Choosing Your Moment: Interruptions in Multimedia Annotation

Christopher P. Bowers, Will Byrne, Benjamin R. Cowan, Chris Creed, Robert J. Hendley, and Russell Beale

School of Computer Science, University of Birmingham
Birmingham, B15 2TT, UK
{C.P.Bowers,W.F.Byrne,B.R.Cowan,R.J.Hendley,
R.Beale}@cs.bham.ac.uk

Abstract. In a cooperative mixed-initiative system, timely and effective dialogue between the system and user is important to ensure that both sides work towards producing the most effective results, and this is affected by how disruptive any interruptions are as the user completes their primary task. A disruptive interaction means the user may become irritated with the system, or might take longer to deal with the interruption and provide information that the system needs to continue. Disruption is influenced both by the nature of the interaction and when it takes place in the context of the user's progress through their main task. We describe an experiment based on a prototype cooperative video annotation system designed to explore the impact of interruptions, in the form of questions posed by the system that the user must address. Our findings demonstrate a preference towards questions presented in context with the content of the video, rather than at the natural opportunities presented by transitions in the video. This differs from previous research which concentrates on interruptions in the form of notifications.

Keywords: Task interruption, multimedia annotation, mixed-initiative annotation.

1 Introduction

Good annotation of multimedia content enables effective search and allows it to be retrieved and consumed again, desirable for anything from archive news footage to YouTube videos. However, the annotation process is difficult and time-consuming, making it an ideal application for a cooperative mixed-initiative system. Here a computer collaborates with the user on a common goal and can initiate and effect dialogue with the user when necessary. The computer and the user may work asynchronously [6]. A good mixed-initiative system makes the best use of the user's time and input and this is reflected in how it carries out interactions with the user. When input is required the system might assess the current state of the overall task, the task(s) the user is currently performing, the importance of the input, and how the user responded to previous interactions, in order to help it decide on the best way to interact with the user.

P. Campos et al. (Eds.): INTERACT 2011, Part II, LNCS 6947, pp. 438–453, 2011.

A mixed-initiative multimedia annotation system might be performing its own analysis of the contents of a document in order to support the user by suggesting annotations to the user or applying its own annotation. Dialogue with the user might take the form of queries designed to help the system with its analysis, perhaps by resolving conflicts between different interpretations of some parts of the content. Responses to these queries will influence the main annotation task. The system as a whole will perform better if queries are answered by the user promptly and accurately, so an effective dialogue should take into account the best times to present particular queries to the user.

However, system-driven interactions can also have a negative effect if they disrupt [12, 17] or irritate [1, 16, 19] the user during their primary task. So as well as a benefit (user input), a query will also have a cost [12, 2, 9, 10] which might be detrimental to the overall outcome of the task.

2 Previous Work

Interruptions in human-computer interaction have received much attention from researchers. Findings show that interrupting users from their primary task can have a negative impact on performance [12, 8, 17], and produce higher levels of frustration, stress and anxiety [19, 1, 16]. Other studies have found that interruptions can also have a detrimental effect on how long it takes to complete a task [14, 7], decision making ability [18], emotional state [19, 1], and the number of errors users make during a task [12].

Researchers have been investigating whether interrupting users at "opportune" moments might reduce some of these negative effects. This approach is largely guided by Miyata and Norman [15] who argued that interruptions should be made at times of lower mental workload. They also argued that these periods typically occur at subtask boundaries of task execution. Bailey and Iqbal [5] also found that interruptions can have lower cost if made during periods of lower mental workload. Pupil dilation (a recognised measure of workload) was measured whilst performing a number of different tasks (route planning, document editing, and email classification) and it was largely confirmed that signs of decreased workload occurred at subtask boundaries. This work suggests that, if possible, interruptions should be deferred to a breakpoint in the task to reduce costs and negative effects.

Iqbal and Bailey [11] also found that presenting notifications at break points in problem solving tasks (i.e. a programming task and diagram editing) reduced frustration and resumption time of the task when compared with presenting the notifications immediately. It was also found that the relevance of the notification content influenced at what time the notification should be presented. In particular, it was found that notifications relevant to the user's current work should be presented at medium or fine breakpoints (i.e. during lower level activities such as editing code or adding shapes). Conversely, it was found that less relevant or generic notifications should be presented at coarser breakpoints (i.e. higher level events such as the user switching to their mail or instant messaging client).

Adamczyk and Bailey [1] also examined the effects of interrupting a user at different stages of a task (document editing, writing summaries of videos and web searching tasks). The selection of the points at which users were interrupted was based on their predicted cognitive load. There were two primary conditions where interruptions were presented: *presumed best* and *presumed worst* times. The other conditions were *random* and *no interruptions*. The *presumed best* times tended to be at the completion of subtasks whilst the *presumed worst* were when the user was performing work on their primary task. The *presumed best* condition resulted in reduced annoyance, frustration, and mental effort, while the *presumed worst* condition resulted in the poorest ratings on each of these measures. Mark et al. [13] examined the impact of interruptions in an email management task and found that in contrast to the results of the other studies, participants who were interrupted completed the task in less time than those who were not interrupted. The authors suggest that people compensate for interruptions by working more efficiently, but also that this increased efficiency comes at an extra cost. Whilst participants that experienced the interruptions completed the tasks faster, they also experienced increased levels of stress, frustration, time pressure, and effort.

In this paper, we describe an experiment to measure the effect of interrupting users performing a video annotation task by asking them questions in a range of contextual conditions, and discuss how our results compare to the existing body of research.

3 The CASAM System - Interruptions and Disruptions

The experimental data was collected during trials with a prototype of the CASAM (Computer Aided Semantic Annotation of Multimedia) system. As well as a user interface in which the user can view video and perform the usual playback functions, CASAM incorporates a video analysis engine that extracts concepts from video and a reasoning engine that constructs and maintains an ontology to manage these concepts. The reasoning engine makes inferences about possible relationships between concepts and generates queries to help it resolve ambiguities. These queries can sometimes seem trivial to the user. However, answering the query might have a significant impact on the content and structure of the ontology and therefore any future inferences made from it. This will, in turn, affect the efficiency of the reasoning component and the future dialogue with the user. The user interface component of the system is responsible for transforming these queries into a human-readable form and deciding how and when to present them to the user. Concepts from the ontology are then used by the system to annotate the video in collaboration with the user. The operation of these analysis and reasoning components is otherwise transparent to the user.

If a system such as CASAM presents a query while the user is engaged with another task or simply concentrating on the video then the user will be interrupted, which is likely to have a negative effect on the user's work, or at least make the user think it does. How disruptive the user finds the interruption depends to a large extent on when it occurs relative to what they are doing [1]. We can identify three broad contextual conditions, with respect to the user's current activity: ***In context***, ***Out of context***, and ***Opportune***.

For the ***In context*** condition queries are presented immediately after the relevant part of the document, which the query refers to, is viewed by the user. In the ***Out of context*** condition queries are presented at a point at which the user is engaged with a part of the video which is unrelated to the content of the query, and the ***Opportune*** condition accounts for those interruptions which take place at times when the user is not engaged with any particular task or section of the video. In a video annotation system this might be at the boundary between two consecutive shots. ***Opportune*** interruptions are generally considered preferable.

Fig. 1. Screenshot of the prototype interface

Bailey et al. [3, 4] find that users are more annoyed when presented with peripheral tasks when they are engaged with another primary task than when they are not, and that users perform slower on an interrupted task than when uninterrupted. They go on to show that interrupting the user at ***Opportune*** moments (that is, when users are not engaged with another task) lessens the disruptive effect.

Experimental data on interruptions may be used to formulate some measure of the cost of those interruptions, both in terms of the user's perception of the disruption the interruptions cause and the measurable effects they have. However, it should be noted that any assessment of cost would ideally need to be linked to some measure of the actual effect of the interruptions on the outcome of the user's task. We do not pursue that in detail here, focusing instead on the perceived and measurable effects of the interruption itself.

4 Experiment Design

The prototype interface used for the experiments consists of a video player and a panel in which queries about the content of the video are displayed and answered. The video player allows the user to replay previous sections of the video, and the query panel allows them to tick check boxes to choose an answer to a query (Figure 1).

Table 1. Time and content of questions presented to the user for Context and Question Type conditions

	Condition	
ID	**Context :Time**	**Query Type: Question**
1	***In context***: 00:06 ***Out of context***: 02:12 ***Opportune***: 00:23	**Important**: Are these flies, bees, or snow? **Trivial***:* What colour is the wall?
2	***In context***: 00:35 ***Out of context***: 02:04 ***Opportune***: 00:41	**Important***:* Is this man a journalist? **Trivial***:* Is this a field?
3	***In context***: 00:45 ***Out of context***: 01:48 ***Opportune***: 00:58	**Important***:* Is this man wearing a protective suit? **Trivial***:* Is this person wearing a hat?
4	***In context***: 01:05 ***Out of context***: 01:31 ***Opportune***: 01:21	**Important***:* Is this Dave Hackenburg? **Trivial***:* Are these boxes in the background?
5	***In context***: 01:25 ***Out of context***: 00:06 ***Opportune***: 01:31	**Important***:* What object is the person in this frame using? **Trivial***:* What is the colour of this person's jacket?
6	***In context***: 01:31 ***Out of context***: 00:35 ***Opportune***: 01:42	**Important***:* Is the person on the left a researcher? **Trivial***:* Does the man on the left have dark hair?
7	***In context***: 01:48 ***Out of context***: 00:45 ***Opportune***: 01:54	**Important***:* Is this is a hive? **Trivial***:* Are there trees in the background?
8	***In context***: 02:04 ***Out of context***: 01:05 ***Opportune***: 02:09	**Important***:* Is this person angry? **Trivial***:* Is this person wearing a shirt?

When a query arrives in the interface the video is paused, interrupting the user, and the query is highlighted using a red border. Once answered, the video continues to play. Queries presented to the user only concern the content of the video and were hand designed to replicate the form, complexity and context of queries that would be generated by the automated reasoning engine of the CASAM system.

Two between subject variables were used in the study. The first (***Question Type***) consisted of two levels; ***Important*** and ***Trivial*** questions. This classified queries according to how relevant they were to a high-level description of the content of the video. For example, an important query might be: "What object is the person in this frame using?" while a trivial query might be: "What colour is this person's jacket?". The second (***Context***) varied the context within which the interruption occurred. The

Table 2. Post experiment questionnaire items

Question	Description	Rating (0 - 100)
Interaction Success	How successful were you in accomplishing what you were supposed to do?	Failure – Perfect
Support	How well did the interface support the annotation task?	Very Low – Very High
Key Points	To what extent do you think the system's questions focused on the key points of the video?	Very Low – Very High
Irritation	How irritated were you with the questions the system asked?	Very Low – Very High
Usability	How usable was the interface?	Very Low – Very High
Expected Behaviour	Did the interface behave in ways which you expected?	Very Low – Very High
Effort	How much effort did it require for you to answer the system's questions?	Very Low – Very High
Overall Perception	What is your overall perception of the interface?	Very Poor – Very Good
Mental Demand	How mentally demanding was the task?	Very Low – Very High
Quality	How well do you think the system will have annotated the video given the nature of the questions asked?	Very Badly – Very Well
Response	How responsive was the interface?	Very Low – Very High
Question Usefulness	How useful do you think the questions were in helping to enhance the system's processing?	Very Low – Very High
Understanding	Did you understand what was going on during the annotation task?	Very Low – Very High
Rushed	How hurried or rushed was the pace of the task?	Very Low – Very High

conditions within this variable differed on whether the question was asked ***In context***, ***Out of context***, or in an ***Opportune*** condition.

In the ***In context*** condition a question must be presented to the user immediately after the relevant information has been presented in the video. In the ***Opportune*** condition a question must be presented to the user at a point of transition between shots in the video. In these cases the timing of the questions are predetermined by the content of the video and the question. For the ***Out of context*** condition, questions are presented at a considerably different time to the relevant content within the video.

In combination this makes a total of 6 ***Question Type-Context*** condition pairs (e.g. ***Trivial – In Context***) within the experiment. Participants were randomly allocated to one of these 6 possible pairings. The video the users were shown contained 8 shot breaks, and each user was asked 8 questions about the video. For the experiment the content and presentation of the questions were both hand coded. The full set of questions and conditions used is shown in table 1. For each query the time it took for the user to respond was recorded. After completing the experiment users completed a questionnaire designed to assess their perception of both the content and usefulness of the queries, their effect on the task and how challenging they were to deal with. The post experiment questions are shown in table 2, and were scored on a visual analogue scale with 100 being very high and 0 being very low. Participants were told before starting that they would be asked questions about the video as it progressed.

92 participants took part in the study and were recruited from the body of students at the University of Birmingham using an email request for participation. The experiment was conducted online. The data for two of the participants was excluded

entirely from the dataset due to high amounts of missing data and the existence of extreme outliers. The video they were shown was a news report about falling bee populations in the United States with a running time of two minutes. The same video was shown to all participants through the same interface. Only the question condition pairings were varied for each participant.

5 Results

5.1 User Perceptions

Two-way (2x2) between subject ANOVAs were used to analyse the effects of the Question Type and Context variables on the user perception dependent variables measured in the post experiment questionnaire. Before reporting the findings of the statistical analysis it is worth noting that due to the amount of analysis conducted there is an increased probability of Type 1 error (a false positive). This has been controlled within each ANOVA analysis by the use of Bonferonni post hoc tests however due to the amount of ANOVA's conducted the readers must interpret the findings whilst being aware of the potential for Type 1 error in this analysis. Additionally due to the amount of analysis conducted, only significant effects are reported in the main body of this paper. A full description the analysis is included in table 12.

Interaction Success

Participants in the ***In context*** condition rated the perception of their success in accomplishing the task as higher than the ***Out of context*** condition. There was a significant main effect of ***Context*** with respect to the perceived success of the interaction [$F(2, 84) = 6.97$, $p<0.01$]. Participants in the ***In context*** condition rated the interaction as significantly more successful than those in the ***Out of context*** condition ($p<0.001$). There was no significant difference between the ***In Context*** and ***Opportune*** ($p>0.05$) and ***Out of Context*** and ***Opportune*** conditions ($p>0.05$). There was no significant main effect of ***Question Type*** on interaction success [$F(1,84) = 0.17$, $p>0.05$]. There was also no significant interaction between ***Question Type*** and ***Context*** on ratings of interaction success [$F(2,84) = 0.11$, $p>0.05$]. The means related to this analysis are displayed in table 3.

Table 3. Mean scores of interaction success by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	90.14	9.968
	Trivial	14	91.50	16.723
	Total	28	90.82	13.527
Out of Context	Important	15	68.40	25.539
	Trivial	15	64.53	35.855
	Total	30	66.47	30.649
Opportune	Important	17	79.82	23.349
	Trivial	15	75.87	28.010
	Total	32	77.97	25.292
Total	Important	46	79.24	22.400
	Trivial	44	76.98	29.693

Support

Participants in the ***In context*** condition perceived that the interface better supported the task compared to those that experienced the ***Out of context*** and ***Opportune*** conditions. There was a significant main effect of ***Context*** on how supported participants felt during the annotation task [$F(2,84) = 11.82$, $p<0.001$]. Participants in the ***In context*** condition felt significantly more supported than those in the ***Out of context*** ($p<0.001$) and ***Opportune*** conditions ($p<0.05$). There was no significant main effect of ***Question Type*** [$F(1,84) = 0.26$, $p>0.05$] or significant interactions of ***Question Type*** and ***Context*** on participants' feeling of support during the task [$F(2,84) = 2.13$, $p>0.05$]. The means for this analysis are displayed in table 4 below.

Table 4. Mean score of support for the annotation task by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	90.14	10.683
	Trivial	14	78.36	25.662
	Total	28	84.25	20.200
Out of Context	Important	15	45.20	28.008
	Trivial	15	58.13	27.417
	Total	30	51.67	28.015
Opportune	Important	17	71.65	21.946
	Trivial	15	62.27	33.184
	Total	32	67.25	27.722
Total	Important	46	68.65	27.905
	Trivial	44	65.98	29.640

Key Points

Those participants in the ***Important*** condition perceived the question to address the key points of the video more so than those in the ***Trivial*** question condition. There was a significant main effect of ***Question Type*** with regard to the extent that participants perceived the questions to focus on the key points of the video [$F(1,84) = 22.15$, $p<0.001$]. Participants in the ***Important*** condition perceived the questions to focus on the key points of the video significantly more when compared with participants in the ***Trivial*** condition ($p<0.001$). There was no significant main effect in relation to the ***Context*** [$F(2,84) = 0.75$, $p>0.05$] and no significant interactions [$F(2,84) = 0.18$, $p>0.05$]. The means for this analysis are displayed in table 5.

Table 5. Mean scores of extent to which questions address the key points of the video by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	50.57	31.072
	Trivial	14	29.21	23.580
	Total	28	39.89	29.169
Out of context	Important	15	47.07	26.980
	Trivial	15	18.20	17.845
	Total	30	32.63	26.845
Opportune	Important	17	46.53	30.268
	Trivial	15	18.60	24.752
	Total	32	33.44	30.823
Total	Important	46	47.93	28.884
	Trivial	44	21.84	22.299

Irritation

Irritation was shown to be lower in those participants in the ***In context*** condition as compared to those in the ***Out of context*** condition. There was a significant main effect of ***Context*** on how irritated participants felt during the interaction [$F(2,84) = 4.40$, $p<0.05$]. Those in the ***In context*** condition felt significantly less irritated than those in the ***Out of context*** condition ($p<0.05$). There was no significant difference between the ***Out of context*** and ***Opportune*** ($p>0.05$) and the ***In context*** and ***Opportune*** conditions ($p>0.05$). There was no significant main effect of ***Question Type*** [$F(1,84) = 0.18$, $p>0.05$] or interaction of ***Context*** and ***Question Type*** [$F(2,84) = 0.48$, $p>0.05$] on how irritated participants felt. The means are displayed in table 6.

Table 6. Mean scores of irritation with the questions by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	15.64	22.595
	Trivial	14	21.43	22.531
	Total	28	18.54	22.337
Out of context	Important	15	45.73	31.226
	Trivial	15	36.67	32.956
	Total	30	41.20	31.880
Opportune	Important	17	29.18	35.202
	Trivial	15	24.60	29.281
	Total	32	27.03	32.127
Total	Important	46	30.46	32.220
	Trivial	44	27.70	28.817

Usability

Usability of the interface was rated higher by those in the ***In context*** condition than those in the ***Out of context*** condition. There was a significant main effect of ***Context*** on how usable participants perceived the interaction [$F(2,84) = 4.39$, $p<0.05$]. Those in the ***In context*** condition rated their interaction higher in usability than those in the ***Out of context*** condition ($p<0.01$). There was no significant difference between usability scores gained in the ***In context*** and ***Opportune*** ($p>0.05$) and the ***Out of context*** and ***Opportune*** conditions ($p>0.05$). There were no significant effects related to ***Question Type*** [$F(1,84) = 0.80$, $p>0.05$]. There were also no significant interaction [$F(2,84) = 0.39$, $p>0.05$]. The means are shown in table 7.

Table 7. Mean scores of usability of the interface by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	86.43	19.057
	Trivial	14	79.29	20.288
	Total	28	82.86	19.654
Out of context	Important	15	66.00	27.586
	Trivial	15	55.20	38.229
	Total	30	60.60	33.213
Opportune	Important	17	69.35	32.336
	Trivial	15	71.07	27.958
	Total	32	70.16	29.890
Total	Important	46	73.46	28.119
	Trivial	44	68.27	30.929

Expected Behaviour

The participants rated the interface as behaving as expected more so in the ***In context*** condition than in the ***Out of context*** condition. There was a significant main effect of ***Context*** in terms of participants judgements of whether the system behaved in ways that the user expected [$F(2,84) = 6.54$, $p<0.001$]. Participants in the ***In context*** condition rated the expected behaviour of the system significantly higher than those in the ***Out of context*** ($p<0.01$) condition. There was no significant difference between ***In context*** and ***Opportune*** ($p>0.05$) and ***Out of context*** and ***Opportune*** ($p>0.05$) conditions. There was no significant main effect of ***Question Type*** [$F(1,84) = 0.28$, $p>0.05$] or interaction effect [$F(2,84) = 1.83$, $p>0.05$] for this dependent variable. The means for the analysis are displayed in table 8.

Table 8. Mean scores of interface behaving as expected by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	89.50	26.593
	Trivial	14	82.29	20.086
	Total	28	85.89	23.415
Out of context	Important	15	48.53	31.062
	Trivial	15	68.27	36.669
	Total	30	58.40	34.866
Opportune	Important	17	75.53	28.496
	Trivial	15	72.73	28.432
	Total	32	74.22	28.039
Total	Important	46	70.98	32.797
	Trivial	44	74.25	29.279

Effort

There was a significant main effect for ***Context*** [$F(2,84) = 3.70$, $p<0.05$]. Participants in the ***In context*** condition rated the effort required as significantly less than those in the ***Out of context*** condition ($p<0.05$). There was no significant main effect of ***Question Type*** [$F(1,84) = 0.0$, $p>0.05$] conditions. There were no significant interactions [$F(2,84) = 0.62$, $p>0.05$]. The means for the analysis are displayed in table 9.

Table 9. Mean perception of effort required by Context and Question Type conditions

Context	Question Type	N	Mean	S.D.
In context	Important	14	14.64	17.456
	Trivial	14	15.29	23.565
	Total	28	14.96	20.352
Out of context	Important	15	26.40	25.925
	Trivial	15	32.67	26.513
	Total	30	29.53	25.961
Opportune	Important	17	20.41	20.893
	Trivial	15	14.20	16.506
	Total	32	17.50	18.928
Total	Important	46	20.61	21.785
	Trivial	44	20.84	23.657

Overall Perception

The overall perception of the interface was rated higher for those in the ***In context*** condition than for those in the ***Opportune*** condition. There was a significant main effect of ***Context*** in overall judgement of the system [$F(2,84) = 3.84$, $p<0.05$]. Participants in the ***In context*** condition rated the expected behaviour of the system significantly higher than those in the ***Opportune*** condition ($p<0.05$). There was no significant difference between the ***Opportune*** and ***Out of context*** ($p>0.05$) as well as the ***Out of context*** and ***In context*** conditions ($p>0.05$). There was no significant effect of ***Question Type*** on participants' judgements on the system overall [$F(1,84) = 0.12$, $p>0.05$]. There was no significant interaction effect [$F(2,84) = 0.96$, $p>0.05$]. The means are displayed in table 10.

Table 10. Mean scores for overall perception of the system by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	81.71	13.736
	Trivial	14	72.21	20.918
	Total	28	76.96	18.026
Out of context	Important	15	59.53	26.016
	Trivial	15	66.80	17.506
	Total	30	63.17	22.099
Opportune	Important	17	63.00	25.647
	Trivial	15	60.20	30.388
	Total	32	61.69	27.542
Total	Important	46	67.57	24.302
	Trivial	44	66.27	23.605

There were no significant main effects for ***Context*** or ***Question Type*** or any significant interactions for any of the remaining questions.

5.2 User Response Times

Measuring query response times is an objective way of quantifying the impact upon user performance. Figure 2 shows the quartile distribution of the response time for each of the test conditions and table 11 depicts the table of means used in the statistical analysis for participant response times. For interruption questions there was a significant main effect of ***Context*** [$F(2,84) = 11.09$, $p<0.001$]**.** There was no significant main effect of ***Question Type*** condition [$F(1,84) = 0.08$, $p>0.05$] and no significant interaction [$F(2,84) = 0.42$, $p>0.05$]. Participants consistently answered queries ***In context*** faster than either the ***Opportune*** ($p<0.01$) or ***Out of context*** ($p<0.001$) cases.

Table 11. Mean scores for participant response time by Context and Question Type

Context	Question Type	N	Mean	S.D.
In context	Important	14	7.24	2.852
	Trivial	14	5.44	1.520
	Total	28	6.34	2.421
Out of context	Important	15	14.32	10.774
	Trivial	15	13.78	6.048
	Total	30	14.05	8.560
Opportune	Important	17	10.93	6.700
	Trivial	15	12.11	5.223
	Total	32	11.48	5.987
Total	Important	46	10.91	7.906
	Trivial	44	10.56	5.871

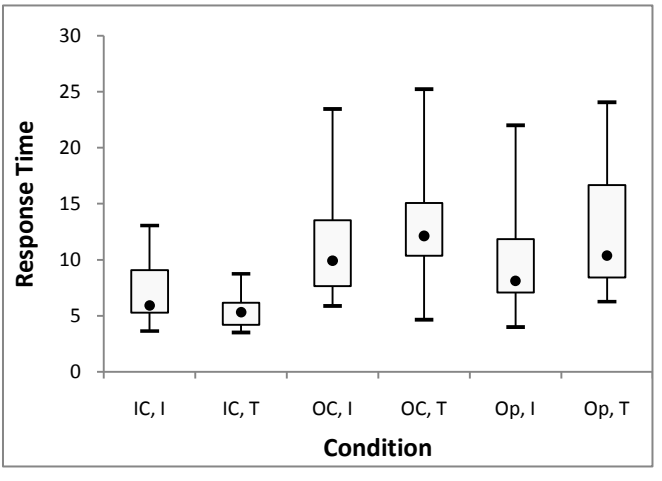

Fig. 2. Quartile distribution of response times for all questions in each condition

6 Discussion

The results show that the time at which the system presented questions to the user had a strong influence on their perceptions and the amount of effort that was required to complete the annotation task. Participants in the important question condition perceived that questions focused on the key points of a video significantly more than those in the trivial question condition. However, there were no significant main effects in the question conditions in all other cases. This suggests that whilst participants could clearly distinguish between the two different types of questions they did not fully understand how the questions influenced the system's processing.

The ***In context*** conditions were generally rated more positively than the other conditions. Participants in the ***In context*** conditions rated the interaction as more

successful, supportive, and usable. They also rated the interaction as being less irritating, requiring less effort and behaving in a way that was closer to what they were expecting.

Similar results are found when looking at user response times. Participants in the ***In context*** condition responded to questions on average more quickly than those in the ***Out of context*** or ***Opportune*** conditions. Based on these findings we propose that, in the context of cooperative, mixed-initiative multimedia annotation, it is better to interrupt a user with questions when they are in context with the multimedia content.

6.1 Comparison to Previous Research Results

This finding differs from much of the related research in the field. In other studies, it has been found that presenting questions at ***Opportune*** moments significantly reduces the negative impact of an interruption [1, 5]. It is instructive to contemplate why we achieved a different result.

The interruptions used in this work are qualitatively different from those typically considered in other work. Typically experiments within the interruptions domain explore interruptions that are gross, often in the form of notifications, and represent a complete change in context to the main task. In the work presented here the interruptions are more subtle and closer to the context of the main task. There are two potential explanations for this observed behaviour:

- Interruptions may be similar to the main task and thus have a reduced impact on cognitive load.
- The user may perceive that responding to an interruption may have a positive impact on their performance on the main task and thus reduce the perceived cost of the interruption.

However, neither of these explanations can be examined within the bounds of the experiment reported here and this is principally due to a number of limitations.

6.2 Limitations and Further Work

Our participants' only task was to watch the video and answer questions: they did not perform any specific annotation themselves. Interruptions may have been more costly if the user had to perform some annotation of their own. Conversely, if the user had been more aware of the impact of responding to questions upon their performance on the task, perhaps through changes to the annotation state, then the perceived cost may have been lower. A more detailed discussion of the cost of interruptions should be based around a more engaging primary task for the user, preferably one for which there is an easily quantifiable outcome. This is because a more demanding task might alter the users' perception of disruption as well as how quickly they respond to it. Moreover, in order to assess an objective cost for a disruption we need to measure the actual impact on the users' task performance as well as their impact perceptions.

It may well be the case that in our more limited scenario cognitive load of users may have been low and thus the differences between the ***Context*** conditions less pronounced. This could account for the lack of impact of the ***Opportune*** condition.

It is also possible that a significant overlap between the ***In-Context*** and ***Opportune***

condition might confound the results. This may occur if queries are presented immediately after the relevant portion of the video and this happens to correlate with a clear shot boundary. However, as can be observed in table 1, there is only one case in which a clear overlap occurs between the ***Opportune*** condition for question 5 and the ***In-context*** condition in question 6.

However, the test performed in this experiment is still representative of the form of interruption in a system like CASAM. It is therefore a useful initial step in understanding the impact of interruptions in such a system. Ideally, the user would not need to do any annotation, and the system would only ask the user questions when necessary. As such, the experiment described in this paper is a useful test to see how users respond.

Other limitations include the use of a single short video. It would be useful in future studies to get participants to interact with the software for much longer period using a range of videos. Additionally, it would be interesting to run some longitudinal studies where participants used the tool for an hour or two on a daily basis. The current prototype would obviously need to be enhanced for longitudinal studies, but this type of test over time would be useful in understanding the true cost of interruptions and how they impact the user's experience.

It would also be interesting to examine how the urgency of a question influences user perceptions. There may be times when the system has an important question that is critical to the automated annotation process. How should the system communicate this urgency to the user? Is it acceptable for the system to immediately interrupt the user from their current task and present the question? Does communicating the importance of the question lower the negative impact of the interruption? Or should the system always wait for the most opportune time to interrupt users?

7 Conclusions

This study demonstrates that, in a video annotation task, users tend to prefer interruptions that are ***In context***, that is, at the appropriate moment in the video, regardless of whether this is an ***Opportune*** moment or not or whether the query is ***Important*** or ***Trivial***. In addition, they also answer them more quickly. This outcome differs from previous research, which may be related to our specific experimental setup, or may be more intrinsically related to the specific task of video annotation. It certainly demonstrates that, if the aim is to have an 'intelligent' system undertaking most of the work, and relatively inexperienced annotators using it whose main function is to respond to the system's questions, then presenting such questions ***In context*** is the most effective approach.

There is also a broader implication of these results. Although we argue that the interruptions presented to the user are qualitatively different from those typically used in the interruptions literature, we also believe that this form of subtle interruption in context with the overall task is representative of a wide range of domains and thus warrants further study.

Table 12. Table of full analysis

Item	Main Effect-Context F(2,84)	Main Effect-Question Type F(1,84)	Interaction	Sig. Post Hoc Comparisons
Interaction Success	6.97^{**}	0.17^{ns}	0.11^{ns}	In Context > Out of Context**
Usability	4.39^{*}	0.80^{ns}	0.39^{ns}	
Expected Behaviour	6.54^{***}	0.28^{ns}	1.83^{ns}	
Overall Perception	3.84^{*}	0.12^{ns}	0.96^{ns}	In Context > Opportune*
Irritation	4.40^{*}	0.18^{ns}	0.48^{ns}	In Context < Out of Context*
Effort	3.70^{*}	0.00^{ns}	0.62^{ns}	
Support	11.82^{***}	0.26^{ns}	2.13^{ns}	In Context > Out of Context*** In Context > Opportune*
Key Points	0.75^{ns}	22.15^{***}	0.18^{ns}	Important > Trivial***
Mental Demand	2.34^{ns}	0.00^{ns}	1.70^{ns}	
Quality	2.49^{ns}	0.41^{ns}	0.67^{ns}	
Response	3.05^{ns}	1.90^{ns}	1.18^{ns}	
Question Usefulness	0.22^{ns}	1.85^{ns}	0.07^{ns}	
Understanding	2.51^{ns}	2.64^{ns}	0.30^{ns}	
Rushed	1.41^{ns}	0.80^{ns}	2.71^{ns}	
Response Times	11.09^{***}	0.08^{ns}	0.42^{ns}	In Context < Opportune** In Context < Out of Context***

* $p<0.05$
** $p<0.01$
*** $p<0.001$
ns $p>0.05$

Acknowledgment. This work was support by EU grant FP7-217061.

References

1. Adamczyk, P.D., Bailey, B.P.: If Not Now When? The Effects of Interruptions at Different Moments Within Task Execution. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2004, pp. 271–278. ACM, New York (2004)
2. Allen, J.F.: Mixed-Initiative Interaction. IEEE Intelligent Systems and their Applications 14(5), 14–23 (1999)
3. Bailey, B.P., Konstan, J.A., Carlis, J.V.: Measuring the Effects of Interruptions on Task Performance in the User Interface. In: IEEE Conference on Systems, Man, and Cybernetics, pp. 757–762 (2000)
4. Bailey, B.P., Konstan, J.A., Carlis, J.V.: The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. In: Proc. of INTERACT 2001, pp. 593–601 (2001)

5. Bailey, B.P., Iqbal, S.T.: Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. ACM Trans. Comput.-Hum. Interact. 14(4), 1–28 (2008)
6. Creed, C., Lonsdale, P., Hendley, B., Beale, R.: Synergistic Annotation of Multimedia Content. In: Proc. of the 2010 Third International Conference on Advances in Computer-Human Interactions, pp. 205–208 (2010)
7. Cutrell, E., Czerwinski, M., Horvitz, E.: Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance. In: Proc. of INTERACT 2001, pp. 263–269 (2001)
8. Czerwinski, M., Cutrel, E., Horvitz, E.: Instant Messaging and Interruption: Influence of Task Type on Performance. In: OZCHI 2000 Conference Proceedings, pp. 356–361 (2000)
9. Horvitz, E.: Principles of Mixed-Initiative User Interfaces. In: Proc. SIGCHI Conference on Human Factors in Computing Systems, pp. 159–166 (1999)
10. Horvitz, E., Kadie, C., Paek, T., Hovel, D.: Models of attention in computing and communication: from principles to applications. Comm. ACM 46(3), 52–59 (2003)
11. Iqbal, S.T., Bailey, B.P.: Effects of Intelligent Notification Management on Users and Their Tasks. In: Proc. of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 93–102 (2008)
12. Latorella, K.A.: Effects of modality on interrupted flight deck performance: Implications for data link. In: Human Factors and Ergonomics Society Annual Meeting Proceedings, Aerospace Systems, vol. (5), pp. 87–91 (1998)
13. Mark, G., Gudith, D., Klocke, U.: The Cost of Interrupted Work: More Speed and Stress. In: Proc. of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 107–110 (2008)
14. Mcfarlane, D.C.: Coordinating the Interruption of People in Human-computer Interaction. In: Proc. of Human-Computer Interaction (INTERACT 1999), pp. 295–303. IOS Press, Amsterdam (1999)
15. Miyata, Y., Norman, D.A.: Psychological Issues in Support of Multiple Activities. In: User Centered System Design: New Perspectives on Human-Computer Interaction, pp. 265–284 (1986)
16. Monk, C.A., Boehm-Davis, D.A.: The Attentional Costs of Interrupting Task Performance at Various Stages. In: Proc. of the Human Factors and Ergonomics Society 46th Annual Meeting, pp. 1824–1828 (2002)
17. Rubinstein, J.S., Meyer, D.E., Evans, J.E.: Executive Control of Cognitive Processes in Task Switching. Journal of Experimental Psychology: Human Perception and Performance 27(4), 763–797 (2001)
18. Speier, C., Valacich, J.S., Vessey, I.: The influence of task interruption on individual decision making: An information overload perspective. Decision Sciences 30(2), 337–360 (1999)
19. Zijlstra, F.R.H., Roe, R.A.: Temporal factors in mental work: Effects of interrupted activities. Journal of Occupational and Organizational Psychology 72, 163–185 (1999)