Available online at www.sciencedirect.com

SCIENCE DIRECT®

Artificial Intelligence 165 (2005) 57–89

Artificial Intelligence

www.elsevier.com/locate/artint

Automatic cinematography and multilingual NLG for generating video documentaries

Charles Callaway, Elena Not, Alessandra Novello, Cesare Rocchi, Oliviero Stock [*], Massimo Zancanaro

ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, Trento, Italy

Received 16 July 2004

Available online 14 March 2005

Abstract

Automatically constructing a complete documentary or educational film from scattered pieces of images and knowledge is a significant challenge. Even when this information is provided in an annotated format, the problems of ordering, structuring and animating sequences of images, and producing natural language descriptions that correspond to those images within multiple constraints, are each individually difficult tasks.

This paper describes an approach for tackling these problems through a combination of rhetorical structures with narrative and film theory to produce movie-like visual animations from still images along with natural language generation techniques needed to produce text descriptions of what is being seen in the animations. The use of rhetorical structures from NLG is used to integrate separate components for video creation and script generation. We further describe an implementation, named GLAMOUR, that produces actual, short video documentaries, focusing on a cultural heritage domain, and that have been evaluated by professional filmmakers.

Keywords: Automatic cinematography; Natural language generation; Multimedia presentations

[*] Corresponding author.

E-mail addresses: ccallawa@inf.ed.ac.uk (C. Callaway), not@itc.it (E. Not), novello@itc.it (A. Novello), rocchi@itc.it (C. Rocchi), stock@itc.it (O. Stock), zancana@itc.it (M. Zancanaro).

URL: http://tcc.itc.it/.

0004-3702/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.artint.2005.02.001

1. Introduction

The use of multimodal output for the purposes of education, tutoring, and instruction has been around for several decades. Beginning with STEAMER [30] in the 1980's, graphics were paired with text descriptions to enhance the learning process. As multimodal techniques have steadily advanced [1,33,48], animated agents and 3D effects have appeared, along with the increasing use of voice synthesis. However, the overall goal has remained the same: to increase the learning capability of users by presenting information to multiple senses simultaneously and keeping them engaged and motivated.

We focus here on a similar task in the context of visitors who are interested in learning more about particular works of art while enjoying a visit to a museum or other cultural heritage institution [47] by automatically creating film documentaries given rudimentary information about them. Specifically, we propose to use existing 2D pictures of artwork along with a knowledge base of facts about that artwork to automatically construct short movie-like multimedia presentations which are similar to professionally made documentaries. The eventual aim is to allow for personalization in the selection of the material to be presented to the user, according to her preferences, interests, and both current and past interactions with the system.

An example film documentary might start with a high-level sentence of spoken text while showing a full-size image of the artwork. It might then focus on an important element of the art, using simultaneous discourse cues and a zoom shot to focus attention on that important element. After holding the image still while the audio explanation continues, a second region of the artwork may be highlighted with a pan motion of the camera as a voiceover explains the relationship between the two regions. After continued explanation and shot transitions, the film might conclude with a reverse zoom to the full-size image again followed by a fadeout as the audio track completes.

Similar applications have been attempted in the past, although for different domains and goals, and without dynamically generated commentary or an underlying theory of film narrative. One of the first case studies of the generation of "motion presentations" is the work of [31], which generated scripts for animation using top-down hierarchical planning techniques. [18] presents a successful attempt to encode several of the principles of cinematography in the *Declarative Camera Control Language*. Another system is CATHI [9], which generated animated presentations in the form of 3D animated clips for the illustration of technical devices. Animated presentations have also been successfully employed in multimodal frameworks for the generation of explanations [22] and in learning environments [4,16]. [5] presents a model for camera planning based on a constraint satisfaction approach. Their system is meant to work in a virtual 3D setting. A camera planning agent for polygonal graphics is described in [28], which uses a genetic algorithm to find the optimal solution for a given set of communicative goals. [26] presents an interesting system, based on non-monotonic reasoning. It produces 3D animated movies from screenplays which have been encoded in a high-level formal language.

Automatically creating a video documentary entails solving several problems: selecting important information from a large repository, ordering the text and film sequences in a coherent and meaningful way, and synchronization and playback of the intended audio and visual effects. These three elements have long been known to be fundamental to the problem of Natural Language Generation (NLG), which uses the phases of content determination, content organization and surface realization to produce paragraphs of text from a knowledge base. It was thus natural to apply techniques that have already proven effective, such as the rhetorical organization of the content (RST [32]), and rules well established in the field of cinematography [36] to solve the problem of image sequencing and animation.

In this paper, we describe an engine to build video sequences from images that are synchronized with synthesized audio generated using deep linguistic representations. The input to the engine, named GLAMOUR, is a series of still images and a knowledge base containing information about those images as well as information about their domain in general. GLAMOUR selects and organizes the content to be conveyed and produces textual descriptions using standard deep NLG techniques, while its video planner, taking into consideration the discourse structure of the commentary, plans film segmentations through the use of shots, camera movements, and transition effects between shots. The output of the engine is a complete script of a “video presentation”, with instructions for synchronizing images and camera movements accompanied by audio commentary, synthesized from the generated text. One of the chief novelties of this work is the use of rhetorical relations to help provide structure to both the image sequences and the spoken part of the script.

In order to test the idea that textual goals and cinematographic goals can be achieved in unison, we have implemented a prototype capable of creating several simple movies lasting on average two minutes. Each film describes a particular renaissance fresco to a museum visitor, discussing the work’s historical context, and pointing out particular details of importance or interest.

Given the themes of exploiting artificial intelligence techniques for improving education and using rhetorical structure for combining dynamic images and deep NLG, we are interested in using the system to answer several fundamental research questions, such as (1) how to best express global documentary constraints as realized in a particular media, (2) if there is an underlying organizational similarity in text and film generation, and (3) if it is possible to employ a modular architecture that can accommodate different techniques.

In particular, cinematographic aspects of GLAMOUR address various research challenges, such as how to select an appropriate image to illustrate a topic, what is the rhythm of the discourse and how to convey it in a movie, how to highlight the flow of the information presented (e.g., draw comparisons with related material), how to avoid unpleasant combinations of camera movements, unless explicitly desired in order to achieve particular effects.

We have thus organized this article in the following manner: the next section deals with the high-level architecture that integrates the cinematography and NLG systems, the available resources, and how they are utilized by the system. Section 3 describes the creation of the documentary script, including both text planning and text realization. Section 4 presents the principles and terminology of cinematography as used by professional filmmakers, and details how the cinematographic planner uses the high-level text structure to create the visual aspects of the film. Finally, Sections 5 and 6 describe respectively the implementation and evaluation of the GLAMOUR documentary film generator.

2. Generation of video documentaries

Cinematography is the art of coupling pictures and sounds in a meaningful way. Audio and video resources become tightly connected, supporting one another and becoming a whole. This synergy is not a mere sum of two different communicative techniques. Audio and video continuously rely on each other to carry the message to the viewer. The construction of a movie thus involves issues ranging from the selection of the appropriate channel (visual, auditory or both) for expressing the message to the planning of synchronous actions (e.g., the actors or a narrator talk while a given scene is being shown), and from the correct positioning of cameras to the setting of scenes and lights.

Throughout the last fifty years film critics have been studying the nature of movies, to verify whether they adhere to certain rules and whether cinematography can be considered on a par with language. According to [36], 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*. When film makers started experimenting with sequences of images, trying to relate them to one another, there was no language of cinematography. The rules they found were the product of trial and error, collected by practice and experimentation. Thus there are no "formal" (in a logic sense) versions of film grammar. Nevertheless, there are conventions which are generally accepted by both directors and filmgoers. These guidelines may be considered a rich resource from which we can borrow principles, heuristics and "rules of thumb" to develop multimedia movie-like presentations.

The steps to create a movie are many and complex: invent the subject, conceive the story, write the script, shoot the video material (production), do the editing or montage (post-production). During the first stages, the scriptwriter selects how the story will evolve, what the actors or the narrator will say and in what order. During the production, the director and his team set up the scene (lights, camera positions and movements) and film one or more takes of a piece of movie. Then in the second phase, the filmed material is analyzed, some shots are dropped, others are cut and transitions between scenes are selected. According to some cinema theorists (mainly [24]), the montage is the key step of the movie creation. Isolated shots, though well filmed and carefully planned, are only raw material. The editing gives sense to the movie, enabling the passage from photography to cinema.

Another set of key principles relates to the "meanings" of camera movements and their possible combinations. Arijon, for example, documents many possible uses of camera movements, providing examples of different scenes and different settings [3]. A discussion of the combinatorics of different camera movements raises a question concerning constraints. There are sequences of movements that are commonly considered as forbidden, since they might be misleading for the viewer. For example, a camera moving forth and back along the same path could lead the viewer to misunderstand the underlying message of a movie section. Therefore, unless the director wants to achieve a really particular, unexpected and non-conventional effect, the use of some sequences of camera movements must be avoided.

Camera movements are typically used to render more explicit the coherence relations that link the various portions of the narration, thus contributing to the overall efficacy of the presentation. E.g., a camera zoom-in might help the viewer focus on the fresco detail currently described, whereas a zoom-out might help to re-establish a general view of the

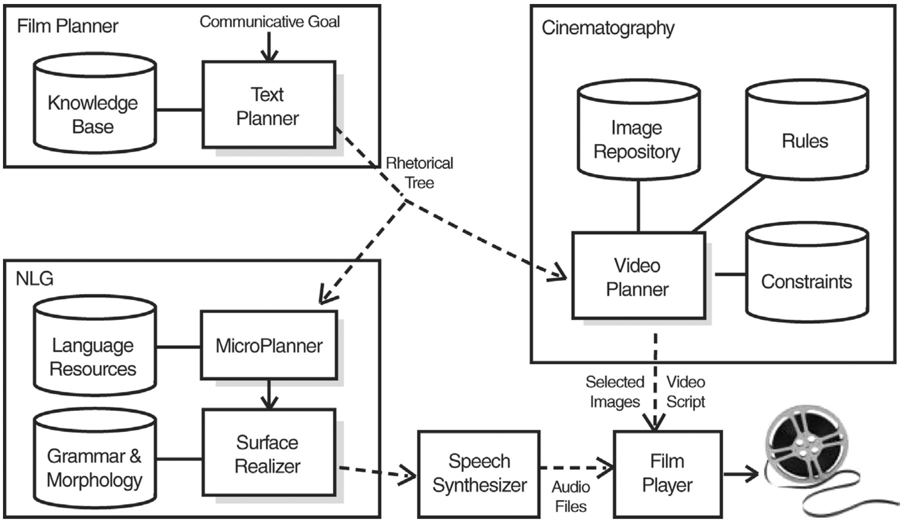


Fig. 1. The GLAMOUR architecture for creating film documentaries.

scene. Principles derived from multimedia and cinematography therefore should be used throughout the different stages of building a video presentation. In the following section, we describe how we designed an architecture for an automatic system that produces video documentaries.

2.1. System architecture

As depicted in Fig. 1, the architecture of the system is based on a bipolar cascade model. The two processes, the planning of the verbal content and the planning of the video script, can be performed in parallel. Parallelism allows for speed (which can be very important if the generation of personalized videos is embedded in an interactive system [45]) and modularity in terms of both reusability of components and the possibility of experimenting with different approaches to cinematography or text generation.

While these are computational issues, other perspectives are also important such as those of documentary filmmakers themselves. For instance, some directors believe that human filmmakers choose among mental images when constructing a film, and only secondarily think about the text of the script [49]. GLAMOUR is a compromise, given that the information needed to construct a film is contained in a knowledge base, not in the 2D images. The system thus uses rhetorical text planning as a bridge, where the cinematography and lower-level NLG systems construct film elements in parallel under a single, overarching plan.

The NLG cascade is organized as the standard architecture proposed in [42]. In the first phase, relevant content of the document is retrieved from the knowledge base (KB) and organized into a discourse plan structured as a rhetorical tree. Then, the microplanner applies pragmatic strategies to decide how to realize nouns (i.e., whether to use anaphora or full descriptive noun phrases, etc.) and verbs (i.e., decide the tense, the aspect and so on). Finally, a surface realizer completes the process by selecting closed-class words and enforcing syntactic and morphology constraints according to the grammar of the target language. In our system, the cascade adds English and Italian speech synthesizers as well in order to produce verbal commentary for each video.

The Video Planner takes a rhetorical tree whose leaves represent the content facts that are to be verbalized by the NLG engine. It decides the shot segmentation, mainly according to the topic progression, and the transitions between shots, mainly according to the type of rhetorical relation between the nodes. The planner uses a database of images in which the visual details are segmented and annotated according to a similar but shallower semantic model employed in the KB.

Since the Video Planner works from the rhetorical tree, it is not aware of the actual duration of the final synthesized speech. This is not an important issue for the synchronization between the camera movements and speech since the actual timing for the chosen camera movements are expressed in terms of relative constraints. Yet, knowledge about the duration of each segment might be useful to choose among different visual strategies. For example, it might be the case that a long elaboration is introduced by a cross-fade, while a short elaboration can simply be introduced by a cut. To allow for these high-level strategies, the text planner estimates the duration of each segment using the number of discourse facts that it contains (see Section 5).

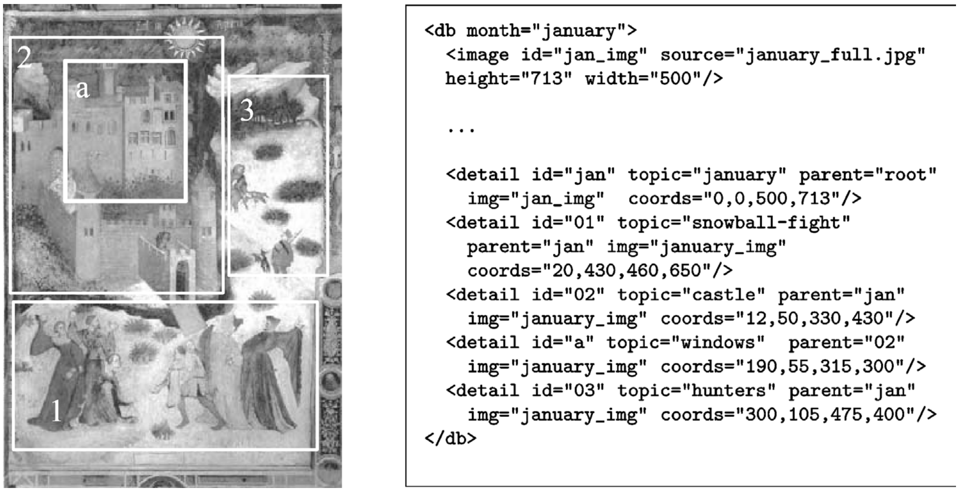
This architecture differs from other similar works. For example, in the architecture employed by [1], the content planner would also comprise the functionalities of the video planner. A single unified planner would need to include rules that contain both cinematic and linguistic conditions and actions. Thus linguistic knowledge governing syntactic constraints would be integrated with rules for determining whether two subimages are near or far away from each other, and shot sequencing rules would have to pay attention to the possible lexicalizations of discourse markers.

The architecture presented here has two advantages. First, it aids efficiency, since both main tasks can be performed in parallel. Second, it enhances the portability of the modules. For example, the verbal commentary might be realized in a simpler way than we describe in Section 3 by adopting a template generator [43] (obviously at the expense of text flexibility).

The hypothesis behind this architecture is that the video planner can effectively plan the script with only rhetorical information and basic knowledge of the topics of each discourse segments. Section 4 introduces our rule-based engine to map shots and transitions from sub-trees of the rhetorical tree produced by the language component. It is worth noting again that underlying both the high-level text and video planning is the parallel use of rhetorical structure trees to ensure cohesion in the final product.

2.2. Image annotation

To enable the choice of images and the planning of camera movements, the Video Planner needs a repository of annotated images. For each image, the relevant details depicted have to be specified both in terms of their bounding boxes and of the topics they represent. For example, Fig. 2 illustrates the details for a 15th century fresco representing activities for the month of January and its annotation. This picture consists of three main details:



```
<db month="january">
  <image id="jan_img" source="january_full.jpg"
  height="713" width="500"/>

  ...

  <detail id="jan" topic="january" parent="root"
    img="jan_img"  coords="0,0,500,713"/>
  <detail id="01" topic="snowball-fight"
    parent="jan" img="january_img"
    coords="20,430,460,650"/>
  <detail id="02" topic="castle" parent="jan"
    img="january_img" coords="12,50,330,430"/>
  <detail id="a" topic="windows"  parent="02"
    img="january_img" coords="190,55,315,300"/>
  <detail id="03" topic="hunters" parent="jan"
    img="january_img" coords="300,105,475,400"/>
</db>
```

Fig. 2. An image and its annotation.

(1) the snowball fight at the bottom, (2) the castle at the top on the right, and (3) the hunting scene beside the castle. Within each detail it is possible to identify further details, as in the case of the castle, which contains further details such as the windows (a). Annotation time is negligible, consisting mainly of recording the coordinates of bounding boxes and assigning an ontological knowledge base concept to each, according to the intuition of the annotator.

3. Planning text in documentary descriptions

Generating text for producing documentary descriptions implies several strategic decisions on (1) what is the most effective content to be conveyed to successfully describe the overall object in focus; (2) how information can be ordered and structured in a coherent way; (3) which lexico-grammatical patterns can be used to manifest the semantic relations between the text units to reinforce the cohesion of the text (e.g., appropriate referring expressions or cue phrases) or its naturalness (e.g., aggregation); and (4) what the final surface realization should look like.

Traditionally, steps (1) and (2) are considered part of the text planning stage, whereas step (3) accounts for microplanning and (4) for tactical generation, even though, from the implementation point of view, these stages are not necessarily implemented in a pure pipeline cascade [10,42,44].

In our scenario, the requirement of generating text to be integrated in a multimedia presentation additionally involves that all the stages above need to take into account how the final text will be rendered (i.e., via speech synthesis) and synchronized with the visual track.[1] This means that the generation system needs to build a narration around the physical and visual structure of the object, with explicit reference to its details.[2]

It should be noted, however, that the major assumption that has motivated our design choices for the overall system architecture is that the text generation component takes its strategic and tactical decisions independently from the specific media used for the visual presentation (be it still images, animations of still images, or real movie clips). What the text generation component can count on is that there will be visual feedback consistent with the content of the commentary. But no knowledge is required at this stage about which camera movements will be used or what shot transitions may apply, allowing for substantial modularity and flexibility.

Generally speaking, video documentaries could be readily generated using standard shallow NLG techniques such as slot-filler templates to create sentences for the audio track. In this work, however, we have adopted the use of deep NLG with a fuller linguistic representation for several reasons: (1) deep NLG allows for a finer-grained control in the generation of referring expressions and other linguistic devices that help improve the naturalness and cohesion of the text (e.g., aggregation), and facilitates the generation of personalized texts which are adapted to the current interaction context and user preferences; (2) as in many other application domains, users in the museum scenario are typically from many countries and speak their own native languages, which makes multilingual generation (a laborious task for templates) very important; and (3) the underlying architecture can be shared with other application tasks over the same domain (e.g., direction giving or report generation for museum visits) or other projects requiring generation, reducing the intensive costs of creating deep linguistic resources and domain models.

Deep generation also presents us with challenges for future work: (4) deep generation of the commentary permits experimenting with a finer tuning of the speech synthesis process given the possibility of automatically annotating the generated text with semantic and prosodic information; and (5) in addition, a finer grained synchronization of the audio and camera movements can be obtained by exploiting the available detailed discourse structure of the text.

3.1. Generation component overview

The architecture we adopted for the generation subsystem adheres to a fairly standard pipeline model [42], as shown in Fig. 3. (The grey boxes in the figure highlight components that are currently not used in the cinematographic scenario but that account for the generality of our system. Their role will be briefly discussed here below, though we will mainly focus our discussion on the technical details of the other components, which directly contribute to the generation of documentary descriptions.)

The input to the generation process is a request for text expressed in terms of the communicative goal to be satisfied. For the generation of documentary descriptions, the

[1] Note, however, that the generation of text annotated with semantic and linguistic information to optimally drive the speech synthesis [11,29] is out of the scope of our current work.

[2] An example of a movie generated by GLAMOUR can be seen here: http://tcc.itc.it/i3p/research/aij05/movie.html.

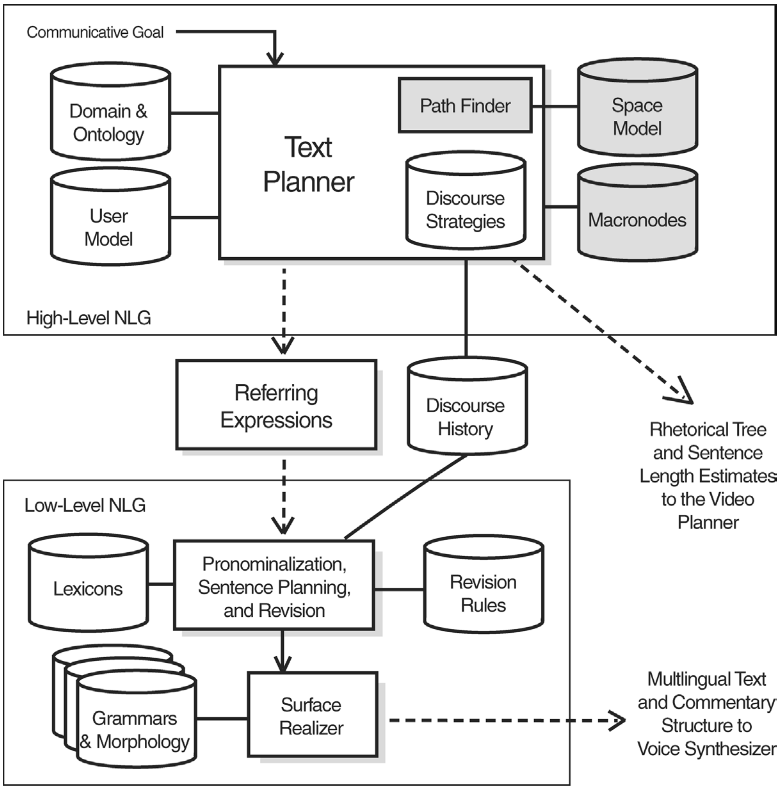
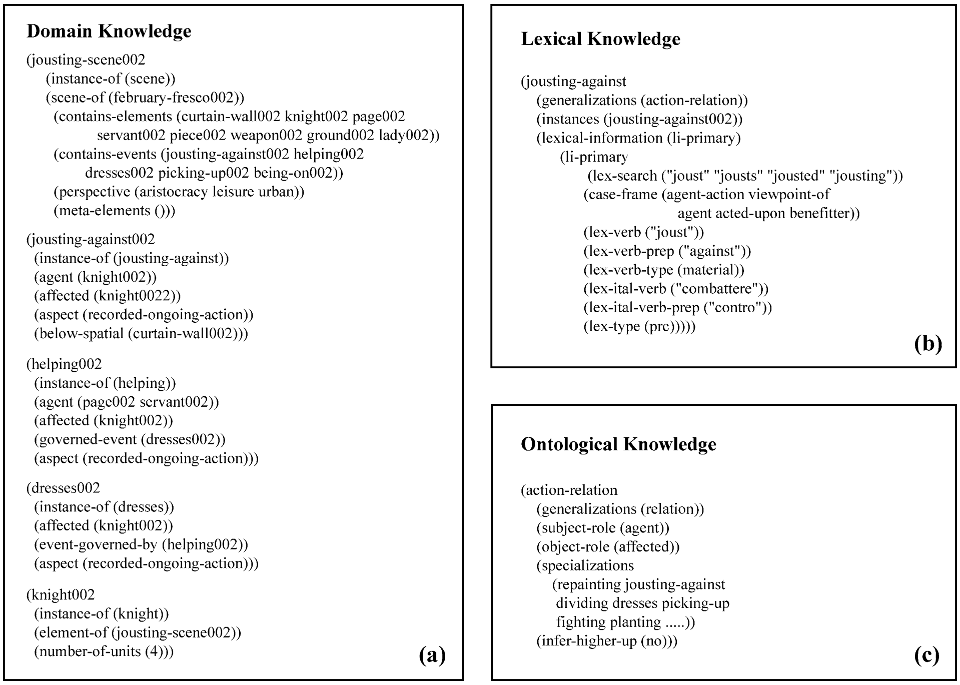


Fig. 3. Architecture for the generation subsystem.

communicative goal tells the *Text Planner* to build an object description, e.g., *describe-object*(*exhibit1*).[3]

The Text Planner is charged with determining the most relevant information to be included in the description and its coherent organization. Information about objects is extracted from a *Domain Ontology and Model* that collects facts and attributes about the objects to be described. For the sake of modularity and reusability, the domain representation is organized into separate partitions that contain (i) specific domain dependent instances describing, for example, which characters and objects are depicted in a fresco, their relative positions, and what activities they are performing (as in Fig. 4(a)); (ii) lexical information required for text realization (see Fig. 4(b)); (iii) domain-independent ontological concepts and relations defining how instances and relations should be semantically interpreted (see Fig. 4(c)). Portions (ii) and (iii) are highly reusable, whereas portion (i) is strictly dependent on the actual objects to be described.

[3] Note, however, that the generation system supports other types of communicative goals (e.g., report generation *summarize-visit*(*log-user1*), direction giving *direction-to-next-door*(*point1*, *point2*), etc.) since the text planner was designed with a multi-purpose scope (see Section 3.1.1).



Domain Knowledge

```
(jousting-scene002
    (instance-of (scene))
    (scene-of (february-fresco002))
      (contains-elements (curtain-wall002 knight002 page002
              servant002 piece002 weapon002 ground002 lady002))
      (contains-events (jousting-against002 helping002
           dresses002 picking-up002 being-on002))
      (perspective (aristocracy leisure urban))
      (meta-elements ()))

(jousting-against002
  (instance-of (jousting-against))
  (agent (knight002))
  (affected (knight0022))
  (aspect (recorded-ongoing-action))
  (below-spatial (curtain-wall002)))

(helping002
  (instance-of (helping))
  (agent (page002 servant002))
  (affected (knight002))
  (governed-event (dresses002))
  (aspect (recorded-ongoing-action)))

(dresses002
  (instance-of (dresses))
  (affected (knight002))
  (event-governed-by (helping002))
  (aspect (recorded-ongoing-action)))

(knight002
  (instance-of (knight))
  (element-of (jousting-scene002))
  (number-of-units (4)))
```

(a)

Lexical Knowledge

```
(jousting-against
   (generalizations (action-relation))
   (instances (jousting-against002))
   (lexical-information (li-primary)
       (li-primary
            (lex-search ("joust" "jousts" "jousted" "jousting"))
           (case-frame (agent-action viewpoint-of
                          agent acted-upon benefitter))
           (lex-verb ("joust"))
           (lex-verb-prep ("against"))
           (lex-verb-type (material))
           (lex-ital-verb ("combattere"))
           (lex-ital-verb-prep ("contro"))
           (lex-type (prc)))))
```

(b)

Ontological Knowledge

```
(action-relation
   (generalizations (relation))
   (subject-role (agent))
   (object-role (affected))
   (specializations
      (repainting jousting-against
       dividing dresses picking-up
       fighting planting .....))
   (infer-higher-up (no)))
```

(c)

Fig. 4. Sample concepts and instances in the domain ontology and model.

Discourse Strategies based on typical patterns of content organization (schemas) and on rhetorical coherence access the Domain representation, extract relevant facts to be conveyed, and guarantee that the discourse plan is properly built (see Section 3.2).

During its decision process, the Text Planner accesses the *User Model*[4] which contains information about (estimated) user interest, knowledge and interaction preferences, and the *Discourse Context* which stores the content conveyed in previous presentations—useful to avoid repetitions and to draw comparisons—and the structure of the presentation currently being built. This latter information is exploited by the *Microplanning* stage to determine the most appropriate referring expressions and lexico-grammatical constraints required to convey the message. The output of the Microplanner is a description of the linearized text plan written in the NSP formalism [15]. The *Surface Realizer* maps this abstract representation onto natural text (including final lexical choice, morphology, and linear precedence) consistent with the output language selected (English or Italian).

The output of the generation process for the cinematographic scenario is not supposed to be plain text, but rather a structured commentary specifying the visual details or concepts the text portions are meant to describe as well as the rhetorical dependencies between them. This output, written in XML notation and compatible with standard tools for RST

[4] Only if the applied context includes one. We have experimented with such a model induced during a museum visit for a separate report generation application.

rhetorical structure visualization [40], is then passed on to the *Documentary Player* for the final synchronization with the video plan, which is generated in parallel by the *Video Planner* working on the intermediate rhetorical tree of domain facts produced by the Text Planner.

3.1.1. Multi-purpose design choices

The design of the text generation subsystem aimed at a high modularity and reusability of components to flexibly support different application tasks and interaction requirements. At the text generation level, this multi-purpose design goal emerged along the following lines:

- Ability to support different communicative goals. The generation component, in fact, is not bound to the cinematographic application task, given that it does not need to take care of shot transitions and camera movements, and can readily be used to generate texts useful in different scenarios, i.e., with a user acquiring information about objects/concepts either off-line (e.g., seeing documentaries at home) and moving in the real environment where objects are found (e.g., browsing information about exhibits in a museum). Apart from the generation of object descriptions which has been coupled with cinematography techniques to generate video documentaries, the generation component supports the generation of summaries for museum visits: at the end of a visitor's tour, a report is generated that includes a basic, personalized narration of her visit, the items and relationships she found most interesting, pointers to additional related online information, and suggestions for future visits to the current and other museums [17]. Preliminary experiments have also been conducted for the generation of wayfinding instructions. This latter scenario motivated the introduction of a *Space Model* and a *Path Finder* component into the system to exploit the potential of deep natural language generation to produce wayfinding instructions, as in the approach adopted by [21].
- Integration of solutions that flexibly allow the integration of pre-existing texts with the deeply generated sentences. During the design and implementation of the text planner, extensions of the base algorithm have been implemented to allow for the inclusion of more shallow NLG techniques, such as templates (e.g., intermixed portions of canned text and deeply generated text, or else generated text alone) and Adaptive Hypermedia solutions. The potential of the latter was investigated in a previous European project, M-PIRO [2], which identified the benefits of a text planner that, besides its abilities to access a knowledge base representing the domain model, is capable of querying a repository of (properly annotated) pre-existing content units (called *Macronodes Repository* in Fig. 3) and appropriately extract and structure the units according to the directives specified in the discourse strategies [38].

  In the current prototype of the GLAMOUR system, we focused on the use of deep NLG techniques alone, but the modularity of the generation component easily allows for experimentation with shallow or hybrid generation techniques, finding the optimal trade-off in the cost/benefit balance according to the various types of application tasks, domains and platform constraints.

- Possibility of experimenting with different surface realizers. The output of the text planner was initially conceived to be at a high level (a rhetorical tree over domain facts) to be passed directly to an off-the-shelf surface realizer provided by the Exprimo multilingual generation system, as adopted in the M-PIRO project [2]. Although the actual integration with Exprimo remained untested, we wanted the ability to use the system with realizers utilizing different approaches. This motivated the explicit introduction of the microplanning phase which allowed us to use the StoryBook low-level generation system, as explained in Sections 3.3 and 3.4 below.

3.2. Content selection and organization

The content selection and organization approach adopted in our text planner is based on the assumption that in descriptive texts/commentaries, the conventions on how information is typically presented play a major role. For example, when describing complex depictions on a painter's canvas, the description (as well as the corresponding image sequence/animation) needs to reflect the spatial organization of details and their salience in motivating the painter's choices.

How the content of the text is actually selected and structured obviously depends significantly on the type of readers addressed and their specific interactions with the system (e.g., difficult terminology needs to be properly explained to children; copyright information for works of art might not always be appropriate as in the case of a commentary generated for a video documentary, etc.). These "patterns of appropriate ordering" (more widely known in the NLG community as *schemas* [34] or the *Generic Structure Potential* (GSP) [27] of a text) have been adopted in many NLG systems to guide the text planner in organizing the text structure.

In our project we have implemented a schema-based text planner in which schemas (which we call GSPs, consistent with the terminology of Halliday and Hasan) are used to declaratively describe how the text chunks should be optimally organized according to the current user profile, discourse history and object to be described. GSPs can be defined at different levels of abstraction and can contain calls to other finer-grained GSPs. Fig. 5 shows a sample GSP. It has an applicability condition, expressed in the Cond field, and contains a conjunction of tests over the discourse context, the user model, and the domain model. For example, the test `[test kb :isa Exhibit FRESCO]` queries the domain model to check whether the input parameter `Exhibit` has been instantiated with a

```
NAME    describe-object
PARAMS  [Exhibit]
COND    [and [test system :nlg on]
             [test kb :isa Exhibit FRESCO]
             [test um :visit-times 0]
LOCVAR  FrescoDepiction [get kb :get-range (Exhibit DEPICTS)]
ELEM    Depiction [schema expand-fact (DEPICTS Exhibit FrescoDepiction)]
ELEM    Detail [schema describe-object [get kb :get-range (Exhibit HAS-SCENES)]]
SCHEMA  [Depiction (Detail* elaboration-detail)]
ROOT    [root Depiction]
```

Fig. 5. Sample GSP.

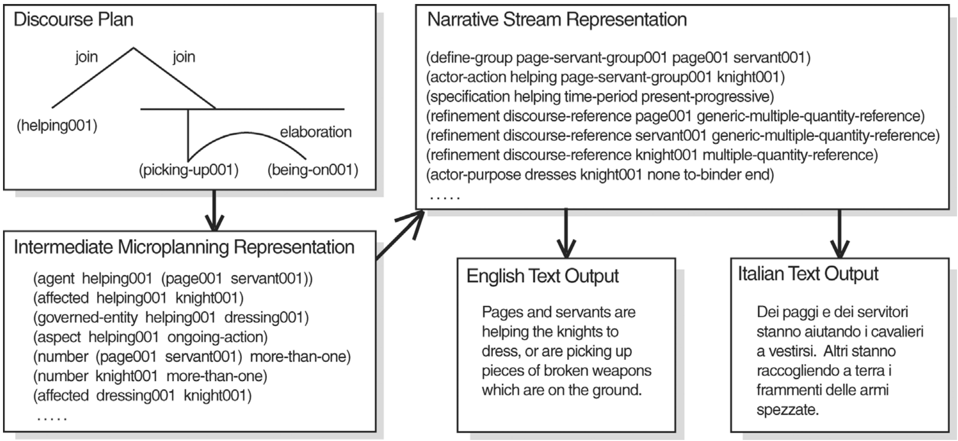

Discourse Plan
join
join
(helping001)
elaboration
(picking-up001)
(being-on001)
Intermediate Microplanning Representation
(agent helping001 (page001 servant001))
(affected helping001 knight001)
(governed-entity helping001 dressing001)
(aspect helping001 ongoing-action)
(number (page001 servant001) more-than-one)
(number knight001 more-than-one)
(affected dressing001 knight001)
.....
Narrative Stream Representation
(define-group page-servant-group001 page001 servant001)
(actor-action helping page-servant-group001 knight001)
(specification helping time-period present-progressive)
(refinement discourse-reference page001 generic-multiple-quantity-reference)
(refinement discourse-reference servant001 generic-multiple-quantity-reference)
(refinement discourse-reference knight001 multiple-quantity-reference)
(actor-purpose dresses knight001 none to-binder end)
.....
English Text Output
Pages and servants are helping the knights to dress, or are picking up pieces of broken weapons which are on the ground.
Italian Text Output
Dei paggi e dei servitori stanno aiutando i cavalieri a vestirsi. Altri stanno raccogliendo a terra i frammenti delle armi spezzate.

Fig. 6. Sample computation flow of internal data structures.

domain entity which is an instance of the type FRESCO. The body of the GSP contains a list of substructures in which the current GSP will be decomposed (ELEM items) and specifies their actual order, optionality and interconnecting rhetorical relations (SCHEMA item). In this case, the GSP tells the text planner that the description of a fresco for a novel user typically begins with the statement of its overall depiction (e.g., “The fresco in front of you depicts a tournament”) and proceeds with several elaborations on its various scenes.

At run time, the text planner chooses the best GSP that satisfies the current communicative goal and performs a decompositional expansion of the GSP calls in its body, ultimately producing a rhetorically annotated discourse tree.

The output of the text planning stage is a rhetorically annotated discourse tree whose leaves represent predicates in the domain model that must be linguistically realized, as shown in the left upper part of Fig. 6 which represents a sample computation flow in terms of internal data structures.

3.3. Microplanning

Microplanning traditionally covers a number of areas in the middle of the NLG pipeline, such as sentence planning, referring expression generation, and revision. The minimum requirement for simple systems is the sentence planner, which is responsible for mapping individual elements of semantic knowledge to grammatical roles. But for text generation systems that need to generate multiple paragraphs of text, the addition of pipeline elements that improve particular linguistic features has been shown to increase the quality of text generation to the higher level required by more natural text [14].

We use the STORYBOOK system [15] for the low-level production of text, which already includes architectural modules typically considered to be part of microplanning: discourse history, pronominalization, sentence planning, revision and limited lexical choice. An important element missing from this list is that of *referring expression generation*, or converting fact-based representations in a knowledge base into concrete descriptions of objects in a particular text. The creation of referring expressions occurs immediately after the discourse plan has been created.

Fig. 6 shows a particular sentential element that represents a leaf node of a discourse plan (helping001) describing a region of an artwork. The leaf node instructs the microplanner that (1) in the KB, there is information about a helping event in the artwork that should be described at this stage of the video documentary, (2) that the actors, objects and relations describing this helping event should be assembled into one or more *referring expressions* and included into the text, and (3) a *narrative stream* should be constructed from this information so that the low-level NLG system can produce the actual text corresponding to this local part of the discourse plan.

Given a concept from the KB to expand in greater detail, the referring expression algorithm must decide which facts would be interesting to the reader or listener, whether they have been mentioned before, and whether to include them in the present description or else wait for a future chance. In contrast to traditional NLG work on referring expressions that tries to create *distinguishing expressions* (i.e., phrases that allow the reader to pick out a particular object from a set of similar objects [44]), we focus on *narrative* or *descriptive expressions*, which are instead concerned with evoking particular images in the reader's mind.

Once the KB-based content has been completely determined, the second major task of the referring expression component is to determine potential grammatical properties for each concrete reference (or *discourse references*) which are related to the linguistic environment and not stored directly in the KB. Thus the microplanning representation for the helping001 event in Fig. 6 must be converted into the narrative stream representation in order to generate correct text. There are four main types of discourse references we have identified in our domain, which are influenced by the fact that the text will be coupled with a visual feedback of what is currently described:

- *Visually-available*: Objects that we can assume the user is currently looking at (i.e., in their field of view). Thus if we know the user is looking at a house, we can say "the door" instead of "a door".
- *Mentally-available*: Objects which haven't been yet mentioned, but which are semantically very closely related to an already mentioned concept. For example, having just described a castle allows us to say "the curtain walls" instead of "curtain walls".
- *Unique*: Objects marked in the KB as being exclusively one value of a finite set. In this case, we can refer to "the top side of the painting".
- *Generic*: Used when making an example out of a generic entity. For instance, "A fresco is painted on a plaster wall while it is still wet".

As said above, the implementation of a referring expression generation stage was required to complement Storybook functionalities. However, we envisage its utility also in a scenario where more shallow generation techniques are adopted for producing the video commentary. The dynamic generation of expressions, in fact, is beneficial also in the assembling and adjusting of pre-existing texts [38], to improve anaphoric and exophoric cohesion.

3.4. Low-level NLG with the STORYBOOK *generator*

STORYBOOK [15] is a low-level natural language generation system intended to function as the tactical component for a variety of strategic front ends, such as text planners, intelligent tutoring systems, dialogue systems, and narrative planners. To this end, it defined a common input formalism called the Narrative Stream, which functions like an API in programming interfaces: a standardized format for generating a wide range of sentences. STORYBOOK integrated a number of additional modules for linguistic processing that had previously been treated in isolation (Fig. 3):

- *Discourse History*: Determines linguistic concepts like givenness.
- *Pronominalization*: Decides when repetitious words should be converted to pronominal or deictic forms.
- *Lexical Choice*: Selecting one of many possible linguistic forms.
- *Sentence Planning*: Creates linguistic structures from purely semantic structures in the KB by assigning them particular grammatical roles.
- *Revision*: Aggregates multiple small sentences.

In addition, STORYBOOK employs the FUF/SURGE surface realizer for English [25] for converting the final linguistic structures directly into readable text. SURGE is one of the most comprehensive generation grammars available, and is a functional unification grammar with a heterogeneous approach: it combines the choice mechanism of a systemic grammar with rules from lexicalist grammars such as HPSG and the Text-Meaning Theory.

3.5. Deep generation of multilingual texts

Multilingual generation is becoming an important aspect of implemented systems that showcases the abilities of deep generation systems, especially in domains such as museum and cultural heritage information presentations where readers can potentially be of any nationality. Multilinguality via deep linguistic representations [6,7,41,46] aims at producing texts in a wide array of domains and genres without the complete recreation of linguistic resources required by shallow multilingual systems.

Different modules in the NLG pipeline have different requirements for multilingualism. In general, higher modules have fewer multilingual constraints and more reusable resources, while lower-level modules are almost completely language dependent. Thus a text planner will need minimal changes, while a pronominalizer may need different rules (e.g., Italian allows zero pronouns while English doesn’t [37], as in example #4 below), a revision component will need different clause aggregation rules [13], and a surface realizer will need to have its grammar and morphology components replaced.

The first important multilingual tasks in our low-level generation system are marking nominal concepts for definiteness and pronominalizing them. While the algorithm for performing these tasks in English has been described in [12], there is no implemented pronominalizer for Italian. We thus created a corresponding set of Italian pronominalization rules, after consulting prior linguistic work [23,37]. As an example, when given the following two sentences:

1. The pages and servants are helping the knights to dress.
   I paggi e i servitori stanno aiutando i cavalieri a vestirsi.
   lit. The pages and the servants stay helping the knights to dress-themselves.
2. The pages and servants are picking up the pieces of broken weapons.
   I paggi e i servitori stanno raccogliendo i pezzi delle armi rotte.
   lit. The pages and the servants stay picking-up the pieces of-the weapons broken.

Given a discourse context, the system would modify the sentences by introducing indefinite determiners and pronouns (or their elided versions) to produce:

3. ϕ Pages and ϕ servants are helping the knights to dress.
   ***Dei** paggi e **dei** servitori stanno aiutando i cavalieri a vestirsi.*
   lit. Some pages and some servants stay helping the knights to dress-themselves.
4. **They** are picking up the pieces of broken weapons.
   ϕ *Stanno raccogliendo i pezzi delle armi rotte.*
   lit. Stay picking-up the pieces of-the weapons broken.

At this point, the sentence planner has produced a functional description for every sentence as specified by the text planner. The revision system then opportunistically looks for ways to combine these short sentences into longer, more natural ones. Thus our two example sentences might be revised by a combination of disjunction with the deletion of a repetitive subject:

5. Pages and servants are helping the knights to dress or are picking up the pieces of broken weapons.
   Dei paggi e dei servitori stanno aiutando i cavalieri a vestirsi, o stanno raccogliendo i pezzi delle armi rotte.

The final stage of text production, surface realization, produces the actual text that is used in the video documentary (Fig. 7). The surface realizer [39] receives the revised functional descriptions and uses language-specific grammars and morphology components to check that a sentence is validly constructed, put the words in the right order, and ensure that morphology such as inflections for number and gender are added to individual words.

For multilingual generation, the surface realizer must be capable of handling language-specific rules. Several differences between English and Italian are evident in our example:

```
<segment id="01" parent="root" topic="tournament" audio="castle.mp3" 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" topic="castle" audio="windows.mp3"
         duration="2"} />
    The differences between the various styles of construction have been reproduced
    extremely carefully.
  </segment>
```

Fig. 7. Sample segments sent to the video player.

the presence of the indefinite plural article “dei” vs. the null determiner for English, the infinitive clause introduced by the closed-class preposition “a” instead of “to” as required by the verb “aiutare”, and the realization of the present progressive with the auxiliary “stare”, instead of the equivalent of “be” (“essere”).

The end result of this process over each sentence is a text which can be used as a script for a film documentary, which can then be sent to text-to-speech synthesis to produce an audio track for the film, or shown as subtitles. Currently, we use Festival for Italian speech synthesis [19] that produces Italian with its own specific pronunciation and prosody [20], and the Festival Lite system [8] to synthesize English. The text is sent to the speech synthesizer and converted to sound files which are then given to the Video Player for synchronization with the video produced with the overall rhetorical structure generated by the text planner.

4. Automatic cinematography

The generation of Video Documentaries encompasses several decisions that range from the selection of appropriate images to illustrate a domain’s elements to the choice of camera movements, and from the selection of transition effects to the synchronization of audio and video tracks. 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.

The most similar documentaries available that we are aware of are stored on the *National Geographic* web site. These videos are a sort of ‘trip report’: pictures taken from professional photographers are placed in sequence with music and audio commentaries, to describe experiences like being in a war or watching animals. Unlike such videos, the documentaries we want to generate are meant for learning purposes. Ideally they should be a kind of compendium of a museum visit. In particular, they are meant to describe frescos, which are much richer in detail than real world pictures.

In this case, it does not seem appropriate to resort to corpus-based techniques to elicit the rules of cinematography. We thus decided instead to follow a ‘principled’, knowledge based approach by exploiting knowledge, one of the most well-known techniques in AI. We conducted a set of informal interviews with a professional documentary maker, who worked on many documentaries for the Italian national television. Our main goal was to identify principles, heuristics, and guidelines on which we could implement our system. There are also many books where the basic elements of cinematography are presented and exemplified. But a documentary is a very specific genre, with well known aims (be simple, be clear, don’t confuse the audience, be salient), and we thought that an expert could also help us find out tips or tricks to follow such aims.

Sometimes documentary makers start from images and then try to write the text, i.e., our audio commentary. Also, if any modification is needed (e.g., for synchronization purposes), they tend to change the text since they cannot re-shoot scenes. Our case is different, in that the starting point is the text, and consequently, the audio script. The video planner task is thus the generation of a video, trying to fit the audio script, with the risk that the worst result is a slide show, i.e., a sequence of pictures shown one at a time. To avoid such a result we thought to exploit the rhythm of the audio/textual script. We collected a small corpus of texts which illustrate a single renaissance fresco. We conducted an analysis of the text at the discourse level and then devised a fairly simple model of discourse rhythm, based on the rhetorical relations holding between text spans. We noticed that:

- `elaboration` and `sequence` relations tend to connect and keep together spans. A block made up of combinations of such relations usually illustrates semantically related or spatially connected topics.
- `background` and `circumstance` relations tend to break the rhythm of the discourse. They signal a sort of pause, a topic shift, or the introduction of a different type of information about the same topic.

Given such a model we implemented a set of rules trying to fit such heuristics in order to insert transition effects (e.g., a fade, see below) when background or circumstance relations occur; another set of rules has been implemented in order to avoid the insertion of transitions when elaborations or sequences are matched. From interviews with experts we identified a set of tips for transition effects and constraints over the combinations of camera movements. For example, during interviews, we have found that:

- A display indicates continuity.
- Fade effects tend to underline a ‘passage’, or a change.
- The camera can not be moved back and forth along the same path.
- A pan communicates the spatiality of a scene and tends to focus the attention on the arrival point and its adjacent space.
- A zoom in helps to focus on specific areas or details which are not usually noticeable.
- A zoom out helps to refamiliarize the viewer with the details and their position in a bigger framework.

We finally coupled our insights about discourse structure with the knowledge extracted from our interviews, giving rise to the detailed rules described in the following sections and to the final system presented in Section 5.

4.1. The terminology of cinematography

To enable the description of our approach and to present the system with its underlying mechanisms we need to introduce some technical terms.

Shots 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 a 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*. During the shot the camera can be fixed or moving on wheels. When fixed it can sweep horizontally or pivot up and down; if attached to a moving vehicle it can move along the x, y, and z axes. In this work we refer only to movements of cameras of the wheeled type, using standard film terminology. Thus the film term *move* is interpreted as movements along the x- and y-axis, whereas the technical term *dolly*, also called a zoom, denotes a movement along the z-axis.

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 next shot 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.

4.2. Heuristics and constraints of cinematography

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. Movements along the x and y axis can be used to reveal spatial relations among objects and to move the viewer's attention from one center of interest to another; a dolly in 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 dolly in towards that aspect can be chosen, unless this choice violates some constraint.

Each move, in fact, has to be "consistent" with respect to the previous one. The viewer, looking at a movie in which the camera moves to one side and then to the opposite one, might experience difficulty following the stream of the presentation. For example, if the previous move is a change of position towards the right, the following effect cannot be a movement towards the left either along the same path or along similar paths. In general, when a camera movement is chosen it imposes constraints on the choice of subsequent movements.

Another important feature of a film 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. This is similar to the ideas of stream-like flow of text found in narrative representations [15]. 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. For example, if the audio commentary is illustrating the object O_1 and the next segment will provide more details about that, we can continue to shoot O_1, or dolly in towards it, but the creation of a new shot would not be recommended.

The creation of a new shot involves the choice of a transition effect at the editing phase, and a transition, though brief, might underline a passage towards a new type of information. Conversely, sometimes we have to impose cut points, that is positions in the discourse where a new shot has to be created. This is appropriate when there is a change of topic, or when the presentation is going to provide a different kind of information about the same topic. For example, if the audio commentary has illustrated the main features of a fresco and is going to provide more details about the painting techniques, a transition might help

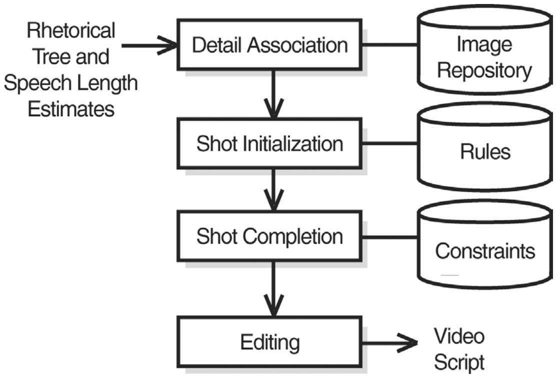


Fig. 8. The Video Planner architecture.

the viewer to realize that the discourse has broken off, and different information is being introduced.

The Video Planner implements a mechanism to automatically compute a full script for a video documentary, starting from a discourse plan annotated according to an RST annotation scheme generated by the Text Planner. In addition, the video planner also takes as input a repository of annotated images, to which camera movements and transitions are applied. Rules and constraints are the core on which the system relies. They encode the rhetorical strategies that are the basic resource for:

(1) selecting appropriate images;
(2) designing the presentation structure;
(3) completing each shot;
(4) synchronizing the visual part with the audio commentary and avoiding the “seasickness” effect (back and forth motion).

The rules, fired by a forward chaining mechanism, are context sensitive, and govern: (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.

4.3. The Video Planner engine architecture

The engine is structured as in Fig. 8. In our application scenario, when a museum visitor requests a video for a given commentary, the engine analyzes the discourse structure of the commentary and selects an appropriate set of images to be presented.[5] The production chain consists of four phases:

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

[5] At the moment, the request is location-based, and is triggered by an infrared sensors system.

```
<movie id="january">
  <shots>
   <shot id="shot603" image="det01">
     <video-track>
       <pause duration="2"/>
     </video-track>
     <audio-track>
       <play audio="january.mp3"/>
     </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.mp3"/>
       <audio-pause duration="1"/>
       <play audio="castle.mp3"/>
     </audio-track>
   </shot>
  </shots>
  <editing>
    <display shot="shot603"/>
    <crossfade shot="shot605" duration="1"/>
  </editing>
</movie>
```

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

```
(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))))
```

Fig. 9. Examples of a script, a shot initialization rule, and a constraint.

Shot initialization and structure planning: a candidate structure for the final presentation is elaborated, taking into consideration the rhetorical structure of the commentary (the result of this phase can be changed, as its processing is iterative);

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

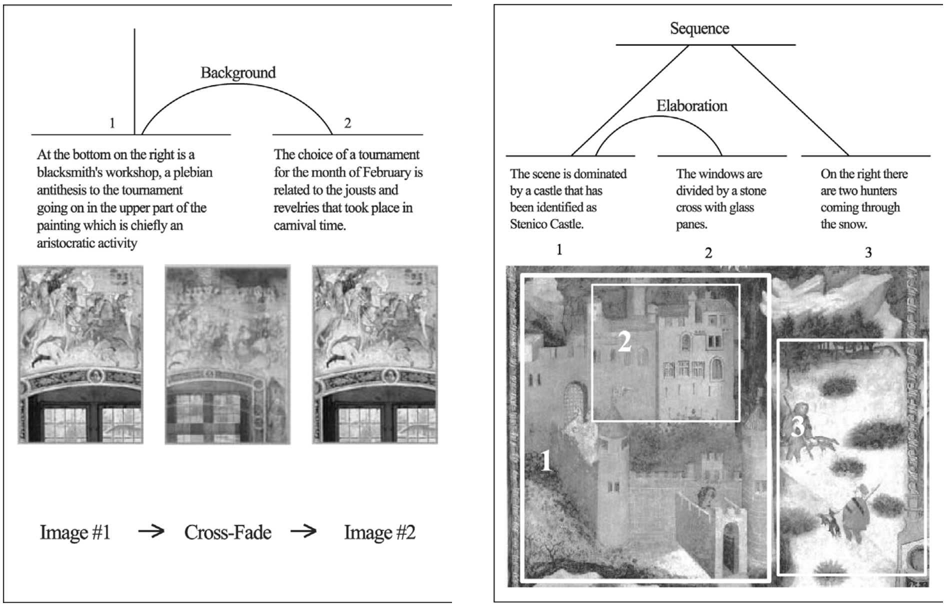
Editing: transitions among effects are selected according to the rhetorical structure of the commentary.

The output is a complete script for the video and audio channels encoded in a renderer-independent markup language (see the left side of Fig. 9).

The generation process is rule-based. An example of a rule is given at the top right of Fig. 9. 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. Camera constraints are conditions that forbid particular combinations of camera movements and are tested according to the type of proposed movement and the sequence of past movements. An example of a constraint is shown in the bottom right of Fig. 9. Potentially each camera movement can lead to an inconsistent sequence. Thus to select a zoom-in movement it is important to know if the previous or penultimate move is a zoom-out; if not, then a zoom-in can be selected.

4.4. 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



At the bottom on the right is a blacksmith's workshop, a plebian 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.

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.

(a) (b)

Fig. 10. The (a) “Tournament” and (b) “Castle” examples.

topic as the segment. The preferred heuristic is to select the detail with exactly the same topic(s) of the segment, since this allows a straightforward identification of the topic. In case this is not possible, we have implemented search rules that look for details which subsume the topics mentioned in the commentary. Referring to Fig. 10, suppose we have a segment which illustrates both the castle (2) and the hunting scene (3). If the system does not find a detail illustrating both, it will choose the whole picture, because that illustrates both topics. To perform this task the Video Planner accesses a subset of the knowledge base used by the other components of the system.

In case the heuristics above do not succeed, the rules for this task are defined in such a way that allow also the choice of images illustrating entities related to the topic. Like a sort of metonymy, the system can choose a picture of the White House if the topic is President Bush even if the repository does not contain any picture of him.

4.5. Phase 2: Shot initialization

In the second phase, shots are initialized taking into consideration the rhetorical structure of the commentary. At the moment, we do not take into account the nucleus/satellite distinction of RST. The result of phase 2 is a candidate structure for the final presentation. The processing is guided by a set of rules that are fired when particular configurations of rhetorical relations are matched (see the top right of Fig. 9).

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 relations tend to highlight the introduction of new information that provides a context in which either the previous or subsequent messages can be interpreted. In addition, they tend to break the flow of the discourse. It is thus preferable to split the segments into 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 4.8).

4.6. Phase 3: Shot completion

This is the phase in which the engine incrementally completes each shot by expanding and 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 that it is suitable according to the list of previous camera movements and the constraints imposed over that category of movement. These 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 pan, the system must 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.

4.7. 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 cuts applies.

These rules have been selected according to the observations about the usual employment of transition effects in the field of cinematography [3]. Fade effects are fit for smooth transition, when there is a topic shift or when the center 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.

4.8. Examples

The first example concerns the rhythm of the discourse (Fig. 10(a)). 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.

The second example concerns the application of constraints in order to avoid an inconsistent sequence of camera movements (Fig. 10(b)). The text first describes the castle in the left side of the image. In this case the system, after a brief pause over 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 the right. However, this combination is forbidden by the constraint on a dolly-out that follows a dolly-in. Thus the engine revises the structure of the movie, declares the current shot completed, initializes a new shot and associates the remaining segments with it.

5. Implementation

GLAMOUR is an end-to-end automatic cinematography system capable of organizing collections of annotated images and facts from a knowledge base to produce documentary films about two minutes long each. GLAMOUR is implemented in Java (Text Planning), Lisp (Text Generation and Video Planning), C (Festival Lite TTS) and Macromedia Flash (Video Player) and runs on a 1 GHz Pentium III. Presentations can be played back on either a desktop or a PDA device. The system uses deep NLG techniques to produce both the text of the documentary and the abstract movie plan, and then combines them to create a script which can be shown by the Video Player.

A typical documentary starts with a fade-in to the most representative high-level image available in the archive. This image is displayed while the spoken text of a single segment of the script is played. When the player estimates the speech has finished, a precomputed transition effect (zoom, fade, etc.) switches to the next image or subimage and the subsequent audio segment is initiated. When the series of images has been exhausted, the documentary returns to the original image and the system is ready to show a new film.

As an application domain we have chosen the *Cycle of the Months* of the Torre Aquila at the Buonconsiglio Castle in the city of Trento (Italy). This work is composed of eleven side-by-side frescos (each one measuring on average 2 meters wide and 3 meters high, and representing a particular calendar month) painted during the 1400 s and illustrates the activities of aristocrats and peasants throughout a full year. Our cinematography prototype constructs complete end-to-end film documentaries for the frescos of January, February, April and November, each of which is represented extensively in a hand-built knowledge base.[6]

As an example, the fresco representing the month of February contains three major scenes (1) a jousting tournament, (2) a castle with ladies looking over the walls at the tournament, and (3) a blacksmith's workshop. Each of these scenes can be further divided into subscenes. For instance, the region of the fresco containing the tournament consists of a set of knights doing battle and their associated pages and servants who are picking up broken weapons. Thus a film about the February fresco utilizes the annotated hierarchy where the entire image is the root, the main scenes are the middle level of the tree, and the individual subscenes represent the leaves. These images and subimages serve as the raw material for the Video Planner.

Fig. 11 shows an excerpt of the generated documentary film. First, the entire image is shown while the audio begins with an introductory phrase (a). Because the next topic concerns the women watching the tournament, a zoom is selected to move the scene to a close-up of that section of the image (b). Also, because the discourse is now focusing on a different region, the generation system selects a localization phrase which is played immediately after the shot transition: "At the top there are twenty young women." After several shot sequences, the topic changes from the court ladies to the knights fighting the actual battle (f). Because the discourse plan encodes this topic shift, the video planner can select an appropriate crossfade to the entire image followed by a zoom in to the tournament.

[6] A typical movie script in our application domain requires a knowledge base containing 800 concepts and 350 relations, and an indexed archive of around 5 images from which subimages may be extracted. Constructing this KB required around 4 person-months of effort. Given this infrastructure, the KB and image annotations for an additional 2 films required only about two weeks. The text for the movie script is generated in around 15 seconds for 2 minutes of speech. Such presentations are usually made of 3 or 4 shots on average where the transitions vary according to rhetorical relations.

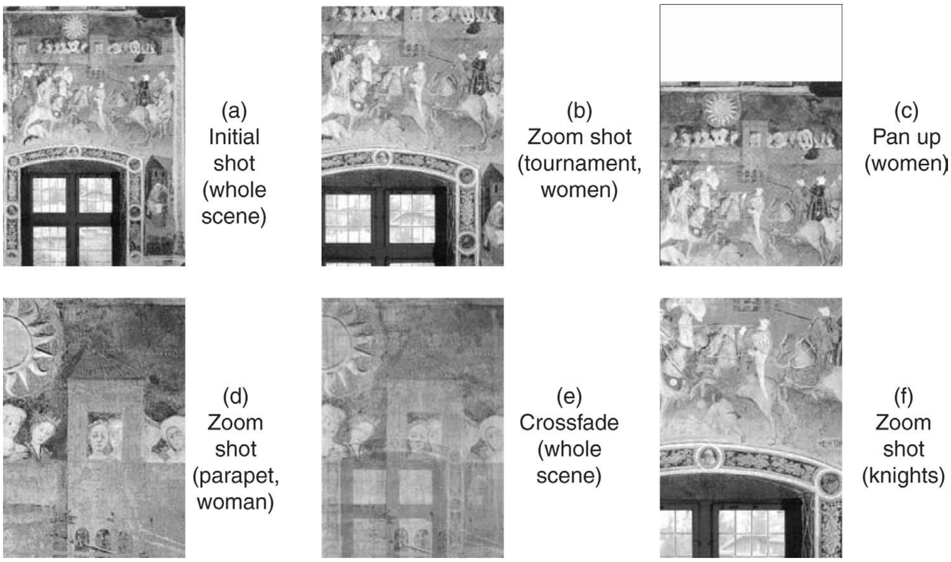

(a) Initial shot (whole scene)
(b) Zoom shot (tournament, women)
(c) Pan up (women)
(d) Zoom shot (parapet, woman)
(e) Crossfade (whole scene)
(f) Zoom shot (knights)

Fig. 11. An example film sequence in the Art Fresco Domain.

Table 1
Correlation of various factors with speech duration

	Characters	Facts	DiscElems	Facts + DiscElems
Average/Sec.	12.72	1.02	1.69	2.71
Adjusted Std. Dev.	1.20 sec	3.15 sec	2.88 sec	2.57 sec
Adjusted Max. Dev.	2.28 sec	5.12 sec	5.28 sec	4.68 sec

Given the level of architectural parallelism that GLAMOUR employs, it is important to be able to estimate parameters which may be influenced by other parallel modules. The chief example in this case is the need of the video planner to know how long each utterance may be when planning shot transitions. But since the actual text is not known until the generation module has completed, some method must be used to estimate within a reasonable error margin. Since the discourse plan is available to the video planner, we used discourse elements to estimate speech duration. Table 1 shows the standard and maximum deviations when predicting from four measures: the number of characters in the final sentence (our gold standard), the number of facts extracted from the knowledge base, the number of discourse element representations (Section 3.3), and the sum of the latter two items. Using this last score gives a reliable-enough estimate of speech duration for use in an interactive environment.

The final result of producing a common discourse plan, generating text/speech, and planning the video aspects of the documentary in parallel is a film documentary that contains the desired information for presentation to the user.

6. Evaluation

Subjective evaluations of generated output by computer algorithms (whether text, speech, or music) have long been considered to be difficult. Most evaluations, when they even have been performed, were focused more towards task effectiveness and other more objective measures of quality. Possible evaluations of GLAMOUR could have included architectural ablation to detect degradation in quality, direct comparison of language quality from its generated scripts, the range of variability that can be produced in its films, overall efficiency of the system, or even how long it would take to produce movies in languages which the system can't currently generate.

A large-scale experiment with a broad audience (i.e., test subjects) for judging the effectiveness of the produced material is not realistic at this point since it is unclear exactly what metrics are applicable, or would be useful, beyond viewer satisfaction. We thus opted instead for an expert-based evaluation of the concept and the prototype. In particular, we wanted to see if planning the cinematographic production by using the topics introduced by the verbal commentary and its rhetorical structure yielded results of higher quality than planning from the topics alone. Implicitly this leads also to an evaluation of the specific cinematic rules in our implementation, rather than the quality of text produced by the low-level generation component, which has been examined previously [15].

In order to evaluate GLAMOUR, we thus decided to keep the NLG contribution constant and vary the cinematic component of the overall short documentaries. We generated three pairs of videos for the individual panels representing the months of January, February, April and May of the Torre Aquila frescos. For each pair, the first video was generated using only information about the topics mentioned in the commentary without using its rhetorical structure. The cinematic rules adopted in this case are quite straightforward: the detail depicting the topic of the current verbal segment is brought to the center of the display following the shortest path and no transitions are planned (which we refer to as the default version). The second video of each pair was instead generated according to the more articulated cinematic rules described previously (the cinematic version).

In the system evaluation we involved three experts: a professional documentary director, a TV director with former experience in multimedia production, and a multimedia designer. The three experts were interviewed separately. Each evaluation session was organized as follows: the expert was first asked to watch all three pairs of videos in a random order. For each video, they were requested to assign an absolute score from 1 (terrible) to 10 (perfect). For each pair, again in random order, they were then requested to directly compare the two videos on a scale from 1 (preference for default version) to 10 (preference for full cinematic version). Finally, an unstructured interview closed the session.

Table 2 reports the absolute scores assigned to the default and cinematographic versions by the experts: the documentary director (Director), the TV director with experience in multimedia (MMDirector) and the multimedia expert (MMExpert). "Jan-C" represents the cinematographic version with full rhetorical structures for the month of January, while "May-D" represents the default, baseline version for the month of May.

Table 3 shows the figures for relative scores of the direct comparisons between versions of the video of each fresco individually (1 = preference for the default version, 10 = preference for cinematic).

Table 2
Absolute scores assigned to the videos by the experts (D = default, C = cinematic)

	Jan-C	Feb-C	May-C	Jan-D	Feb-D	May-D	Totals
Film Director	5	5	3	5	1	4	23
MMDirector	8	7	5	4	4	4	32
MMExpert	4	5	6	3	7	5	30
Totals	17	17	14	12	12	13	

Table 3
Direct comparisons between cinematic and default versions

	Jan	Feb	May	
Director	5	10	4	19
MMDirector	7	7	5	19
MMExpert	7	3	7	17
Total	19	20	16	55
Average	6.33	6.67	5.33	

The limited available number of expert subjects prevents a deep statistical investigation, yet some interesting issues did emerge from this study. The two directors agreed on scoring February better on the cinematic version; surprisingly, the multimedia expert scored February the lowest. The videos of May are somewhat worse than the others; this could be due to the characteristics of that individual fresco that made it particularly difficult to visualize.

The post-scoring interviews revealed that the three experts had different genres in mind. While all the three clearly recognize a difference on the two versions in terms of structure (in particular, they noticed the use of transitions), they didn't always agree with the stylistic choices.

The documentary director mostly appreciated the cinematic version for February because of a few non-trivial decisions in the camera movements; yet he also deplored the use of the fade transitions that he determined to be not adequate for the documentary genre. Upon further inquiry, he admitted that many directors used that kind of transition, but he personally dislikes them.

On the contrary, the TV director, although considering good fades very difficult to obtain, prized the use of transitions since believing that they liven up the narration. Both complained about the lack of music or ambient sounds, and recommended taking the soundtrack into consideration in the future. When asked, the documentary director said that he always plans the visual part of his documentaries starting from a pre-prepared commentary while the TV director claimed that he prefers to start planning the visual part and then asking the script writer to prepare the commentary taking into account the visual part.

The documentary director nicknamed the default version the “dog-to-dog” style: meaning that when the commentary mentions a dog, the visual part also displays a dog. Both directors agreed that although naïve, sometimes this style may be effective in documentaries. They claimed that naïve directors should resort to this style to avoid confusing the documentary with too many visual effects. In any case, they both recognize that “true” directors would never resort to such a simple style.

As might be expected, the multimedia expert had a different perspective. As can be seen in Table 3, his scores tended to be contrary to those of the directors. During the interview it emerged that, although he recognized the greater richness of the cinematic version, he found the “dog-to-dog” style closer to his MM experience. He pointed out the absence of effects such as blurring to emphasize details, although these are more typical of the MM tradition than of the cinematographic tradition. Like the two directors, he also complained about the lack of music and ambient sounds.

This study showed that expert evaluation in this domain might prove more useful than evaluation with naïve users. Yet the experts are often biased by their particular style, making numeric scores of little use. In fact, even in a very specific genre such as documentaries (and even if we have camera movements on fixed images) styles can vary quite substantially. Disagreement among the experts indicates that the field needs to find an appropriate metric before larger evaluations can be undertaken. Interviews are instead very powerful means to assess the results and help improve the system behavior. Concerning this specific evaluation, the minimum that can be said is that our system, on the basis of its rich structural input, can accommodate different levels of sophistication in the visual presentation. The current implementation of the rules yields a cinematic contribution that on average seems reasonably better than the default version.

7. Conclusions

We have presented a prototype system for dynamically producing video documentaries by combining research in automatic cinematography and natural language generation. The system takes initial resources such as annotated still 2D images and a knowledge base, and combines them with state-of-the-art pipelined multilingual NLG system and film planning techniques to produce two minute documentary films about renaissance frescos.

The system exploits the ability of RST to structure output in a meaningful way, whether for the high-level text plan or the organization of cinematography. Additionally, a deeply represented NLG system allows the video commentary to be produced in either English or Italian directly from the underlying knowledge representation.

7.1. Questions ahead

This work was initiated with the purpose of integrating a language-based dynamic guide for cultural tourism with video documentaries produced on demand. For this concrete purpose, as well as for allowing the cinematography system to proceed without concern for interference, we adopted an open structure that allows flexibility and modularity while guaranteeing efficiency. The language-based planner produces RST-based data, typically used as input for subsequent language generation processes, which together with the specification of the various semantic entities involved constitutes the input for the cinematographic part. The production of the documentary proceeds without additional information exchanges between the two sides until after speech synthesis.

The initial planner makes its decisions on the basis of whatever external information it may have at its disposal. Thus in the case of the visitor guide, all contextual information may be relevant for influencing the eventual result of the filmmaking process: location, focus of attention of the visitor, material he/she has been previously presented with, history of the visit, the user model, etc.

With the goal of adding flexibility to the system, possible solutions could include (1) let contextual information directly influence the behavior of the dynamic video producer or (2) introduce a blackboard architecture.

Contextual information, such as user preferences, may induce the system to opt for a certain style of video presentation, or for the best suited cinematographic *genre*. Information about specific choices made at some level of the language generation process could be useful for determining video realizations. As an example, suppose the microplanning component decides to use a figurative expression such as “to flaunt oneself like a peacock”. In the Eisenstein vein [24], the video producer may then decide to reflect this in its realization by actually showing a peacock spreading its tail. Similarly, a video producer decision, for instance to opt for a slow pan over a painting that includes a scene of courting, may influence the decision of the generator to use evocative wording even if the courting theme was not at the focus of its realization.

The case of referring expressions, and the related control by the system of the—analogical—movements of the (possibly virtual) camera on the scene, is one obvious situation that requires finer multimodal synergy. When is redundancy, especially in deictic expressions, good and when is it to be avoided? When, in general, does a scene *take the place* of words? What is the best balance and rhythm between words and silence, in the development of a certain visual presentation? These are some of the questions that are in front of us, which could eventually move us from documentary styles to true narrative.

7.2. Applications

As indicated earlier, the application for guidance and illustration purposes in the context of visits to a museum has a realization on both mobile and stationary devices. The motivation for dynamic film production integrated with language presentation in our case lies mainly in the need for flexibility. There are many situations where applications of this kind will become useful. We can think in particular of the educational domain and of all manner of instructional systems where a specific user’s state and needs will be addressed by the personalized multimodal dynamic presentation system.

Additionally, the domain of personalized news can benefit from advancements in this area. Starting from freshly released images identified by accompanying comments, a system can produce a summary or an integration of news. Several projects are currently being developed on this theme on the language side [35]; according to the view we present here, we can also plan the integration of adapted video material.

Finally, the presented work does not consider one early phase of filmmaking: the actual shooting. With current advances in robot technology it becomes realistic to conceive of robots that have the task of shooting in the real world, according to specific rules of operation that combine optical flexibility, camera movements from a fixed position, and movements of the camera position. At that point, in a circumscribed context (e.g., where at least some subjects are easily recognizable by an intelligent vision system), the whole process of filmmaking could become automated. One example is the production of a short video of a stereotypical event such as a marriage. More generally, the field of multimodal report production opens up many prospects, as costs of all kinds of devices go down and the need for recording material and accessing it in a personalized, synthetic and effective way grows constantly in our society.

Acknowledgements

This work was developed within the PEACH project,[7] funded by the Autonomous Province of Trento.

References

[1] E. André, The generation of multimedia documents, in: A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text, Marcel Dekker, New York, 2000, pp. 305–327.

[2] I. Androutsopoulos, J. Calder, E. Not, F. Pianesi, M. Roussou, MPIRO: Coherence, relevance and appropriateness in information presentation, in: Proceedings of the 2002 CLASS Workshop, Verona, Italy, 2002.

[3] D. Arijon, Grammar of the Film Language, Hastings House, 1976.

[4] W. Bares, J. Grégoire, J. Lester, Realtime constraint-based cinematography for complex interactive 3D worlds, in: Proceedings of the Tenth National Conference on Innovative Applications of Artificial Intelligence, Wisconsin, MA, 1998, pp. 1101–1106.

[5] W.H. Bares, S. Thainimit, S. Mcdermott, A model for constraint-based camera planning, in: Proceedings of the AAAI Spring Symposium on Smart Graphics, Stanford, CA, 2000, pp. 84–91.

[6] J. Bateman, S. Sharoff, Multilingual grammars and multilingual lexicons for multilingual text generation, in: ECAI Workshop on Multilinguality in the Lexicon-II, Brighton, UK, 1998, pp. 1–8.

[7] J. Bateman, E. Teich, I. Kruijff-Korbayová, G.J. Kruijff, S. Sharoff, H. Skoumalová, Resources for multilingual text generation in three Slavic languages, in: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC), Athens, Greece, 2000, pp. 1763–1768.

[8] A. Black, K. Lenzo, Flite: A small fast run-time synthesis engine, in: 4th Speech Synthesis Workshop (ICSA), Scotland, 2002.

[9] A. Butz, Anymation with CATHI, in: Proceedings of the Ninth Innovative Applications of Artificial Intelligence Conference, 1997, pp. 957–962.

[10] L. Cahill, R. Evans, C. Mellish, D. Paiva, M. Reape, Towards a reference architecture for natural language generation systems, Technical Report, Information Technology Research Institute, Brighton, UK, August, 2000.

[11] J. Calder, A. Melengoglou, C. Callaway, E. Not, I. Androutsopoulos, C. Spyropoulos, G. Xydas, G. Kouroupetroglou, M. Roussou, Multilingual personalized information objects, 2004, submitted for publication.

[12] C.B. Callaway, Pronominalization in discourse and dialogue, in: Proceedings of the 40th Meeting of the Association for Computational Linguistics, Philadelphia, PA, 2002, pp. 88–95.

[13] C.B. Callaway, Multilingual revision, in: Proceedings of the 9th European Workshop on NLG, Budapest, Hungary, 2003.

[7] For more information see: http://peach.itc.it.

[14] C.B. Callaway, J.C. Lester, Evaluating the effects of natural language generation techniques on reader satisfaction, in: Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society, Edinburgh, Scotland, 2001, pp. 164–169.

[15] C.B. Callaway, J.C. Lester, Narrative prose generation, Artificial Intelligence 139 (2) (2002) 213–252.

[16] C.B. Callaway, B.H. Daniel, J.C. Lester, Multilingual natural language generation for 3D learning environments, in: Proceedings of the 1999 Argentine Symposium on Artificial Intelligence, Buenos Aires, Argentina, 1999, pp. 177–190.

[17] C. Callaway, T. Kuflik, E. Not, A. Novello, O. Stock, M. Zancanaro, Personal reporting of a museum visit as an entrypoint to future cultural experience, in: Proceedings of IUI-05, 2005.

[18] D.B. Christianson, S.E. Anderson, Li We He, D. Salesin, D.S. Weld, M.F. Cohen, Declarative camera control for automatic cinematography, in: Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, Portland, OR, 1996, pp. 148–155.

[19] P. Cosi, F. Tesser, R. Gretter, A. Cinzia, Festival speaks Italian, in: Proceedings of EuroSpeech 2001, Aalborg, Denmark, 2001.

[20] P. Cosi, A. Cinzia, F. Tesser, R. Gretter, F. Pianesi, A modified “PaIntE” model for Italian TTS, in: Proceedings of IEEE Workshop on Speech Synthesis, Santa Monica, CA, 2001.

[21] R. Dale, S. Geldof, J.-P. Prost, Generating more natural route descriptions, in: Proceedings of the 2002 Australasian Natural Language Processing Workshop, Canberra, Australia, 2002, pp. 41–48.

[22] B. Daniel, W. Bares, C. Callaway, J. Lester, Student-sensitive multimodal explanation generation for 3D learning environments, in: Proceedings of the Sixteenth International Conference on Artificial Intelligence, Orlando, FL, 1999, pp. 114–120.

[23] B. Di Eugenio, Centering theory and the Italian pronominal system, in: Proceedings of the 13th International Conference on Computational Linguistics (COLING-90), Helsinki, Finland, 1990, pp. 270–275.

[24] S.M. Eisenstein, S. Eisenstein, The Film Sense, Harvest Books, 1969.

[25] M. Elhadad, FUF: The universal unifier user manual version 5.0, Technical Report #CUCS-038-91, Dept. of Computer Science, Columbia University, 1991.

[26] D.A. Friedman, Y.A. Feldman, Knowledge-based cinematography and its applications, in: Proceedings of ECAI, Valencia, Spain, 2004, pp. 256–262.

[27] M.A.K. Halliday, R. Hasan, Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective, Deakin University Press, 1985.

[28] N. Halper, P. Oliver, CamPlan: A camera planning agent, in: Proceedings of the AAAI Spring Symposium Workshop on Smart Graphics, Stanford, CA, 2000.

[29] J. Hitzeman, A. Black, C. Mellish, J. Oberlander, M. Poesio, P. Taylor, An annotation scheme for concept-to-speech synthesis, in: Proceedings of the European Workshop on Natural Language Generation, Toulouse, France, 1999, pp. 59–66.

[30] J.D. Hollan, E.L. Hutchins, L.M. Weitzman, STEAMER: An interactive, inspectable, simulation-based training system, in: G. Kearsley (Ed.), Artificial Intelligence and Instruction: Applications and Methods, Addison-Wesley, Reading, MA, 1987, pp. 113–134.

[31] P. Karp, S. Feiner, ‘Automated presentation planning of animation using task decomposition with heuristic reasoning, in: Proceedings of Graphics Interface, 1993, pp. 118–127.

[32] W. Mann, S. Thompson, Rhetorical structure theory: A theory of text organization, Technical Report, USC/Information Sciences Institute, Marina del Rey, CA, ISI/RS-87-190, 1987.

[33] M.T. Maybury, Intelligent Multimedia Interfaces, AAAI Press, Menlo Park, CA, 1993.

[34] K.R. McKeown, Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text, Cambridge University Press, Cambridge, 1985.

[35] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, C. Sable, B. Schiffman, S. Sigelman, Tracking and summarizing news on a daily basis with Columbia’s NewsBlaster, in: Proceedings of the 2002 Human Language Technology Conference, San Diego, CA, 2002.

[36] C. Metz, Film Language: A Semiotics of the Cinema, Oxford University Press, Oxford, 1974.

[37] F. Namer, Subject erasing and pronominalization in Italian text generation, in: Proceedings of the Fourth Conference of the European ACL, Manchester, UK, 1989, pp. 2025–2032.

[38] E. Not, M. Zancanaro, Building adaptive information presentations from existing information repositories, in: Proceedings of the International Workshop on Information Presentation and Multimodal Dialogue,Verona, Italy, 2001.

[39] A. Novello, C. Callaway, Porting to an Italian surface realizer: A case study, in: Proceedings of the 9th European Workshop on NLG, Budapest, Hungary, 2003.

[40] M. O'Donnell, RSTTool 2.4—A markup tool for rhetorical structure theory, in: Proceedings of the International Natural Language Generation Conference, Mitzpe Ramon, Israel, 2000, pp. 253–256.

[41] C. Paris, K. Vander Linden, M. Fischer, A. Hartley, L. Pemberton, R. Power, D. Scott, A support tool for writing multilingual instructions, in: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montréal, Canada, 1995, pp. 1398–1404.

[42] E. Reiter, Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?, in: Proceedings of the Seventh International Workshop on Natural Language Generation, Kennebunkport, ME, 1994, pp. 163–170.

[43] E. Reiter, NLG vs. templates, in: Proceedings of the Fifth European Workshop on Natural Language Generation, Leiden, The Netherlands, 1995.

[44] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, 2000.

[45] C. Rocchi, O. Stock, M. Zancanaro, M. Kruppa, A. Krüger, The museum visit: Generating seamless personalized presentations on multiple devices, in: Proceedings of the 2004 Conference on Intelligent User Interfaces, Madeira Island, Portugal, 2004, pp. 316–318.

[46] D.R. Scott, The multilingual generation game: Authoring fluent texts in unfamiliar languages, in: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999.

[47] O. Stock, Language-based interfaces and their application for cultural tourism, AI Magazine 22 (1) (2001) 85–97.

[48] W. Wahlster, E. André, W. Finkler, H.-J. Profitlich, T. Rist, Plan-based integration of natural language and graphics generation, Artificial Intelligence 63 (1993) 387–427.

[49] W. Wenders, On Film, Faber and Faber, London, 2001.