Is One Second Enough? Evaluating QoE for Inter-Destination Multimedia Synchronization using Human Computation and Crowdsourcing

Benjamin Rainer, Stefan Petscharnig, Christian Timmerer, and Hermann Hellwagner
Institute of Information Technology
Alpen-Adria-Univeristät Klagenfurt
Austria, Klagenfurt 9020
Email: first.lastname@itec.aau.at

***Abstract*—Modern-age technology enables us to consume multimedia for enjoyment and as a social experience. The traditional way to consume multimedia together (e.g., with family or friends in the living room) is being superseded by a location-independent scenario where geographically distributed users consume the same content while having a real-time communication channel among each other. Inter-Destination Multimedia Synchronization (IDMS) is the tool of choice in order to enable users a high-quality multimedia experience. In this paper, we investigate the influence of asynchronism when consuming multimedia content together while being geographically distributed. In particular, we adopt the concept of human computation and developed a reaction game which we used to conduct a crowdsourced subjective quality assessment in order to evaluate a threshold for multimedia synchronization within an IDMS scenario. Our results show a significant decrease in overall Quality of Experience (QOE) at an asynchronism level of 750ms. At the same time, we were able to show that asynchronism at a level of 400ms does not have significant differences regarding the QoE when compared to the synchronous reference case.**

I. Introduction

With the upraise of social networks traditional scenarios like watching TV with friends and/or the family drifts more and more towards a distributed experience where the participating users are geographically distributed while chatting (text, voice, video) using real-time communication tools. In order to provide the same Quality of Experience (QoE) as if the users were in the same room in front of the TV, the playback of the multimedia content among the geographically distributed participants needs to be synchronized which is referred to as Inter-Destination Multimedia Synchronization (IDMS) [1][2][3].

A previous study has shown that asynchronism impacts the QoE during a *social TV* session with friends [4]. For other application scenarios like online quiz shows, synchronization plays an even more important role than in a social TV scenario. In such quiz shows asynchronism may not only cause a certain level of annoyance but will also affect the fairness of the quiz because most quiz shows are designed as competitive reaction games. Geerts et al. assess the lower asynchronism threshold in the context of a *social TV* scenario [4] whereas in this work we focus on a more general case, like in online quiz shows. Therefore, our hypothesis is that the threshold of asynchronism for a competitive game is lower than reported in [4] which may already cause unfairness. These factors may decide whether a system that claims to provide IDMS is accepted. The research question that we address with the work presented in this paper is as follows:

> **What is the lower asynchnronism threshold for IDMS in general?**

In order to answer this research question we adopt concepts from *human computation* to develop a reaction game which allows us to conduct a *crowdsourced subjective quality assessment* (henceforth referred to as SQA). We will introduce and describe the design of the reaction game that provides the possibility to mimic the scenario of online quiz shows. Using this game, we evaluate the impact of asynchronism on the overall *QoE*, *togetherness*, *fairness*, and *annoyance*. Quality of Experience denotes the overall delight of the users with the game. Togetherness refers to the perceived quality of being together. Fairness aims at measuring the users' perception whether they are treated equally. Annoyance refers to the users state of being annoyed [4]. These are subjective measures and shall reflect the opinion of the users, they are measured on a scale from 0–100. The results of the SQA indicate the lower asnychronism threshold for which the QoE decreases significantly. The novelty in this research is the adoption of human computation – specifically, game with a purpose (GWAP) – for evaluating the QoE of IDMS use cases.

The remainder of the paper is organized as follows. A review of related work on human computation, specifically on games for a purpose, and IDMS is presented in Section II. Section III describes the methodology including game design, crowdsourced SQA, and the adopted platform. The results of a statistical analysis of the data obtained by the crowdsourced SQA are presented in Section IV. Finally, Section V provides a discussion of the results and the conclusions.

II. Related Work

The idea of utilizing human processing power in order to solve problems that computers cannot yet solve was initially introduced by von Ahn [5]. One subcategory in human

978-1-4799-8958-4/15/$31.00 ©2015 IEEE

computation are games with a purpose (GWAP). GWAPs aim at collecting data from people while they do something they enjoy: playing a game. A well-known example from the variety of GWAPs is the ESP Game [6] whose task is to collect labels for images. The design of these GWAPs has influence on how the gathered data is verified. Ahn and Dabbish [7] categorize the game design as follows:

Output agreement games, the players are given the same input. The users evaluate the output of each other by agreeing on a common output. The players win if the outputs match. An example for such games is the ESP game [6] in which two users see the same picture and are given the instruction "write what the other one would write" .

Input agreement games give the players some input. Only the game itself knows whether these inputs are equal. The goal for the players is to successfully guess whether they were given the same input. In order to do this, they produce some output like chat messages. Based on this output, the users decide whether they were given the same input. An example for this sort of games is tag-a-tune [5], which produces tags for audio files.

Inversion problem games introduce the opportunity to perform different tasks. For this kind of games, player 1 is given an input. This player produces some output for player 2. The players win, if player 2 is able to guess the input player one was given. An Example for this sort of games is Verbosity [8], which aims at collecting common sense facts.

Gamification is the use of game elements in non-game contexts. Mekler et al. conducted a study which shows that points, levels and leader-boards are capable of driving user behaviour in the short term [9]. Anderson at al. [10] as well as Grant and Betts [11] suggest encouraging user behaviour with game elements like badges or achievements.

The authors of [4] investigate the impact of asnychronism on the togetherness, fairness, and annoyance in a social TV scenario using a system that provides IDMS among the participating entities. The results of the presented SQA show that for active text chatters the upper threshold is at about *two seconds* until a significant difference is noticed by the users. For active voice chatters the threshold is at about *one second* until the asynchronism is significantly perceived. These thresholds may vary if we take another IDMS scenario such as online quiz shows or any other interactive (competitive) multimedia scenario. This explains also the title "is one second enough?"

III. Methodology

For assessing the asynchronsim threshold for IDMS we introduce a reaction game following the GWAP principle. The game was designed based on an in-lab study allowing us to conduct crowdsourced SQA for the actual evaluation as we believe it is very difficult that strangers will team up and actually communicate during a SQA. This section describes the game design, SQA methodology (incl. stimuli), and the platform used for the SQA.

A. Reaction Game

The reaction game which has been designed for conducting the SQA follows a simple principle such as quiz shows do.

TABLE I: Videos used for evaluation

video name	length [mm:ss]	#events
training	00:54	3
Knack	01:50	4
Famson 1	01:46	6
Famson 2	01:58	8

In our game two players have to collaborate in order to achieve the highest possible score. Therefore, we introduce time-bounded events during which both players have to click onto the canvas that is displaying the video sequence. Figure 1 depicts the time sequence beginning by the start of a video sequence, occurrence of an event and the different possibilities to click during a given time window. An event is signalled by displaying an attention symbol in the upper left corner of the video sequence (cf. Figure 2a). If the multimedia playback of both players is synchronous the time window for being awarded with additional (bonus points) score equals to the total duration of the occurred event. With an increase in asynchronism the time window for the bonus score decreases and introduces a certain level of difficulty as it would be the case in online quiz shows. In the case that the multimedia playback of both players is not synchronous and, therefore, the occurrence of the events differ in time, the bonus time window shrinks according to the introduced asynchronism (total time window - asynchronism). If a player clicks too early or too late (depending on the player and the sign of the asynchronism) but still in the time window of the event the player is awarded 100 points. If the players manage to click during the bonus window, both players are awarded 200 points.

The game provides the players with a visual feedback if they were able to achieve bonus score (cf. Figure 2c), if the player was able to click during the event without clicking during the *overlapping/bonus window* (cf. Figure 2b), and if they failed to react during the given time window (cf. Figure 2d). If players just click into the video sequence even if no event has occurred, minus scores are given.

Since pairing participants online may introduce unpleasant waiting times, we decided to introduce an artificial intelligence (AI) as the second player. In order to provide an AI that mimics the behavior of a real player, we invited students to take part in an in-lab SQA where they played together in pairs. With the gathered statistics (described in Section III-C) from this SQA we simulate a player in the SQA using crowdsourcing. Since we have gathered real user data, we were able to use a simple algorithm for the simulated player. The underlying assumption was, that the opposite player always clicks if an event occurs. We parsed the reaction times of the successful clicks and built a knowledge base of reaction times. At every event, one of these reaction times is picked at random (uniformly) and acts as second input for the original two-player game.

B. Assessment Methodology and Stimuli

In order to assess the impact of asynchronism on *QoE*, *fairness*, *togetherness*, and *annoyance*, we selected a single stimulus with hidden reference as recommended by the ITU [12], [13]. We further used Microworkers [14] to hire participants for the SQA from the U.S.A, Canada, Australia, New Zealand, West Europe, and East Europe. The duration

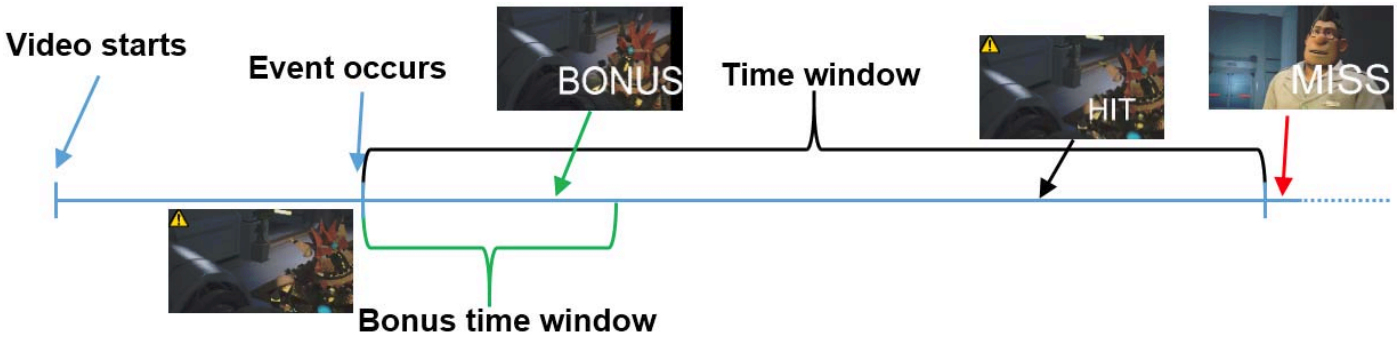


Fig. 1: Temporal sequence of the game events.

(a) A game event occurs and is indicated in the upper left corner.

(b) Indicate that the player has not passed the event but did not click within the bonus time window.

(c) Indicate that both players clicked within the bonus time window.

(d) The player missed an event or clicked even if no event has occurred.

Fig. 2: Reaction game states and signalling of events.

TABLE II: Test cases for the crowd sourced evaluation

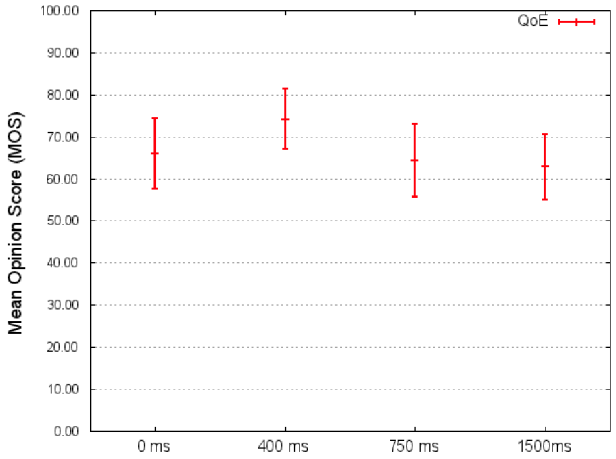
case	description	asynchronism [ms]	window length [ms]	bonus window length [ms]
0	training video	0	2000	2000
1	synchronous	0	2000	2000
2	small asynchronism	400	2000	1600
3	medium asynchronism	750	2000	1250
4	big asynchronism	1500	2000	500

of the whole SQA is about 15 minutes. For each successful participation we pay $0.50.

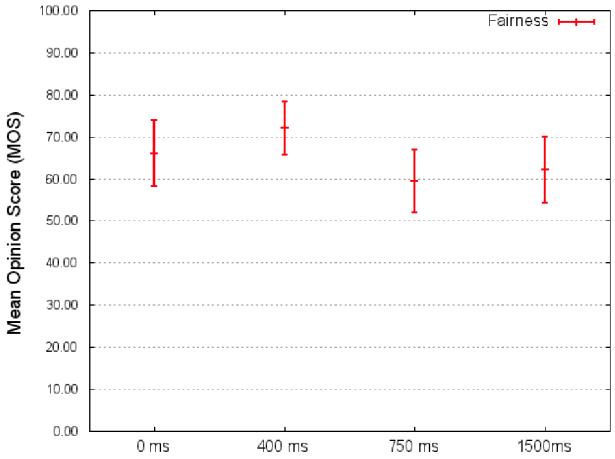
In order to gather appropriate multimedia content that allows for assessing IDMS using a reaction game, the multimedia content should provide enough possibilities to introduce game events. Therefore, we recorded in-game scenes of video games. We selected non-violent in-game scenes from the two games *inFAMOUS Second Son* [15] (henceforth denoted by Famson) and *Knack* [16] (henceforth denoted by Knack). Table I depicts the length and the number of events for the selected video sequences. All the video sequences comprise audio.

The SQA is split into the following five parts:

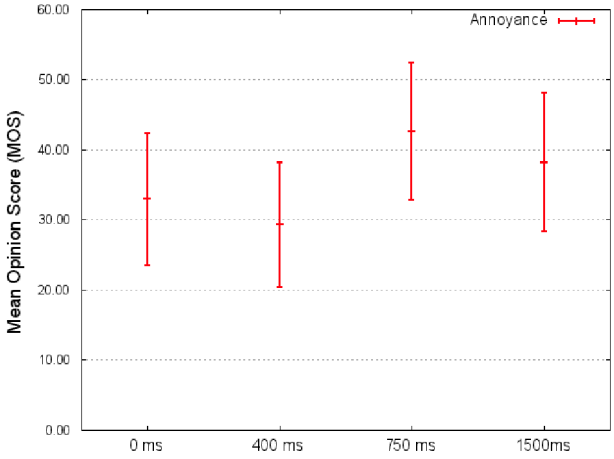
Introduction. In the beginning an introduction is presented to each participant which explains in detail the experiment and the actual task. In particular the introduction explicitly states that the audio devices should be turned on and the audio volume should be adjusted accordingly. We further instruct the participants to switch off their mobile devices. The reaction



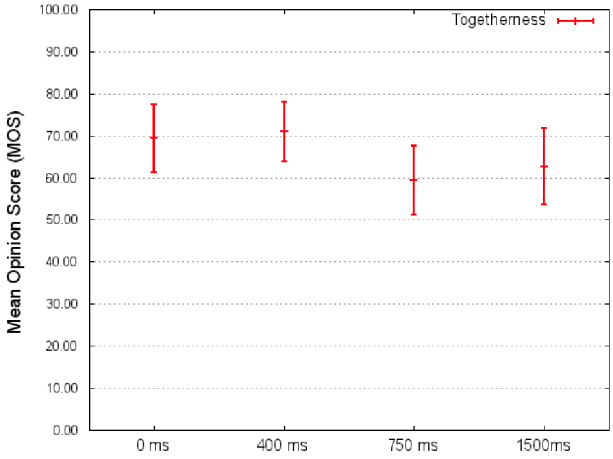
(a) MOS and 95% CI for the overall QoE (higher is better).



(b) MOS and 95% CI for fairness (higher is better).



(c) MOS and 95% CI for annoyance (lower is better).



(d) MOS and 95% CI for togetherness (higher is better).

Fig. 3: MOS for the assessed variables with 95% confidence intervals (CI).

game and the rating possibility are explained in detail such that no questions are left open, including figures that show how events are indicated. At the end of the introduction an before the actual SQA start each participant has to agree to a disclaimer.

Pre-Questionnaire. After agreeing to the disclaimer the participants have to fill out a pre-questionnaire. The pre-questionnaire is used to gather demographic information about the participants, i.e., age, gender, nationality, and country of residence. The data is used to cross check the preferred regions from which participants are hired for the SQA using crowd-sourcing. During the pre-questionnaire the video sequence are cached. Only if the video sequenced are cached successfully the participants are allied to pass on to the training and main evaluation.

Training. In order to provide the participants the possibility to become familiar with the reaction game we introduce a training phase to the SQA. For the training phase we use an in-game scene from Famson with a duration of approximately 54 seconds (cf. Table I). We further use the opportunity of a training phase to present the voting possibilities to the participants which are used to assess the following variables: overall QoE, fairness, togetherness, and annoyance. The voting is done by adjusting a slider on a numerical scale from 0 to 100 for each of the variables, where 0 indicates a very low QoE, togetherness, fairness, or annoyance and 100 indicates a very high QoE, togetherness, fairness, or annoyance. We do not restrict the duration of the voting phase. We asks the participants for the same variables as in [4].

Main Evaluation. As already mentioned, the main evaluation adopts a single stimulus with hidden reference as recommended by the ITU [12], [13]. We selected a single stimulus with hidden reference because the participants shall not know which one is the reference condition during the SQA. This allows to investigate whether there is a significant difference between the reference (in our case both players are synchronous) and the cases where we introduce asynchronism. For the main evaluation we present the video sequences randomly (uniform) assigned to the test cases and each participant has to play the reaction game using the test cases depicted in Table II. The main evaluation comprises only test case one to four. The first test case (cf. Table II case 1) depicts the hidden reference. The second test case (cf. Table II case 2) introduces an asynchronism of 400 ms which results in a bonus time window for each event of 1600 ms. The third test case (cf.

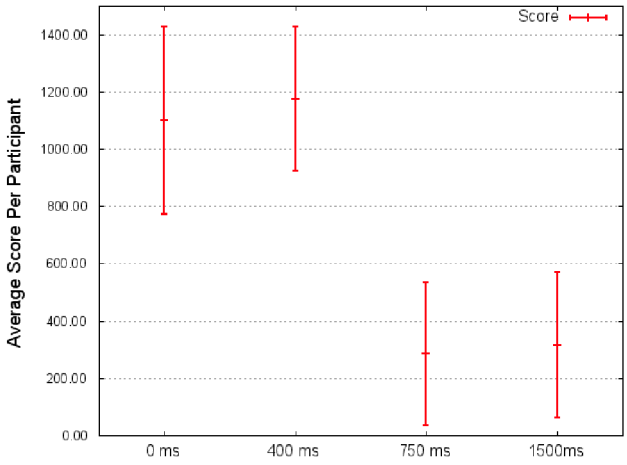


Fig. 4: Average score achieved by a participant with 95% confidence interval.

Table II case 3) introduces an asynchronism of 750 ms which corresponds to a bonus window of 1250ms. The fourth test case (cf. Table II case 4) introduces an asynchronism of 1500 ms which corresponds to a bonus window of 500 ms. We selected these test cases in order to see whether the assumptions deduced in [4] holds and to assess if the threshold lies below one second. The test cases and the assigned video sequences are randomly presented to the participants during the main evaluation. After each stimulus presentation the participants are asked to vote the overall QoE, togetherness, fairness and annoyance as discussed before.

Post-Questionnaire. Finally, in the end of the SQA the participants are asked to fill out a post-questionnaire. This provides the participants the possibility of providing general feedback. We further ask if the participants have already participated in a similar SQA.

C. Platform and gathered Statistics

For conducting the SQA using crowdsouring we use a web-based quality assessment platform provided by [17]. Besides the possibility of selecting the assessment methodology the platform allows to easily extend the stimulus presentation. In addition to the stimulus presentation time we measure some more variable regarding the reaction game in order to investigate the behavior of every participant. This is done to identify participants that do not *honestly* take part in the SQA, to identify participants that did not understand the actual task, or to identify participants that try to cheat by reducing the stimulus presentation time [18].

Therefore, we measure the following variables regarding the reaction game: reaction time (time between the occurrence of an event and the first click after it occurred), number of stalls or pauses of the multimedia playback, number of browser window focus changes, the audio volume, the number of clicks during time window, and the total amount of clicks during a video sequence. By the use of these statistics we filter the participants. How many participants are filtered is given in Section IV.

IV. Statistical Analysis of the Results

In total 89 persons participated in the SQA. After the screening 44 participants were accepted for the final statistical analysis. The participants have been screened accordingly to the following rules: *i*) **Browser focus change**, 27 participants were screened because they changed the Browser focus during the stimulus; presentation. *ii*) **Total number of clicks**, 16 participants were screened because they did not click a single time during a stimulus; presentation. *iii*) **Number of clicks during an event**, 2 participants were screened because they never clicked during an event.

Figure 3 depicts the results for the overall QoE (cf. Figure 3a), fairness (cf. Figure 3b), annoyance (cf. Figure 3c), and togetherness (cf. Figure 3d) for each test case (cf. Table II). For all the test cases we assume that the samples follow a normal distribution according to Shapiro-Wilk tests [19] and by investigating Q-Q plots. Before conducting a Student's t-test we verified homogeneity of the variance by conducting F-Tests.

Figure 3a depicts the Mean Opinion Score (MOS) for the overall QoE for an asynchronism of 0 ms, 400 ms, 750 ms, and 1500ms. A Student's t-test revealed significant differences for the following pairs of test cases with t-,p- and α-values as follows: (400 ms, 750 ms) $t = 1.73$ $p - value = 0.087$ $\alpha = 0.1$, and (400 ms, 1500ms) $t = 2.1$ $p - value = 0.039$ $\alpha = 0.05$.

Figure 3b depicts the MOS for fairness. The participants had to vote how fair (in their opinion) each test case was with respect to the reaction game. A Student's t-test revealed the following significant difference between the means of the following test cases: (400 ms, 750 ms) $t = 2.51$ $p - value = 0.014$ $\alpha = 0.05$, and (400 ms, 1500ms) $t = 1.93$ $p - value = 0.057$ $\alpha = 0.1$. For the test cases (0 ms, 750 ms) and (0 ms, 1500 ms) the p-value is slightly above $\alpha = 0.1$.

Figure 3c depicts the MOS for annoyance. The annoyance states how annoyed the participants were after the test cases. As the figure indicates there is clear increasing tendency with an increase in asynchronism. A Student's t-test revealed the following significant difference between the means of the following test cases: (400 ms, 750 ms) $t = -1.31$ $p - value = 0.049$ $\alpha = 0.05$. For the other pairs of test cases no significant differences in the means could be found for $\alpha \leq 0.1$.

The last variable we measured is the togetherness. Figure 3d depicts the MOS for togetherness. The results for this variable follows the same principle as the the other three. A Student's t-test revealed the following significant difference between the means of the following test cases: (0 ms, 750 ms) $t = 1.68$ $p - value = 0.096$ $\alpha = 0.1$, and (400 ms, 750 ms) $t = 2.08$ $p - value = 0.03988$ $\alpha = 0.05$.

Finally we report the average score a participant has achieved for each of the test cases. Figure 4 depicts the average score with 95% confidence interval. Again, the same tendencies can be observed. For the test cases below an asynchronism of 750 ms the participants are able to achieve high scores. If the asynchronism increases to 750 ms the scores suddenly drop below 400 points on average.

V. Discussion and Conclusion

In [4] it is already assumed that for active voice chatters the lower asynchronism threshold is at about one to two seconds in a social TV scenario. In this paper we reported about research on assessing the lower asynchronism threshold for IDMS that is valid in a social TV scenario and in more complex scenarios where asynchronism does not only affect the QoE but also other factors like the fairness. The fairness of a game is very important and especially if it is a competitive one. Quiz shows are always competitive and asynchronism may introduce unfairness among the participating entities. Therefore, it is crucial for a system that tries to preserve IDMS among the participating entities that it maintains at least the lower bound of asynchronism.

The results of the SQA clearly show that one second as the lower asynchronism threshold is not enough. The results show that all the measured variables significantly decrease (QoE, togetherness, fairness) or increase (annoyance) if the asynchronism reaches 750 ms. Therefore, we assume that the asynchronism shall not exceed 400 ms. We have selected very coarse asnychronism steps for the test cases. This allowed us to cover a wide range of asnychronism even if the last test case was too difficult. This difficulty has the cause that not many participants were able to hit the bonus time window. This is confirmed by Figure 4 which depicts the average score. Nevertheless, our aim was to find a first approximate guess for the lower asynchronism threshold and to show that it is below one second. Another limiting factor was the duration of the SQA. SQAs using crowdsourcing shall not exceed a certain duration.

One major problem is the reliability of the participants that are acquired using crowdsourcing platforms. In order to achieve a certain reliability we have measured many variables during a gaming session of each participant. It is crucial for the quality of the results to filter the participants accordingly because the SQA is conducted in an uncontrolled environment. Nevertheless, not every bias can be neglected by screening or filtering participants by the gathered data. There are always (environmental) factors that cannot be controlled (e.g., brightness and contrast of the screen, size of the screen, resolution, etc.).

The results provide even more insights. By taking a look at the sub-figures of Figure 3 it can be observed that there is strong relationship between the fairness, togetherness, annoyance and the overall QoE. Indeed, the QoE may be modelled by the other three variables. We declare this to future work.

Acknowledgment

This work was supported in part by the Austrian Science Fund (FWF) under the CHIST-ERA project CONCERT (A Context-Adaptive Content Ecosystem Under Uncertainty), project number *I1402* and partly performed in the Lakeside Labs research cluster at Alpen-Adria-Universität.

References

[1] M. Montagud, F. Boronat, H. Stokking, and R. Brandenburg, "Inter-destination multimedia synchronization: schemes, use cases and standardization," *Multimedia Systems*, vol. 18, pp. 459–482, 2012.

[2] F. Boronat, J. Lloret, and M. García, "Multimedia group and inter-stream synchronization techniques: A comparative study," *Information Systems*, vol. 34, no. 1, pp. 108–131, 2009.

[3] B. Rainer and C. Timmerer, "Self-Organized Inter-Destination Multimedia Synchronization For Adaptive Media Streaming," in *Proceedings of the 22st ACM International Conference on Multimedia*, ACM, Ed. New York, NY, USA: ACM, nov 2014.

[4] D. Geerts, I. Vaishnavi, R. Mekuria, O. van Deventer, and P. Cesar, "Are we in sync?: synchronization requirements for watching online video together." in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 311–314.

[5] E. Law and L. von Ahn, "Input-agreement: A new mechanism for collecting data using human computation games," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 1197–1206. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518881

[6] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '04. New York, NY, USA: ACM, 2004, pp. 319–326. [Online]. Available: http://doi.acm.org/10.1145/985692.985733

[7] L. v. Ahn and L. Dabbish, "Designing games with a purpose," *Commun. ACM*, vol. 51, no. 8, pp. 58–67, Aug. 2008. [Online]. Available: http://doi.acm.org/10.1145/1378704.1378719

[8] L. v. Ahn, M. Kedia, and M. Blum, "Verbosity: A game for collecting common-sense facts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '06. New York, NY, USA: ACM, 2006, pp. 75–78. [Online]. Available: http://doi.acm.org/10.1145/1124772.1124784

[9] E. D. Mekler, F. Brühlmann, K. Opwis, and A. N. Tuch, "Do points, levels and leaderboards harm intrinsic motivation?: An empirical analysis of common gamification elements," in *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, ser. Gamification '13. New York, NY, USA: ACM, 2013, pp. 66–73. [Online]. Available: http://doi.acm.org/10.1145/2583008.2583017

[10] S. Grant and B. Betts, "Encouraging user behaviour with achievements: An empirical study," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 65–68. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487085.2487101

[11] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Steering user behavior with badges," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 95–106. [Online]. Available: http://dl.acm.org/citation.cfm?id=2488388.2488398

[12] "Rec. ITU-R BT.500-11," Tech. Rep. [Online]. Available: http://www.dii.unisi.it/~menegaz/DoctoralSchool2004/papers/ITU-R_BT.500-11.pdf

[13] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Geneva, Switzerland, Tech. Rep., Apr. 2008.

[14] Microworkers, "http://www.microworkers.com."

[15] inFAMOUS Second Son - Sukker Punch, "http://infamous-second-son.com/."

[16] Knack - SCE Japan Studio, "http://us.playstation.com/ps4/games/knack-ps4.html."

[17] B. Rainer, M. Waltl, and C. Timmerer, "A Web based Subjective Evaluation Platform," in *Proceedings of the 5th International Workshop on Quality of Multimedia Experience*. Los Alamitos, CA, USA: IEEE, jul 2013, pp. 24–25. [Online]. Available: http://www.qomex2013.org

[18] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2013.

[19] S. S. Shapiro; M. B. Wilk, "An Analysis of Variance Test for Normality," *Biometrika, Vol. 52, No. 3/4, pp. 591-611.*, 1965.