CREWS Report 99-06

Interrelating Goal Models and Multimedia Scenes: An Empirical Investigation

by

Peter Haumer, Jürgen Rack, and Klaus Pohl

Lehrstuhl Informatik V, RWTH Aachen
Ahornstraße 55, 52056 Aachen, Germany
+49 (0)241 80 21 501

{haumer,rack,pohl}@informatik.rwth-aachen.de

Interrelating Goal Models and Multimedia Scenes: An Empirical Investigation

Peter Haumer, Jürgen Rack, Klaus Pohl

Lehrstuhl Informatik V, RWTH Aachen
Ahornstraße 55, 52056 Aachen, Germany
{Haumer,Rack,Pohl}@Informatik.RWTH-Aachen.DE

Abstract. *Conceptual goal models are used to express intentional aspects of the system under development. Among others, goal models facilitate stakeholder discussions and agreement about main system aspects during early requirements engineering phases. As experiences from participatory design indicates, the use of multimedia representations (especially videos) leads to better stakeholder involvement and, as a consequence, the produced conceptual (goal) models and specifications respectively are of higher quality.*

In this paper, we report on our empirical investigation which shows that the use of associations between goals and video parts documenting goal achievements and goal failures improve the performance of typical requirements engineering tasks. More precisely, they lead to more correct and complete results.

1 Introduction

Multimedia representations of current system usage (real world scenes), are widely used in video-supported participatory/user-centred design techniques [3], [6], [9], [13], ethnography [4], [8] and workplace culture [5] for understanding existing systems and working environments, but also for the envisionment of future systems [10] and different system changes [12]. Experiences indicate the use of video leads to a better stakeholder involvement and thus to a better understanding of the usage domain, enforces focused observation of (temporally and/or spatially) distributed aspects, enables repeatability of results and late reflections. As a consequence, the produced conceptual models and specifications, respectively, are of higher quality.

Particularly in early RE phases, goal models are more and more frequently used [1],[2],[11],[14]. In contrast to more solution-oriented models such as behavioural or structural models, goal models represent intentional knowledge about the system. A desired system often has to attain many of the functional and non-functional goals of existing systems. Many of those goals can be elicited by observing, documenting and analysing current system usage, e.g., documented in videos.

Despite the advantages of using videos, participatory design techniques do not support a tightly integrated management of videos and conceptual models. For instance, observations of current system usage captured on videos are used to elicit and validate abstract concepts (e.g. [9]), but the influence of the videos on the definition of those concepts is not documented. In other words, important inherent relationships between video scenes and concepts are not persistently recorded and therefore are likely to be easily forgotten. To overcome this shortcoming, we developed the

PRIME-CREWS[1] environment, which, among others, supports the requirements engineer to record these interrelations [7].

In this paper, we describe the results of an empirical investigation, which shows that fine-granular interrelations between goal model components and video excerpts lead to a better performance of requirements engineering tasks. We examined this by performing controlled experiments with two groups. One group had to use unrelated videos and conceptual models to perform typical requirements engineering tasks (Group A, Fig. 1). The other group (Group B, Fig. 1) had the additional opportunity to make use of the fine-grained interrelations between the conceptual model components and the parts of the video scenes. The results of both groups were compared according to the completeness and correctness achieved.

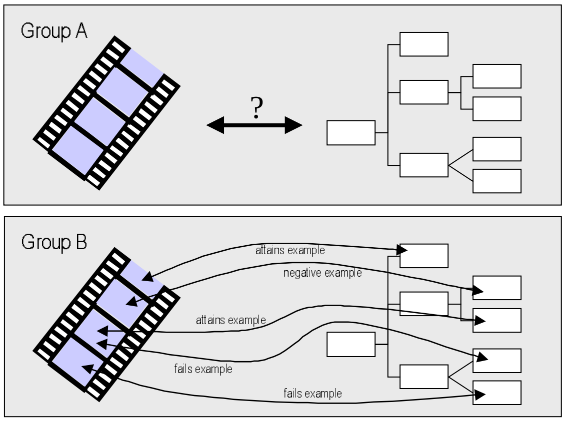

Fig. 1: Information given to Group A and Group B for performing the experiment.

The remainder of the paper is structured as follows. After presenting the hypotheses and explaining the nature of RE task examined (Sect. 2), we will present the experimental design (Sect. 3) followed by an overview of how we conducted the experiments and the obtained results (Sect. 4). Finally, we interpret the results, summarise and discuss our findings (Sect. 5).

2 Theory for Evaluation Experiments

2.1 Focus of the Evaluation

The focus of the study was to determine the difference in completeness and correctness of the results of requirements engineering tasks achieved by Group A and B (Fig. 1) with respect to the following three hypotheses:

$H1_1$: *System aspects influenced by a goal can be determined more correctly and completely by Group B than A.*

$H1_0$: *System aspects influenced by a goal cannot be determined more correctly and completely by Group B than A.*

$H2_1$: *Goal attainment in the current system can be determined and described more correctly and completely by Group B than A.*

$H2_0$: *Goal attainment in the current system cannot be determined and described more cor-*

---

[1] PRIME: PRocess Integrated Modelling Envionment; CREWS: Cooperative Requirements Engineering with Scenarios

rectly and completely by Group B than A.

H3$_1$: *Goal refinements in respect to their AND/OR relationships can be validated more correctly by Group B than A.*

H3$_0$: *Goal refinements in respect to their AND/OR relationships cannot be validated more correctly by Group B than A.*

We tested this in a controlled experiment which simulated the situation that a stakeholder just joined an ongoing project and had to get familiar with the current specification consisting of a goal model and a set of videos which document typical usage situations of the actual system. To provide the stakeholders with an overview of the actual specification, we performed an introductory meeting in which the case study's domain was introduced and all videos were shown. The stakeholders were asked to make notes of the introductory meeting to be used during the experiment later on. During the experiment, they were asked to perform a set of requirements engineering tasks. For Group A, the goal model and videos were completely unrelated (cf. top of Fig. 1). Thus, they had to rely on their notes and memory of the introductory meeting. For Group B, the goals were interrelated with video parts relevant for the goal (cf. bottom of Fig. 1). The interrelations were typed to express the nature of the relation between the video part and the goal (attainment and failure). The typed interrelations were created during the creation/modification of the model.

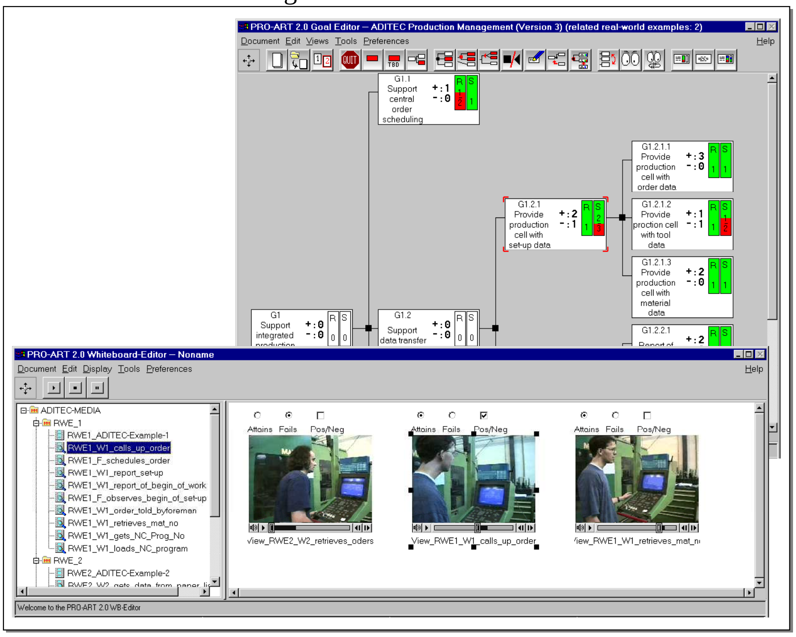

Fig. 2: Retrieving real world example fragments for a specific goal concept.

2.2 Tool Support

Fig. 2 depicts the tools used for the experiment (cf. [7] for a detailed description of the whole tool environment). The upper window shows our goal editor presenting a scrollable goal hierarchy and the lower window represents the multimedia editor which allows to (dis)play, edit and manage different types of media especially video.

For Group A, the goal hierarchy (without the coloured annotations of the goals, see below) and the unrelated video scenes were accessible and browseable within these tools. For Group B, goals were annotated with numbers expressing how many interrelations exist for a goal and which groups of interrelation types they represent (see [7] for details). The related parts of the videos were accessible directly from the goal editor. For instance, in Fig. 2 one can see that Goal G1.2.1 has been selected in the goal hierarchy and the related video parts are displayed on the multimedia editor with the types of the interrelation as textual comments. In addition, the physical positions of the video parts within the video are displayed in the lower slider bars. Thus, test persons of Group B had direct access to the relevant parts of the videos, but could also watch the whole videos if desired.

3 Designing the Experiments

In this section, we discuss the design of the controlled experiments we performed to evaluate our hypotheses. After presenting the definitions for the dependent variables of the experiment (Sect. 3.1), we discuss major influence factors to be considered for the design (Sect. 3.2), the rationale for the experiments tasks (Sect. 3.3) followed by the description of qualitative questionnaires used (Sect. 3.4) and finally an outline for the course of the experiments (Sect. 3.5).

3.1 Dependent Variables

We defined correctness and completeness rates as the dependent variables for the hypotheses. For each of the hypotheses, we delineated a task to be solved that consisted of a set of subtasks expressed as questions. For each subject let C_j be the number of correct answers and W_j the number of wrong answers for Task *j*. We defined the main dependent variables as follows:

– The ***completeness rate*** indicated the ratio C_j / *(# correct answers*$_j$*)* between given and correct answers.
– The ***correctness rate*** $C_j / (C_j+W_j)$ expressed how much of the given answers were correct.

Since, we could not consider the completeness rate in the 3rd task (cf. Sect. 3.3 for explanation), we considered five dependant variables (correctness rates for hypotheses H1 – H3 and completeness rate for hypotheses H1 – H2).

3.2 Influence Factors

In this section, we discuss major influence factors for the dependent variables. The most relevant independent variable was ***group membership.*** As outlined in Sect. 2.2 the two groups were provided with different tool support. Thus, as assumed by the hypotheses, being a member of a particular group implied the major influence on the results. To avoid correlation between the subjects' capabilities and their group membership, we randomly allocated the subjects to the two groups that were of the same size. The decision to confront only individuals with the goal model had the advantage that side effects due to group dynamics could be eliminated. In addition to group membership, we identified several factors that had an influence on the results of the experiment. In the following, we describe these factors and how we handled them in the experiment's design.

3.2.1 Goals and Interrelated Video Parts

To ensure validity for Requirements Engineering of the experiments and to show scalability of the approach under evaluation, we used observation material and goal models gained from a case study we performed in advance to the experiments at a machine manufacturing company in Aachen (Germany) named ADITEC. This company uses loosely integrated production planning, management and control systems, which we analysed in respect to specify improvements of the current integration. After performing several small trail applications of our real world scene based elicitation technique (on which we reported in [7]), this case study was much bigger in scale and provided a realistic context for the experiments. Consequently, the video material used for the experiments was captured in close cooperation with the ADITEC personnel and the used goal models were elicited by external RE experts from Ericsson Eurolab and ORACLE Germany.

Videos: The content of the video had an influence on the elicitation of the goal models and the interrelations. If we would have recorded arbitrary usage situations, we could have influenced the experiment in respect to our desired results. However, the actual scenes used for the experiment were determined by the case study, which was performed in cooperation with the ADITEC personnel.

Goal model and interrelation structure: As mentioned above, the goal model and the interrelation structure were elicited by industrial experts based on the captured video material which eliminated the possibility of us influencing the goal model and the interrelations towards the desired outcome of the experiments. However, the complexity and structure of the goal model and the interrelation structure as well as understandability of the textual goal descriptions had effects on the results, which we tried to control with modelling guidelines for the experts who created them, e.g. by using goal verbs and sentence structures which already were used successfully in other projects [2].

3.2.2 Qualification of the Subjects

Mother tongue: The language used in the videos, the goal model, and the questionnaires were German. Subjects having problems to understand the language would have influenced the results. Therefore, we checked if German was their mother tongue. If not, the subject had to rate his language capabilities on an ordinal scale (reaching from 1 to 7). Subjects who judged their language capabilities to be bad (1-4) were excluded from the experiment. If they judged their capabilities to be good (although German was a foreign language), they participated in the experiment. Even in this case, we compared the results with the results of native speakers. If the results differed significantly, the results of the non-native speaker were not considered in the evaluation.

Background knowledge: The students were recruited from a database lecture teaching conceptual modelling. We expected the population to consist mainly of computer science students in their fourth or fifth year of study. Additional specific knowledge in computer science was not necessary (besides usage experience with computers, cf. below).

Usability problems: Basic experience in using a graphical user interface such as MS-Windows had to be ensured. This was important, because people not being familiar with buttons, sliders etc. would have needed significantly more time for the task (which was restricted, cf. Sect. 3.5). To cope with usability problems concerning our tools, we assigned one attendant to each subject. The attendants were instructed only to provide help using the tools and not for solving the tasks to avoid any external influence.

Motivation: To increase the motivation of the subjects, we provided attractive prizes for the best results. Those prizes were awarded independent to each group so that being member of a certain group was no advantage.

3.2.3 Modelling Practice

General skills: Being students of a database lecture, we expected all subjects to have basic skills in conceptual modelling. We excluded subjects that had additional practical experience in the experiment's domain (e.g. being a programmer in a similar manufacturing company), because they would have achieved incomparable better results.

Goal modelling skills: To ensure that all subjects had the same education in conceptual goal modelling, we conducted an extra lecture on goal modelling one week before the experiments.

3.2.4 Tasks

Timeframe for each task: The time available to solve the tasks had a big influence. Since in practice time always seems to run short, we designed the tasks to include time pressure. Testing our experiment's design performing trial experiments with different subjects, we adjusted the number of questions belonging to each task, such that answering all questions took slightly more time than available[2]. Additionally, we kept records of the used time within the experiment for the design of future experiments.

Mutual influence of the tasks: Since we wanted to consider the hypotheses separately, we had to ensure that performing one task had no influence for the solution of another task except for a training effect while working with the model which we could not prevent. Therefore, we designed each task in a way that the questions dealt with different goals and interrelated video parts.

3.2.5 Working Conditions

Used hard- and software: Each subject worked with an individual computer. To guarantee that the computer's performance had no impact on the results, we used computers with identical hard- and software.

Mutual influence of subjects: We could not provide the possibility for all subjects to work in a separate room. In cases when they had to share rooms, they were seated in a way that they did not have the possibility to see the other subject's screen and results. In addition, we provided head sets for each computer.

3.3 Design of the Tasks

3.3.1 General Structure of Task

For each of the three tasks the subjects had 20 minutes time. To ensure time pressure, as discussed in Sect. 3.2.4, each task was adjusted to an individual number of subtasks. Thus, the first task was comprised of three subtasks, the second task of two subtasks, and the third task of seven subtasks.

Task 1 and 2 dealt with the identification of systems and solutions in respect to identify influencing scenes for goals and cases of solution for goals in the systems. (For instance, one question belonging to the first task asked for all systems producing print-jobs in relation to a goal concerning printing capabilities for the system.) The subjects were asked to write down their answers giving a small explanation to exclude

---

[2] The available timeframe did not include the time to read the tasks first and write down remarks at the end of the respective task.

correct answers we did not think about in the design. For each of these questions, we determined the set of correct answers. Combining the answers of each task each subject achieved a number C_j and W_j of correct and wrong answers respectively.

Task 3 three dealt with the identification of defects. Since the goal model we used in the experiment was too complex (153 goals) to ask general questions, such as “Where in the goal model can you identify wrong AND-OR-refinements?”, we decided to conduct this part of the experiment as a multiple choice test. The subjects had to assess if certain refinements within the goal model were correct or not. Due to the multiple-choice character of these questions, we assumed that all subjects answered all seven questions. Therefore, we did not compute a completeness rate for Task 3.

3.3.2 Statistical Evaluation

We used an unpaired t-test to compare the empirical mean of the results of both groups. To accept our hypotheses we demanded a probability p not exceeding 0.05.

3.3.3 Task-Related Qualitative Information

To be able to explain irregularities in the results, we captured additional information that was not used for the quantitative analysis. This information comprised of remarks by the subjects and the attendants. For each task, the subject was asked, if he had any problem to understand the questions. The attendant's job was (besides helping in case of usability problems) to observe and write down the time spent on questions and which videos were used.

3.4 Qualitative Evaluation of Experiment

To capture an overall impression on how the subjects judged the experiment and the usefulness of the tested approach for Group B, they were asked to fill out a questionnaire immediately after the experiment. This form included several statements which had to be rated on an ordinal scale reaching from 1 to 7, where 1 stood for “complete rejection of the statement” and 7 for “complete agreement”.

The statements included issues like

- support for the understanding of abstract goals by the videos (generally, without interrelation)
- support of the access to the videos by the interrelation structure
- possible extensions
- declaration of how much they used the video and their notes to solve the tasks
- etc. (see the whole list in Sect. 4.4)

To obtain reliable results some facts were covered by more than one statement (e.g., different statements addressing the usefulness of videos). We analysed resulting rates with statistic methods such as empirical standard deviation and empirical mean of the distribution.

3.5 Outline of the Evaluation Study

Summarising, we depict the outline of the evaluation study in Fig. 3.

In a preparation phase, we collected the video material in a case study performed in cooperation with the ADITEC personnel. We then used this material to elicit the goal models and the interrelations to video parts in an expert session.

In an introductory meeting, we showed the videos to all group members, because for members of Group A it would have been very difficult to find appropriate scenes within the videos in a comparable manner. We therefore decided to exclude aspects of the short-term memory to present all necessary parts of the videos to all subjects, one week before performing the actual experiment and the subjects were asked to make notes which they were allowed to use in the experiment.

In the actual experiment, each subject got a form containing the task descriptions and space for the answers. Once a task has been declared as finished by the subject, (s)he was not allowed to jump back to it to be able to determine the actual time used. Adherence to time and order of tasks was checked by attendants.

Preparation	
	Interviews and video capture at ADITEC company
	Expert sessions eliciting goal models and video interrelations
Introductory meeting	
	Introduction in goal modelling (cf. Section 3.2.3)
	Introduction into the ADITEC domain (cf. Section 3.2.3)
	Presentation of all important parts of the videos (cf. Section 3)
Experiment (one week later)	
	Fill in questionnaires concerning personal data and previous knowledge
	Brief refresh of the knowledge about the ADITEC domain
	Explanation of all tasks to avoid understanding problems
	Explanation how to use the tools
	Performing task 1
	Performing task 2
	Performing task 3
	General assessment using a questionnaire (cf. Section 3.5)

Fig. 3: Resulting Experiement.

4 Conducting the Experiments

In cooperation with the ADITEC staff we recorded four videos (duration: 19 min, 12 min, 48 min and 26 min) in November 1998. The goal model elicited by experts from Ericsson and ORACLE based on videos comprised of 153 goals, 117 interrelations allocated to 89 video parts (not for all goals interrelations were created; some video parts were related to several goals).

The introductory meeting took place on December 7th, 1998. To ensure, that all test persons participating in the workshop had seen the videos before, we used an attendance list. This was very important, because differences concerning the familiarity with the videos probably would have caused a strong impact on the results. The experiment itself took place one week later.

4.1 Results of the Tasks

Tab. 1 presents the measured correctness and completeness rates separately for each task. The subject identifier includes information about group (A or B) the subject was assigned to. The subjects of each group were ranked for each task. Then, to obtain an overall experimental ranking, we calculated the arithmetic mean of these three ranking values, which led to the overall ranking depicted in Tab. 1.

Comments on results of Task 1: Subjects of Group B found more correct answers, although nobody found all the correct ones. All subjects (of both groups) gave at least one wrong answer. Members of Group B achieved a better ratio of correct answers.

Comments on results of Task 2: Members of Group B found noticeable more of the correct answers than members of Group A. Members of Group B made not one single mistake. Three of Group A’s members gave only correct answers. Remarkably, subject A-07 did achieve no correct answer (although he had worked on the task and found some wrong answers). He assessed his motivation to be medium and stated that he had no problems to understand the tasks, but also said (in contrast to most other participants) that the video did not support him.

Comments on results of Task 3: Due to the multiple-choice character of the questions of the third task, only the correctness rate was considered. All subjects answered all questions belonging to this task.

Tab. 1: Results of the Tasks

		Task 1		Task 2		Task 3	
	Subject	Correctness Rate	Completeness Rate	Correctness Rate	Completeness Rate	Correctness Rate	Overall Ranking
Group A	A-01	0,375	0,3	0,429	0,33	0,57	7
	A-02	0,714	0,5	0,375	0,33	0,29	6
	A-03	0,833	0,5	1,000	0,56	0,86	1
	A-04	0,667	0,2	0,286	0,22	0,29	8
	A-05	0,800	0,4	1,000	0,22	0,71	3
	A-06	0,833	0,5	0,500	0,33	0,86	2
	A-07	0,500	0,5	0,000	0,00	0,86	5
	A-08	0,286	0,2	1,000	0,44	0,71	4
Group B	B-01	0,800	0,8	1,000	0,89	0,57	7
	B-02	0,875	0,7	1,000	1,00	0,86	3
	B-03	0,889	0,8	1,000	0,89	0,43	4
	B-04	0,889	0,8	1,000	0,89	0,57	5
	B-05	0,818	0,9	1,000	0,89	0,57	6
	B-06	0,857	0,6	1,000	0,56	0,86	8
	B-07	0,900	0,9	1,000	0,89	0,86	2
	B-08	0,900	0,9	1,000	1,00	0,71	1

4.2 Qualification of Subjects

Two people whose mother tongue was not German participated in the experiment. Both of them rated their capability to understand and speak German with 5 using an ordinal scale from 1 to 7 (1 for weak, 7 for excellent). B-08 achieved good results and the attendant confirmed that the subject had no problems with the language. On the contrary, A-04 achieved poor results (ranking 8). After the experiment the subject added in his remarks, that he probably had overestimated his language capabilities, which was also confirmed by the attendant. Thus, we decided to exclude A-04 completely from the evaluation.

No subject had any influencing experiences such as domain knowledge etc.

4.3 Evaluation of Results

Tab. 2 contains the empirical mean and empirical standard deviation for the five dependent variables, calculated separately for each group.

Tab. 2: statistical evaluation of the results

		group A	group B
Correctness Rate for Task 1	empirical mean	0,620	0,866
	empirical deviation	0,230	0,038
Completeness Rate for Task 1	empirical mean	0,414	0,800
	empirical deviation	0,122	0,107
Correctness Rate for Task 2	empirical mean	0,615	1,000
	empirical deviation	0,393	0,000
Completeness Rate for Task 2	empirical mean	0,318	0,875
	empirical deviation	0,175	0,139
Correctness Rate for Task 3	empirical mean	0,694	0,679
	empirical deviation	0,209	0,167

As explained in Sect. 3.3.2, we used an unpaired t-test to compare the empirical mean of the dependant variables of group A and B. This yields to the following results:

- Group B achieved a significantly higher correctness rate concerning Task 1 ($p = 0{,}03215$)
- Group B achieved a significantly higher completeness rate concerning Task 1 ($p = 0{,}00251$)
- Group B achieved a significantly higher correctness rate concerning Task 2 ($p = 0{,}0411$)
- Group B achieved a significantly higher completeness rate concerning Task 2 ($p = 0{,}00002$)
- The analysis failed to reveal a significant difference between the correctness rate of both groups concerning Task 3

4.4 Evaluation of the Questionnaire

The results of the questionnaire the subjects filled out at the end of the experiment are depicted in Tab. 3. In addition to the rating values, we show the empirical mean as well as the empirical standard deviation calculated separately for each group. We covered the fifth and seventh issue with a second semantically equivalent statement and combined the ratings to strengthen the exactness of the results.

Tab. 3 : Results of the Questionnaire

	Question	A-01 B-01	A-02 B-02	A-03 B-03	A-05 B-04	A-06 B-05	A-07 B-06	A-08 B-07	 B-08	mean (A) mean (B)	deviation (A) deviation (B)
1	I mainly used my notes of the introduct.	2	1	1	3	1	2	1		1,57	0,79
	meeting to perform the tasks.	2	1	1	1	3	1	2	1	1,50	0,76
2	I mainly used my memory of the	4	4	3	5	7	6	2		4,43	1,72
	introduct. meeting to perform the tasks.	3	3	4	3	3	2	6	3	3,38	1,19
3	The introduction and information available	5	2	4	6	7	5	3		4,57	1,72
	was sufficient to perform the tasks.	6	5	7	6	6	5	7	6	6,00	0,76
4	Videos have been a support to understand	6	2	2	2	7	2	3		3,43	2,15
	goals and answer questions.	5	6	6	5	5	5	7	5	5,50	0,76
5	Textual descriptions are sufficient to	3,5	3,5	3,5	3	1,5	5	2,5		3,21	1,07
	understand goals.	4,5	1,5	1,5	2,5	3,5	2,5	2,5	2,5	2,63	0,99
6	Multimedia is appropriate support	6	7	5	6	7	6	4		5,86	1,07
	to explain abstract facts.	6	7	7	6	6	6	7	4	6,13	0,99
7	Interrelations improve access	6,5	6,5	7	--	7	6,5	5,5		6,50	0,55
	to the videos.	7	5,5	7	6	6	6,5	6	5,5	6,19	0,59
8	It is necessary to extend the approach	7	6	3	6	4	7	6		5,57	1,51
	with annotated interrelations.	5	7	5	6	5	4	6	2	5,00	1,51

In the following, we briefly discuss the answers to the questions of Tab. 3:

Question 1: Neither Group A nor Group B used the notes of the introductory meeting (empirical mean: 1,57 and 1,50). The empirical standard deviation of both group is nearly the same (0,79 and 0,76).

Question 2: The mean of Group A is about 1 higher than the mean of Group B (4,43 vs. 3,38). The deviation of both groups is quite high (1,72 and 1,19). No influence from the subject's ranking on the rating of this statement can be detected.

Question 3: Subjects of Group B assessed the information available as sufficient (mean: 6,00, deviation: 0,76). In Group A the empirical mean was lower (mean: 4,57), but the rating varied strongly (deviation: 1,72). No influence from the subject's ranking can be detected.

Question 4: Members of Group B judged video to be helpful for understanding goals and answering questions (mean: 5,50, deviation: 0,76). In Group A only the subjects A-01 (ranking: 7) and A-06 (ranking: 2) rated the statement high.

Question 5: There is no significant difference between the ratings of Group A and B (mean: 3,21 and 2,63). The empirical deviation of both groups is about 1 (1,07 and 0,99).

Question 6: Most of the members of both groups judged that multimedia is appropriate to explain abstract facts (mean: 5,86 and 6,13). Only one subject of each group rated this statement with an average value.

Question 7: There was agreement (deviation: 0,55 and 0,59), that the interrelations improves the access to videos (mean: 6,50 and 6,19). Since Group A could not use interrelations during the experiment, one subject (A-05) gave no answer.

Question 8: The empirical mean of both groups is nearly the same (5,57 and 5,00). The standard deviation of both groups is high (1,51 in both groups). The subjects (A-03 and B-08) rating the necessity to be small, achieved rank 1 in their group.

5 Conclusions

As the main result of the experiment, the null hypotheses $H1_0$ and $H2_0$ (for each two sub-hypotheses: completeness and correctness) could be rejected at 0.05 level. We therefore can accept the hypotheses H1 and H2.

We were not able to show Hypothesis H3. The main reason for this we concluded not the hypothesis to be wrong, but a lack in the design of the experiment (surely, we might be wrong). To test H3 we defined binary (yes/no) questions and thus the subjects were able to guess the answers with a probability of 0.5. That the subjects have guessed some answers is confirmed by the attendants who reported that both groups did not make much use of the video in this task. We therefore conclude that to test hypothesis H3, it would be better to let subjects review a subpart of the overall goal model instead.

Task 2 showed the most unequivocal results. We interpreted this in the way that it generally requires less interpretation effort to find a solution for a goal than determining the mutual influences of goal and systems (Task 1).

The fact that most of the members of Group A did not manage to answer all questions of Task 1 and Task 2 indicate that they would have needed more time. Additionally, they made more mistakes than Group B, even if we only consider questions, which have been answered by both groups.

Notes of the introductory meeting were rarely used. Member's of Group A relied more on their memories than Group B. This is quite natural, since the members of Group A had more problems to find the video parts containing relevant information.

One reason for the high deviation of Group A in the Tasks 1 and 2 could probably be the differences in short-term memory among the test persons to solve their tasks. (This would also explain why there have been some quite good results of some individuals in Group A.) Group B had a smaller deviation, which clearly shows that our approach improved their task performance and that personal skill factors such as their short-term memory were less needed for answering the questions.

In the final comments, there was agreement that multimedia provides an excellent means to explain abstract facts (Question 6) and that textual descriptions are often not sufficient. Nevertheless, test persons from both Groups suggested to provide additional textual annotations for the videos and goals (Question 8, Group A) and the interrelations (Question 8, Group B, notable exception: the winner of Group B apparently seemed to be satisfied with the existing tool support).

In addition, the following conclusions can be drawn from the experiment:

– *The interrelations helped in recalling knowledge from memory*. An important observation was that during the experiment, subjects of Group B did not watch the whole video parts presented for a goal but just listened to the first few sentences. Typically, they watched a video part only for a short time, which was enough to recall knowledge acquired during the introductory meeting or during earlier tasks of the experiment. As one of the subjects pointed out after the experiment "The video parts were of great help to recall what I have seen earlier" (acquired in the introductory meeting or during the experiment).
– *Need for a video structure*. The initial strategy for Group A subjects was typically to first try to do a binary search in the videos to locate scenes dealing with a certain goal. After having worked on several goals some (the attendants reported of at least two) of the subjects' changed to a more goal centred strategy. First, they spent more time contemplating the goal's position in the overall hierarchy (i.e. trying to understand the super- and sub-goals) to get more insights of the goal itself and then used this information to recall video scenes they already had seen in this context restricting their search to scenes relevant for the actual task. Thus, they inherently tried to set up a similar association structures in their minds as it was explicitly available to Group B.
– *Unstructured videos cause frustration*. At the beginning of the experiment, all subjects made much use of the videos. After a while (most of them within Task 2), four of subjects of Group A tried less and less to find appropriate scenes. This is reflected by their final remarks which state that the videos were not of much help to solve the tasks (also cf. Sect. 4.4 Question 4). One could therefore assume that videos used for requirements elicitation are only useful in ongoing development stages, as the one we tested in the experiment, if an appropriated access structure is offered.
– *Results applicable for additional RE situations*. The results of the experiment show that stakeholders joining an ongoing project can perform certain requirements engineering tasks better (more complete and correct) if they are provided

with fine-granular interrelations between goals and video parts of recorded system usage. It is quite likely that other system development activities with similar tasks as the tested ones such as formal reviews, system maintenance or the integration of change will also benefit by our approach. Moreover, we do not expect different results if the content of the video is extended towards future system usage situations produced by applying role play, virtual reality techniques etc. as well as in cases when the associated video fragments document relevant group discussions or other important background information.

In addition to the results of the experiments, the sessions with the industrial experts provided valuable insights for our real world scene-based elicitation technique [7] for goal elicitation and semi-automatic interrelation of video parts, which is supported by the PRIME-CREWS environment. The sessions led to a much more diverse goal models than expected and showed us new ways of using the environment. For instance, we could extend our guidelines for goal identification to support the detection of non-functional goals in addition to functional ones.

To improve the empirical evidence, we will perform further experiments with different and refined tasks for hypothesis H1 and H2. For H3 we want to learn from our mistakes and create a new design for the tasks in which areas of the goal model have to be examined for defects. In general, we want to show in future experiments the advantages of our approach for validation of models, e.g. in formal reviews. We also learnt from individual talks after the experiment that a more elaborate qualitative analysis of the experiment, e.g. in extra discussion sessions, would provide valuable insights. Another objective for the refined experiments is that we want perform it with different types of test persons, e.g., real software practitioners and typical domain experts such as mechanical engineers, to examine differences in using the videos concerning their background knowledge.

Acknowledgements. This work was funded by the European Community under ESPRIT Reactive Long Term Research 21.903 CREWS (Cooperative Requirements Engineering with Scenarios, http://Sunsite.Informatik.RWTH-Aachen.DE/CREWS/). The authors would like to thank Lutz Bauer, Stephan Jacobs, the ADITEC team, especially Silvia Schlaphof and Franz-Josef Stepprath, as well as all test persons and attendants.

6 References

[1] A.I. Antón, "Goal-based Requirements Analysis", Proc. Int'l Conf. Requirements Eng. (ICRE'96), IEEE Computer Soc. Press, Colorado Springs, Colorado, USA, pp. 136- 144.

[2] A.I. Antón and C. Potts, "The Use of Goals to Surface Requirements for Evolving Systems", Proc. of Int'l. Conf. on Software Eng. (ICSE'98), Kyoto, Japan, 19-25 April 1998.

[3] H. Beyer and K. Holtzblatt, *Contextual Design: Defining Customer-Centered Systems.* Morgan Kaufmann Publishers, 1997.

[4] J. Blomberg, "Ethnographic Field Methods and their Relationshipto Design", in *Participatory Design: Principles and Practices.* D. Schuler, A. Namioka (eds.), Lawrence Earlbaum Associates Inc., Hillsdale, New Jersey, 1993, pp. 123-155.

[5] K. Bødker and J.S. Pedersen, "Workplace Cultures: Looking at Artifacts, symbols and Practices", in *Design at Work: Cooperative Design of Computer Systems.* J. Greenbaum, M. Kyng (eds.), Lawrence Earlbaum Associates Inc., Hillsdale, New Jersey, 1991, pp. 121-136.

[6] F. Brun-Cottan and P. Wall, "Using Video to Re-present the User", *Communications of the ACM*, Vol. 38, No. 5, Mai 1995, pp. 61-71.

[7] P. Haumer, K. Pohl, and K. Weidenhaupt. "Requirements Elicitation and Validation with Real World Scenes", in *IEEE Transactions on Software Engineering*, Vol. 24, No. 12, Special Issue on Scenario Management, 1998, pp. 1036-1054.

[8] J. Hughes, J. O'Brien, T. Rodden, M. Rouncefield, I. Sommerville. "Presenting Ethnography in the Requirements Process.", *Proc. 2nd IEEE Int'l. Symp. on Requirements Eng.*, York, England, IEEE Computer Society Press, 1995, pp. 27-34.

[9] K. McGraw and K. Harbison, *User-Centered Requirements: The Scenario-Based Engineering Process*. Lawrence Erlbaum Associates, Inc., Publishers, Mahwah, New Jersey, 1997.

[10] N. Millard, P. Lynch, and K. Tracey. "Child's Play: Using Techniques Developed to Elicit Requirements from Children with Adults", *Proc. of 3rd Int'l. Conf. on Requirements Eng. (ICRE '98)*, Colorado Springs, USA, April 6-10, 1998.

[11] J. Mylopoulos, L. Chung, E. Yu, "From Object-Oriented to Goal-Oriented Requirements Analysis", Communications of the ACM, Vol. 24, No.1, Jan. 1999, pp. 31-37.

[12] A.-W. Scheer, S. Leinenbach, and M. Luzius, "Nutzenpotentiale neuer Medien im Geschäftsprozeßmanagement: Virtual Reality und Multimedia zur Visualisierung und Simulation betrieblicher Abläufe", in: Karrenbauer, R.; Lauer, T.; Weisgerber, D. (Hrsg.): 3. *SaarLorLux Multimedia-Kongreß 1997*, Shaker Verlag, Aachen 1997, pp. 43 - 48 (in German).

[13] L.A. Suchman and R.H. Trigg, "Understanding Practice: Video as a Medium for Reflection and Design", *Design at Work: Cooperative Design of Computer Systems*, J. Greenbaum and M. Kyng, eds., Lawrence Erlbaum Associates, Inc. Publishers, New Jersey, 1991, pp 65-89.

[14] E. Yu and J. Mylopoulos, "Why Goal-Oriented Requirements Engineering", *4th Int'l. Workshop on Requirements Eng.: Foundation for Software Quality (RESFQ'98)*, Pisa, Italy, Jun., 1998, pp. 15-22.