline conclusions and future work.

2 Related Work

One of the first case studies of the generation of "motion presentations" is the work of (Karp and Feiner, 1993). Their system generates scripts for animation using top-down hierarchical planning techniques. (Christianson et al., 1996) presents a successful attempt to encode several of the principles of cinematography in the *Declarative camera control language*.

Similar systems are BETTY (Butz, 1994) and CATHI (Butz, 1997). BETTY is an animation planner, which generates scripts for animated presentations. The CATHY system generates on-line descriptions of 3D animated clips for the illustration of technical devices, in the context of a coordinated multimedia document.

Animated presentations have been successfully employed also in multimodal frameworks for the generation of explanations (Daniel et al., 1999) and in learning environments (Bares and Lester, 1997).

The novelty of our approach lies in the use of rhetorical structure of the accompanying audio commentary in planning the video. In particular, knowledge of rhetorical structure is extremely useful in taking decisions related to the punctuation of the video, in order to reflect the rhythm of the audio commentary and its communicative goals. In our view, the verbal part of the documentary always drives the generation of the visual part.

3 Relevant concepts and terminology

According to Metz (1974), cinematic representation is not like a human language, which is defined by a set of grammatical rules. It is nevertheless guided by a set of generally accepted conventions. These guidelines may be used for developing multimedia presentations that can be best perceived by the viewer. Following, we briefly summarize the basic terminology of cinematography.

3.1 Shot and camera movements

The shot is the basic unit of a video sequence. In the field of cinematography a shot is defined as a continuous view from single camera without interruption. Since we only deal with still images, we define a shot as a *sequence of camera movements applied to the same image*.

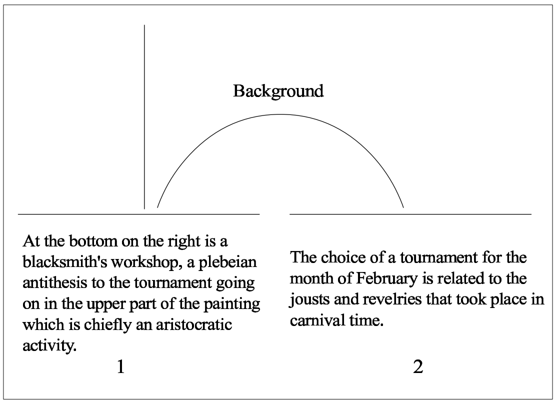
The basic camera movements are *pan*, from "panorama", a rotation of the camera along the x-axis, *tilt* a rotation along the y-axis and *dolly*, a movement along the z-axis.

3.2 Transition effects

Transitions among shots are considered as the punctuation symbols of cinematography; they affect the rhythm of the discourse and the message conveyed by the video. The main effects are *cut* - the first frame of the shot to be displayed immediately replaces the last frame of the shot currently on display; *fade* - a shot is gradually replaced by (fade out) or gradually replaces (fade in) a black screen or another shot and *cross fade* (or dissolve) which is the composition of a fade out on the displayed shot and a fade in applied to the shot to be shown.

4 Rhetorical Structure Theory

Rhetorical Structure Theory (Mann and Thompson, 1987) allows the analysis of discourse structure in terms of dependency trees, with each node of the tree being a text span. Each branch of the tree represents a relationship between two nodes. One node is called the nucleus and the other is called the satellite. The information in the satellite relates to that found in the nucleus in that it expresses an idea related to what is said in the nucleus. For example, a *background* relation holds when a satellite provides a context to the information expressed in the nucleus. Figure 1 shows an example of a portion of a rhetorical tree. The second paragraph provides details with respect to the content expressed in the first paragraph. This additional information acts as a sort of reinforcement for what has been previously said in the first paragraph and consequently facilitates the absorption of information. In the original formulation by Mann and Thompson the theory posited twenty



At the bottom on the right is a blacksmith's workshop, a plebeian antithesis to the tournament going on in the upper part of the painting which is chiefly an aristocratic activity.

The choice of a tournament for the month of February is related to the jousts and revelries that took place in carnival time.

Figure 1: An example of Rhetorical Tree.

different rhetorical relations. From this original repository we borrowed a set of relations (elaboration, background, sequence and circumstance), which are commonly used in descriptive text, like those we have analyzed (see Section 6.1).

5 Heuristics and constraints of cinematography

Directors and film critics have identified several heuristics for making good movies. In designing a shot, it is important to consider the message that it has to convey and the (semantic) relations with the previous and following messages. Camera movements can be used to signal some of these semantic relations. For example, according to Arijon (1976), panning and tilting can be used to reveal spatial relations among objects and to move the watcher's attention from one center of interest to another; dollying can be employed to focus the attention on a particular zone or object previously displayed. For example, if an object is currently displayed and the following message deepens one aspect of it, a zoom on that aspect can be chosen.

Besides rules for movement selection, cinematographers have also identified a set of constraints on possible camera movement combinations, in order to ensure a pleasant presentation. In particular, each camera movement has to be "consistent" with respect to the previous movements. The watcher, looking at a movie in which camera moves to one side and then to the opposite one, can misunderstand the underlying message and experience some difficulties in following the stream of the presentation. For example, if the previous move is a pan towards the right the following effect cannot be a pan towards the left neither along the same path nor along similar paths. In general when a camera movement is chosen it constrains the choice of the following movements.

Another important feature of a movie is *cohesion*. A video sequence has to be a *continuum*, an uninterrupted stream in which each piece is connected to the others and is part of a whole. To achieve cohesion in designing the visual part of a presentation it is worth considering the relations among the new information to be delivered and those already given (*discourse history*) and to provide rhetorical strategies to build the presentation. The combination of rules and constraints encode some of the basic "principles of cinematography", which have been identified with the help of an expert director of documentaries (see also Arijon 1976).

Rules and constraints are the core on which the system relies. They encode the rhetorical strategies that are the basic resource for: (i) selecting appropriate images, (ii) designing the presentation structure, (iii) completing each shot, (iv) synchronizing the visual part with the audio commentary and avoiding the "seasickness" effect. Rules are formalized in a context sensitive "presentation grammar", are fired by a forward chaining mechanism and are relative to: (i) rhetorical relations among the text spans; (ii) the geometric properties of images selected from the information repository and (iii) the topics matching among segments and images. An example of rule is given in Figure 2. The rule applies when a segment has a relation of type background or circumstance; in that case the segment is assigned to a new shot.

Constraints are conditions that forbid particular combinations of camera movements and are tested

```
(defrule split (segment)
  (conditions
    (or (has-relation segment background)
        (has-relation segment circumstance)))
  (actions
    (init-shot shot)
    (add-segment segment shot)))
```

Figure 2: An example of rule for shot initialization.

```
(defconstraint zoom-in
  (var mv (get-previous-movement))
  (var mv2 (get-previous-movement mv))
  (and
      (not (equal mv zoom-out))
      (not (equal mv2 zoom-out))))
```

Figure 3: An example of constraint.

according to the type of movement proposed by the engine and the sequence of past movements. An example of constraint is shown in Figure 3. Potentially each camera movement can lead to an inconsistent sequence. To select a zoom-in movement it is worth considering whether the previous move or the penultimate one is a zoom-out; if not, then a zoom-in applies.

6 The Video Planner Engine

The engine is structured as in Figure 4. When a video for a given commentary is requested, the engine analyses the discourse structure of the commentary and selects an appropriate set of images to be presented. The generation chain consists of four phases:

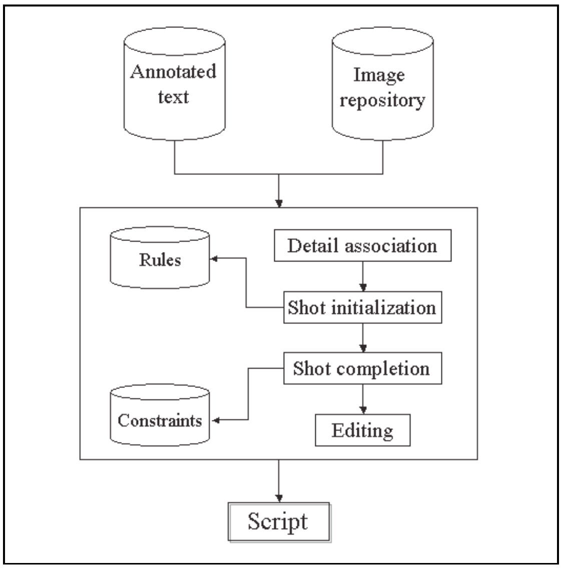


Figure 4: The system architecture.

Detail Association: a detail is associated with each segment of the commentary;

Shot initialization and structure planning: a candidate structure for the final presentation is elaborated, taking into consideration the rhetorical structure of the commentary;

```
<movie id="january">
  <shots>
    <shot id="shot603" image="det01">
      <video-track>
        <pause duration="2"/>
      </video-track>
      <audio-track>
        <play audio="january.wav"/>
      </audio-track>
    </shot>
    <shot id="shot605" image="det01">
      <video-track>
        <pause duration="1"/>
        <zoom duration="4" scale="4"/>
        <pause duration="2"/>
      </video-track>
      <audio-track>
        <audio-pause duration="3"/>
        <play audio="snowball-fight.wav"/>
        <audio-pause duration="1"/>
        <play audio="castle.wav"/>
      </audio-track>
    </shot>
  </shots>
  <editing>
    <display shot="shot603"/>
    <crossfade shot="shot605" duration="1"/>
  </editing>
</movie>
```

Figure 5: Example of script.

Shot Completion: camera movements between details are planned. Constraints are considered in order to avoid “inconsistencies”;

Editing: transitions among effects are selected according to the rhetorical structure of the commentary.

The output is a complete script for the video and the audio channels encoded in a renderer-independent markup language (see Figure 5).

6.1 Resources

The video engine requires access to information about the structure of the data and a certain amount of knowledge about the domain.

As domain of application we have chosen the *Cycle of the Months* of Torre Aquila at the Buonconsiglio Castle in the city of Trento (Italy). This fresco is composed of eleven panels (each one representing a month) painted during the 1400s and illustrates the activities of aristocrats and peasants throughout the year.

As a case study we have collected a set of text that have been annotated by means of RST (see below). The nature of these texts, taken from a guide

```
<segment id="01" parent="root" relname="none"
topic="tournament" audio="castle.wav"
duration="3" >
At the bottom on the right is a blacksmith's
workshop, a plebeian antithesis to the
tournament going on in the upper part of the
painting which is chiefly an aristocratic
activity.  </segment>
<segment id="02" parent="01"
relname="elaboration" topics="castle"
audio="windows.wav" duration="2" />
The differences between the various styles of
construction have been reproduced extremely
carefully.
</segment>
```

Figure 6: Enriched RST annotation of a text.

of Torre Aquila, is descriptive and the prevailing rhetorical relations are elaboration, sequence, circumstance and background. At the moment we have favoured a sentence-by-sentence segmentation and the average size of the resulting trees ranges from seven to ten nodes.

The domain knowledge is encoded in a simple taxonomy, that is a set of keywords - called *topics* - representing entities, such as characters and animals, and processes, such as hunting and leisure activities. At this phase of the work, only one relation between topics is defined, the *member-of* relation, that denotes that a topic belongs to a particular class. For instance, the topic *fox_hunting* is in a *member-of* relation with the topic *hunting*, which means that *fox_hunting* is a form of *hunting*. At the moment, even if simple, knowledge representation is rich enough to accomplish our purposes.

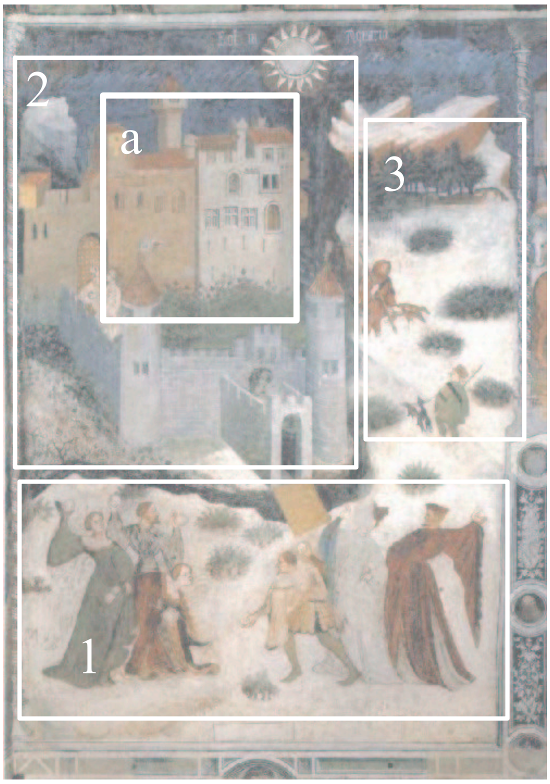


Figure 7: Details for the picture of January.

The main input of the engine is a textual representation of the commentary annotated according to its rhetorical structure (see Figure 6). Additionally, the main concept of each segment is specified as well as the duration in milliseconds of the segment when played (although in Figure 6 the transcription of the commentary is shown, it is never used). Finally, the engine employs a database of images. For each images, the relevant details depicted have to be specified both in terms of their bounding boxes and in terms of the topics they represent. For example, Figure 7 illustrates the details for the panel of the month of January, annotated as in Figure 8. This picture consists of three main details: the snowball fight at the bottom (1), the castle at the top on the right (2) and the hunting scene (3), beside the castle. Within each detail it is possible to identify further details, as in the case of the castle, which contains the detail of windows (a).

6.2 Phase 1: Detail association

In this phase the system assigns one or more details to each segment of the commentary. This operation is performed by searching the image repository for details with the same topic of the segment.

```
<db month="january">
  <image id="jan_img" source="january_full.jpg"
  height="713" width="500"/>
  ...
  <detail id="01" topic="january" parent="root"
    img="jan_img"  coords="0,0,500,713"/>
  <detail id="02" topic="snowball-fight"
    parent="01" img="january_img"
    coords="20,430,460,650"/>
  <detail id="03" topic="castle" parent="01"
    img="january_img" coords="12,50,330,430"/>
  <detail id="03a" topic="window1"  parent="03"
    img="january_img" coords="190,55,315,300"/>
  <detail id="04" topic="hunters" parent="01"
    img="january_img" coords="300,105,475,400"/>
</db>
```

Figure 8: Annotation of the image in figure 7.

6.3 Phase 2: Shot initialization

In this phase, shots are initialized taking into consideration the rhetorical structure of the commentary. At the moment the nucleus/satellite distinction is not taken into account. The result of phase 2 is a candidate structure for the final presentation. The processing is guided by a set of rules, which are fired when particular configurations of rhetorical relations are matched (see Figure 2). For example a relation of type elaboration or sequence signals a smooth transition from the current topic to new information that is strictly related to it; it is thus preferable to aggregate segments in the same shot and to exploit camera movements. Background and circumstance tend to highlight the introduction of new information that provides a context in which the following or the previous messages can be interpreted. They tend to break the flow of the discourse. It is thus preferable to split the segments in two different shots so that, in the next phase, it is possible to exploit proper transition effects in order to emphasize that change of rhythm. There are cases in which the structure planned in this phase is revised during successive stages of computation. For example, to avoid the ”seasickness” effect the system can apply constraints and then modify the previously planned structure by adding new shots (see examples in section 7).

6.4 Phase 3: Shot completion

This is the phase in which the engine incrementally completes each shot by illustrating each of its segments. In performing this task the engine traces the camera movements already planned. When a candidate move is proposed the system verifies whether it is suitable or not according to the list of past camera movements and the constraints imposed over that type of movement. Constraints encode the cinematographer’s expertise in selecting and applying camera movements in order to obtain “well-formed” shots. For instance, when a panning movement is proposed where the previous movement is also a panning, the system has to check if the resulting sequence is suitable. Simple constraints include:

- When the previous movement is a dolly-out
  a dolly-in cannot be applied;

- When the previous movement is a dolly-in
  a dolly-out cannot be the subsequent movement;

- When a panning or a tilting is along a similar path and in the opposite direction of the previous movement
  that panning or tilting cannot be applied.

Constraints encode *schemes* of forbidden movements and when one of them is not satisfied the proposed move is rejected. In this case the engine initializes a new shot, declares the previous one completed and associates the remaining segments to the new shot.

6.5 Phase 4: Movie Editing

This is the phase in which the engine chooses the “punctuation” of the presentation. Movie editing is achieved by selecting appropriate transitions among shots. In order to reflect the rhythm of the discourse, the choice of transition effects is guided by the rhetorical structure of the commentary. The system retrieves the last segment of the shot displayed and the first segment of the shot to be presented and plans the transition according to the following rules:

- If two segments are linked by a relation of type elaboration
  a short cross fade applies;

- If two segments are linked by a relation of type background or circumstance
  a long cross fade applies.

- If two segments are linked by a relation of type sequence
  a cut applies.

- If a relation of type enumeration holds among two or more segments
  a rapid sequence of cut applies.

These rules have been selected according to the observations about the usual employment of transition effects in the field of cinematography (Arijon, 1976). Fade effects are fit for smooth transition, when there is a topic shift or when the center

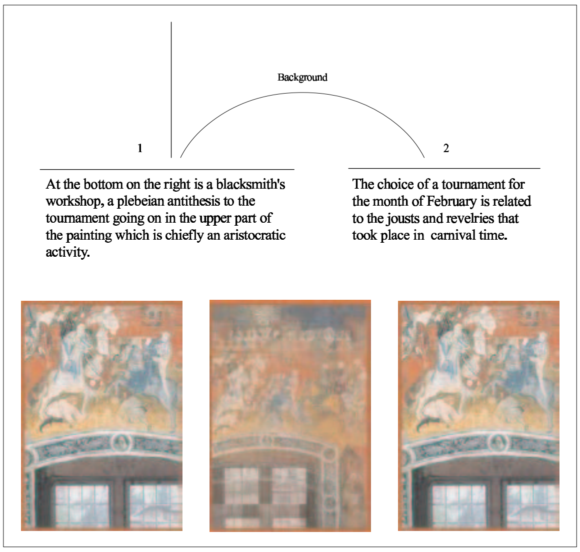

Background
1
2
At the bottom on the right is a blacksmith's workshop, a plebeian antithesis to the tournament going on in the upper part of the painting which is chiefly an aristocratic activity.
The choice of a tournament for the month of February is related to the jousts and revelries that took place in carnival time.

Figure 9: The “Tournament” example.

of interest changes but the new topic is related to the old one, as in the case of elaboration or background. Cut is more appropriate for abrupt and rapid changes, to emphasize the introduction of a new concept, as in the case of sequence. A special case holds when the verbal commentary enumerates a set of subjects or different aspects of the same object; in those cases a rapid sequence of cuts can be used to visually enumerate the elements described.

7 Examples

The first example concerns the rhythm of the discourse (Figure 9). Since the topic of both segments is the same, the text could be visually represented by displaying the same image during the playing of both the first and the second audio commentary. In this case a cross fade effect helps the user to understand that background information is going to be provided. In fact, the second segment provides contextual information to support the user in understanding the information presented in the first paragraph. The first image is thus presented while the audio of the first segment is played; then, when the audio switches to the second segment, the image is enlarged to cover the entire panel and finally refocused on the detail once the audio has stopped. By adopting this strategy the system generates a movie that reflects the discourse structure of the text and the rhythm of the discourse, supporting the same communicative goals of the verbal part of the presentation.

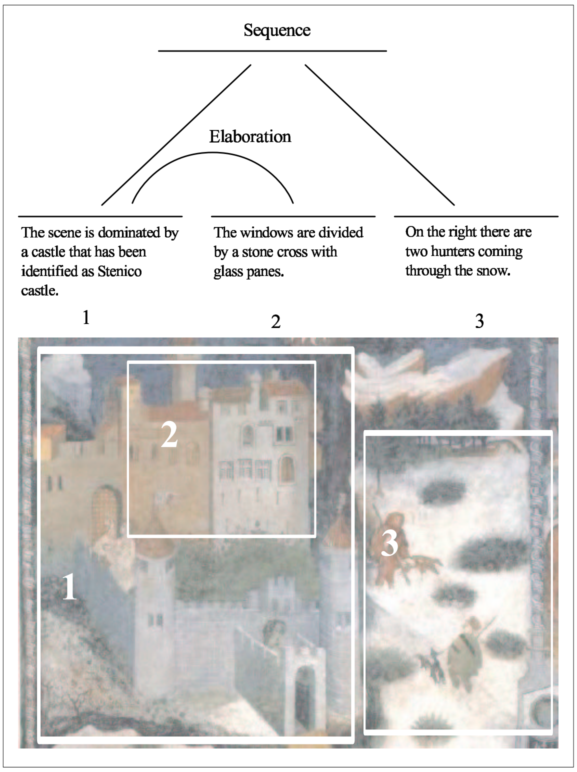

Sequence
Elaboration
The scene is dominated by a castle that has been identified as Stenico castle.
The windows are divided by a stone cross with glass panes.
On the right there are two hunters coming through the snow.
1
2
3
1
2
3

Figure 10: The “Castle” example.

The second example concerns the application of constraints in order to avoid an inconsistent sequence of camera movements (Figure 10). The text first describes the castle on the left. In this case the system, after a brief pause on the whole scene, selects a dolly-in movement, magnifying the detail of the castle (1). Then a second dolly-in is applied to focus on the castle’s windows (2). Finally, in order to focus on the hunting scene (3) the camera should dolly out and then move towards right, but this combination is forbidden by the constraint on dolly-out. In this case the engine revises the structure of the movie. It declares completed the current shot, initializes a new shot and associates the remaining segments with it.

8 Conclusions and future work

In this paper we have presented an engine to generate video sequences starting from an audio commentary. First, we have identified a set of cinematic techniques that are the basic resources to plan the presentation. Second, we have shown how the resources (knowledge on the rhetorical structure of the commentary, knowledge about the domain and the repository of images) are annotated. Third, we have illustrated the architecture of the engine and the four steps of computation. Finally we have presented some examples, which show how the system employs rules and constraints to generate engaging presentations.

At the moment the system relies on a set of fifteen rules and ten constraints. Improvements are envisaged in particular to take into consideration the time needed to complete the movements (in this moment we assume a constant speed of the camera in movements) and more elaborated strategies to replan forbidden sequences of camera movements.

We have noted that the annotation of the resources (especially text) is time-consuming. In the future, in order to speed-up this task, we intend to investigate the possibility of a (semi-)automatic annotation of the discourse structure.

The application of the video clips in a mobile museum guide is currently under study (Zancanaro et al., 2003) and we are now experimenting with the techniques described here to automatically produce user-tailored videos.

Acknowledgments

This work has been supported by the PEACH and TICCA projects, funded by the Autonomous Province of Trento.

References

Elisabeth André. 2000. The Generation of Multimedia Documents. In *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, pages 305–327. Marcel Dekker Inc., New York.

Daniel Arijon. 1976. *Grammar of the Film Language*. Silman-James Press, Los Angeles, CA.

William. H. Bares and James. C. Lester. 1997. Realtime Generation of Customized 3d Animated Explanations for Knowledge-based Learning Environments. In *AAAI97 Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 347–354, Rhode Island, July 27-31.

Anreas Butz. 1994. BETTY: Planning and Generating Animations for the Visualization of Movements and Spatial Relations. In *Proceedings of Advanced Visual Interfaces*, Bary Italy.

Anreas Butz. 1997. Anymation with CATHY. In *Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference*, volume 1, pages 957–962, Providence, Rhode Island, July 27-31.

David B. Christianson, Sean E. Anderson, Li We He, David Salesin, Daniel S. Weld, and Michael F. Cohen. 1996. Declarative Camera Control for Automatic Cinematography. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference*, volume 1, pages 148–155, Portland, Oregon, August 4-8.

Brent H. Daniel, Charles B. Callaway, William H. Bares, and James C. Lester. 1999. Student-sensitive Multimodal Explanation Generation for 3d Learning Environments. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, volume 1, pages 114–120, Orlando, Florida, July 18-22.

Peter Karp and Steve Feiner. 1993. Automated Presentation Planning of Animation using Task Decomposition with Heuristic Reasoning. In *Proceedings of Graphics Interface*, pages 118–127.

William C. Mann and Sandra Thompson. 1987. Rhetorical Structure Theory: a Theory of Text Organization. In *The Structure of Discourse*. Ablex Publishing Corporation.

Mark T. Maybury. 1993. *Intelligent Multimedia Interfaces*. AAAI Press.

Christian Metz. 1974. *Film Language: a Semiotics of the Cinema*. Oxford University Press.

Wolfgang Wahlster, Elisabeth André, Wolfgang Finkler, Hans-Jrgen Profitilich, and Thomas Rist. 1993. Plan-based Integration of Natural Language and Graphics Generation. *Artificial Intelligence*, 63:387–427.