MEASURING VIDEO QUALITY IN THE NETWORK: FROM QUALITY OF SERVICE TO USER EXPERIENCE

Nicolas Staelens[1], Margaret H. Pinson[2], Philip Corriveau[3], Filip De Turck[1], and Piet Demeester[1]

[1]Ghent University - iMinds; [2]NTIA/ITS; [3]Intel Corporation

ABSTRACT

Consumers demand video delivery to a wide variety of multimedia capable playback devices in dissimilar contexts. Monitoring the quality of the network provides a preliminary indication of the video quality end-users can expect. However, more advanced measurements are needed to reliably predict perceived quality. In this paper, we provide information on how to measure video quality, from leveraging fundamental and pure network measurements all the way to modeling and measuring perceived video quality. Currently, video quality research is migrating towards measuring Quality of Experience and User Experience, which is known to be highly subjective and dynamic. Therefore, we also present and discuss the research challenges resulting from this change in focus in more detail.

1. INTRODUCTION

The amount of online video is exploding and already accounts for more than half of all mobile Internet traffic globally [1]. Consumers also own a growing number of multimedia-capable devices, from mobile phones to High Definition (HD) and even 4k-capable television sets. Widespread Internet connectivity enables users to watch online video anytime, anywhere. Nowadays, online video consumption refers not only to web-based video (e.g. YouTube or Vimeo) but also to enhanced video broadcasting services, such as Internet Protocol Television (IPTV) and Video on Demand (VoD). These services deliver live content to end-users at numerous resolutions. Key to the success and acceptability of these new digital service offerings are the end-users' overall experience and customer satisfaction.

However, during the transport of video streams over error-prone IP-based network infrastructures, impairments such as packet loss and jitter can severely deteriorate the audiovisual quality as perceived by the end-users. Furthermore, depending on the streaming technology, network impairments will result in different kinds of visible distortions. For example, in the case of RTP-based video streaming, packet loss will typically result in blockiness artifacts [2], whereas network impairments during HTTP Adaptive video Streaming (HAS) can result in video stalling [3]. Despite the fact that best effort packet-based IP networks were not originally designed for the distribution of high-quality video, service providers who use such networks must ensure that their customers receive adequate quality at all times.

The problem of understanding and ensuring a good user experience has produced a variety of methods for measuring video quality:

- **Quality of Service (QoS)**: Objective metrics that measure the performance of IP-based networks and services.
- **Subjective video quality tests**: Subjective experiments that quantify people's perceptions of video quality.
- **Objective video quality**: Metrics that predict subjective video quality.
- **Quality of Experience (QoE)**: "The degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state" [4].
- **User Experience (UX)**: How a person feels about the experience delivered. UX is one factor in QoE.

In this article, we provide an overview of how to measure video quality, both from a network point-of-view and from the end-user point-of-view. We describe how to accurately measure end-users' perception of quality by means of subjective experiments and review existing state-of-the-art objective video quality metrics. Next, we introduce the user experience concept and how it links directly to all of this research by providing an overview of ongoing efforts and challenges concerning measuring QoE and UX. Where needed, the reader is referred to additional resources for more in-depth information on the topics discussed in this article.

2. QoS-BASED VIDEO QUALITY PREDICTION

From a network monitoring point-of-view, video quality should be measured at several demarcation points along the delivery chain. This enables fault detection and fault isolation, which in turn allows service providers to identify possible bottlenecks during video delivery.

QoS metrics use available network diagnostic parameters, such as traffic jitter, arrival time, and packet loss. These metrics are easy to measure at any point inside the delivery network. QoS metrics do not examine packet payloads, and therefore have no access to the audio or video data being streamed.

2.1. Measuring network delivery quality

QoS can be measured objectively by the method described in ITU-T Rec. Y.1540, using calculated values such as packet transfer delay, loss ratio, and service availability. These values can be crosschecked against the IP performance objectives established in ITU-T Rec. Y.1541, in order to guarantee and validate the quality of the delivered services from a network point-of-view.

In order to ensure that end-users receive adequate quality at all times, performance requirements (from the end-users' point-of-view) for DSL networks have been defined by the Broadband Forum in Technical Report TR-126 and in ITU-T Rec. G.1080. In the case of HD video streaming, these requirements specify that, for example, at most one visible impairment can occur for each four hours of video playback.

The Internet Engineering Task Force (IETF) proposed the Media Delivery Index (MDI) as a QoS diagnostic monitoring tool in RFC4445. MDI can be used as an indicator for networks carrying streaming media to assess how well the network can transport (real-time) video. MDI is based on two QoS measurements: Delay Factor (DF) and Media Loss Rate (MLR). DF is a measure of the amount of jitter present, and MLR represents the number of lost or out-of-order media packets. Originally, DF was calculated with the assumption of constant bitrate video streams. The European Broadcasting Union (EBU) has proposed the Time Stamped-DF (TS-DF) [5], which is calculated in a similar way to RTP jitter and works for both constant and variable bitrate videos.

The RTP Control Protocol Extended Reports (RTCP-XR) framework provides a set of more advanced metrics for monitoring VoIP (Video over IP), and IPTV networks. In the case of video, extensions to RTCP-XR have been defined to include measurements such as the proportion of impaired I-frames, the loss rate within BP frames, A/V delay, playout interrupt count, etc. To generate these RTCP-XR reports, network packets need to be parsed up to the level of the video headers.

2.2. Estimating video quality from network parameters

QoS measurements, such as the ones mentioned above, can be used to estimate QoE by means of different, typically non-linear, QoS/QoE mapping functions. An in-depth survey of different mapping functions and techniques is provided by Alreshoode et al. [6]. However, this remains a challenging task as QoE is by its nature highly subjective. The authors conclude that, despite the fact that there are many QoS/QoE correlation models, there is still a need to further investigate the many aspects of QoE and how the two interact with each other.

In 2012, the International Telecommunication Union (ITU) approved ITU-T Recs. P.1201.1 and P.1201.2 for monitoring audio, video, and audiovisual quality of IP-based video services based on packet header information. These metrics are commonly referred to as objective parametric (audio) visual quality metrics. As these metrics only rely on access to the packet header data, they are especially interesting in the case of DRM and encrypted video content. Furthermore, as no decoding is involved, they can easily be placed at several demarcation points inside the delivery network for real-time monitoring.

2.3. Limitations

The advantage of QoS-based metrics is that these models are designed to be easily deployed at any point in the distribution network. The disadvantage is that QoS models do not have access to two pieces of important information: (1) how the video originally looked and sounded, and (2) what the end-user sees and hears. This limits the ability of QoS models to predict quality as perceived by the end user. By parsing the video headers, some additional high-level information can be obtained and, to some extent, be used to better fine-tune quality prediction as is done by the parametric video quality models.

3. MEASURING END-USERS' QUALITY PERCEPTION

QoS is network-oriented. QoS metrics cannot fully quantify end-users' perception of quality due to missing information: the video as seen by the end-user. Subjective video quality tests and objective video quality metrics provide established techniques for end-user point-of-view.

3.1. Subjective video quality experiments

Subjective video quality experiments offer an accurate and repeatable method to estimate people's opinions of short video sequences. Subjects rate each video sequence on a perceptual quality scale such as "excellent, good, fair, poor and bad". These experiments attempt to measure people's immediate impression of a video sequence's

technical quality. Repeatability and accuracy are maximized by eliminating uncontrolled variables. This intentionally excludes the perceptual impact of the video sequence's artistic quality, the monitor, UX and QoE. The focus is the visual impact of distribution network issues (e.g., coding artifacts, or the visual impact of transmission errors).

Table 1 lists ITU Recommendations (Rec.) that describe how to subjectively measure people's immediate opinion of the quality of video and audio.[1] These methodologies provide detailed guidelines on how to set up and conduct different types of subjective quality evaluation experiments. The most appropriate ITU Rec. depends on the question to be answered. Janowski and Pinson [7] provide a tutorial.

Table 1. Subjective Assessment Methodologies Standardized By The ITU

Issue	*ITU Recommendations*
Picture/video quality	ITU-R Rec. BT.500 ITU-T Rec. P.910 ITU-T Rec. P.913
3D video quality	ITU-R Rec. BT.1438
Audio/sound quality	ITU-R Rec. BS.1116 ITU-R Rec. BS.1284 ITU-R Rec. BS.1534
Speech quality	ITU-T Rec. P.805 ITU-T Rec. P.830 ITU-T Rec. P.835 ITU-R Rec. P.880
Audiovisual quality	ITU-T Rec. P.913 (1-way communication) ITU-T Rec. P.911 (1-way communication) ITU-T Rec. P.920 (2-way communication)

3.2. Objective video quality metrics

Objective metrics model subjective quality tests, and thus link perceived video quality to the service provided. These metrics can use the decoded video as seen by the end-user—and thus are typically placed as close as possible to the end-user. There are five types of objective video quality metrics:

- **Full Reference (FR)** metrics estimate perceived quality of a degraded video signal by performing pixel-by-pixel comparisons with the high quality original video. This is the most accurate possible approach, because the metrics "know" exactly what the video should look like and what the end-user sees. FR metrics are typically used out-of-service.
- **Reduced Reference (RR)** metrics are like FR metrics but rely upon low bandwidth features to allow in-service, real-time deployment. RR metrics can only be deployed when the high quality original video is readily accessible. Unfortunately, the location of the head end is often unknown (e.g., due to automatic switching in response to advertising needs). RR metrics can be nearly as accurate as FR metrics.
- **No Reference (NR)** metrics estimate perceived quality using only the decoded video. NR metrics are ideal for industry because they can be deployed on any video stream. However, good NR metrics are extremely difficult to design. They require accurate models of human vision, object recognition, and quality judgment. No such tools exist today—at least, none that have been independently validated and standardized.
- **Bitstream** metrics estimate perceived quality by parsing the network bitstream up to the level of the video payload without a reconstruction of the pixel data. Bitstream metrics are easy to deploy at any point in a distribution network. The disadvantage is that bitstream metrics must be separately trained for each type of coding algorithm. Increased accuracy results from training for each vendor's unique implementation of that coding standard. Parametric video quality metrics (see Section 2.2) can be regarded as a special kind of bitstream metrics.
- **Hybrid perceptual/bitstream (Hybrid)** metrics merge a bitstream model with either an FR, RR, or NR model. The goal is to improve accuracy by using the network bitstream, while maintaining resilience to diverse codec implementations by using the decoded video. The Hybrid-NR model is arguably the most interesting, due to problems with NR models.

The question with any objective video quality metric is whether the metric is accurate enough to be trusted. The only reliable proof of a metric's accuracy is independent validation by a standards organization. Table 2 lists the objective video quality metrics that have been independently validated and standardized.

Table 2. Objective Video Metrics Standardized By The ITU

Issue	*ITU Recommendations*
FR	ITU-T Rec. J.144 & ITU-R Rec. BT.1683 ITU-T Rec. J.247 & ITU-R Rec. BT.1866 ITU-T Rec. J.341 ITU-T Rec. J.340
RR	ITU-R Rec. J.246 & ITU-R Rec. BT. 1867 ITU-T Rec. J.342
Bitstream/Parametric	ITU-T Recs. P.1201 & P.1201.1 & P.1201.2 ITU-T Rec. P.1202 & P.1202.1
Hybrid	ITU-T Rec. J.343 (pending approval)

FR, RR, NR and Hybrid models are validated by the VQEG. Information on the design, implementation, and findings of VQEG validation tests is available at http://www.vqeg.org/. Bitstream models are validated by

[1] Other ITU Recs. exist for related subjective methods, such as telemeetings and video recognition tasks.

ITU-T Study Group 12 (SG12), and the validation test details are not publically available.

4. TOWARDS QUALITY OF EXPERIENCE AND USER EXPERIENCE

The limitation of subjective video quality tests (and by extension objective metrics) is the elimination of uncontrolled variables. While this produces well defined techniques and closed form solutions, it also eliminates factors that impact QoE and UX. Research is being conducted into the open question of how to properly measure and model QoE and its influencing contextual factors during video quality assessment.

4.1. Assessing Quality of Experience

QoE depends on user expectations which, in turn, can be influenced by the person's context and current state. By contrast, the subjective quality assessment methods listed in Table 1 pose stringent limits on the video presentation and eliminate context. Therefore, there is a need for modifying the subjective video quality test methods to include context and uncontrolled variables in order to explore QoE further.

Subjective tests measure peoples' opinion of the current video quality, typically by using short duration video sequences (~10 seconds). Frölich et al. [8] compared subjective quality ratings of video sequences with durations ranging from 10 sec up to 4 min. The study indicated a relatively small influence of duration on subjective scores. There was a small, yet significant improvement in scores for longer durations, particularly for high quality sequences. The high quality and long duration seemed to allow subjects to become immersed in the content and thus less sensitive to impairments.

Staelens et al. [9] investigated the importance of immersion on quality perception. Impairments were embedded within a full length movie that each subject watched at home. The goal was to measure the context impact of the home television viewing experience. During face-to-face interviews, subjects indicated that they did not tolerate impairments that broke playback fluidity. Thus, most subjects found blockiness impairments less annoying than frame freezes, when compared to their ITU-T Rec. P.910 ACR ratings. De Moor et al. [10] also concluded that when an impairment interrupts the video flow, subjects become disengaged and perceived quality drops.

Staelens et al. [11] investigated the importance of task and focus on quality perception. This audiovisual subjective quality experiment included two sets of subjects. The first set contained non-experts, who evaluated lipsync. The second set contained expert interpreters, who both evaluated lipsync and simultaneously translated the video from English to Dutch. The simultaneous translation task significantly influenced the experts' ability to detect lipsync issues.

This is consistent with research by psychologists on the impact of attention and effort, summarized by Kahneman [12]. Active cognitive processing is required to perform two tasks simultaneously. In the absence of sufficient motivation, people tend to focus their attention on one task and let their automatic thought system perform the other. The automatic system occasionally substitutes easier questions for hard questions. The survey study by Matulin et al. [13] also acknowledges that, in general, subjects are more forgiving to quality impairments in the case of video consumption in real-life environments.

These studies show that context impacts the severity and rank ordering of some types of impairments. Potential causes of these differences include the task being performed, the real-life environment, immersion, attention, and memory (e.g., recently viewed video and previously viewed events are judged differently). This indicates that new subjective methods are needed in order to better understand and assess different aspects of QoE.

In [10] and [14], the authors propose the use of a living lab in order to get more user-centric and context-specific insights compared to what is possible using the traditional quality assessment methods and environments. A living lab serves as a controllable home environment for investigating user interaction and user adaption of new technologies in (simulated) real-life situations. Compared to the traditional subjective test environments, the living lab is closer to real-life and is able to better reflect the user's own specific experience of several services [15].

The use of crowdsourcing is also gaining attention for conducting subjective experiments [16]. These web-based video quality experiments target a wide variety of subjects located all over the world. Crowdsourcing tests include the impact of myriad environments and playback devices on QoE. However, this leads to a highly uncontrolled assessment environment. Hence, special measures must be taken into account to identify subjects who cheat or do not performing the task adequately. In [17], the authors list several methods for estimating subjects' reliability and consistency when using crowdsourcing. For example, questions related to the video content might be used to detect subjects not paying attention to the videos.

As QoE is influenced by context and current state, the Experience Sampling Method (ESM) [18] can be used to gain insights into end-users' quality perception. The ESM defines a methodology to question/survey users about their experience (with a particular service). This surveying can be automatically triggered depending on the location of the test subject, at specific time intervals, after particular events, or completely at random. For example, the authors in [19] used the ESM to question users after

using a mobile application in order to evaluate QoE for different applications in different contexts and user environments.

Pinson et al. [20] proposes a new subjective test method for immersive audiovisual quality assessment. This method uses characteristics from the test methodologies in Table 1 yet allows for immersion and testing under more natural viewing experiences. The main differences between the test methods listed in Table 1 and the new proposed methodology are that each source video is only shown once to the test subjects, longer duration video sequences are used (e.g., 1 min duration), and distractor questions are asked after each sequence. The distractor questions focus subjects on the video as a whole and shift the attention towards acceptability of the stimuli for the particular application being tested. As such, this novel method allows for engaging the subjects in the video content while maintaining characteristics from the existing test methods such as experiment repeatability.

4.2. Measuring User Experience

In addition to the influence of expectations, context, and current state, QoE also encompasses the *degree of delight or annoyance of a person experiencing an application, service, or system*. This is UX, or in other words, the measurement of emotions [21]. In traditional approaches, video QoE is quantized by means of the well-known Mean Opinion Score (MOS), which measures perceived quality. Different rating scales are needed to assess UX and thus better understand QoE as a whole.

In [9] and [11], the authors conducted in-depth face-to-face interviews with the test subjects before and after participation in a subjective video quality assessment experiment. These interviews aided in contextualizing the experiments and gathering information beyond MOS. Such interviews reveal important information, but they put an additional workload on the researcher.

UX and emotions can also be quantified by means of specific rating scales. Two commonly used scales are the Self-Assessment Manikin (SAM) [22] and the Differential Emotions Scale (DES) [23]. SAM, as depicted in Figure **1** [24], can be used to measure pleasure, arousal and dominance by means of a pictorial rating scale. DES allows respondents to rate the intensity of their emotions on a 5-point scale with respect to ten different discrete categories: joy, surprise, anger, disgust, contempt, shame, guilt, fear, interest, and sadness.

De Moor et al. [10] conducted a living room lab experiment in order to investigate the relationship between traditional QoE measures and alternative subjective emotions measures including SAM and DES. The authors found a significant positive correlation between overall quality and two self-reported measures, “pleasure” and “enjoyment.” Furthermore, results also indicate that the overall quality (e.g., MOS) is not a good measure of QoE in terms of delight.

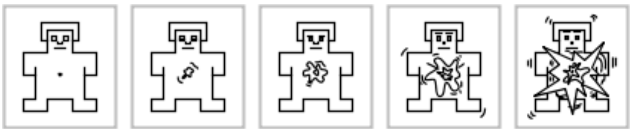

Figure 1. Example 5-grade pictorial Self-Assessment Manikin (SAM) for measuring arousal [24]

Even one step further, Laghari et al. [25] use neurophysiological monitoring tools to measure real-time QoE aspects such as cognition, attention, emotion, and fatigue. Based on three different pilot studies, the authors illustrate that correlations exist between neurophysiological activities and subjects’ arousal, pleasure, attention, and preference during multimedia QoE evaluation.

The interested reader is referred to [21] for more information on the relationship between QoE and UX.

5. CONCLUSION

QoS, subjective tests and objective metrics provide proven tools for measuring video quality. Instead of just letting a system provide a best-effort service, we can pragmatically apply these tools to start down the path of enabling a good experience. With the proper tools in place, the service provider moves from a passive role to an active role of driving good experience.

QoE and UX are relatively new fields. A variety of new subjective methods have been proposed that evaluate the impact of context, environment, application, and mood on the video watching experience.

Subjective quality assessment is the most accurate method for obtaining human judgments on video quality. However, there are some drawbacks. Subjective experiments are time-consuming, expensive, and require specialized expertise. Pragmatically, subjective quality assessment cannot produce real-time quality ratings throughout a distribution network. As QoE and UX research matures, objective metrics will be needed.

All of this pushes the industry towards building robust QoE models. The concept of a QoS model is important, but in reality the current idea of QoS is flawed. Reframing or coupling QoS and QoE is critical to linking engineering and experience. Without a proper definition of the requirements and a way to measure them, there is no way to communicate if a desired experience level was ever met. The QoE framework makes a basic assumption that there is a minimum level of experience that a consumer expects. If the provider cannot meet that level, there is no point in starting the streaming service. The coupling of subjective assessments and objective measurements

allows the two paths of QoS and QoE to be merged in such a way that users' needs are addressed.

The ultimate goal is to achieve a state where the right experience is enabled in the right context at the right time for all consumers. The need is for altered QoS metrics and tools that can proactively drive QoE anywhere in the video ecosystem, from encoding to rendering on the end display.

6. REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2013–2018," *White Paper*, June 2014. Available online: http://www.cisco.com. [Last visited 07 Oct. 2014]

[2] A. R. Reibman et al., "Quality monitoring of video over a packet network," *IEEE Transactions on Multimedia*, Volume 6, Issue 2, April 2004

[3] M.-N. Garcia et al., "Quality of Experience and HTTP Adaptive Streaming: A Review of Subjective Studies," *Proceedings of the 6th International Workshop on Quality of Multimedia Experience*, September 2014

[4] A. Raake, and S. Möller, "Quality and Quality of Experience," in *Quality of Experience: Advanced Concepts, Applications and Methods*, pp. 11-33, Springer, 2014

[5] European Broadcasting Union, "A Proposed Time-Stamped Delay Factor (TS-DF) Algorithm for Measuring Network Jitter on RTP Streams," *EBU Tech Report 3337*, January 2010

[6] M. Alreshoodi, J. Woods, "Survey on QoS/QoE Correlation Models for Multimedia Services," *International Journal of Distributed and Parallel Systems*, Vol. 4, No. 3, May 2013

[7] L. Janowski and M. Pinson, "Video Quality Assessment," *IEEE Signal Processing Magazine*, publication pending, January 2015

[8] P. Frölich et al., "QoE in 10 seconds: are short video clip lengths sufficient for quality of experience assessment?" *Proceedings of the 4th International Workshop on Quality of Multimedia Experience*, July 2012

[9] N. Staelens et al., "Assessing Quality of Experience of IPTV and Video on Demand Services in Real-life Environments," *IEEE Transactions on Broadcasting*, Vol. 56, Issue 4, pp. 458-466, December, 2010

[10] K. De Moor et al., "Evaluating QoE by Means of Traditional and Alternative Subjective Measures: An Exploratory 'Living Room Lab' Study on IPTV," *Proceedings of the 4th Workshop on Perceptual Quality of Systems*, September 2013

[11] N. Staelens et al., "Assessing the Importance of Audio/Video Synchronization for Simultaneous Translation of Video Sequences," *Springer Multimedia Systems*, Vol. 18, Issue 6, pp. 445-457, November 2012

[12] D. Kahneman, *Thinking, Fast and Slow,* New York: Farrar, Straus and Giroux, 2011, pp. 1-105

[13] M. Matulin, and Š. Mrvel, "State-of-the-Practice in Evaluation of Quality of Experience in Real-Life Environments," *Promet – Traffic & Transportation*, Vol. 25, No. 3, pp. 255-263, May 2013

[14] J. Song et al., "QoE Evaluation of Video Services Considering Users' Behavior," *IEEE International Conference on Multimedia and Expo Workshops*, July 2014

[15] M. Eriksson et al., "State-of-the-art in Utilizing Living Labs Approach to User-centric ICT Innovation - A European Approach," *Working paper*, Luleå University of Technology, 2005

[16] T. Hossfeld et al., "Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment," *Proceedings of the 16th International Workshop on Multimedia Signal Processing*, September 2014

[17] T. Hossfeld et al., "Best Practices for QoE Crowdtesting; QoE Assessment with Crowsourcing," *IEEE Transactions on Multimedia*, Vol. 16, No. 2, pp. 541-558, February 2014

[18] J.M. Hektner et al., "Experience Sampling Method: Measuring the Quality of Everyday Life," Sage Publications Inc, 2007

[19] K. Wac et al., "Studying the Experience of Mobile Applications Used in Different Context of Daily Life," *Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack*, August 2011

[20] M.H. Pinson, M. Sullivan, and A. Catellier, "A New Method for Immersive Audiovisual Subjective Testing," *Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 2014

[21] I. Wechsung, and K. De Moor, "Quality of Experience Versus User Experience," in *Quality of Experience: Advanced Concepts, Applications and Methods*, pp. 35-54, Springer, 2014

[22] M.M. Bradley, and P. J. Lang, "Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential," *Journal of Behavior Therapy and Experimental Psychiatry*, Vol. 25, Issue 1, pp. 49–59, March 1994

[23] C.E. Izard, *The Psychology of Emotions*, Springer Science & Business Media, NY, 1991

[24] H. Irtel, *PXLab: The Psychological Experiments Laboratory, Mannhein*, Version 2.1.11. Mannheim: University of Mannheim. Available online: http://www.pxlab.de [Last visited 07 Oct. 2014]

[25] K. Laghari et al., "Neurophysiological Experimental Facility for Quality of Experience (QoE) Assessment," IFIP/IEEE International Symposium on Integrated Network Management, May 2013