NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELLECOMMUNICATIONS

THESIS

Assessing Quality of Experience of Video Streaming Applications via Crowdsourcing

Dimitrios N. Kyriazanos

Supervisors **Lazaros Merakos,** Professor
Eirini Liotou, Researcher

ATHENS

SEPTEMBER 2016

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αξιολόγηση της Ποιότητας Εμπειρίας σε Εφαρμογές Βίντεο Συνεχούς Ροής μέσω Crowdsourcing

Δημήτριος Ν. Κυριαζάνος

Επιβλέποντες **Λάζαρος Μεράκος,** Καθηγητής
Ειρήνη Λιώτου, Ερευνήτρια

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2016

THESIS

Assessing Quality of Experience of Video Streaming Applications via Crowdsourcing

Dimitrios N. Kyriazanos
S.N.: 1115200500048

SUPERVISORS: **Lazaros Merakos,** Professor
Eirini Liotou, Researcher

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αξιολόγηση της Ποιότητας Εμπειρίας σε Εφαρμογές Βίντεο Συνεχούς Ροής μέσω Crowdsourcing

Δημήτριος Ν. Κυριαζάνος

Α.Μ.: 1115200500048

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Λάζαρος Μεράκος,** Καθηγητής
Ειρήνη Λιώτου, Ερευνήτρια

ABSTRACT

For decades, Quality of Service has been the dominant metric by which the capabilities of a network are determined. Recently though another metric, Quality of Experience, has gained significant traction and is steadily becoming the standard, as it has been proven to depict the satisfaction rate of the subscribers much more accurately.
Naturally, challenges have risen relating to his new methodology. The means to produce credible results in a financially viable way is the main one, as some traditional techniques such as controlled experiments in a laboratory can only be utilized in a limited capacity. A possible solution to this problem is crowdsourcing, which essentially is the process of conducting experiments via special online platforms. Test subjects around the world willingly participate in these experiments for a small monetary compensation.
This paper focuses on presenting the principles of Quality of Experience along with the key differences with Quality of Service, describes in detail the proper use of crowdsourcing and examines whether it is a suitable source of user feedback, by conducting several test cases regarding specific aspects of the quality of a network, as experienced by the subscribers.

SUBJECT AREA: Networks
KEYWORDS: Quality of Experience, Crowdsourcing, Stalling, Adaptive Streaming

ΠΕΡΙΛΗΨΗ

Για δεκαετίες, το Quality of Service έχει υπάρξει η κυρίαρχη μετρική με την οποία οι δυνατότητες ενός δικτύου μετριούνται. Πρόσφατα όμως μια νέα μετρική, το Quality of Experience, έχει ανοδική πορεία και σταθερά γίνεται η πρωτεύουσα, καθώς έχει αποδειχτεί ότι απεικονίζει το βαθμό ικανοποίησης των συνδρομητών με πολύ πιο ακριβή τρόπο.
Όπως είναι φυσιολογικό, προκλήσεις έχουν προκύψει που αφορούν αυτήν τη νέα μεθοδολογία. Τα μέσα με τα οποία παράγονται αξιόπιστα αποτελέσματα με έναν οικονομικά βιώσιμο τρόπο είναι το κύριο από αυτά, αφού κάποιες παραδοσιακές τεχνικές όπως τα ελεγχόμενα πειράματα σε εργαστήρια μπορούν να χρησιμοποϊηθούν μόνο σε περιορισμένο βαθμό. Μία πιθανή λύση σε αυτό το πρόβλημα είναι το crowdsourcing, που πρακτικά είναι η διαδικασία της εκτέλεσης πειραμάτων σε ειδικές διαδικτυακές πλατφόρμες. Συμμετέχοντες στα πειράματα από όλον τον κόσμο παίρνουν μέρος εθελοντικά για μια μικρή χρηματική αποζημίωση.
Αυτή η εργασία επικεντρώνεται στην παρουσίαση των αρχών του Quality of Experience μαζί με τις βασικές διαφορές με το Quality of Sevice, περιγράφει λεπτομερώς τη σωστή χρήση του crowdsourcing κι εξετάζει αν αυτή αποτελεί μια κατάλληλη πηγή ανάδρασης των χρηστών, πραγματοποιώντας διάφορα πειράματα που αφορούν συγκεκριμένα χαρακτηριστικά της ποιότηττας ενός δικτύου, όπως αυτά βιώνονται από τους χρήστες.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Δίκτυα
ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ποιότητα Εμπειρίας, Crowdsourcing, Καθυστέρηση, Προσαρμοστική Ροή

CONTENTS

LIST OF TABLES

LIST OF GRAPHS

PRELUDE

This project was carried out from January to August 2016 in Athens, Greece. It is my thesis for the Department of Informatics & Telecommunications of the National and Kapodistrian University of Athens. I would like to especially thank my direct supervisor, researcher Eirini Liotou for helping me pick the subject as well as guiding me through the whole process. Additionally, a lot of credit should be given to my professor Lazaros Merakos who played a key role in materializing this idea. I hope that my original research will prove handy to the scientific community and that it generally is an interesting read. The structure is as described below.

The first chapter focuses on explaining the meaning of quality of experience in networks, its differences from the traditional quality of service and currently available methods of measuring it. Additionally, a framework is suggested that implements most of its methodologies.

In the second chapter, the concept of crowdsourcing is explained including an introduction to the term, all the proper techniques that should be used when conducting such an experiment and a thorough look in some of the most important online crowdsourcing platforms that dominate today's market.

The third chapter includes a specific crowdsourcing experiment which attempts to measure users' quality of experience when watching videos online. Factors such as http adaptive streaming and stalling are simulated and test subjects are asked to rate and compare video clips with different kinds of impairment. At the end of the chapter, the results of the experiment are presented along with the respective comments.

The fourth chapter summarizes the conclusions extracted from all previous chapters.

The fifth chapter consists of all images referenced with the proper captions.

1. QUALITY OF EXPERIENCE

1.1 Introduction

A number of metrics have been used over the years in order to adequately measure the capabilities of a given network. Every network has a plethora of technical features that need to be tested, measured and assessed but the overall performance of it is judged upon other factors as well. Few networks are self-sufficient nowadays and the coalition between them is a key element in making a safe assumption about their level of quality. Moreover, results may vary for every user even if they are provided with exactly the same network characteristics due to subjective criteria, often categorized in relevance to gender, age, financial status, nationality or location.
There are two general approaches with which networks' quality can be quantified, Quality of Service (QoS) and Quality of Experience (QoE). Even though a lot of times the Quality of Experience is considered a part of the Quality of Service, it has been proved in recent years that there are fundamental differences between the two, as they tend to assess things from totally different scopes.

1.2 A Brief History

During the 1990's, the notion of Quality of Service (QoS) was dominating the world of network assessment and had shaped the communications landscape to a substantial extent. However, its essence is perceived differently among various organizations, something that can be easily comprehended by noticing their respective definitions of the term QoS:

- "*The definition and assessment of service quality, class and grade.*"
- *"The specification of a contract between a customer and a service provider."*
- *"Architectures of networks for controlling quality and improving performance."*
- *"The collective effect of service performance which determines the degree of satisfaction of a user of the service."*

The confusion is clear as to whether this includes exclusively intrinsic, purely technical terms, or the final performance of the network, as perceived by the user, is also measured. As years passed and QoS was implemented on most major networks, it became obvious that the focus was on objectively measurable network and service performance factors. This is evident by the definitions used at that time by the ITU-T and the IETF.
Some years later, the need to reintroduce user-centricity to quality assessment emerged and could no longer be ignored. The first terms that were used at that time to reflect the new trend were Subjective QoS and User-Perceived QoS. The current term, Quality of Experience (QoE), was originally introduced by Aad van Moorsele in 2001 in the context of web-based services. In his paper, he clearly distinguishes QoE from QoS, by stating that QoE has subjective elements to it, while QoS does not. From that point forward, the notion of QoE was quickly adopted to other domains, such as mobile communications and audio/video quality assessment and has since taken off to many more.

1.3 Definition of QoS

The QoS metric is defined by ITU-T as "*the ability of a network or network portion to provide the functions related to communications between users*", while IETF uses the following: "*a set of service requirements to be met by the network while transporting a flow"*. It is a totally intrinsic term used to describe the purely technical capabilities of a network. Essentially, QoS describes the technical specifications that the user can expect and lets them decide whether or not they fulfill their needs, implying that each user has a basic understanding of the features of the network. These specifications often include bit rate, throughput, transmission delay, availability, jitter, packet loss and so on.
Although QoS has been widely used to test the performance of networks for decades, the need for a more user-centric approach has been obvious and has recently come to fruition in the form of QoE.

1.4 Definition of QoE

ITU-T defines QoE as "*the overall acceptability of an application or service, as perceived subjectively by the end-user*". Simply put, it describes whether a subscriber is annoyed, delighted or anything in-between from a certain service or application within the network. In other words, it depicts a level of satisfaction of the user that is the result of a number of factors. Firstly, the infrastructure and technical factors of the network contribute massively. The quality of the source signal is significant but then each system element affects the overall performance, including the servers, nodes, equipment, codecs, techniques, protocols and terminals. Secondly, the QoE score is influenced by subjective factors, mainly environmental, psychological and sociological. These include user expectations, experience with similar products, pricing policies and characteristics of the particular location where the service is received. Those are crucial and completely overlooked by the QoS. Thirdly, the price the customer paid for a specific service is always in the back of his mind when judging it. Users tend to accept some quality degradation if a service is free but raise the quality standards and get easily disappointed by a service that costs them money.
As a conclusion, QoE takes into consideration both objective and subjective factors. Objective factors can be technical such as end-to-end network quality, service coverage or terminal functionality or non-technical such as ease of installation, service content, customer support. Generally, the meaning of the objective categorization lies within the network provider's ability to control them. On the contrary, subjective factors cannot be manipulated and vary widely among the users. These subjective factors include user emotional status, experience and expectation. Another categorization of these factors, both objective and subjective, is whether they are relatively stable during a user 's session or are dynamically morphing.

1.5 Dependencies

Four type of dependencies can be detected when examining QoE, each of which affects the final result in a different way [1].

User Dependency

Even when the service provided is identical to several users, the reaction it gets usually differs among them. This can be attributed to three, mainly, reasons. Firstly, the users may show different preferences towards their sessions established over the network. Some factors are more important to some subscribers than others. Secondly, due to major variations in subjective emotion, experience and expectation, users may evaluate services with the same objective QoS, much differently. Finally, the aforementioned subjective factors are not even concrete over the course of the service but can change from session to session.

Application Dependency

Each application has a different impact on the user 's QoE score. This happens because users rank their applications in regard with their respective significance, differently. Moreover, these applications may have diversified technical requirements. Generally, the applications that are mostly provided through a network can be classified by the means they are transporting as data, voice and video. Voice applications are more sensitive to delay and jitter than data while video applications are focused on transmission rate and perceived resolution. This is of course a simplified classification and many more subcategories with unique requirements need to be considered.

Terminal Dependency

Nowadays most applications can be utilized in numerous terminal devices. Each device though, has its own technical capabilities such as resolution, color or screen size. The technical limitations of the devices may blur the perceptual difference between network provisioned functionalities and terminal-enabled ones. This can happen either in the form of higher QoE evaluation due to a high-end device that satisfies the user, or the dissatisfaction of the user over the fact that they cannot fully take advantage of the capabilities of a device due to insufficient QoS from the network.

Time Dependency

Some factors that contribute to the overall user experience are time sensitive and are virtually impossible to control. Transport functions and application-layer configurations may be unable to control newly-fluctuated subjective factors. Furthermore, a decline in signal strength, due to multi-path propagation or shadowing, greatly affects the user's satisfaction, resulting in a lower QoE.

1.6 Assessment

It has already been established that the overall quality of the users' experience is influenced by a lot of factors and is, ultimately, subjective. The way to measure the degree of their satisfaction -or lack thereof- can also be divided in two categories, subjective and objective [5]. It is important to make clear at this point that both sets are used to quantify the same experience as it has occurred from both the technical aspects of the service provided, as well as the personalized criteria of each user. The difference between them is the means that are being utilized in order to reach a final result that can sufficiently depict the QoE of the subscribers.

Subjective Quality Assessment Methods

The subjective assessment methods imply that the data are gathered directly from human assessors. These test subjects contribute to the final result in a number of different ways.
The most frequent case is the gathering of a group representative of users, placing them in a controlled laboratory and exposing them to various stimuli. This is not an easy task as careful planning needs to precede the experiment. A lot of variables need to be controlled, measured and monitored. For starters, the selection of test subjects ought to be on point, meaning that there has to be diversity in regards to gender, age, race, social status, financial status, and many more sociological and psychological factors that reflect the target audience as well as possible. Secondly, any malicious intent from the selected group needs to be detected and dealt with. This can either mean dishonest testers who only care about the paycheck or people who genuinely are not able to complete such a test. Some ways to counter these potential problems are task performance (quality and speed of goal completion), physiological indicators (heart rate, skin conductance) or general user behavior (cancellation rates, viewing time). Finally, the testers need to be subjected to a number of quality levels which lead to some form of explicit or implicit response.
After the test subjects have been exposed to network services with various levels of quality, they assess their experience with the product or service in a pre-determined scale of satisfaction. This typically consists of five grades: “bad”, “poor”, “fair”, “good”, “excellent”. The aforementioned absolute scale is the most popular one, although comparative metrics are sometimes employed as well, when searching for the better of a number of options. Afterwards, the mean opinion score (MOS) is calculated among all approved participants which expresses the average verdict of the group, regarding how a specific factor influences the overall satisfaction of the potential customer.
Another method that has evolved in recent years is crowdsourcing. In crowdsourcing, test participants around the world are hired online and evaluate a particular service from the comfort of their home. This results in a much broader group of people, both in numbers and in range. Crowdsourcing is a very interesting development and will be thoroughly examined in a following chapter.
Apart from these tests, another method is utilized sometimes towards the same goal. Most networks providers provide a platform where subscribers can rate their experience with various services or applications whenever they want. The main problem with this approach is that users tend to overlook this specific feature unless they have recently experienced non-satisfactory network performance. In other words, a user is far more

likely to spend time to report their dissatisfaction with an aspect of the network, than a delighted user who has not encountered a single problem. This obviously leads to lesser than concrete results.
As a conclusion, the subjective quality assessment methods have the critical advantage of including the human aspect to the final result, meaning that they take into consideration psychological and sociological factors. They are considered the most accurate way to measure perceived quality from the subscribers' scope. On the other hand, they require a lot of planning in order to yield credible results and are often highly expensive and time consuming.

Objective Quality Assessment Methods

The purpose of the objective quality assessment methods is to bypass the feedback given by human users, by simulating their behavior. They are designed to automatically predict QoE at high accuracy by algorithmically assessing the level of quality of a service, given a set of current network parameters. Their success rate relies upon closely resembling the results of their subjective counterparts. As a result, a critical component of any such method is the derivation of quality models that map quantifiable factors to predicted MOS scores. These models are constructed using data provided by actual subjective methods and then formulate the necessary model functions that are required to provide an optimum fit with human quality perception. A number of possible classifications of these approached can be made [3]. The major ones are listed below.

- **Targeted Service** focuses on the service type of a network (VoIP, IPTV, Web, mobile).
- **Model Type** utilizes a reference signal in order to reach a conclusion. This class can be further divided to full reference (FR) metrics, which need both the original source and the transmitted signal of interest, no reference (NR) metrics, that take into consideration only the output signal and reduced reference (RR) takes into account just the input signal, or partial information. Practically, when designing a network, only NR models can be used, due to the fact that no signals are available during that phase. On the contrary, FR models are typically used in laboratory settings where the target is high accuracy and all signals are available for monitoring.
- **Application** divides the objective quality assessment methods in regards to which exactly aspect of the network is being tested.
- **Model Input** emphasizes on the parametric description of the processing path, additional payload information from bit-stream and reconstructed signal.
- **Modeling** approach examines the differences between psycho-physical and empirical methods. The first ones describe and model the human perceptual system whereas the latter ones are based on extracting characteristic system features by conducting experiments.

It is worth noting at this point that despite the philosophy of every category of methods, the input data that are required to each method are crucial and comes in three ways. The methods can either operate on signal-based models which assess the quality of a signal and comparing it to a reference, or on parametric planning models which predict quality by taking into account the planning parameters of the technical system, or finally packet-level and bit-stream models that are mainly used for monitoring purposes and are based on data extracted from the bit-stream with little or no decoding.

Concluding, the objective quality assessment methods are cost-effective, can occur in real-time but are heavily linked with the QoE functions that model human behavior which are currently not very accurate.
The verdict of the subjective vs objective quality assessment methods is that when subjectively-oriented models are available, they should be prioritized. The objective ones should most of the times be used in a supplementary capacity, reinforcing the results of their subjective counterparts.

1.7 A Conceptual Framework

In this chapter, a framework proposed in [2] is presented. It attempts to implement the QoE philosophy in mobile cellular networks, including its building blocks, their functionalities and interactions. The core of the framework is in the form of a central QoE management entity, strategically placed in a central location in regards to the network. Logically, it is designed to be a level above the network itself meaning it is able to send and retrieve specific information in order to communicate with it, measure its performance and modify it accordingly. It consists of three main building blocks, strongly connected to each other, which will be thoroughly described next [img1].
The QoE controller is basically the interface between the aforementioned entity and the network that is under evaluation. It materializes a communication system between them that allows a smooth flaw of data in both directions. The controller is in charge of the data acquisition process which translates to decisions regarding the nature of the input collected. There is a variety of possible ways to achieve this, such as collecting from various distributed nodes located in the core and access network that capture service degradations, or from agents installed locally at end-devices that capture more subjective QoE influence factors. The type of data that is being collected is also crucial to the final result. This can be raw network data, real time measurements, statistical information or information at the operator's possession that can be obtained through a number of sources such as probes on distributed network elements, sensors on user devices, user feedback and possible databases owned by the operator. In regards to the communication in the other direction, from the controller to the network, there should be a capability that allows for dynamic configuration of the data generation by manipulating the sources of the input. An important aspect of this feature is the periodicity adjustment of the data collection when necessary, as it greatly affects the total overhead and the timeliness of the input.
The QoE monitor is in charge of estimating the quality of experience per user's session. This is possible when some specific steps are followed. To begin with, the data that has been collected by the QoE controller is examined and the monitor classifies it to a respective category that complies with the subscriber 's usage, basically determining the type of the application to evaluate. Then, a QoE estimation model is applied in order to measure the quality of the particular service at hand. Obviously these models are the most crucial part of the whole system as they quantify the actual QoE of each user for each type of service. They have to be very specialized and accurate in order to produce credible results, comparable to the conclusions of human testers. The models are generally classified based on their evaluation method, as media-layer models, packet-layer models, parametric models and QoS-to-QoE mapping models. At this point it is important to mention that all models are installed in the framework before it runs for the first time.

Consequently, they have to be expertly designed beforehand, using some of the techniques mentioned earlier in this paper. A plethora of such models that mimic human behavior can be found in standardization bodies' recommendations.
The QoE manager is the third main building block and is responsible for conducting any practical modifications to the network parameters with the purpose of maximizing the overall QoE of the users. It takes into account the input provided by the controller, the calculated QoE scores through the monitor and some operator-specific information such as network policies or service level agreements. The next step is the calculation of effective measures that when imposed, several quality problems will be faced and dealt with. Decisions are taken per flow or catholically respecting user preferences and network resources management. The main advantage of this whole framework is that up to this point, there was not a viable way to evaluate the offered QoE in real time, but all assessments had to happen previously, in the design phase. But now, with this approach, the possibilities are eye-opening. To begin with, the QoE manager should be able to record and monitor real-time quality estimations per session. This can assist operators in better comprehending and managing the long term satisfaction of their subscribers, as well as in offering more personalized services that can undoubtedly increase their loyalty. Another innovative feature is the improvement of a current flow or the optimization of the sum/average QoE of all users. A quality improvement may be requested proactively or reactively. In the first case, this means predicting network problems via QoE-based alarms and in the second, handling problems that are already present. The manager 's capabilities extend to the point that it can track the effectiveness of these decisions and therefore adapt itself with the purpose of self-improvement over time. To finalize, using the QoE manager efficiently can have a positive outcome to the network economy, saving potential resources from getting wasted. There are a lot of cases where the perceived experience of a subscriber would not be lowered if less network resources were to be offered to them. This practically means that with real-time measurement of each user 's experience and the ability to quickly adjust some network parameters, a more acceptable balance can be achieved and the average QoE will certainly spike.

1.8 QoS vs QoE

At this point it should be clear what QoE represents, its advantages and drawbacks. But the question remains: is it the obvious choice? Although it has been established that QoE represents a more interactive, user-centric and ultimately accurate depiction of the users' degree of satisfaction than its counterpart, the importance of QoS should not go unnoticed. In most cases, both notions come hand-to-hand and acceptable QoE is the byproduct of high quality QoS. Of course this does not apply on all cases, as different applications require different technical aspects of a network and each user rates their experience in their own personal way and in regards to their own standards. However, a stable, high-end QoS usually translates to a good overall QoE.
The parallelism that can be made is that of a tree [7]. Quality of Service would be the trunk and branches of said tree: a sturdy foundation and intricate network through which the nutrients/packets are transmitted. Quality of Experience on the the other hand, represents the leaves. Although they might not be as big or imposing as the trunk, the leaves are the source of photosynthesis and if the quality of that process is not good

enough, the tree will not get the nutrients required to sustain itself, regardless of how efficient the trunk is or how expansive its network of branches is.
To finalize, if a short conclusion needed to be made about their relationship and degree of accuracy, it could be this: They are both significant, they compliment each other when used in coalition but the ultimate goal of the network should be the highest possible level of satisfaction for their customers and this is only depicted sufficiently by their Quality of Experience.

2. CROWDSOURCING

2.1 Introduction

Crowdsourcing is a modern term coined in 2006 and is defined by the Webster dictionary as *"the process of obtaining needed services, ideas or content by soliciting contributions from large group of people and especially an online community, rather than from traditional employees or suppliers"*. It has emerged recently as the most interesting method of subjectively assessing the users' perception in relation to a network and its features. Crowdsourcing is used to evaluate the degree of their delight or annoyance, aka Quality of Experience, when remotely using various services and applications. Specifically, this method encourages people to conduct subjective tasks over the internet, calculates their scores and reaches to useful conclusions. The output of this process is a series of scale values, assigned by the test subjects (workers), that depict their reaction on a set of stimuli, all varying to their underlying attributes. It basically expands the old-fashioned experiments conducted in labs, in the sense that the number of test subjects is exponentially increased and their physical presence is no longer required.

The advantages of this evolution are significant [8]. Firstly, the much broader group of testers automatically results in more accurate mean scores as is imposed by the laws of statistics. The test subjects are not only greater in number, but in diversity too. International, geographically distributed users provide results that are not tied to a specific location but are more universal. This is a major challenge for smaller-scale, supervised experiments in the lab where the end results are often considered location-specific. Another factor that favors crowdsourcing is the environment in which the tests are conducted. The test subjects are not forced to participate at a specific time or place, nor do they feel that they are being watched for the duration of the experiment. They choose when to take the test from the comfort of their homes and this is key to the direction of simulating realistic conditions, as external factors come into play such as state of mind, mood and possible distractions. Last but certainly not least, the reimbursement costs of the participants is way lower than the respective one for the lab test subjects which makes this method more cost-effective.

Naturally, crowdsourcing faces a number of challenges as well. The transmission of the tests themselves is not a trivial process and a lot of practical matters occur. The users' terminals may not be on par with the requirements of a test or some participants may not fully comprehend a portion of the experiment and rate something other than asked. Other factors should be considered as well such as the influence of incentives, payment schemes and general dishonest behavior. All of the above contribute to results that are not completely reliable and certain counter-measures ought to be employed. These include specific strategies in the test design as well as in the actual test campaign, while statistical methods are required to identify reliable user ratings, eradicate the noise and ensure the highest possible quality of the data.

2.2 Proper Techniques

Crowdsourcing is a relatively new concept and an experimental field. However, it gained friction quickly and thousands of test campaigns have already been completed. Some research towards the optimization of the process has been published, recommending specific strategies and techniques to be taken into consideration when designing and executing an experiment using crowdsourcing [6]. A summarized list of these recommendations follows.

Web-based implementation

Although crowdsourcing tests can be developed in virtually every programming environment, it is strongly recommended to be presented in the form of a web page or application. In this way, no extra software is required to be installed by the participants, which would be a discouraging factor. It has been discovered that when elaborate software is needed to interpret the test data, the number of test subjects willing to make the effort decreases significantly, as opposed to conducting the tests via their web browser. What is more, they might express their dissatisfaction over the fact in their ratings and thus create unreliable results. Another important advantage of the web-based implementation is the easy updating of the tests. Whereas a stand-alone executable would need to be deleted and then replaced by its successor, requiring extra installation, a web page can effortlessly be updated to accommodate the most recent changes made by the developer, with users not even noticing.

Simplicity of the questions

Due to the nature of crowsourcing, direct contact between the designer of the test and the participants is not easy. Thus, any confusion over some of the questions is not easy to be clarified and must be avoided. This means that relatively simple vocabulary should be used and the context of each question should be definite, not subject to interpretation. It should be noted that people from different backgrounds and education levels take part in the tests in an unsupervised environment, so it would be wise to facilitate them all.

Duration of the task

Workers in crowdsourcing platforms are able to choose from a large variety of tasks to participate in, making them a lot more selective than test subjects in a lab. The main factor in their choice is the duration of the tests or the time/payment ratio. Most users get frustrated with long, tedious tasks, opting either to not participate in them, quit without finishing them or in the worst case scenario, hurrying to complete them without paying the proper attention. The current rule of thumb states that a task should be about five minute long in order to avoid upsetting the users.

Inclusion of training sessions

It has already been established that workers who participate in crowdsourcing experiments, do it in an unsupervised fashion. Therefore, some training is required before

the task at hand is tackled, that can ensure better understanding of the philosophy of the test, they way it should be perceived and executed. This may come in the form of specific guidelines at the beginning of each task, a demonstration of frequently asked questions answered or some practice before the actual test takes place. In this way, the workers are more prepared to take part and the results they yield have proven to be a lot more trustworthy.

Integration of a feedback channel

Even with the aforementioned ways of reducing the possibility of confusion of the test subjects, there could still be some questions regarding the interpretation of some of the content. As a result, the need for a communication line between the workers and the experimenter is fundamental in order to achieve optimum results. The channel should fulfill some requirements in some aspects of the communication including bidirectional messaging, the ability for a question to be transmitted during the test and not exclusively after completing it and the reservation of all user rights as provided by each crowdsourcing platform, especially privacy.

Event logging

An efficient way to evaluate whether a task was performed smoothly or not, is to integrate automatic event logging. On one hand, the logs indicate the devotion shown by the participants by measuring user behavior for the duration of the task such as clicking patterns, switching of tabs, windows resizing and others. Usually, when test subjects are not fully committed, the respective task results are deemed irrelevant and discarded. On the other hand, it is useful for assessing hardware or software specifications during the test which could possibly affect the user experience in an indirect way.

Reliability checks during test design

Unfortunately, the essential anonymity that accompanies crowdsourcing can be exploited by the participants. Sometimes, in an effort to maximize their income, they tend to be sloppy or cheat when carrying out a task, so they can move on quickly to the next one. Consequently, an estimation of trustworthiness is required and it comes in the form of reliability checks that occur during a test is taken. These tests include verification tests to exclude automated bots from participating, consistency tests that assure workers do not answer randomly, content questions about the test and others. When malign use is detected, the results of this particular user are rejected and the user should be penalized.

Reliability checks during the test

Similarly to the previous, a variety of real-time checks can be performed during an actual task in order to quantify the degree of focus of the workers. These checks are not related to the content of the task, but are universal. They are applied to distinguish honest, devoted testers from untrustworthy, random clickers. Some of these checks appear in the form of questions about invisible elements appearing on the screen, reversed rating scale and more.

Reliability checks after the test

After a task is completed, the answers provided by the participant are compared to the mean scores calculated by the whole community and if found to differ by a factor greater than a specific threshold, they are considered to be the byproduct of dishonest use and are rejected. Another way to judge whether a set of results is produced after meaningful participation or not, is the task execution time. When the time spent on a task is way less than the average, it usually means that the respective worker skipped some portions rather quickly, whereas in the case of it being way more than average, it probably translates to users being distracted by off-screen stimuli.

Adaptation of research from lab testing

Although there are major differences between supervised lab testing and large-scale crowdsourcing, at their core, both methods attempt to quantify user perceptions and preferences. As a result, most of the literature published in order to optimize the testing process in the lab, can also be applied to crowdsourcing. The main challenge in this adaptation is the time factor, in the sense that experiments carried out in laboratories are way longer in duration than those performed in crowdsourcing platforms and the splitting of the tasks is not a trivial process but requires careful modifications to the original algorithms.

Appropriate use of scaling

Most answers that depict the degree of the workers' satisfaction of a stimuli, come in the form of a rating. Some problems may occur with the scale when it is not carefully designed. Firstly, the test designer should be aware that participants only rarely rate an experience with fringe values, either positive or negative. The amount of options in a scale should be generous but not confusing and most critically, when words are used rather than numerical values to describe a test subject's perception, it should be abundantly clear what each term represents.

Balancing the monetary incentives

Workers globally are mainly motivated by the reimbursement they get for their efforts. Payment is crucial to them and is the main criterion by which they choose tasks to carry out. However, it has been proved that larger payments attract more unreliable users whose data is eventually rejected due to one of the reasons explained earlier. Furthermore, tasks with a significant reimbursement tend to be extremely popular, causing massive participation in the early stages and are sometimes completed before people from other time zones have a chance to notice them, limiting the global aspect of crowdsourcing. On the other hand, extremely small payments make the tasks undesired which in many cases fail to be executed by the desired amount of workers. To summarize, increasing or decreasing the payment for a task by a large factor is often counter-productive and should be avoided.

Motivation of the workers

Although, as already stated, payment is the most important incentive for the workers, it is not the only one. Research has shown that test subjects often take part in the experiments in order to kill time or have fun. Therefore, an intelligently designed task is much more likely to attract casual workers who most of the times are the most reliable ones. Due to the fact that the monetary reward is of secondary significance to them, they don't rush or click randomly during the tests, making them especially wanted.

On the next chapter, several crowdsourcing platforms are going to be examined and showcased. The emphasis will be on whether or not they are suitable to accommodate QoE-related tests, their unique features, strong and weak points.

2.3 MicroWorkers

Official Description

“MicroWorkers is an innovative, international online platform that connects Employers and Workers from around the world. Our unique approach guarantees Employers that a task paid is a task successfully done, while Workers that successfully complete a job get paid.
The tasks assigned to Workers and paid for by Employers are simple and quick, mostly completed in a few minutes, thus they are called "microjobs”. These tasks include simple sign ups, social bookmarking tasks, forum participation, website visits, rating contents, adding comments, suggesting leads, creating backlinks, writing reviews or articles, downloading applications, testing websites and so much more.
Joining MicroWorkers is free, and as an international site, anyone from any country can be a member.”

Introduction

MicroWorkers is one of the most popular crowdsourcing platforms. The person willing to set up a test campaign is defined as an employer, while people carrying out tasks in said campaign are called workers. However, when registering to the site, no such distinction has to be clarified, as anyone can be at the same time an employer and a worker on unrelated campaigns. Double profiles are strictly prohibited and the procedure used to prevent it is the dispatch of a personalized PIN code via mail to every registered member's home.

Notes On Workers

MicroWorkers is particularly user-friendly and the browsing from the workers' point of view is fairly easy. Every time a worker logs in, they can select the “Jobs” tab and immediately, all available tasks will appear [img2]. Each job has a name, the potential payment amount if successfully completed, the success rate so far, the number of days needed for the employer to assess a completed task as a success or a failure, the estimated number of minutes needed for a worker to carry it out and the percentage of already successfully

completed instances. It should be noted that some jobs have specific requirements, usually regarding geographical aspects, but are invisible to users who are ineligible to execute them.
Some filters are available so the participants can easily detect tasks of their preference. Firstly, there are four main filters that sort all tasks in relation to various properties. These include the potential payment to be awarded, how recently they have been created, the current successful completion ratio and the time required for the pass/fail rating from the employer to occur. Secondly, the tasks are classified in numerous categories regarding the nature of the required actions by the workers. Some indicative categories are google, youtube, sign-up, bookmark, facebook, twitter, write a review, surveys, mobile applications, among others.
After applying the filters of their choosing, the workers may select a task that appeals to them, read the detailed description and choose whether they are going to perform it or not. Every task states clearly the necessary steps that need to be followed if a completed test is to be accepted by the employer, as well as the essential proof that needs to be provided by the worker [img3]. If the terms are acceptable, the worker declares that he will carry out the task at hand and a form appears below ready for their input of proof. Proof of completion often comes in the form of a url that appears in the latter steps of the task, or a screenshot with crucial information, or some username created in the process.
After the participant has entered the required proof, the task is considered completed and they can move on to the next one. The outcome of the employer's assessment usually becomes known after a couple of days. If the work put on by the worker is considered satisfactory, the agreed upon amount of money is transferred to the worker 's account and their personal rating rises. All the information regarding payments can be accessed in the “Tasks I finished” tab [img4]. Workers generally must have a rating in excess of 75% in order to perform the majority of tasks offered. They can withdraw their earned money to their bank account at any moment, providing that the amount to be withdrawn is over 9$.

Notes On Employers

When creating a campaign, employers have to make a number of decisions regarding various available features, so that the final result will reflect their preferences as well as possible. The primary one is whether their campaign is a “basic”, or a “hire group” one. The first one indicates that all workers are eligible to participate, with the exception of possible geographical restrictions, while the latter implies that certain criteria take place in order to filter workers and assign tasks only to able ones [img5]. These criteria include overall rating, profession, certified qualifications and others. Generally, the hired group type of campaign is applied when the tasks are specialized and a certain expertise is required in order to be successfully completed.
Another important decision is which template is going to be used for the test campaign presentation. Employers can choose among a number of available templates or even create their own using HTML source code, in which case it is possible to maintain their brand and logos. Another possible course of actions is to modify an existing template and make it more personalized with the editor provided. Most employers opt for a predefined template due to cost efficiency and the fact that their variety and specialization is adequate [img6]. It is worth noting that there is a functionality to preview every custom template before selecting it.
In order to achieve the desired effectiveness, several methods that ensure worker

dedication can be applied. These include certain captcha questions that detect random replies and sets of predefined answers to questions, to exclude random typing. Data verification can be two-way, being performed by both the respective employer and a MicroWorkers administrator [img7]. It is possible to reward an extremely dedicated worker who carried out even more workload than required with a bonus, during the rating phase.
Every aspect of the campaign is, of course, customizable including the potential payment, the number of completed tasks required, the time to rate, the task instructions and the proof needed from the workers' part. There are specific guidelines for acceptable tasks in regards to what is asked. For example, writing articles or reviews, promoting in social media, commenting in forums, signing up, linking to websites or voting in contests are allowed but disclosing personal information, using credit cards, creating fake negative reviews or sending material to personal e-mail accounts are prohibited. Every campaign must gain approval from a MicroWorkers administrator before it goes live.
Even during the actual testing phase, employers are still under total control of their campaigns, in the sense that they can pause and later resume them, cancel them and only pay for completed tasks at that point, or adjust the desired number of executed tasks. Rating the workers' efforts also begins simultaneously with the submission of proof by the participants and it can be done individually or in groups, which is usually the case.

2.4 CrowdFlower

Official Description

“At CrowdFlower, we break down large-scale projects into microscopic online Tasks to be completed. In other words, through the power of the Internet, we connect companies that have work to be done with people who are looking for work and opportunities to be compensated. So far, we have completed over one billion Tasks by five million Contributors from all over the world! CrowdFlower was founded in 2009 in San Francisco, CA, USA. Founded as “'Dolores Labs” in 2007, CrowdFlower made its public debut at TechCrunch50 in 2009 and was a finalist for the TechCrunch50 award.”

Introduction

CrowdFlower is possibly the most commonly used crowdsourcing platform on the Internet. Subscribers to the site are divided during the registration phase to Customers, who are companies or individuals willing to create task campaigns and Contributors, who are people looking to carry out tasks in exchange for monetary compensation. The sum of both parties is defined by the administrators as the CrowdFlower Community. User interfaces are considered very easy to use but the process of browsing for support could be further optimized.

Notes on Contributors

After logging in, a Contributor needs to connect their CrowdFlower account with their Facebook one before starting executing tasks, as a measure of extra validation of their personal information. This is a mandatory step that may force users who don't have an

account on the social network to create one if they want to proceed with working in CrowdFlower. When they have done so, they can browse available jobs through the “Tasks” tab where they are prompted to see available work on a new tab [img8]. Each task campaign is listed along with its attributes such as id, title, requirements, reward, maximum amount of completions and mean satisfaction rate.
A key feature of CrowdFlower is the leveling system depicting the efficiency of each Contributor. Every new signee is assigned with an initial level of zero value and is up to them to increase it by successfully executing tasks. When viewing jobs, there is a clear distinction between currently available jobs and potential ones, which greatly relates to the Contributor's current level. Customers who create the task campaigns select the minimum level that is acceptable in order for a Contributor to take part in their microjobs.
At first, only a handful of jobs are marked as available, due to the fact that most Customers require their Contributors to have reached at least the second level of efficiency. To counter this, CrowdFlower has implemented an entrance exam for participants who want to proceed from the zeroth level to the first. The entrance exam consists of a variety of questions, tests and other requirements that are often encountered in real test conditions and given that the Contributor's results are satisfactory, they are deemed capable of carrying out jobs and therefore greenlit for reaching level one. This exam does not offer any compensation but should be regarded as an essential investment for the future.
After achieving their first progression, users can start doing regular, paying microjobs in order to receive some money and perhaps more importantly, reach the latter levels where the high-profile tasks can be found. Level progression is related to three factors: number of questions answered, variety of job types selected and success rate. The maximum level that can be achieved is the fourth.
Except for the desired levels, another criterion for possible exclusion from a specific job is the geographical restrictions. Some Customers prefer to conduct their experiments locally or with a targeted audience originated from a specific location, which makes the majority of the Contributors ineligible for the particular task. However, this happens rarely and almost every test campaign is distributed globally. It should also be noted that a number of available tasks can be taken multiple times, as indicated in their description.
Contributors can view the ratings of their work at the “Job History” tab, where the total amount earned is also presented [img9]. CrowdFlower uses the services of PayPal in order to manage their payments when they choose to make a withdrawal.

Notes On Customers

After signing in, Customers are immediately prompted to create a new task campaign by selecting its type [img10]. CrowdFlower gives extraordinary emphasis on campaign classes and it provides detailed information and guidelines for each classification. A brief description of these classes follows below [img11].

Sentiment Analysis

The purpose of crowdsourced sentiment analysis is to gain feedback for any kind of content. It has been proved that human test subjects respond quite differently to certain stimuli than their automated counterparts, regardless of their level of sophistication. Artificial language processors often fail to identify sarcasm, misspelling and other subjective aspects of the human way of communicating. Therefore, workers are employed

to comprehend, interact and evaluate specific content that is designed to target other people, before it goes live to the public. Finally, this class also includes tasks that are created to support machine learning algorithms by training them in comparison to actual, precise user feedback.

Search Relevance

When developing a search engine, a lot of challenges occur that can make the end product inefficient. Results may not reflect the search terms or their order may not be appropriate, issues that disappoint the users and discourage them from further exploring the specific webpage. An effective solution is asking the crowdsourcing community to test and rate the search results during the development phase of the algorithm. The Contributors will rank how well the queries surface the best results while indicating possible limitations, through careful examination of a variety of input. Then, the search models can easily be fine-tuned and optimized by either per-result relevance, whole-page relevance or any other metric, with methods provided in the platform.

Content Moderation

Recently, the volume of data generated in certain domains is so massive that it is practically impossible to be thoroughly monitored. This is exploited by malicious users or software, with the purpose of damaging the image of a website. CrowdFlower offers a solution to this problem by employing Contributors who, given specific instructions, tirelessly scan for content and flag it, if it does not comply with the standards provided. This includes forum posts, comments, social media replies, profiles, text, audio and video. It is obvious that crowdsourcing works much better than the alternative automatic methods that for example could never trace a racey image.

Data Collection

Data is crucial to every company, whether it relates to clients, competitors or suppliers. CrowdFlower offers services to enrich that data and give its customers an edge in the market. This can be achieved by having the Contributors perform the tedious tasks of either examining existing data that may be outdated and in need of altering, or search online for desired information that can help a business grow. In both cases, the data can translate to sales lists, contact information, urls, business addresses, investor information and others. It should be noted that only public information is allowed to be asked for by the Customers.

Data Categorization

This class includes all tasks in relation to data categorization and its various forms. Since this process can be very time consuming for a company, CrowdFlower provides the necessary tools to achieve the desired effect through crowdsourcing. Capable Contributors have no problem going through existing data and classify it as required. Usual cases include image tagging with keywords, branch listing of corporations, service tickets management, business sites categorization or domain organizing. After a successful campaign, the results can be fed to machine learning applications that can

carry on similar tasks from that point on.

Transcription

It is virtually impossible for an algorithm to accurately transcribe an image or an audio file into usable text. Therefore, working force is required to take on such tasks and when the number of files to be transcribed is large, it is only natural to resort to crowdsourcing. CrowdFlower has specialized methods for transcription in its disposal and combined with its massive community of workers, the results are guaranteed to be satisfactory. After the Contributors have transcribed each file, the end product is calculated by taking into account what the majority of them extracted for every line of the text, thus discarding any noise that comes from misinterpretations.

The previously mentioned categories are some of the most frequent use cases of CrowdFlower. Each has its own subcategories and a variety of available templates for each one, while there is also a collection of successful examples that have taken place in the past. However, the platform encourages its Customers to contact an administrator if their crowdsourcing need does not fit in any of the above classes with the reassurance that a custom solution will be found and presented.

After creating the campaign and setting all the adjustable attributes, Customers can monitor the progress in real time. The fees to be paid are monthly in regards to the Contributors and yearly in regards to the platform itself. There is also an option for a pro package which provides some extra features.

2.5 Amazon Mechanical Turk

Official Description

“Amazon Mechanical Turk is a marketplace for work that requires human intelligence. The Mechanical Turk service gives businesses access to a diverse, on-demand, scalable workforce and gives Workers a selection of thousands of tasks to complete whenever it's convenient.
Amazon Mechanical Turk is based on the idea that there are still many things that human beings can do much more effectively than computers, such as identifying objects in a photo or video, performing data de-duplication, transcribing audio recordings, or researching data details. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce (which is time consuming, expensive, and difficult to scale) or have gone undone.”

Introduction

Amazon Mechanical Turk is an Amazon subsidiary company. As usual, specific terms are used to describe the familiar concepts of individuals carrying out tasks and the entities that design them. In this case, they are called Workers and Requesters respectively. The main difference with its competitors is the selective manner in which the registration process occurs. There is a number of strict standards that both parties need to fulfill in

order to be approved by an administrator after their registration. This process takes roughly 48 hours and the potential member is notified by an e-mail whether their application has been accepted or not. It is important to note that the criteria are not available for the public to view and consequently, most turned down users do not even know the exact reason that lead to their rejection. The purpose of this strategy is shaping the community according to what the administrators see as efficient but on the other hand, keeps its Workers number limited compared to other crowdsourcing platforms. To finalize, the registration process is separate for Workers and Requesters as different personal information is required for each type.

Notes on Workers

Every individual job on Amazon Mechanical Turk is called a HIT (Human Intelligence Task). After logging in, Workers can move to the “HITs” tab and view all available work [img12]. There is a plethora of filters that can be used to limit the hundreds of thousands jobs that normally exist on the platform. Workers can search for a keyword that is included in the job title, the amount of the potential payment to be earned and others. Every HIT clearly states its attributes which include category, Requester name, expiration date, maximum time alloted and potential payment. After selecting it, further information is revealed such as detailed description and required qualifications.
Qualifications are often a prerequisite for some jobs. They can be obtained by successfully completing the respective tests that are provided by Amazon Mechanical Turk to certify certain skills [img13]. After completing qualification tests, more advanced ones become available progressively that require a similar skill. For example, after successfully completing a transcribing challenge, Workers can move on to unlock the next level of the transcribing qualification, thus making themselves eligible for more works that require a high level on the specific skill.
Another way of making more HITs visible, is the mastering of a type of work. This is possible when a user repeatedly carries out jobs that fall under the same category and their success rate is exceptionally high. The combination of a plethora of successfully completed tasks with very high satisfaction rates from the Requesters' point of view counts towards a Worker being marked as a master in that line of work. Therefore, they can be assigned to do high level work which is a lot more rewarding that the regular ones. It should be noted that the master status, in contradiction with the qualification achievements, is not permanent and can be lost if the quality of the work declines overtime.

Notes on Requesters

When a Requester is willing to set up a new test campaign, they can use one of three options, depending on their level of computer programming expertise. Firstly, a user-friendly web interface is available on the website. It requires minimal technical capabilities and offers a pleasant experience with its graphical environment. Secondly, the intermediate level consists of a robust API, along with various SDKs that can facilitate a variety of programming languages. Finally, the command line tools represent the lowest-level task creation form on the platform. They target experienced programmers and operating systems experts who need exceptionally personalized results.
Available categories for microjobs are: data collection, image moderation, sentiment

feedback, survey, survey link, image tagging, transcription from audio/video, transcription from text and writing. Requesters usually merge a large volume of tasks to a project which can later enrich with more batches of jobs. During the campaign creation phase, apart from setting the usual attributes that every microjob has, possible qualification requirements are selected as well as possible exclusivity to master Workers.
A fact that sets this crowdsourcing platform apart from its competition is the permission to post HITs that contain explicit or offensive material. Certain notifications need to be added asking for user discretion before a Worker can accept the job but still, tasks of this kind are not allowed on other platforms.
Results from the Workers begin to flow as soon as they start submitting them, giving the Requesters the opportunity to assess and utilize them before the desired number of completed instances is reached. The payment is sorted out once a job has been accepted as successful. The fees added from the service consist of 20% of the payment to the participants and an extra 5% in the case of employing master Workers.
The main drawback of the Amazon Mechanical Turk crowdsourcing platform is that significant limitations affect Requesters residing outside of the United States of America. Some of them are lifted if a USA billing address is provided but it still causes a lot of frustration.

3. The Experiment

3.1 Introduction

In this chapter a specific experiment will take place that materializes the majority of the aforementioned techniques related to QoE and crowdsourcing. The purpose of the experiment is to sufficiently quantify the quality of experience of users in regards to watching a video clip on one of the major online video hosts. To this end, a series of videos were modified to simulate the online viewing experience with its limitations and then participants in a crowdsourcing platform were asked to watch, rate and compare them. The quality loss factors of an online video host are mainly two.

Firstly, the relatively new http adaptive streaming algorithm plays a key role in user experience when watching a video [4]. To explain briefly, this method converts any video uploaded by a user to several layers of descending picture quality and saves them all. When another user wishes to watch said video, the appropriate quality layer is served to them, taking into account the user's quality of internet connection. However, this can dynamically change during the video playback if a change in their network connection quality is detected. Of course, it can be either an improvement or an impairment which respectively translates to serving a higher or lower layer of the video. The maximum frequency of layer switches occurs every 2 seconds but usually the number of detections of change in the users' connection specifications is quite small. The users' quality of experience is affected by a variety of factors such as the number of switches, time spent on the higher layers and the sudden drops in quality of two or more layers at once.

Secondly, old-school stalling during video playback is still an issue nowadays. Albeit, it occurs much less frequently than a couple years earlier, due to the effectiveness of the adaptive streaming but should still be taken into consideration. A stalling event happens when a rapid change in the user's internet connection takes place which the adaptive streaming mechanism is momentarily unable to handle, or when the server is overloaded and its response is delayed for a few seconds, or when a user explicitly requests a higher level of video quality than his internet connection can handle. As a result, playback of the video freezes as the server looks for a solution which at most cases is the switch to a lower quality layer. The frequency of the stalling events during playback is equally important to their duration and the optimization of this combination is being researched in this paper.

To summarize, the experiment will consist of simulating the quality loss factors in a pre-defined manner on some video clips by using video editing, uploading them and then having test subjects of a crowdsourcing platform to watch them and answer a questionnaire regarding their viewing experience. The feedback should be a reliable quantification of user QoE. The following chapter is a detailed presentation of the video editing techniques and respective applications that were used to simulate both the http adaptive streaming switches and the stalling events.

3.2 Video editing

The crowdsourcing experiment that will be conducted in the next chapter requires excessive video editing. A series of video clips will be uploaded with various combinations of quality loss factors, mainly stalling events and unstable picture quality. In this chapter, the methods used to replicate the decrease of video quality are presented, along with the respective applications that were used, which are all free to download from the internet.

Downloading source video files

The first step of the process is to acquire a video clip in all possible resolutions that are offered by YouTube and other video hosts. To this end, YTD Video Downloader is proposed, which can be found at http://www.ytddownloader.com/download.html. After successfully installing the application, an ultra high definition video clip should be found and its url copied and pasted in the input box. Virtually every major video host is supported. Then, the desired quality of the video to be downloaded ought to be selected and the downloading procedure begins [img14]. This can be repeated for every quality of picture available, with the end result being a series of video clips with the exact same content but with very different technical aspects that affect the viewing experience.
For our experiment, a video clip showcasing the wild life in Costa Rica was chosen which can be found at https://www.youtube.com/watch?v=iNJdPyoqt8U. It benefits from YouTube's 4K resolution support for excellent picture quality and is offered in every other available quality as well. Consequently, all seven layers were selected for download so that maximum maneuverability can be achieved in regards to simulating the http adaptive streaming in latter steps. In our case, the video clip was locally stored in 7 layers of descending quality. The exact technical specifications for each layer are presented in the table below and depict the clip's total duration of 5:08.

Table 1: Quality Layers

Layer	Width (pixels)	Height (pixels)	Bitrate (Kbps)	Size (Mb)
1	3840	2160	15600	579
2	1920	1080	2801	108
3	1280	720	1515	61
4	854	480	783	44
5	640	360	396	20
6	426	240	242	14
7	256	144	140	9

Trimming and merging

In order to convincingly simulate a video that switches quality layers resembling the typical http adaptive streaming algorithm, firstly we need to cut small pieces of the video clip in

each quality level. For example, splitting each video file to smaller ones every n seconds would result in duration/n shorter clips that if played in succession would produce the original video clip. This process is carried out for all quality levels. Now, if pieces from several layers are combined and merged together in a single video file, it would seem that the picture quality changes through the duration of the playback, while the flow of the content carries on normally, which is the desired effect.
The proposed application for the aforementioned part of the procedure is Freemake Video Converter, located at http://www.freemake.com/free_video_converter/. The trimming process consists of opening the original video file in the program [img15], clicking on the scissors button on the right side of the user interface, selecting the parts of the video that need to be discarded [img16], cutting them in order to leave only the portion of interest intact, clicking ok to return to the main screen and then saving the trimmed video clip by clicking on the "to MP4" button and then selecting the quality to be same as the source and selecting two-pass encoding for better results [img17]. Consequently, there is now a video file stored locally with the exact same technical characteristics, but different start and end points than the source. Precision in timing is key when it comes to splitting video clips with the purpose of joining some of them together. The last frame of a clip should be the first one of the next, so that viewers do not experience frozen screens or disruptions at the critical points of intersection.
The merging process is significantly simpler. A number of video files are added in the application and the desired order is arranged. The "Join Files" button on the top right must be activated and the conversion is ready to begin [img18]. Using the same functionality as before, by selecting the "to MP4" button, same quality as the source video and two-pass encoding, a video file is created that is, essentially, the sum of the smaller parts trimmed in the previous part. Therefore, if the start and end points were carefully selected to not overlap and a variety of quality levels were used, the final video file should sufficiently simulate the http adaptive streaming technology with several switches between quality layers.

Audio editing

Regardless of the level of precision applied on the trimming process, when smaller clips are joined together, the background music is unfortunately momentarily cut. As a result, the transition between two clips is made obvious and the illusion of the incessant flow is ruined. The proposed solution for this challenge is to substitute the audio track of the produced video file with a clean one without disruptions. This is a two-part process that includes extracting the audio from a source video file and then adding it to the product of the merging process.
In order to save the audio track of a video file, another application is needed. Format Factory by PCFreeTime does it quite efficiently and can be downloaded from http://format-factory.en.softonic.com/. After installing and executing it, the Sound tab should be selected and the "-> MP3" button clicked. From that point, the source video file needs to be selected and opened, the start and end points for the duration of the audio clip established and the desired quality chosen. The next step is starting the conversion which produces a quality audio file with no disruptions and the duration of the user's choosing. For our experiment, the audio track was extracted from the video file with the best quality.
The next course of actions focuses on muting the audio component of the video file that was created by merging, which contains noise and then embedding the newly made audio

file from the previous step. These tasks can be accomplished with Microsoft's Movie Maker, which can be found at http://windows.microsoft.com/en-us/windows/get-movie-maker-download. After opening the application, the video file to be edited is added and then from the “edit” tab, the existing audio track can be muted with the “video volume” button [img21]. The audio track should be dragged and dropped next and the start and end points set from the “options” tab. To finalize, the video file can be saved by clicking on the “save movie” button and then selecting “for high-definition display” so that the picture quality does not drop due to the conversion [img22].
This concludes the predetermined http adaptive streaming simulation of the video clip. The visual part flows smoothly in regards to the content, while the resolution changes constantly in a controlled manner. The audio part was independently embedded to mask the incision points' distortions.

Adding stalling events

The simulation of stalling events is largely similar to the one of the adaptive streaming, in the sense that methods of trimming and merging again play a pivotal role in the process. The main difference is the use of static images to replicate the effect of the stuck playback. A description of the necessary steps towards that direction follows.
Firstly, the source video needs to be opened by any media player. When in full screen mode, playback should be paused at the exact moment when the stalling event is to be inserted. Then, a screenshot must be captured using the specific key on the keyboard and pasted on any image viewer. In our case, IrfanView was employed which can be downloaded from http://www.irfanview.com/. The image is saved with a 100% lossless quality .jpg conversion.
The next step is to split the source video in two halves, at the exact point where the stalling event is to occur. Freemake Video Converter is once again used for this type of operations. Once successfully split, the partial clips are added again in the same application with the purpose of rejoining them. However, in this particular case, the image is also added and placed between the clips [img19]. The duration of the image display can be set by clicking on the properties button on the right side and adjusting the time interval [img20]. Of course, no audio should be played during stalling, which is the default case of the program. The “join” button must be activated and the overall compilation saved as described previously.
Again, precision is of the utmost importance when splitting, capturing images and joining the parts. When performed carefully, the resulting clip should closely resemble a stalling event at some point as the frozen image is played for a number of seconds. The content of the image is the same as both the last frame of the first clip and the first one of the second clip, thus giving the sense of continuity content-wise.
The process for more than one stalling events is very similar. However, FreeMake Video Converter does not support this feature and Movie Maker should be used instead. Obviously, a screenshot is required for each stalling event and the respective images need to be strategically inserted in the compilation.

3.3 Implementation

Having already presented the technical aspects of the experiment, it is time to proceed to the specifics. Eight test cases were performed, where in each one a modified video clip was paired with another and the users were asked to compare and rate them. In each sub-experiment, the same loss of quality factor was applied to both videos because it was presumed that the opposite approach would be the cause of confusion and inconclusive results [7]. In other words, four of the experiments include two videos with different stalling events patterns whereas the other four include two videos with different but comparable switches between quality layers.

In every experiment, measures to prevent improper use by the participants' part were employed. The first question was always related to the content of the video clip, in order to ensure that they have watched them. The time that took every individual to complete the test was also measured and cases where the total time was less than the sum of the clip durations were deemed useless. Finally, the users were asked to rate both clips in a scale of one to five after having declared which video offered them a better experience, with the purpose of detecting random clickers from inconsistent answers. Only users who passed all three tests were paid and their results were taken into consideration. The rest set of answers were discarded at no cost and replaced by fresh ones which were again tested for proper use.

The MicroWorkers crowdsourcing platform was chosen to conduct our experiment for its simplicity and efficiency. The custom templates were also a factor that was taken into account. The experiments obviously fell out the “Video Quality Rating” category which is included in their “Content Moderation” class. The “Hire Group” option was selected in each case and the “Data Collection” group of workers was proposed by the platform and selected. It consists of participants that are proven to be able to perform this kind of tests. It should be noted that an experiment was open to public for testing purposes with highly disappointing results. Well over half the sets of answers were performed in a matter of a few seconds, they were random and inconsistent. Results coming from the aforementioned “Hire Group” however, were much more honest and reliable.

The videos were not uploaded to any online video host, because the possible added quality loss would ruin the desired predictability of the actual user viewing experience. Instead, they were uploaded on the server of the university using the “video” html tag and the “preload” argument to avoid further stalling events than needed. All clips were kept to a small size of less than 20 megabytes to ensure that the playback is smooth in most cases with a decent internet connection. This was achieved by cutting the clip duration and avoiding using the first two layers.

As already mentioned, each of the eight tests consisted of two video clips with a different version of the same deficiency, followed by a questionnaire. The duration of each clip varied slightly from 0:40 to 1:15, so that the whole test duration would not exceed the five minutes mark. A fair payment amount was calculated at 0.25$ for each successful completion and every test needed 40-50 reliable sets of answers in order to be terminated. Two different templates were created in the platform via their rich text editor for each category of video impairment, one for measuring stalling events dissatisfaction and one for assessing adaptive streaming layer switches. The only change between tests of the same category is the links which lead to different video clips. They are both presented in [img23] and [img24].

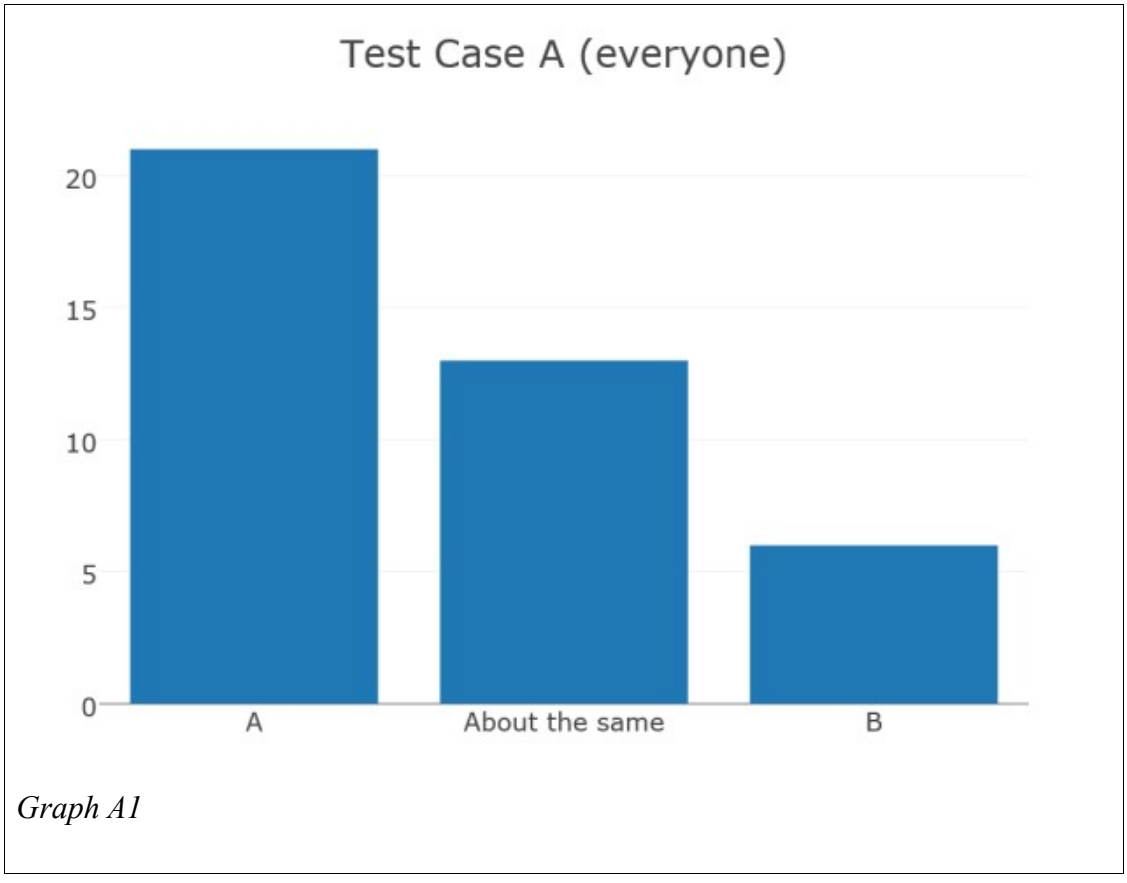
3.4 Test Cases and Results

Below, the test cases are presented, analyzed and their results explained. The first four are stalling based, while the latter four examine the http adaptive streaming effect.

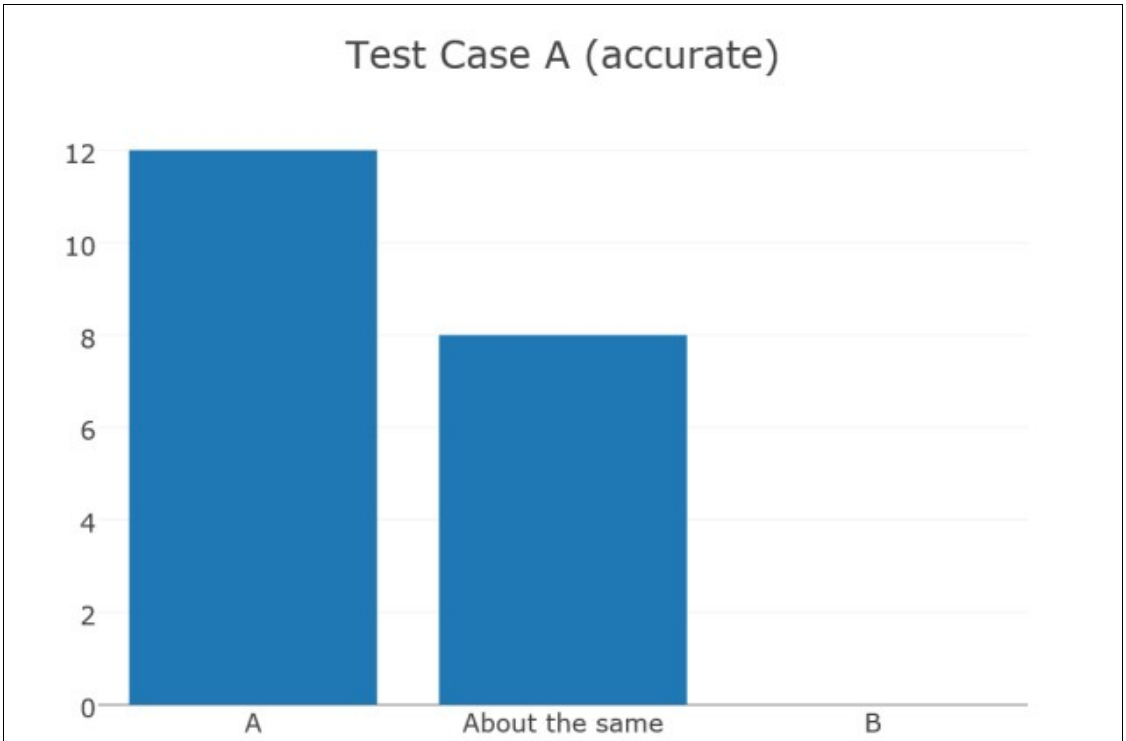
A) 1 stalling of 6 seconds vs 2 stallings of 3 seconds each

Links: http://gain.di.uoa.gr/kyr/s2b/qoe.html http://gain.di.uoa.gr/kyr/s1c/qoe.html

In the first case, we examine the quality of experience perceived by users when watching a video clip that includes a stalling event that lasts six seconds in comparison with another video clip which is impaired by two smaller stalling events of three seconds each. More specifically, the first video clip has a total duration of 71 seconds with the stalling occurring at 0:30-0:36 whereas the second has a total duration of 72 seconds with the stallings occurring at 0:18-0:21 and 0:49-0:52.
40 valid sets of results were required for the test to be completed. 4 unsatisfactory ones were detected and replaced with others. Only 20/40 of the workers answers correctly in the first three questions, which are the objective ones and their results are depicted separately due to their enhanced importance.
***Important note**: Two graphs are presented for each test case. The first one represents all acceptable sets of answers, meaning those that were the product of honest work. The second one demonstrates the results from the subset of users whose viewing experience was exactly as intended, as shown by their replies to the first three questions.*



Graph A1



Graph A2

Remarks

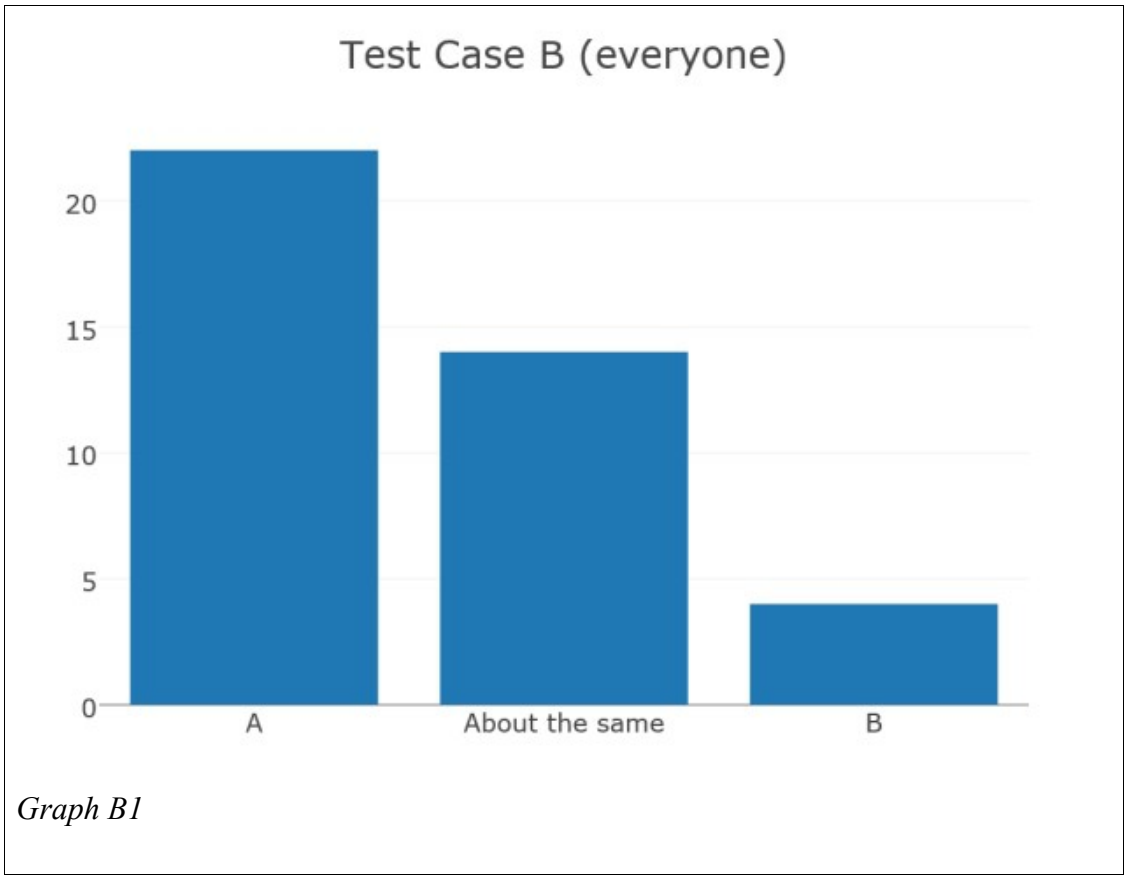
The preference towards the first clip is overwhelming. Users seem to get frustrated by the number of stalling events, rather than their duration. The second graph demonstrates that of all participants who experienced the right amount of stallings, meaning that none was added due to network issues, not a single one preferred the second video clip over the first.

MOS Scores: 4.075 for video A, 3.900 for video B.

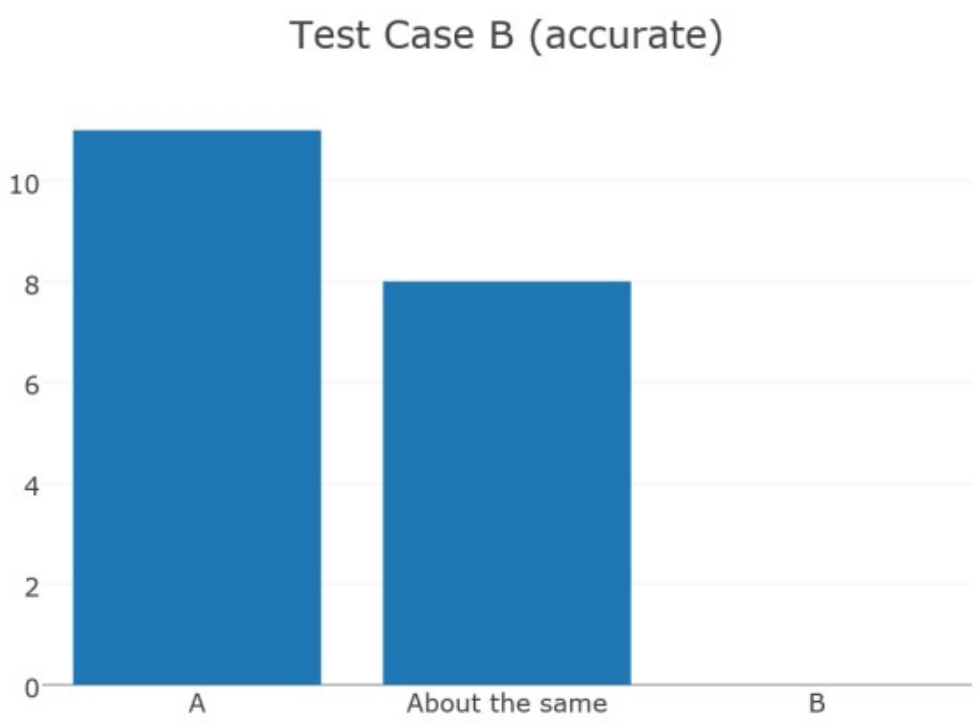
B) 1 stalling of 12 seconds vs 4 stallings of 3 seconds each

Links: http://gain.di.uoa.gr/kyr/s2c/qoe.html http://gain.di.uoa.gr/kyr/s1d/qoe.html

In this case, the users' quality of experience is measured in the event that a video clip has a massive stalling event of 12 seconds, while the other has an equal total amount of stalling split in 4 shorter ones. Specifically, the first video last a total of 76 seconds with the stalling event occurring at 0:30-0:42. The duration of the second video is 78 seconds and its stalling events can be noticed at 0:14-0:17, 0:37-0:40, 0:49-0:52 and 1:07-1:10.
40 satisfactory instances of work were required for this experiment and only 2 of the initial ones were by random clickers and replaced. Of the remaining 40, 19 were absolutely accurate in experiencing the desired stalling events and their results can be viewed separately below.



Graph B1



Graph B2

Remarks

Similarly to the first test case, the clip with the single, long stalling event prevails. Shorter, more frequent delays seem to irritate the test subjects way more. This is obvious from the graph which shows that every participant who experienced the desired experiment effect, either showed their preference towards the first clip, or thought their quality is about the same.

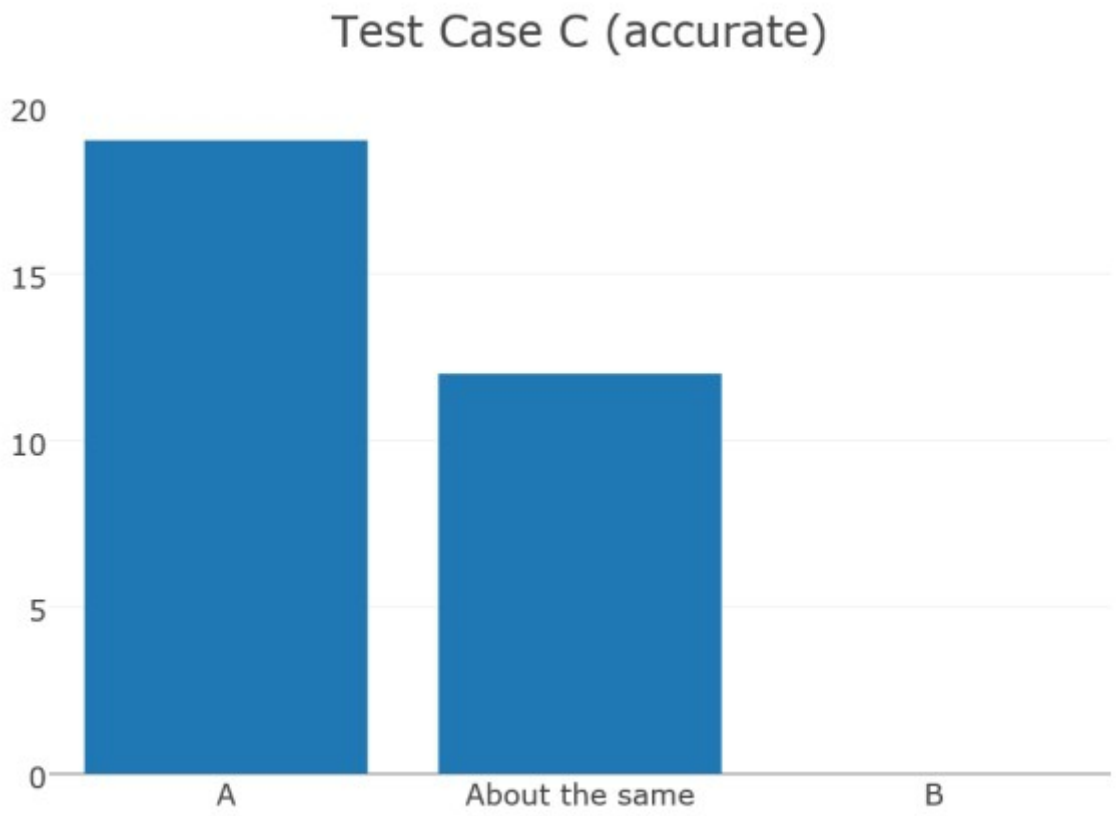
MOS scores: 4.075 for video A, 3.625 for video B.

C) 2 stallings of 3 seconds each vs 4 stallings of 3 seconds each

Links: http://gain.di.uoa.gr/kyr/s1c/qoe.html http://gain.di.uoa.gr/kyr/s1d/qoe.html

For this experiment, a video clip which includes 2 stalling events of 3 seconds was compared to another that freezes its playback 4 times for the same duration. Although the optimum choice between the two is obvious, the distance among them was measured through users' rating. In detail, the first video has a total duration of 72 seconds and features two stalling events at 0:18-0:21 and 0:49-0:52 while the second one is the same as in the previous test B.
Just one worker answered without paying the proper attention and after their replies were replaced by fresh ones, 40 acceptable sets were gathered and the test concluded. The vast majority of the workers who watched the videos had the desired experience as shown by the correct answers in the crucial first three questions, with only 9/40 declaring that the noticed the wrong number of stalling events.

Graph C1



Graph C2

Remarks

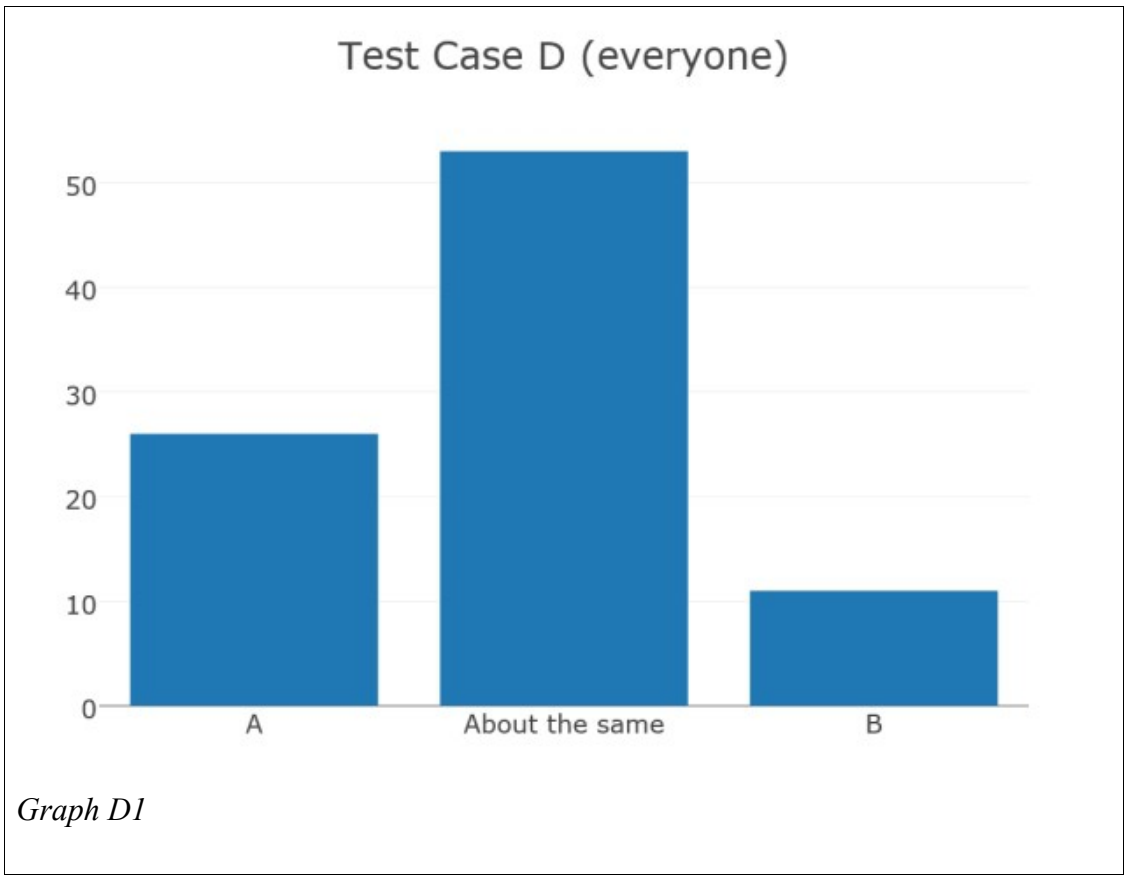
As perhaps predicted, the majority of people asked replied that the clip with the fewer stalling events offered the better viewing experience among the two. Less than half declared that the playback was similar in both cases, while not a single participant rated the second clip better than the first. Interestingly enough, accurate users produced the same results as the whole community, meaning that slight variations of the stalling pattern due to connection problems played little part in this case.

MOS scores: 3.85 for video A, 3.475 for video B.

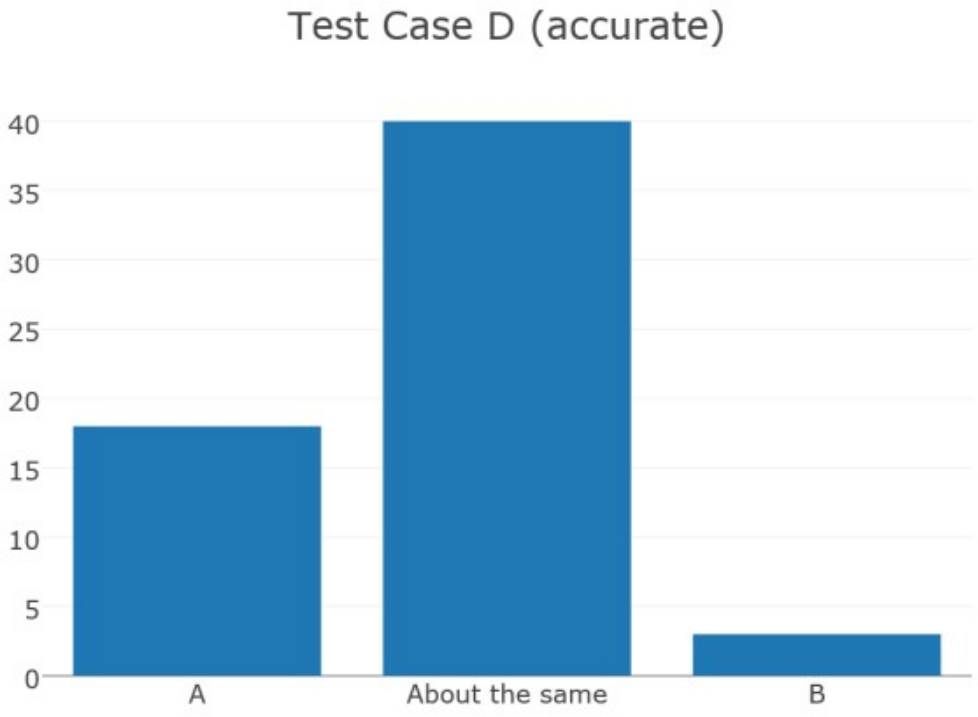
D) 1 stalling event of 1 second vs 1 stalling event of 3 seconds

Links: http://gain.di.uoa.gr/kyr/s2a/qoe.html http://gain.di.uoa.gr/kyr/s1b/qoe.html

Two videos clips are put to the test where the first one has a slight stalling of 1 second whereas the second has a longer one that lasts 3 seconds. Again, the choice between the two is obvious but the purpose of this test was to evaluate the degree of difference among them from the viewer's point of view and conclude whether or not they are considered similar. The first video lasts a total of 67 seconds and the stalling appears at the 30th second while the second has a duration of 68 seconds including a stalling event at the 0:30-0:33 mark.
For this case, 90 proper sets of answers were required and 9 were deemed useless and replaced at no cost. 61/90 successfully detected the correct amount of times the playback froze, which is considered a high percentage that yields reliable results.



Graph D1



Graph D2

Remarks

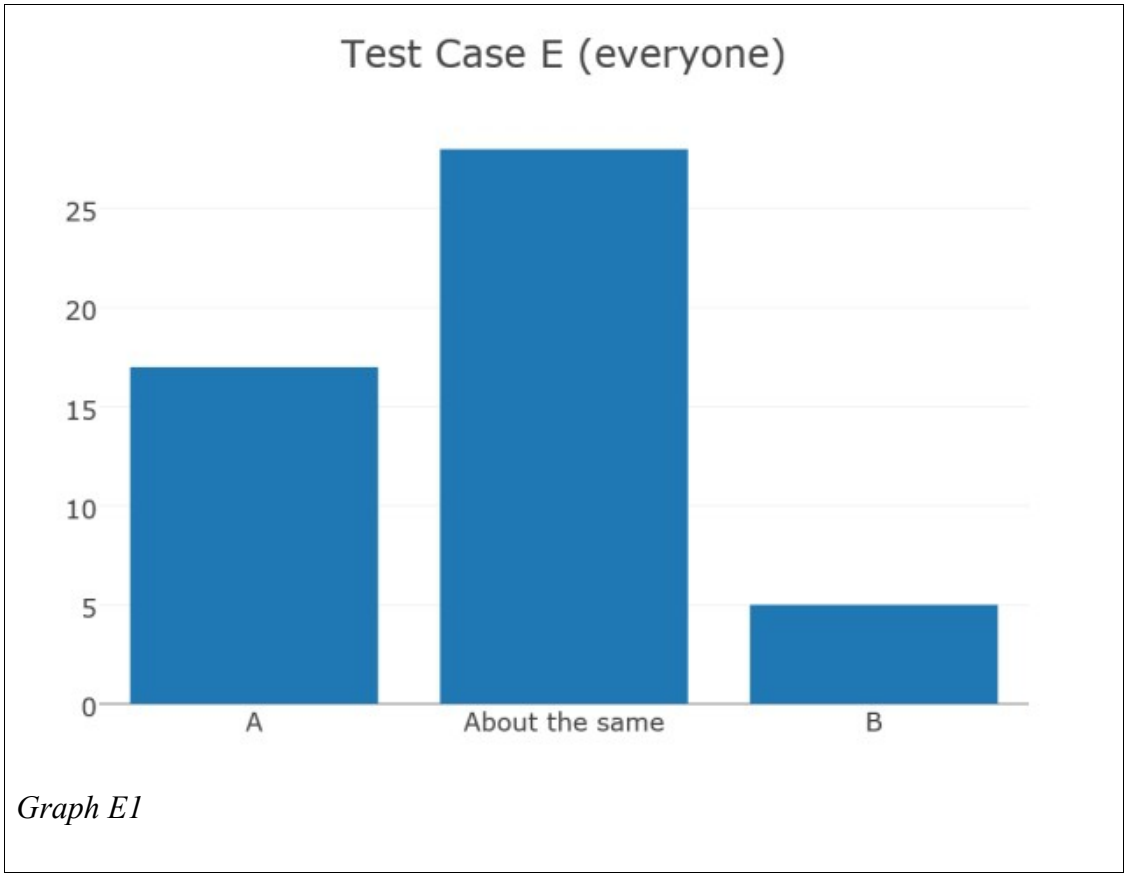
The logic of this experiment is quite similar with the previous one in the sense that there is objectively a smaller and a bigger impairment respectively in these clips, as they both feature one stalling event with a different duration in each case. However, the result differ significantly from the previous case. This time, the vast majority of the test subjects rated the clips similarly, probably because neither a stalling event of one second, nor another of three seconds irritated them. Of course, some people showed their preference towards the first clip with the slightly shorter delay but the final conclusion is that when experiencing the same amount of stalling events, users can not tell the difference between them when their duration is rather short.

MOS scores: 3.99 for video A, 3.84 for video B.

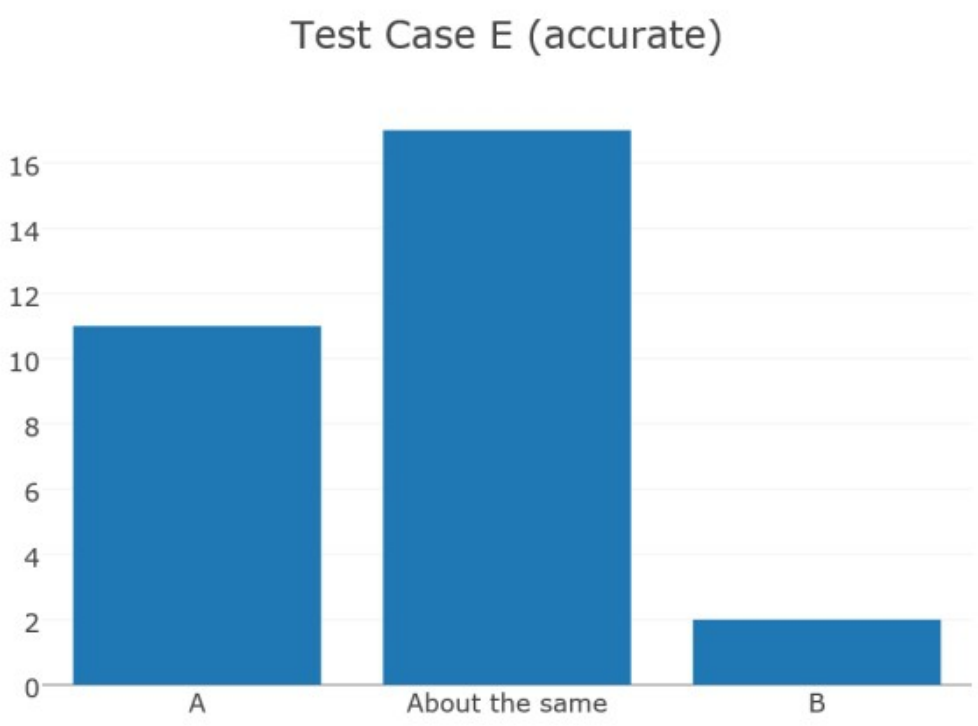
E) Descending vs ascending quality

Links: http://gain.di.uoa.gr/kyr/357/qoe.html http://gain.di.uoa.gr/kyr/753/qoe.html

From this point and on, the experiments research preferences related to adaptive streaming and not stalling. The number of the layers that are being heavily mentioned below have already been explained and their technical aspects can be found in Table 1. For the first one of the batch, two video clips of the same duration (00:43) are being compared that feature the same switches among layers, but in the opposite order.
To be specific, video A initially offers high image quality on the third layer before it switches to the medium one of the fifth at 00:15. Then at 00:30, the quality gets even lower when the seventh layer is featured until the end of the playback. On the other hand, video B switches in an ascending way, moving from the seventh layer to the fifth and then the third at the exact same time marks as video A.
The purpose of this test is to assess whether users are more negatively influenced by their first impression of the clip or by the residual effect that the last part of the clip left them. 50 successful completions of this test were required and it was achieved after 13 sets were unacceptable and replaced. The percentage of test subjects who accurately identified at which point each video was at its peak quality were 30/50.



Graph E1



Graph E2

Remarks

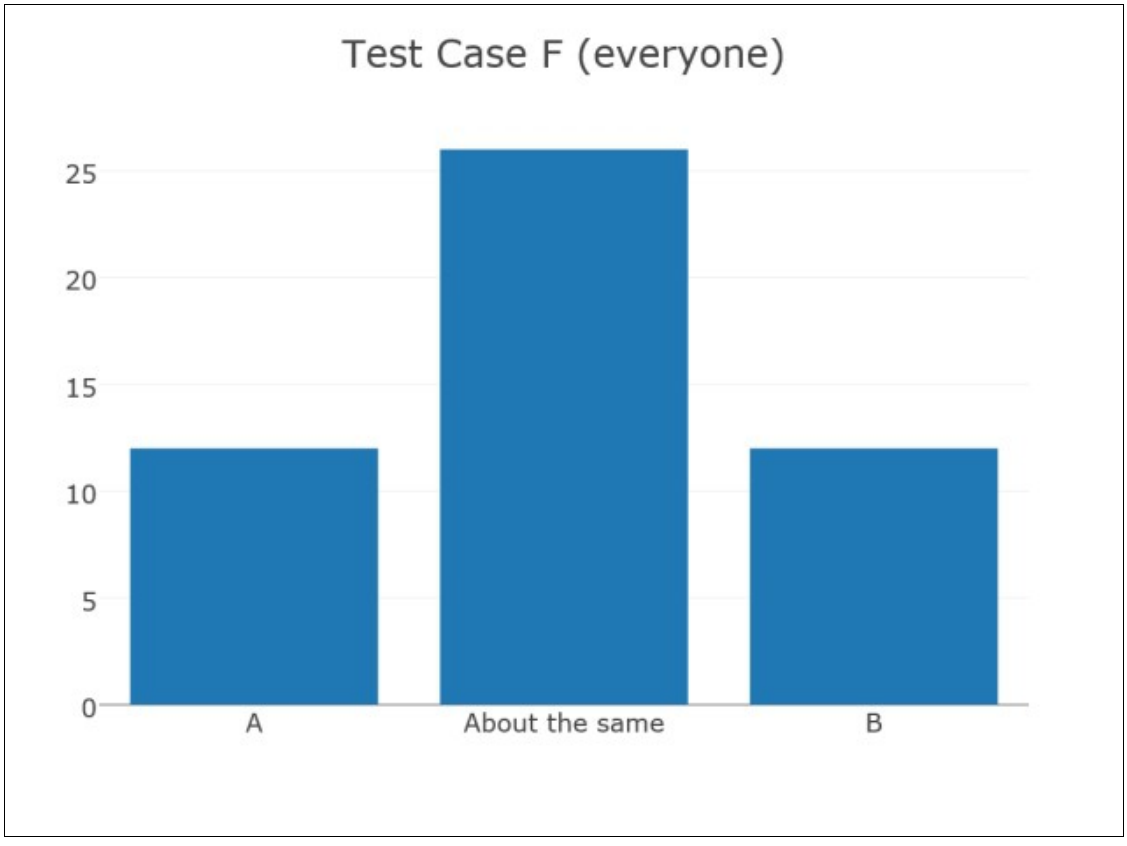
This is a highly subjective test case as the only factor that differentiates the two clips is the order in which they switch from one layer to another. As the results indicate, the majority of the participants thought that their viewing experience was about the same for both video clips. However, video A was a not so distant second, indicating that users tend to be influenced from their first impression much more than the way the clip ended.

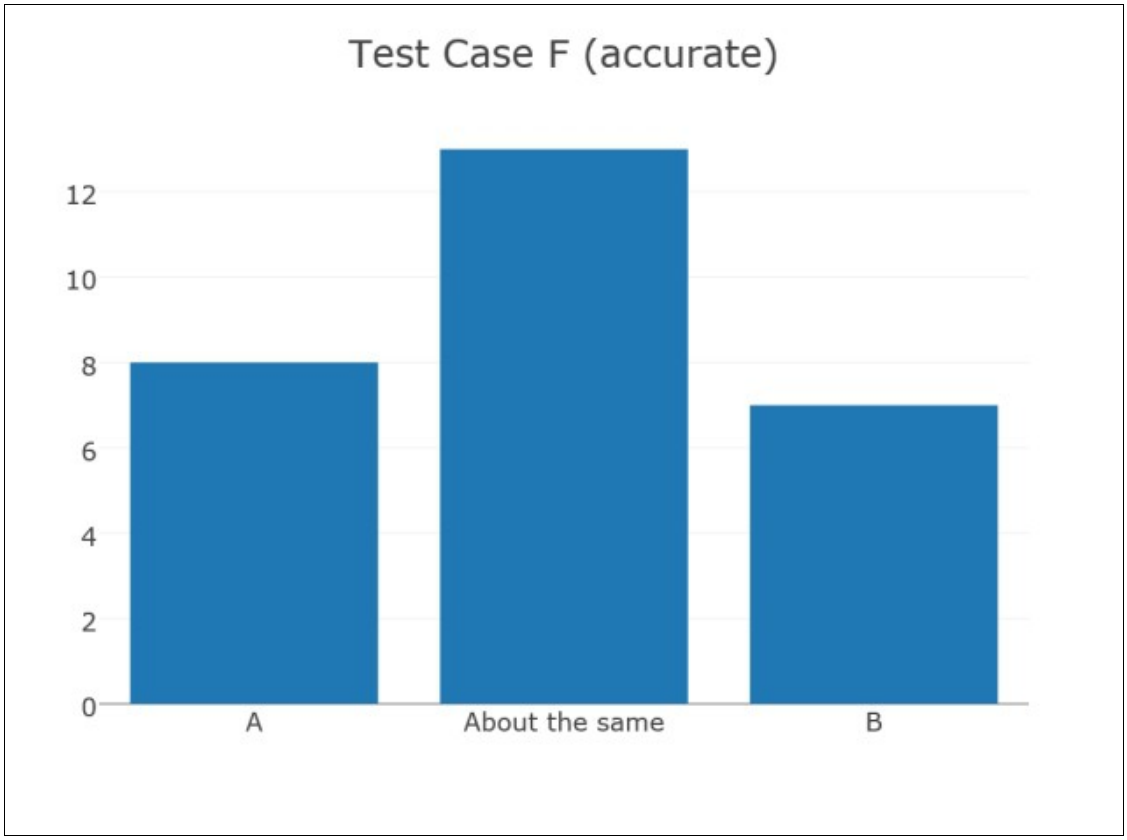
MOS scores: 2.86 for video A, 2.64 for video B.

F) Smooth vs steep transitions

Links http://gain.di.uoa.gr/kyr/357/qoe.html http://gain.di.uoa.gr/kyr/37/qoe.html

In this test case, the effect of smooth transitions is being researched. Two videos of the same duration and, as always, content are compared by the test participants. They both last 43 seconds, begin on the third layer and finish on the seventh. However, there is a significant difference between them.
Video A switches from layer 3 to layer 5 at the 00:15 mark, as it was already described in the previous test case, and then to layer 7 at 00:30. Video B, on the other hand, performs only one abrupt switch from the third to the seventh layer which occurs at 0:22. Obviously, the steep transition is much more noticeable than the smoother ones that took place in the first clip. This case aims at quantifying the difference between the two perceptions.
For this experiment, a slight change was attempted in the setup of the test campaign by employing anyone willing to participate rather than proved workers from highly esteemed groups. The initial results were catastrophic with the vast majority (41/50) of the test subjects randomly clicking anything to submit their answers and move on. Of course, none of them was compensated and the whole test case had to be redone in order to reach the desired amount of usable results, which was 50. 8/50 answers needed to be replaced on the second phase of the experiment that was set up properly and from the final 50 sets that were kept, 28 of them successfully detected that both video clips peaked in their beginning.





Remarks

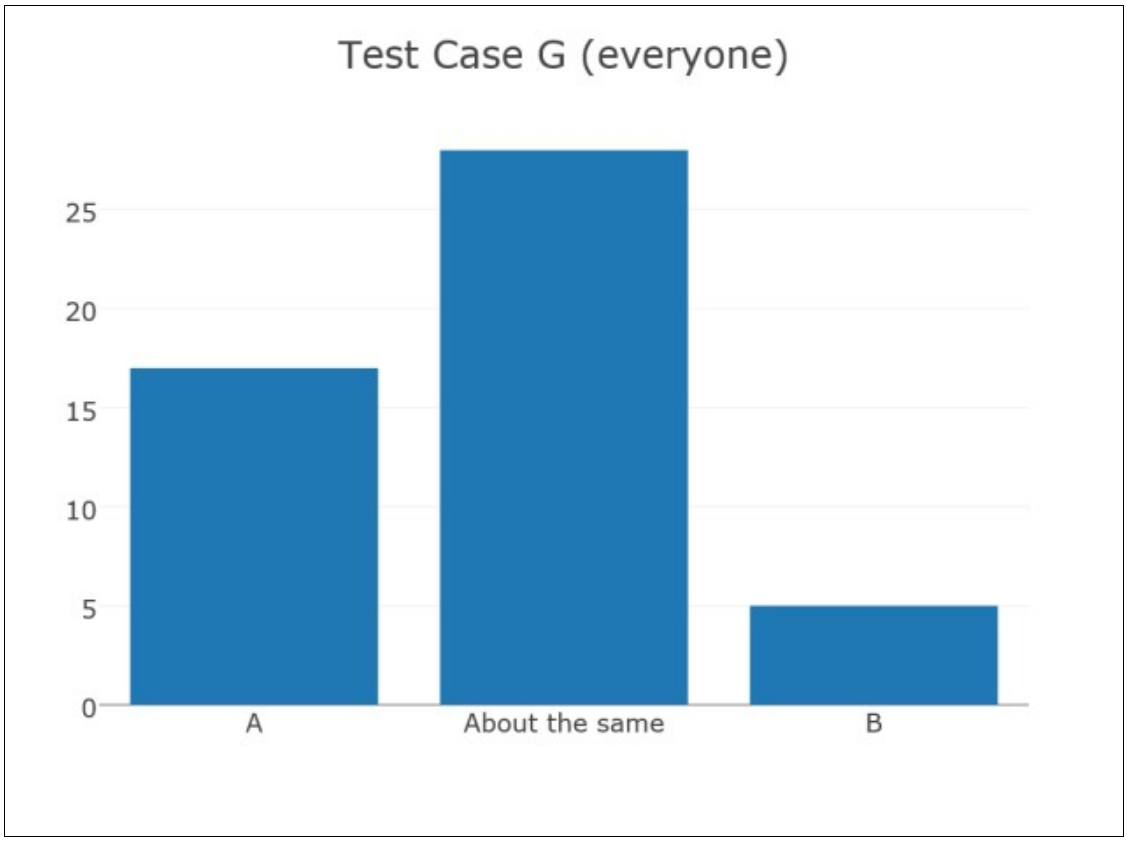
The results in this particular test case are perhaps unpredictable. The ratings for the first clip where the layer switches were executed smoothly were almost identical with the ones for the second clip which featured an extremely abrupt transition. Most of the participants felt like both videos were of similar overall quality, while the percentages of people preferring one clip over the other are practically the same for both of them. This can be attributed to the fact that a portion of the subjects gets annoyed by sudden drops of quality while others only care about the time spent on each quality layer.

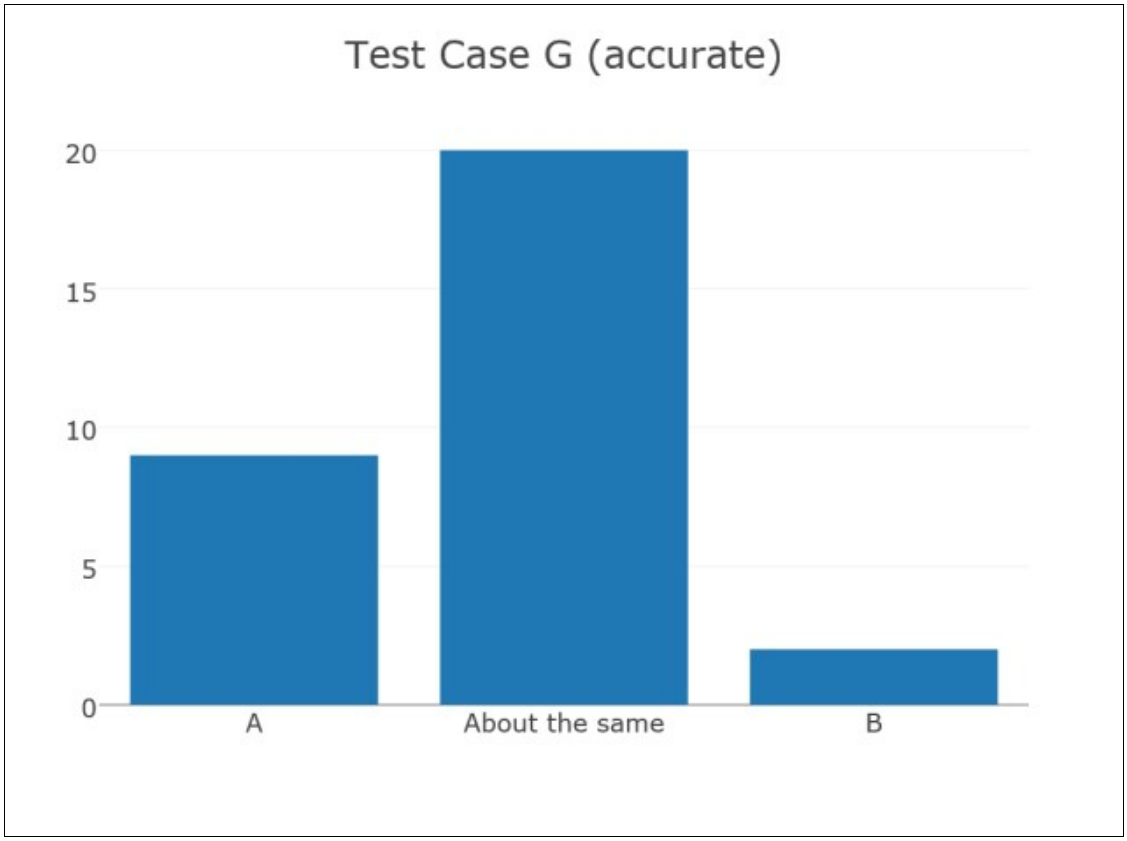
MOS scores: 3.32 for video A, 3.22 for video B.

G) Random vs content-aware transitions

Links: http://gain.di.uoa.gr/kyr/357/qoe.html http://gain.di.uoa.gr/kyr/357c/qoe.html

This experiment examines two videos that feature the same layer switches in the same order. The only difference is that in the first clip, the switches occur virtually randomly while in the second, they have been inserted in a more surgical way, meaning that the decreases of the image quality take place when the scene changes, making them less noticeable. This content-aware approach takes into consideration the keyframes of the clip and makes the switch right before them rather than dropping the resolution suddenly mid-scene, as was happening until now in our test cases.
Both video clips last 0:42. The exact timings of the switches for the first one have already been presented. The second clip drops its picture quality from the third layer to the fifth at 0:18 and then further down at the seventh layer at 0:28. Both instances represent a major change in the video content-wise where the scenes dramatically change. Additionally, the timings of the transitions were kept as close as possible to the first clip, in order for the duration of playback in each layer to be about the same.
The target for this case was 50 usable sets of results and it was achieved very easily with only one of them getting discarded. What is more, 31/50 workers accurately experienced the impairments, which is definitely a satisfying percentage.





Remarks

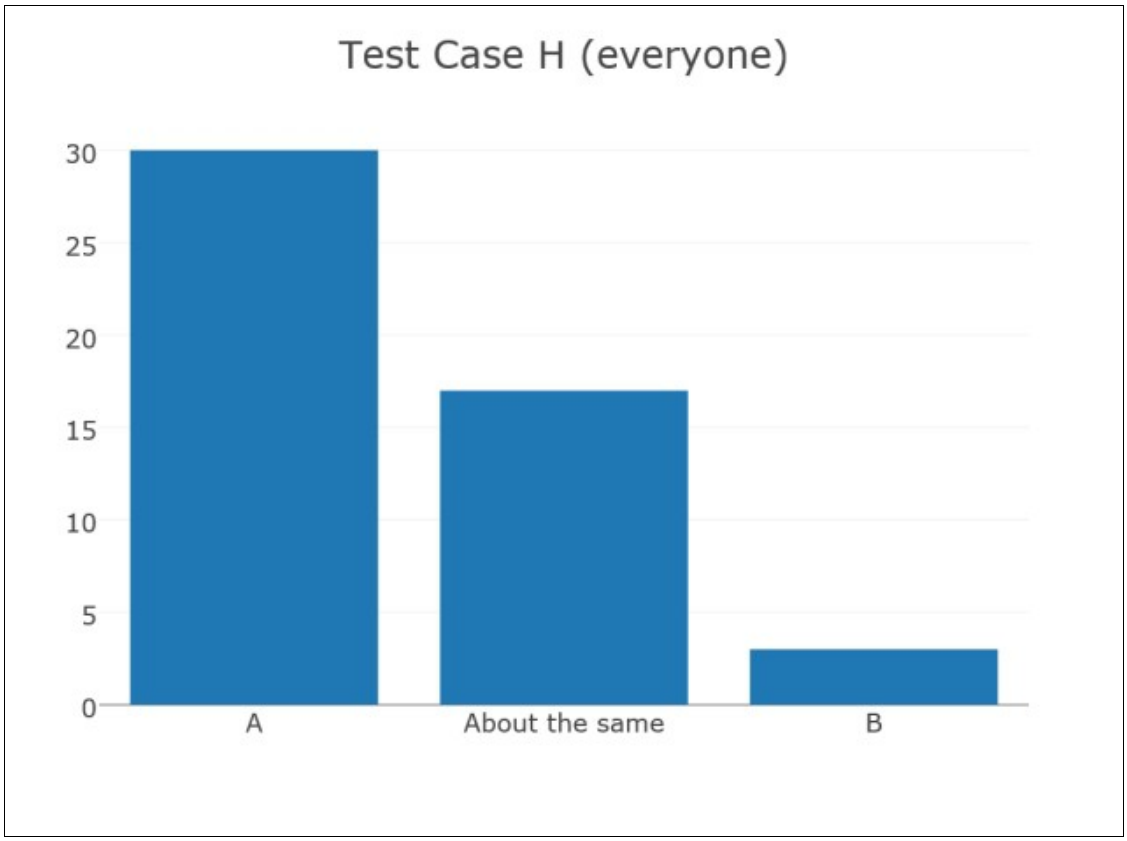
This is another highly subjective experiment. Users had to select if a video clip that drops its quality mid-scene is better than another one which switches to a lower layer between scenes. Predictably, the vast majority thought that the end result was about the same. Perhaps unpredictably, however, more participants declared that the first clip offered better quality than the second. The explanation for this is not obvious but a possible reason could be that viewers thought that the clip was a product of merging videos of variable quality, which is generally a bad technique.

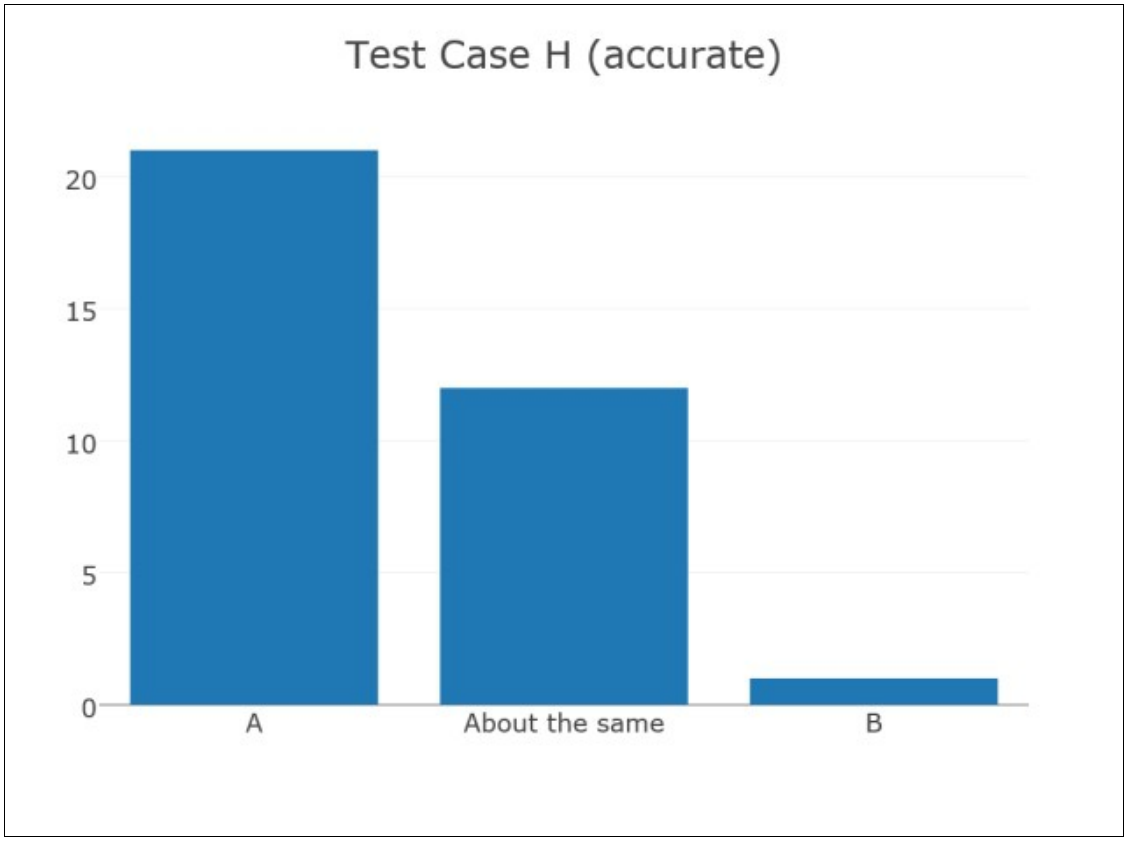
MOS scores: 3.3 for video A, 3.16 for video B.

H) Early transition to medium layer vs late transition to low

Links: http://gain.di.uoa.gr/kyr/355/qoe.html http://gain.di.uoa.gr/kyr/337c/qoe.html

In this test case, we asked the participants to compare and assess two videos that last 0:44 and have one layer switch. The first one starts its playback on the third layer but 15 seconds after, drops to the fifth where it remains until the end. The second clips also begins to play on the third layer but it stays there longer before it transitions abruptly to the seventh at 00:30.
This is a very interesting case because it compares two basic techniques used to manage the limited network resources. Video A sacrifices a longer duration on the high level but does not drop lower than the medium one, while video B offers high quality for two thirds of its duration but then makes a steep transition to low quality for its final part. To summarize, this case attempts to resolve the preference of the average user, between staying in the medium layers but avoiding the lower ones, or a prolonged use of the higher layers that leads to the lower ones at the end.
50 acceptable sets of answers were required for this case and were acquired after replacing 19 ones submitted by problematic workers. It is worth noting that a large percentage (52%) of participants did not notice the layer switch in the first video, probably because the picture quality for the whole duration was kept to acceptable levels.





Remarks

In the last case, the results are absolutely conclusive. Users rate clips that avoid the lowest quality layers way highly, even if that means a lesser amount of time spent on the highest ones. Over half workers stated that their experience was better with the first clip, with a considerable percentage not showing a specific preference and very few of them leaning on the way the second clip was structured.

MOS scores: 4.08 for video A, 3.62 for video B.

4. CONCLUSION

The main question that this paper attempted to answer, is whether a crowdsourcing platform can be utilized in order to accurately measure the quality of experience perceived by users. Given the subjective nature of the whole concept, it is difficult to reply definitely. However, after successfully conducting a series of experiments with useful results, the author is inclined to answer positively.

In order for a crowdsourced experiment to simulate a controlled one, a series of precautions must take place, along with careful filtering of the initial results. If both these measures take place, then the remaining results are virtually of the same quality as the ones gathered in a laboratory. What is more, in our case they are much more cost effective and easily scalable in any extent.

Both these features make crowdsourcing experiments rather attractive to testing organisations, especially after taking into consideration that the global online workforce is constantly expanding. As far as network testing is concerned, the process organically works great, given the fact that it all takes place online and naturally tests the connection at every step of the way. It is logical to assume that crowdsourced testing in relation to network metrics will continue to grow and become the main method used to quantify the users' quality of experience.

5. IMAGES

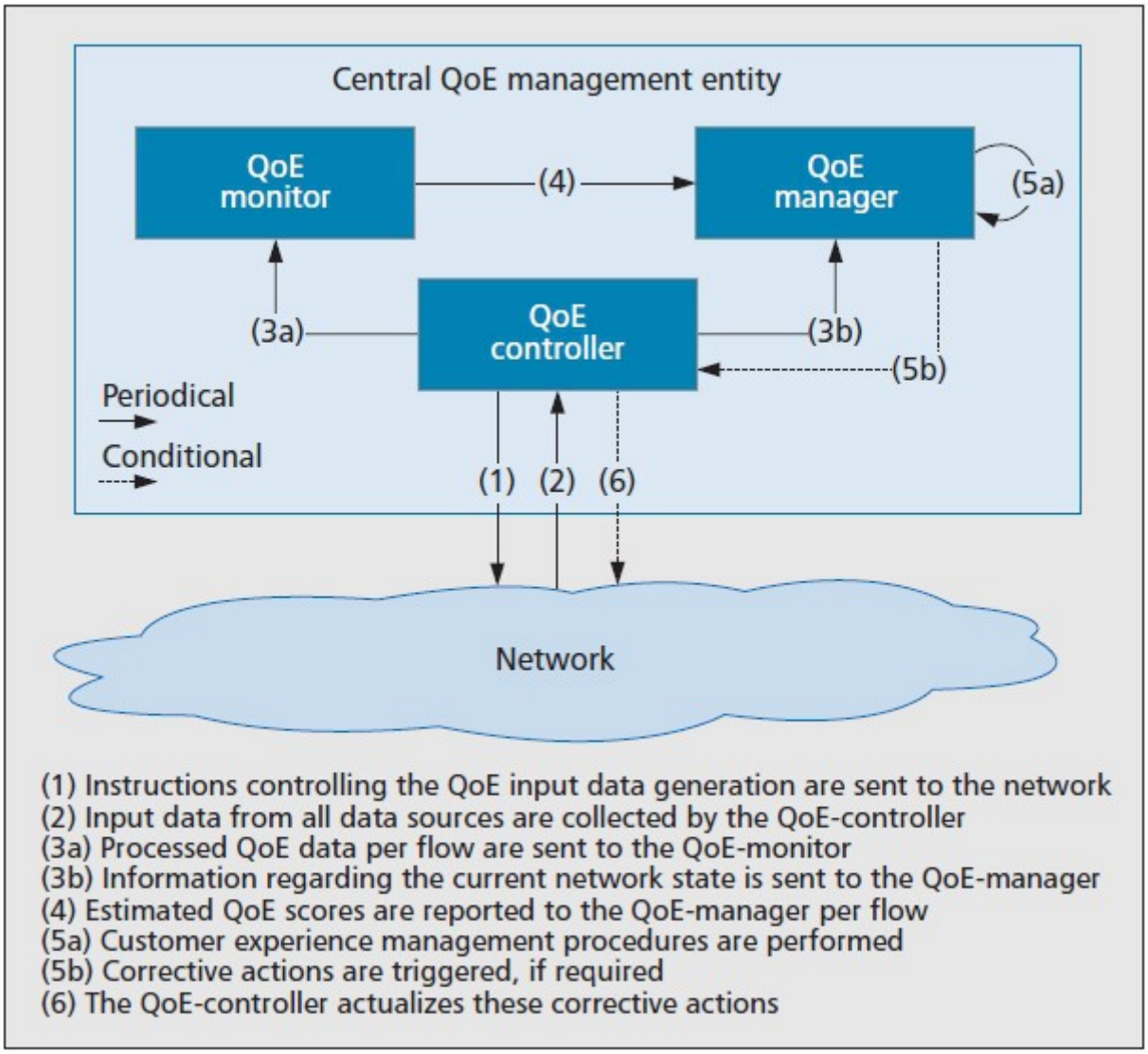


Image 1: A conceptual QoE framework

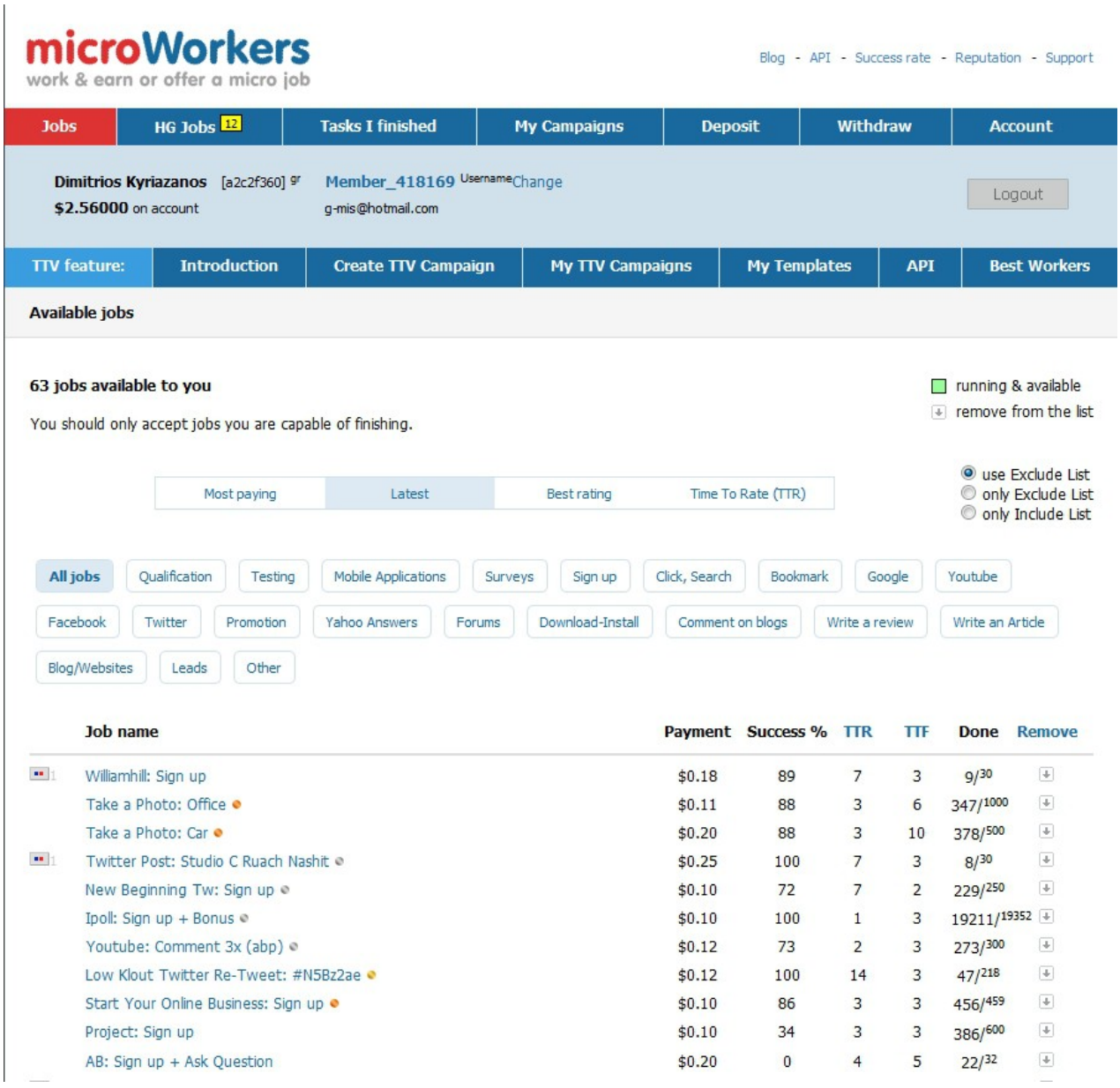


Image 2: Available work at MicroWorkers

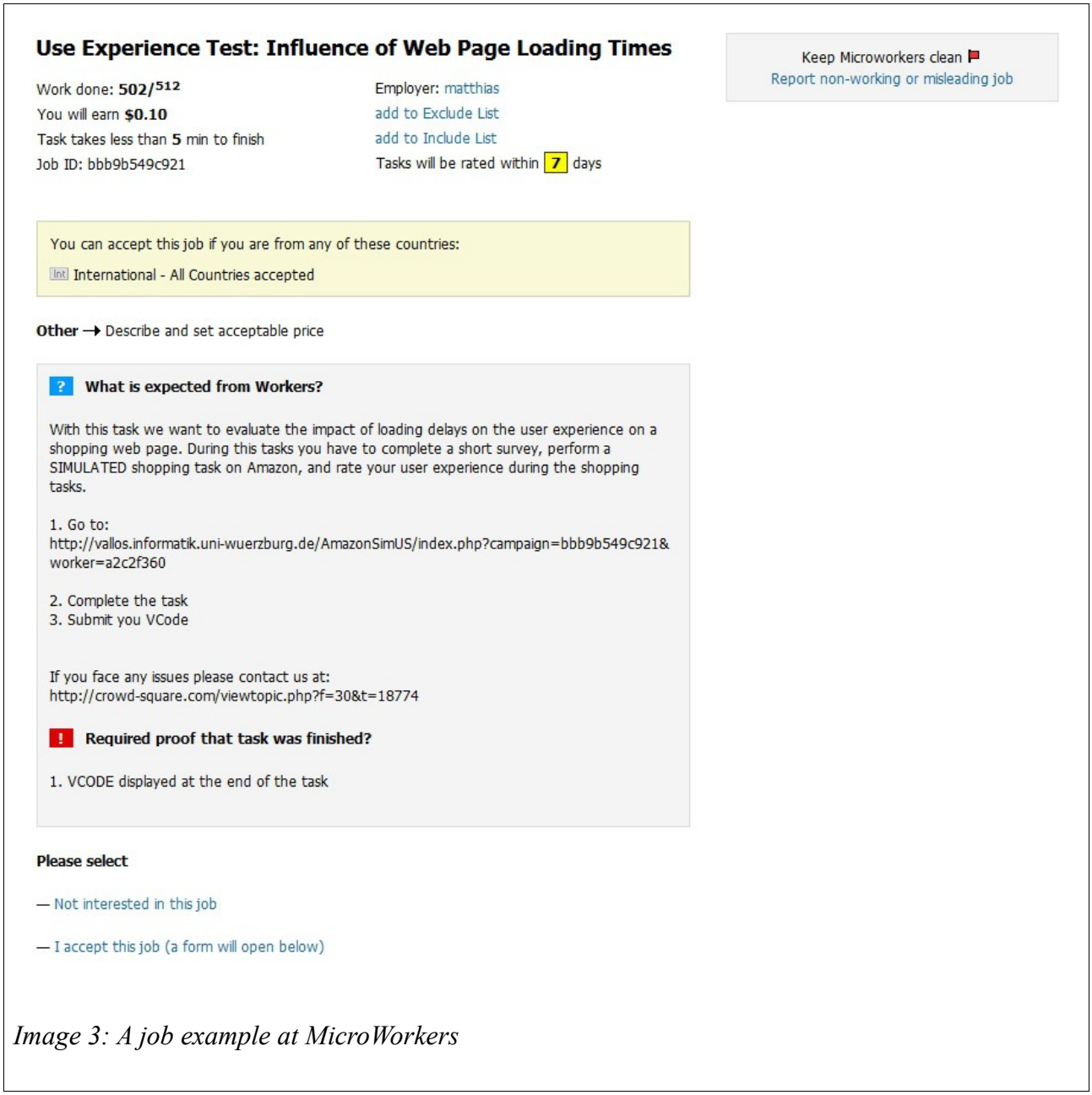


Image 3: A job example at MicroWorkers

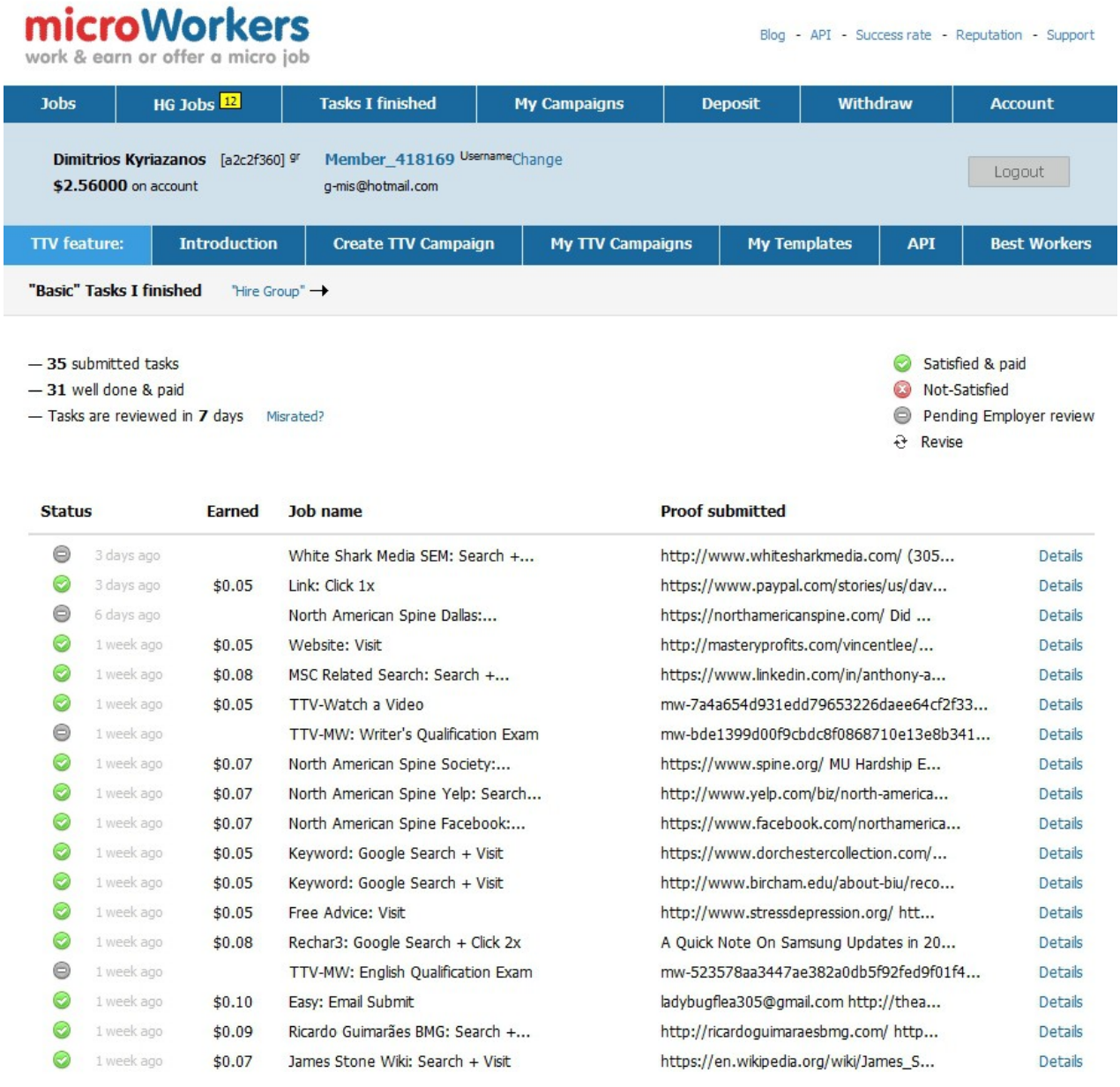


Image 4: Finished jobs at MicroWorkers

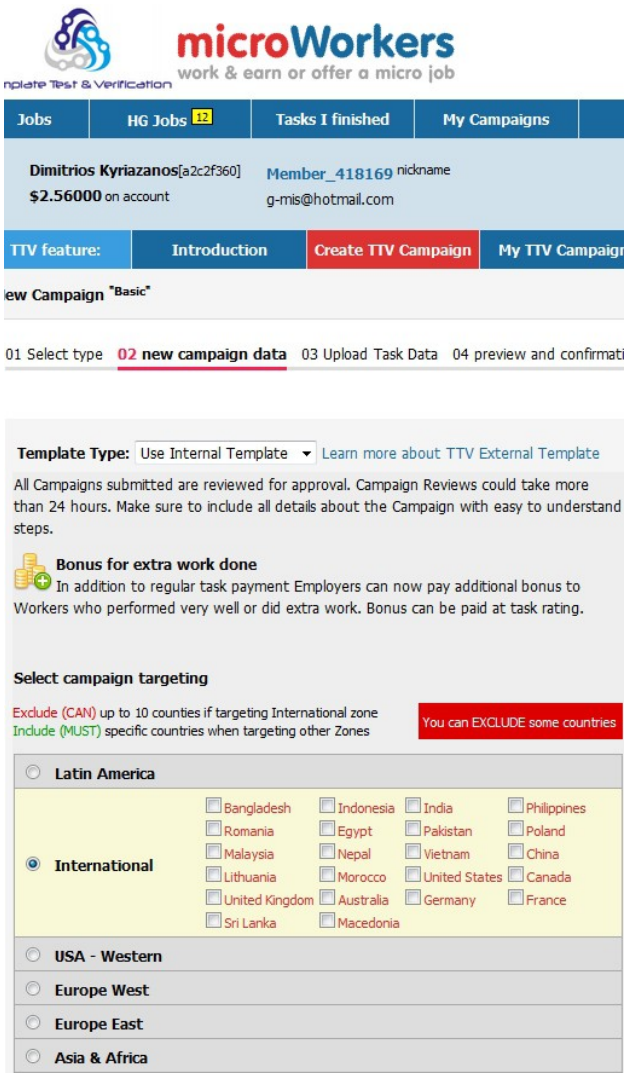


Image 5: Creating a job at MicroWorkers 1

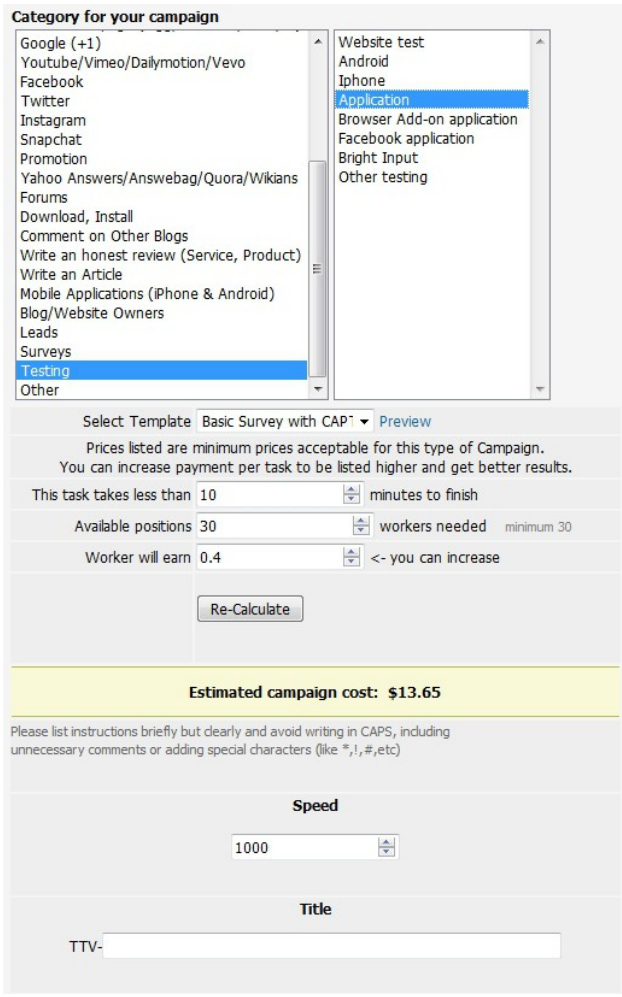


Image 6: Creating a job at MicroWorkers 2

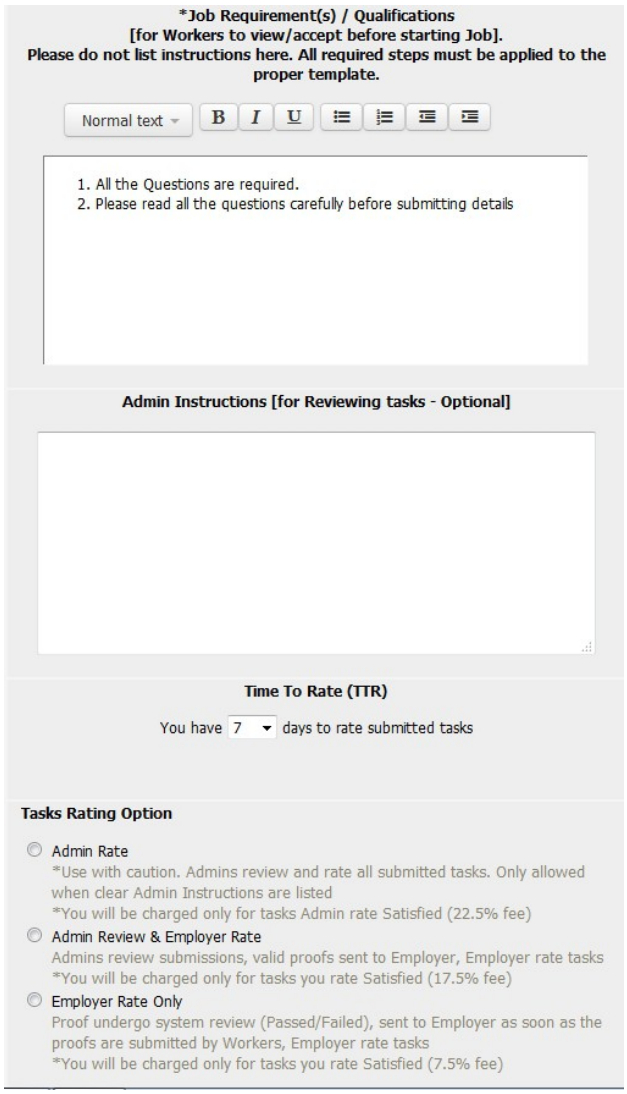


Image 7: Creating a job at MicroWorkers 3

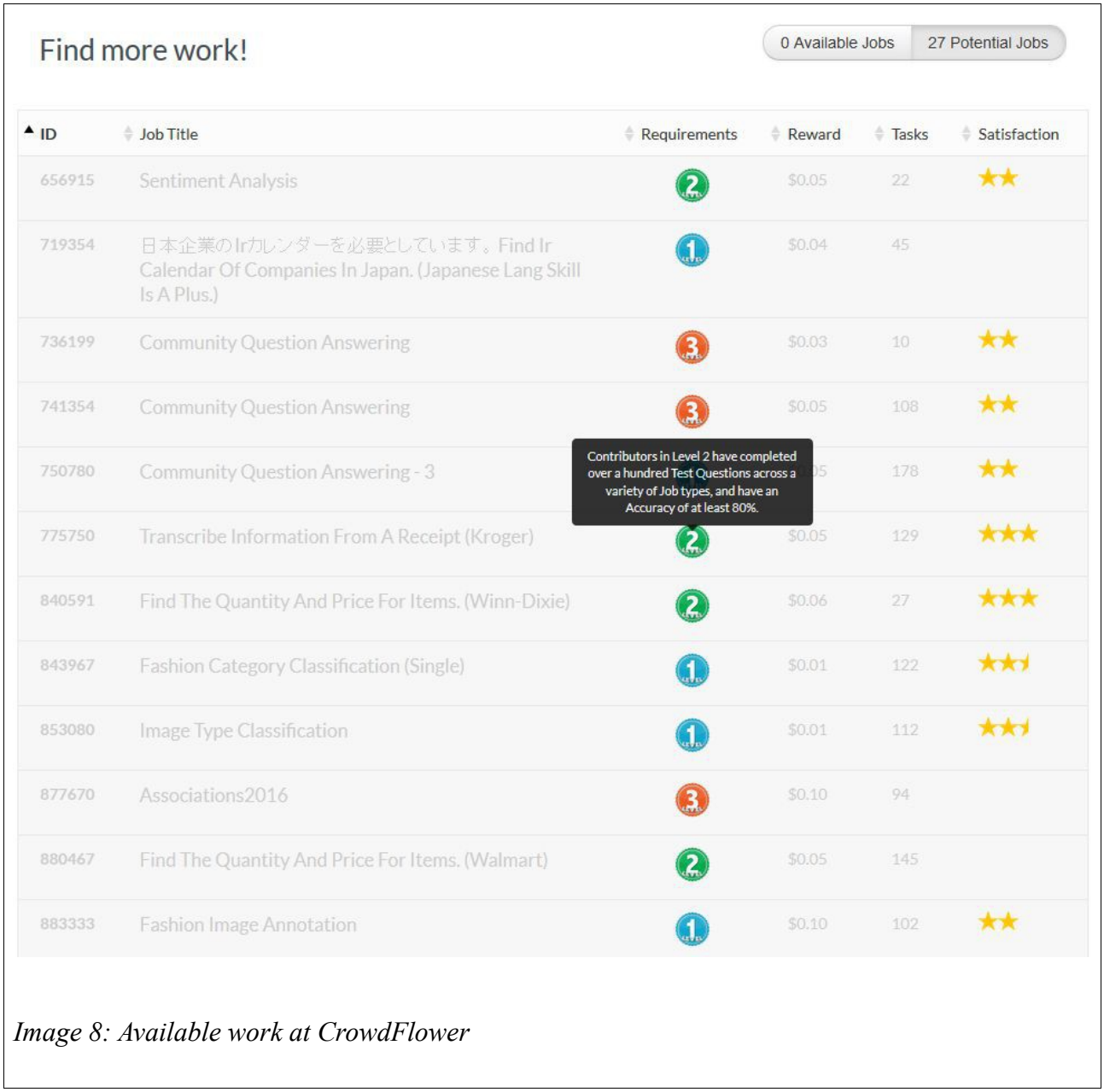


Image 8: Available work at CrowdFlower

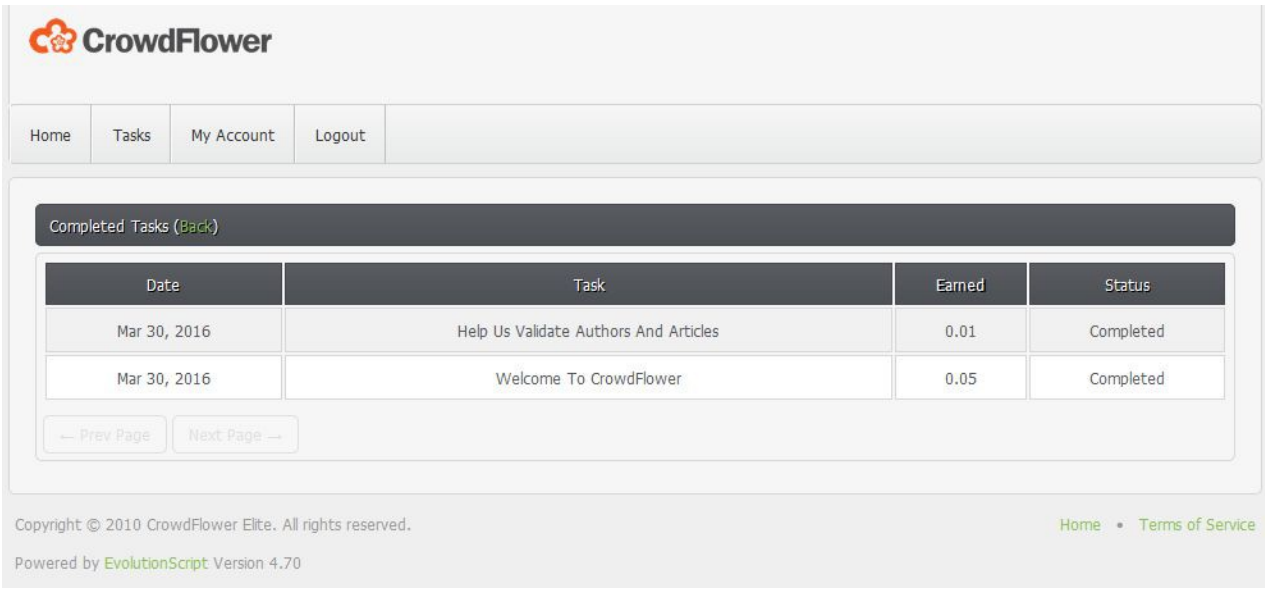


Image 9: Completed jobs at CrowdFlower

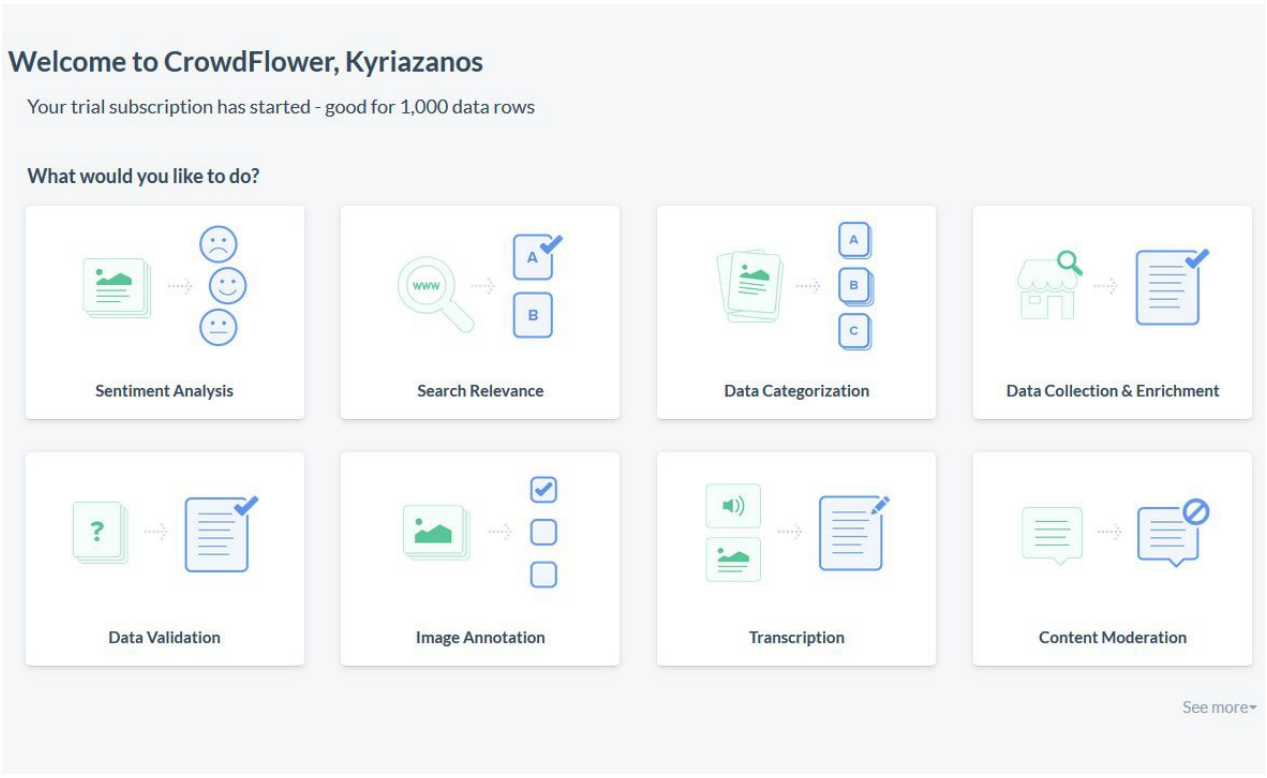


Image 10: Job categories at CrowdFlower

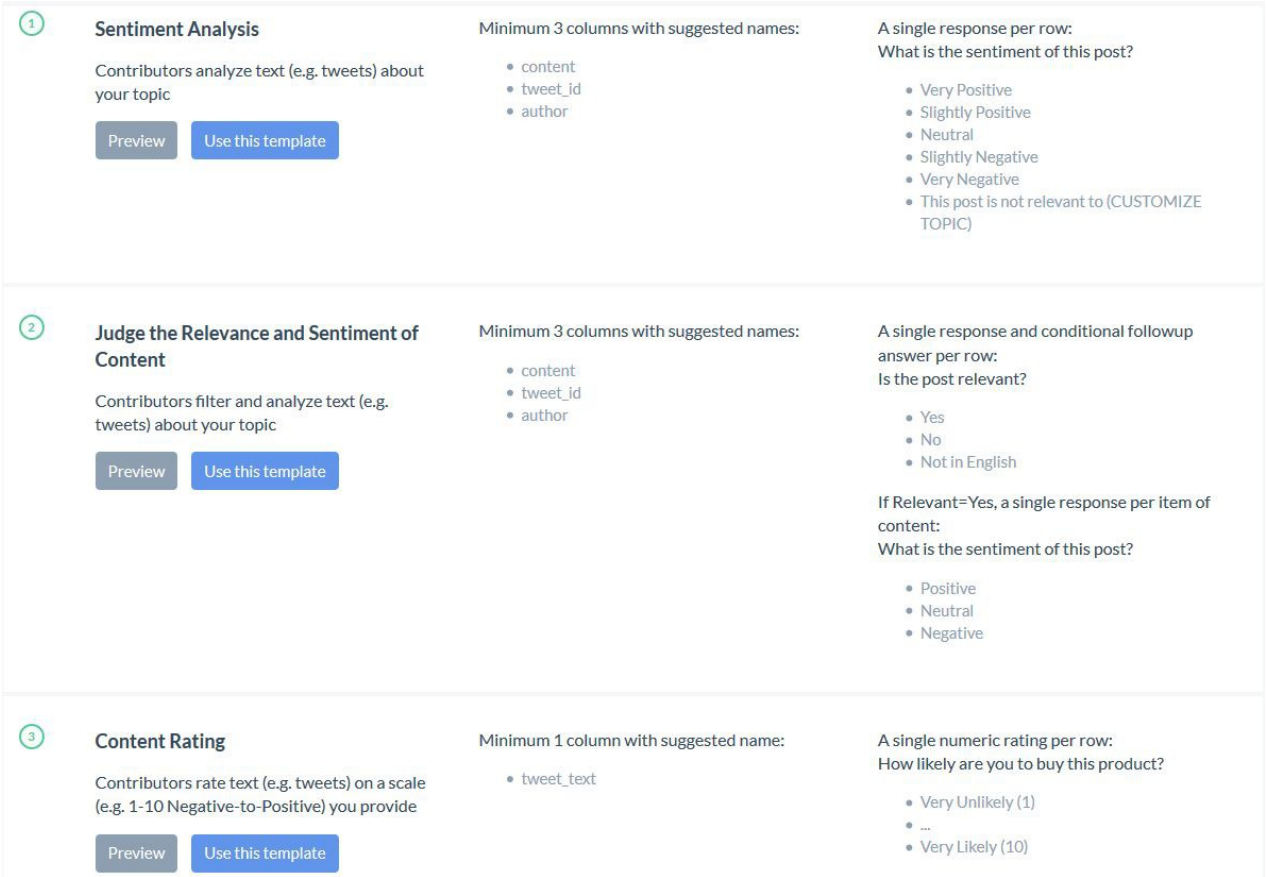


Image 11: Description of categories at CrowdFlower

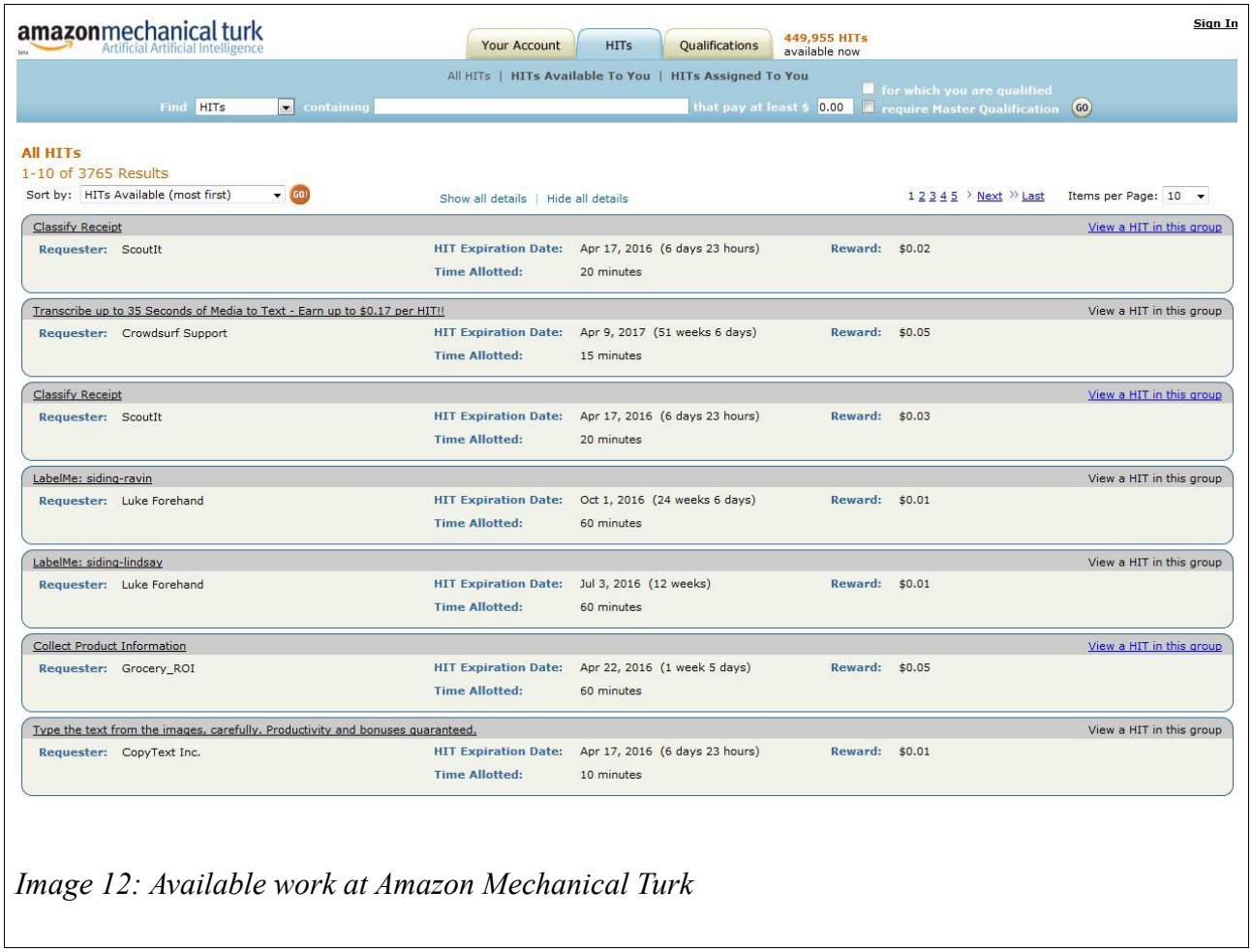


Image 12: Available work at Amazon Mechanical Turk

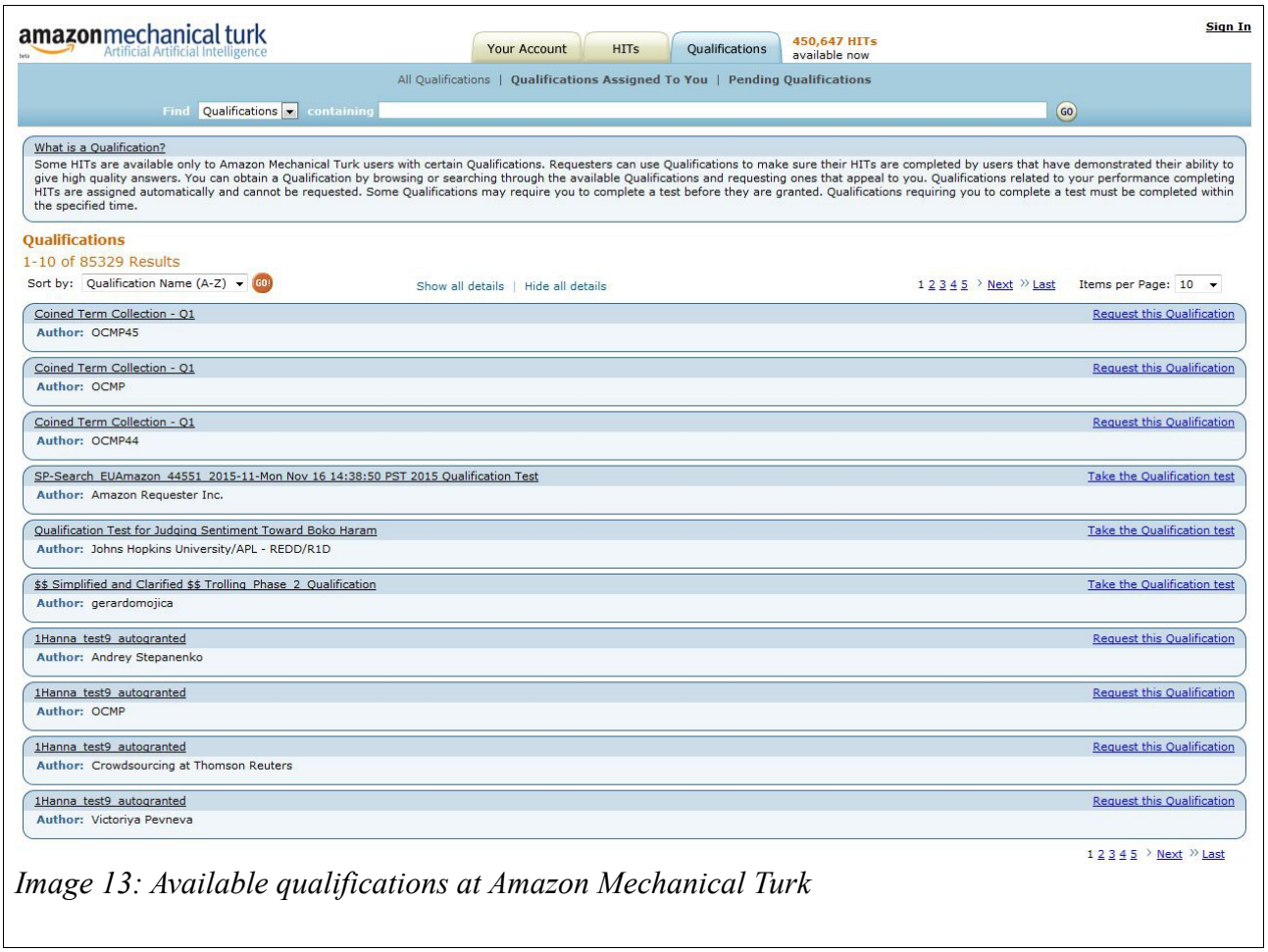


Image 13: Available qualifications at Amazon Mechanical Turk

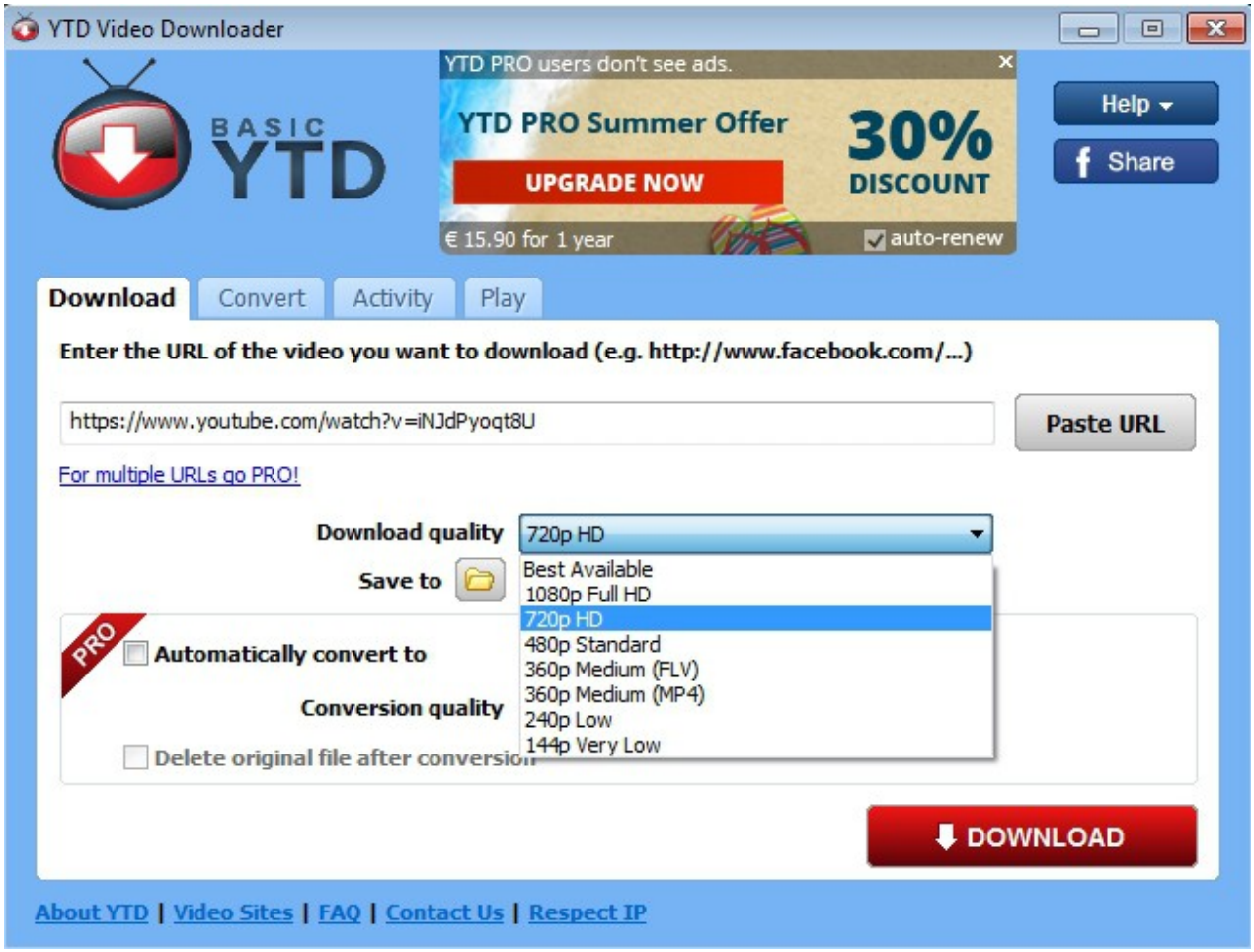


Image 14: YTD Video Downloader

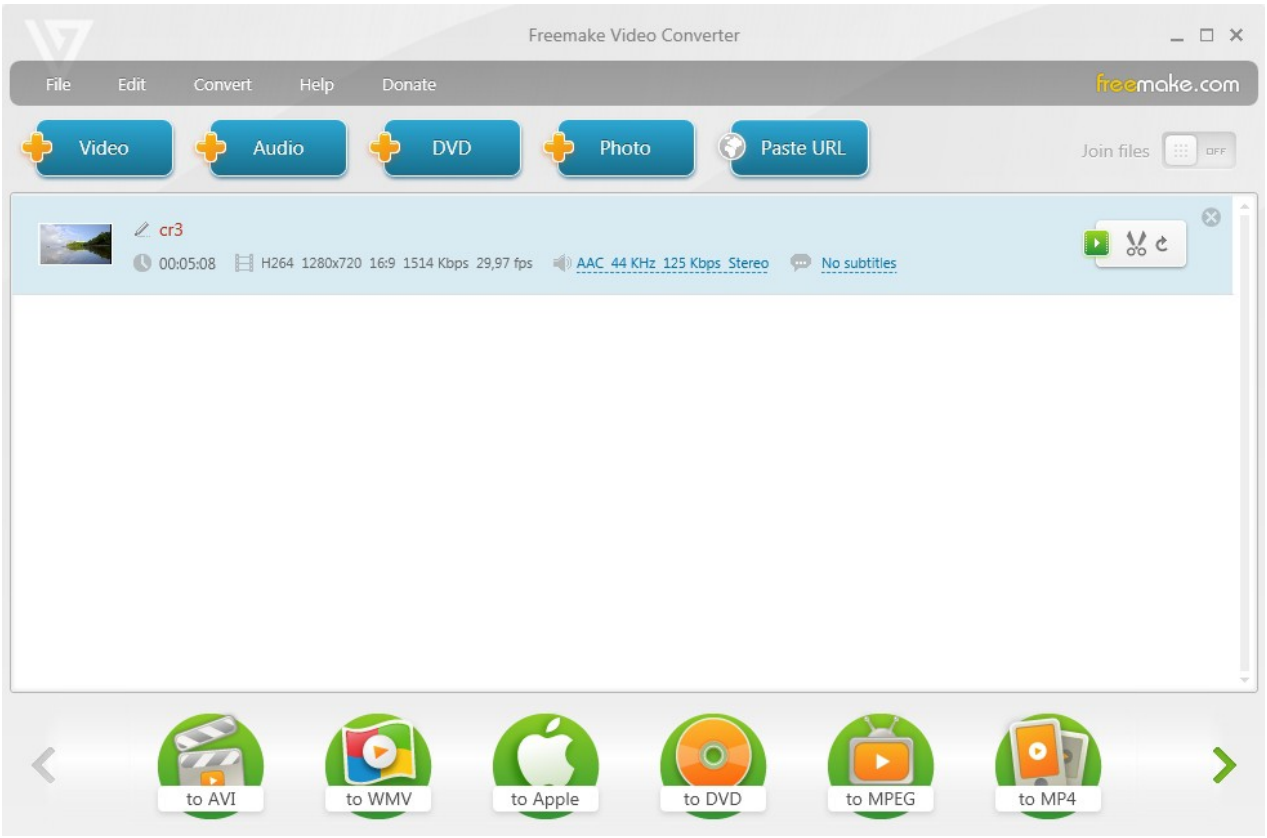


Image 15: Selecting a video at Freemake Video Converter

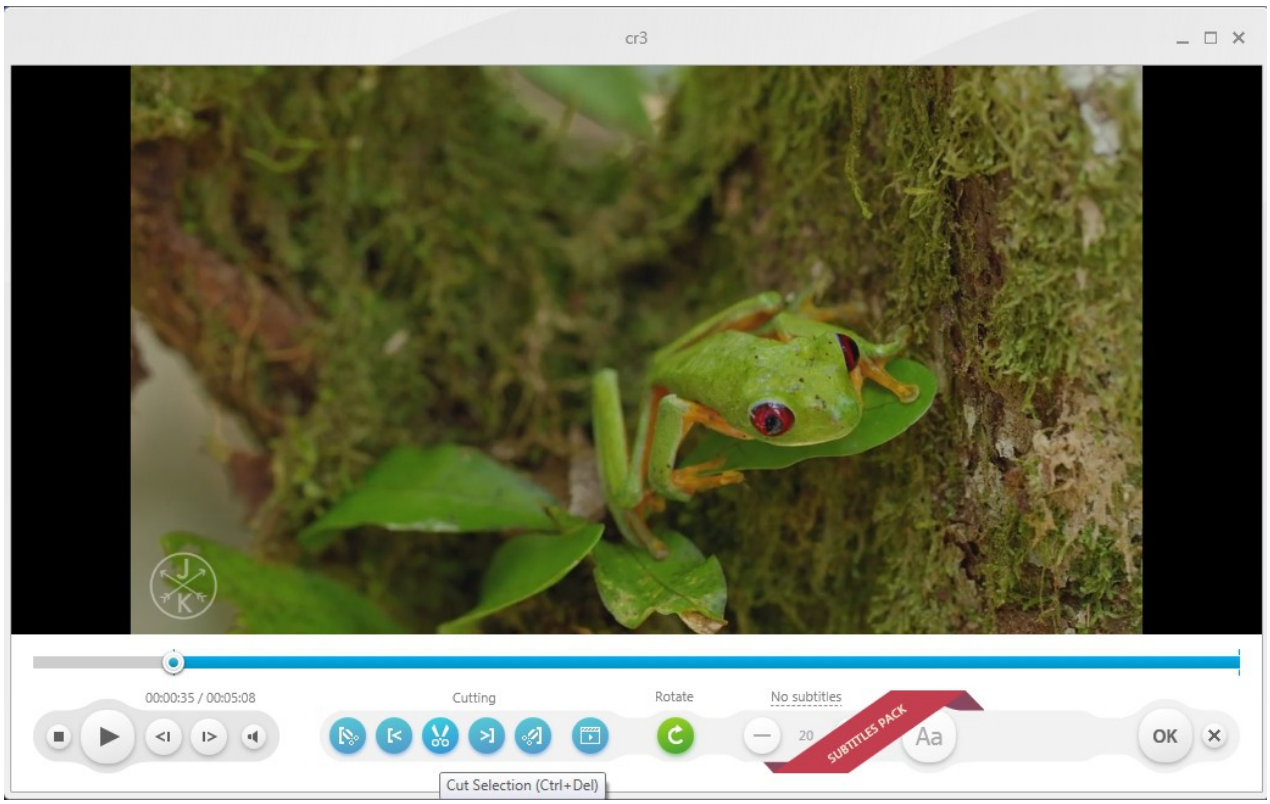


Image 16: Selecting the part to be cut at Freemake Video Converter

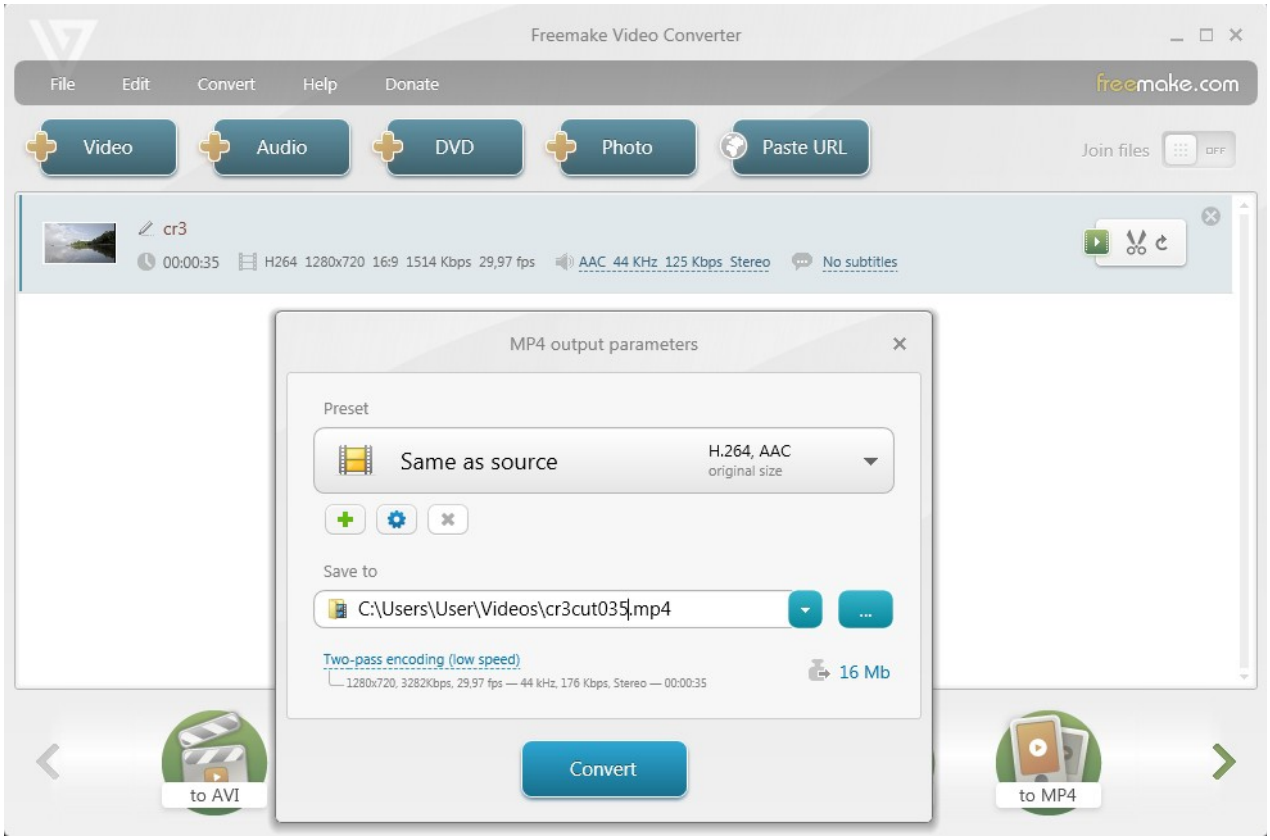


Image 17: Saving the trimmed video with Freemake Video Converter

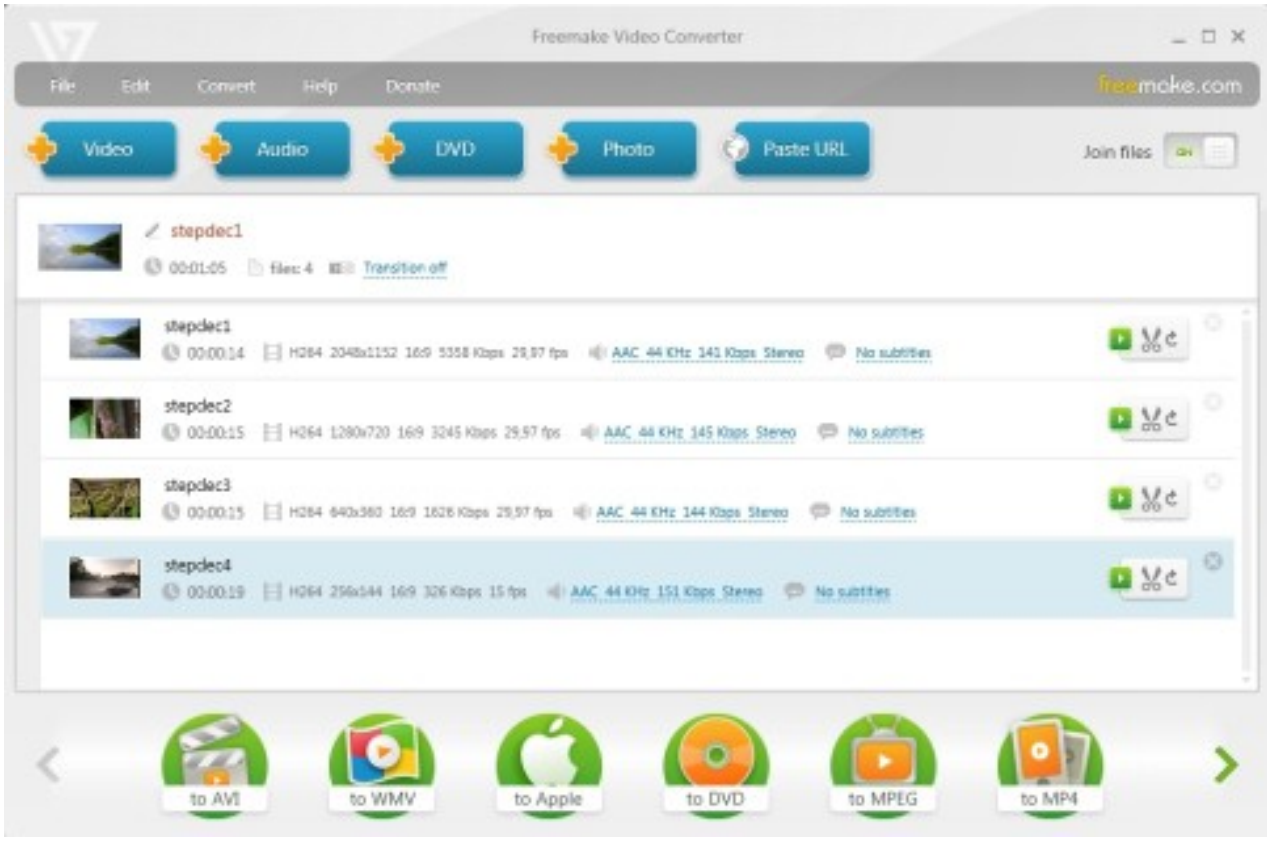


Image 18: Merging videos with Freemake Video Converter

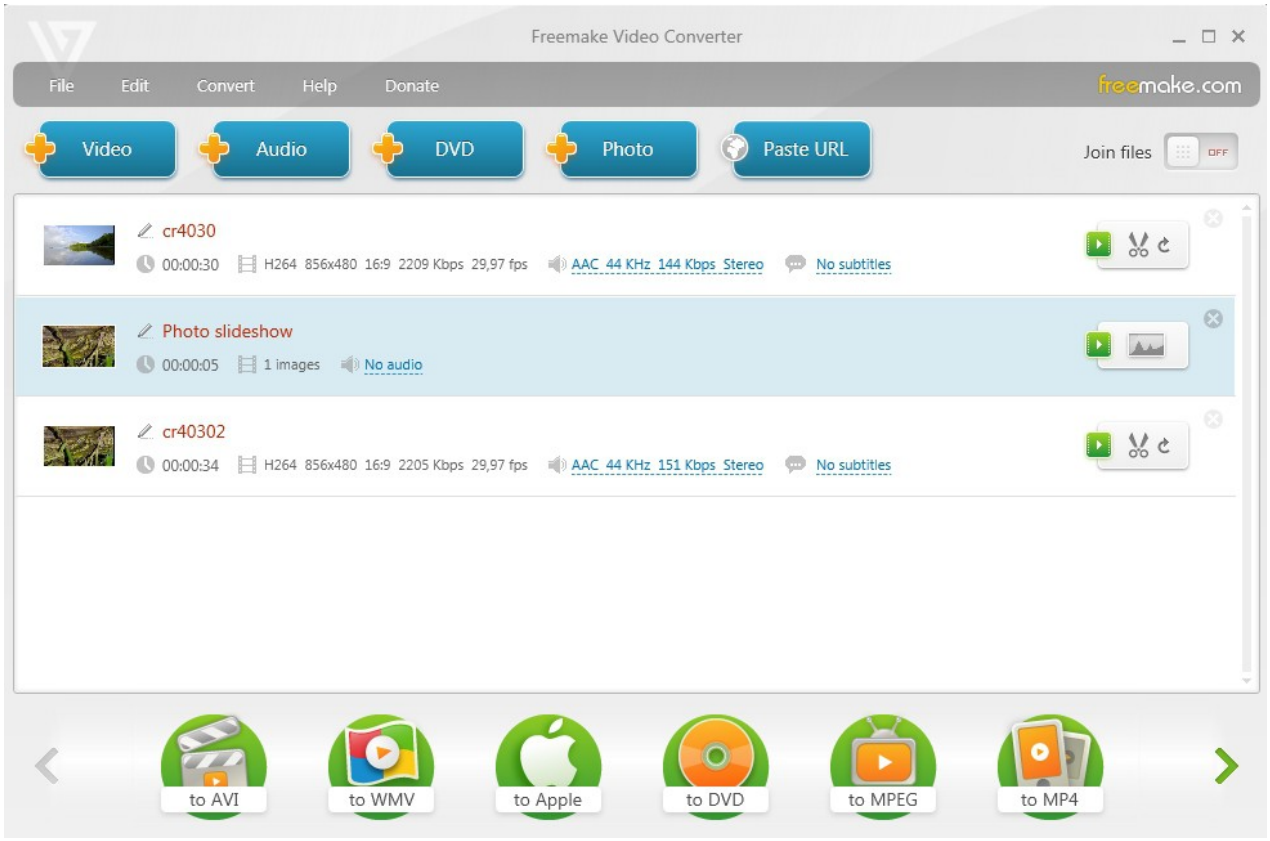


Image 19: Adding a photo to the compilation with Freemake Video Converter

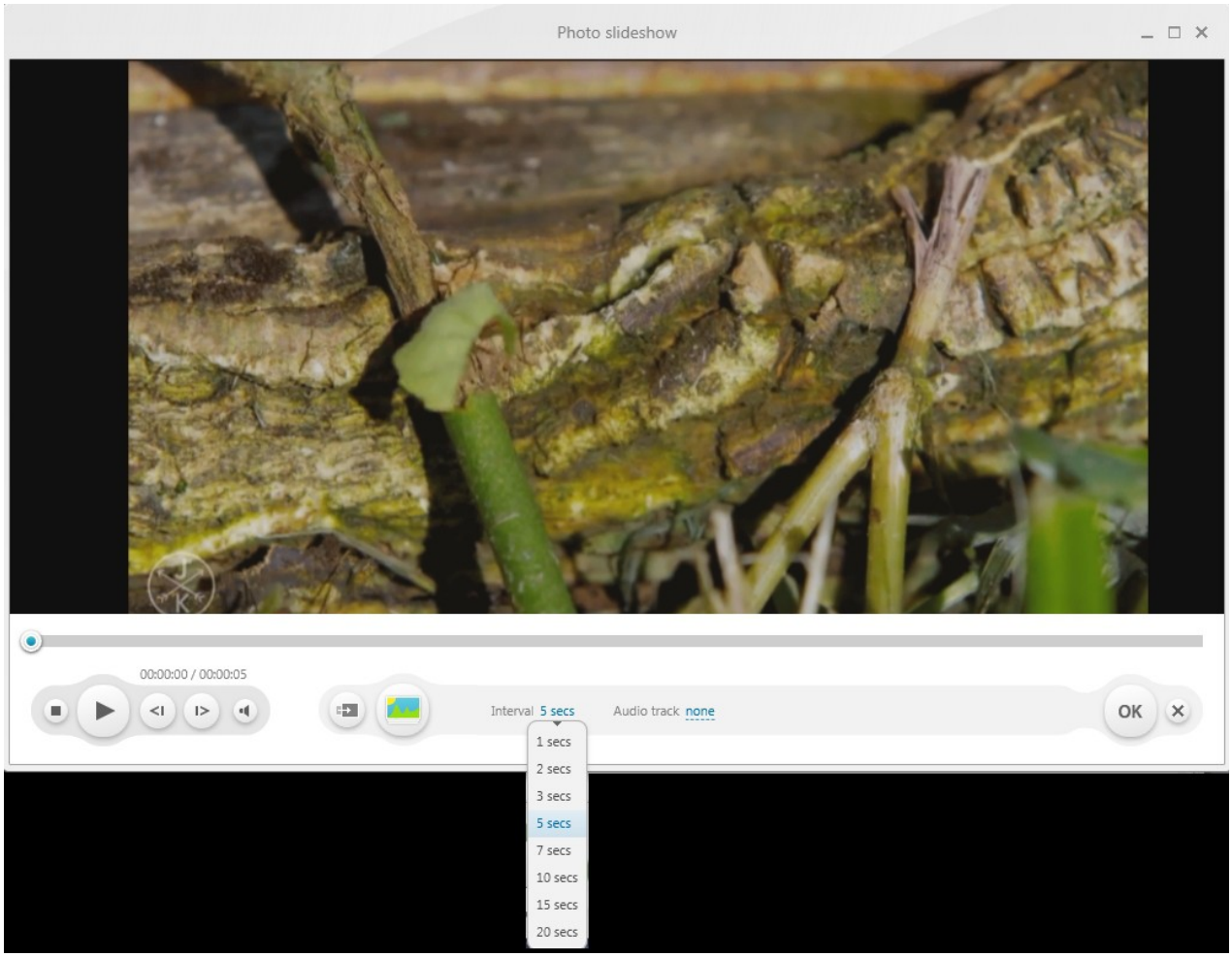


Image 20: Selecting stalling duration with Freemake Video Converter

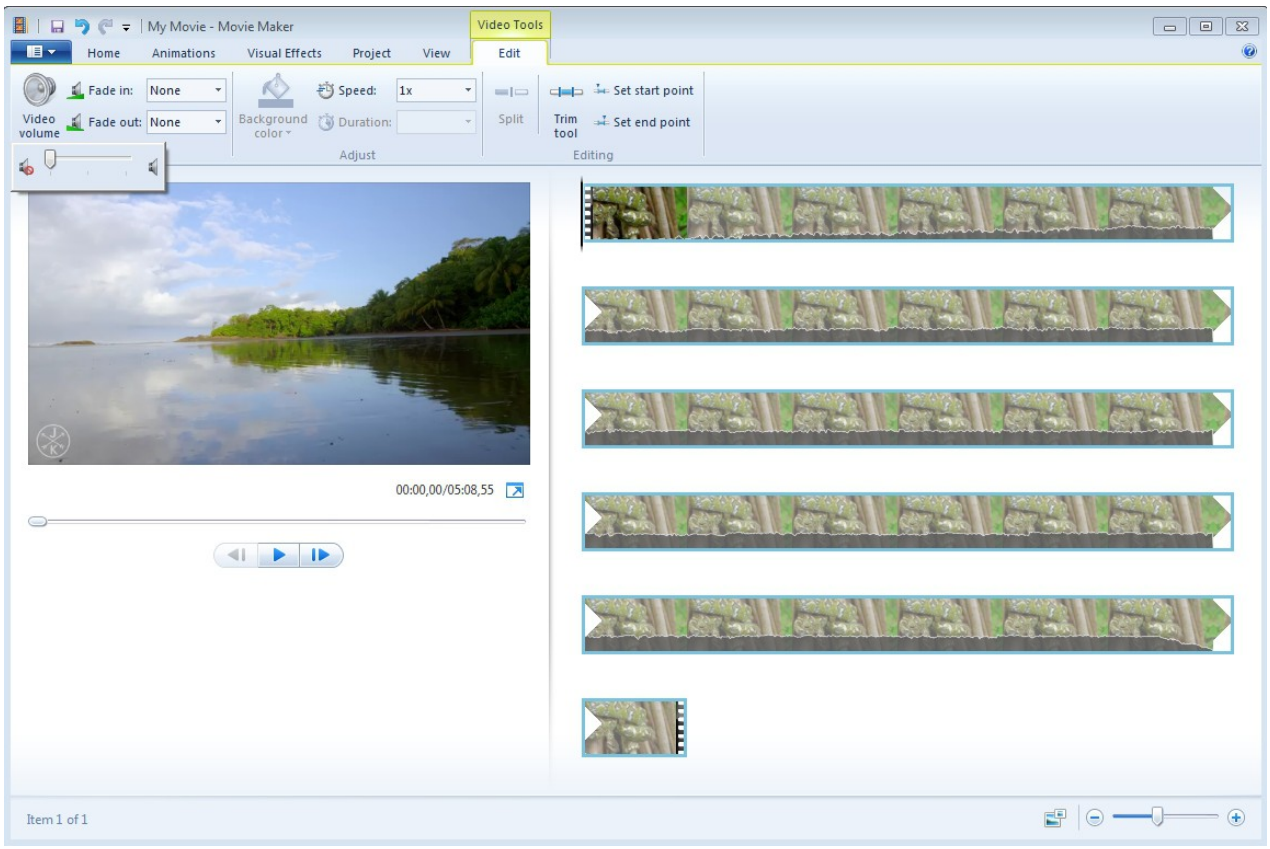


Image 21: Muting original audio with Windows Movie Maker

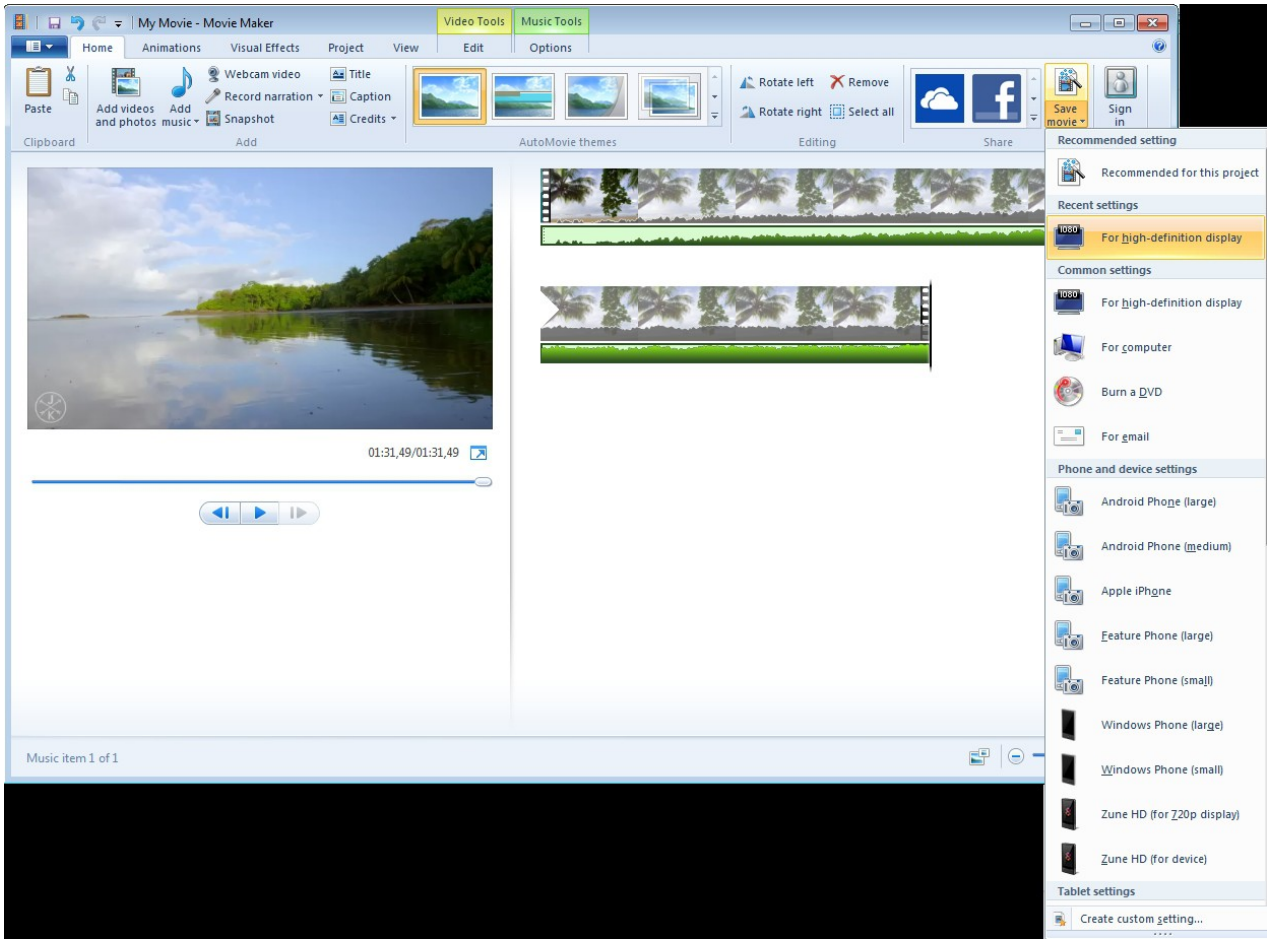


Image 22: Adding new audio and saving with Windows Movie Maker

Image 23: Template for stalling experiments

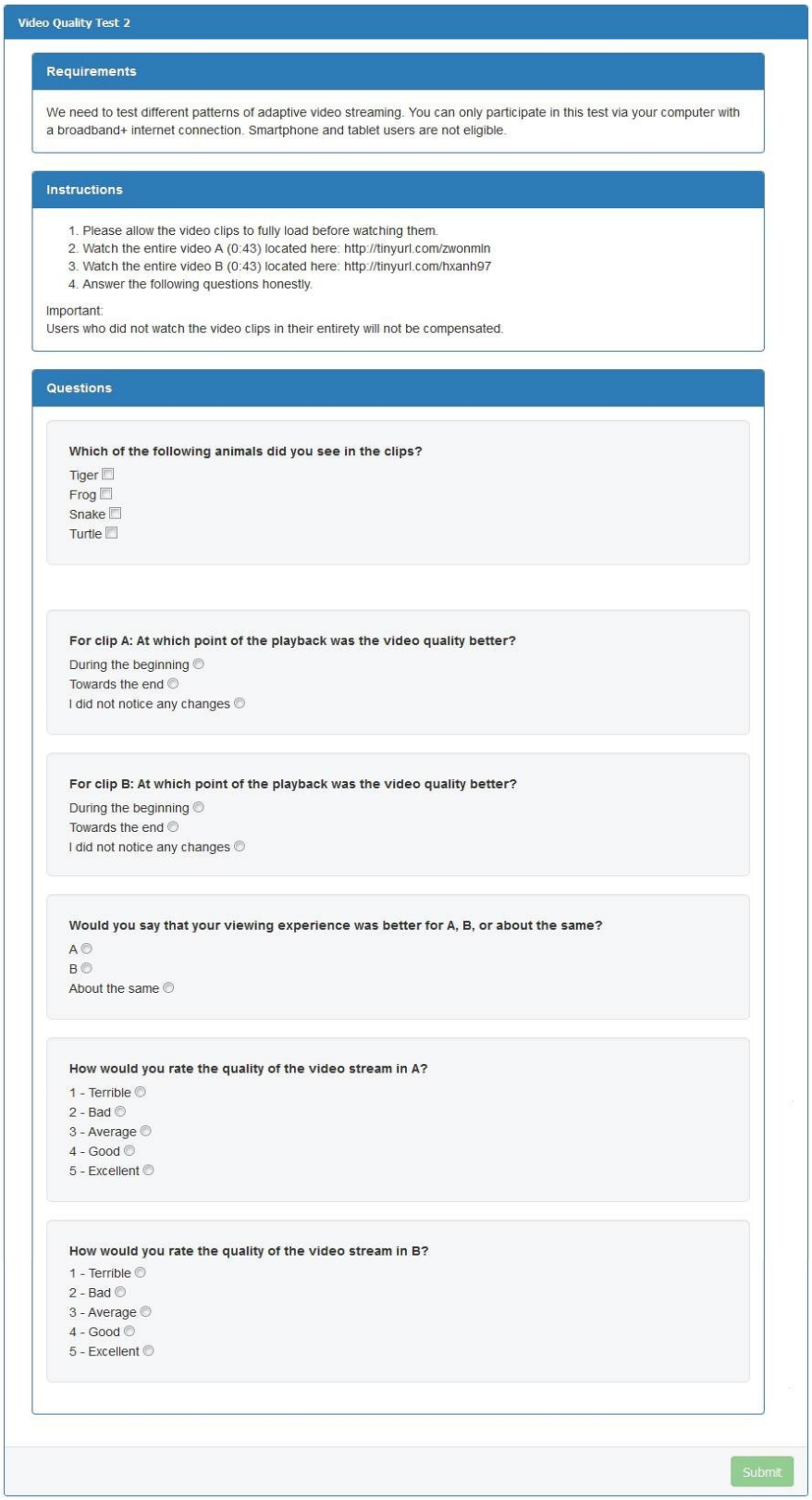


Image 24: Template for adaptive streaming experiments

ABBREVIATIONS

QOE	Quality Of Experience
QOS	Quality Of Service
ITU-T	Internation Telecommunication Unit - Telecommunications
IETF	Internet Engineering Task Force
VOIP	Voice Over Internet Protocol
IPTV	Internet Protocol TeleVision
HTML	Hyper Text Markup Language
HIT	Human Intelligence Task
API	Application Programming Interface
SDK	Software Development Kit
MOS	Mean Opinion Score

REFERENCES

[1] J. Zhang and N. Ansari, *On Assuring End-to-End QoE in Next Generation Networks: Challenges and a Possible Solution*, IEEE Communications Magazine, 2011.
[2] E. Liotou, D. Tsolkas, N. Passas and L. Merakos, *Quality of Experience Management in Mobile Cellular Networks: Key Issues and Design Challenges*, IEEE Communications Magazine, 2005.
[3] R. Stankiewicz, P. Cholda, and A. Jajszczyk, *QoX: What is it Really?*, IEEE Communications Magazine, 2011.
[4] T. Hobfeld, M. Seufert, C. Sieber, T. Zinner, *Assessing Effect Sizes of Influence Factors Towars a QoE Model for HTTP Adaptive Streaming*, University of Wurzburg, Germany, 2013.
[5] R. Schatz, T. Hobfeld, L. Janowski and S. Egger, *From Packets to People: Quality of Experience as a New Measurement Challenge*, Telecommunications Research Center Vienna, Austria, 2013.
[6] T. Hobfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, C. Keimel, *Best Practices and Recommendations for Crowdsourced QoE*, QUALYNET, 2014.
[7] T. Hobfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, *Quantification of YouTube QoE via Crowdsourcing*, IEEE International Symposium on Multimedia, 2011.
[8] T. Hobfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, P. Tran-Gia, *Best Practices for QoE Crowdsourcing: QoE Assessment with Crowdsourcing*, IEEE Transactions on Multimedia, 2014.