Assessing Unreliability in OTT Video QoE Subjective Evaluations Using Clustering with Idealized Data

Jie Jiang *, Petros Spachos§, Mark Chignell**, Leon Zucherman**
*Technology Strategy and Operations, TELUS Communications Company, Toronto, Canada
§School of Engineering, University of Guelph, Guelph, ON, Canada
**University of Toronto, Toronto, ON, Canada

***Abstract*—In this paper, we describe an Over-The-Top (OTT) video Quality of Experience (QoE) subjective evaluation experiment that was carried out to examine variations in the way subjects assess viewing experiences. The experiment focuses on different level of impairment and failure types, using 5-point measurement scales. Clustering is used to differentiate between unreliable and reliable participants, where reliability is defined in terms of criteria such as consistency of rating and ability to distinguish between qualitative differences in level of impairments. The results show that clustering a data set that is augmented with unreliable pseudo-participants can provide a new and improved perspective on individual differences in video QoE assessment.**

I. Introduction

Monitoring and control of quality is an important aspect of many services including Over-The-Top (OTT) video. Although Quality of Service (QoS) can be defined in terms of the physical properties of video transmission, the quality of actual human experience of online video may diverge from the level of quality predicted by service features. For instance, customers might feel that one or two instances of freezing of the video are acceptable, but perceive a large decrement in experienced quality if further instances of freezing occur. Since there are no obvious algorithms for predicting Quality of Experience (QoE) in the face of video impairments, it seems natural to rely on subjective ratings of Technical Quality (TQ), which covers the technical aspects of the signal quality that can be controlled by the network provider. One popular approach is the Mean Opinion Score (MOS) developed by the ITU (ITU-T, 1996, ITU-T Rec 1201) and associated researchers.

Individual subjective ratings are affected by errors. Thus, the judgments of a number of participants are typically averaged to obtain estimates of the "true" values of the construct being judged. However, there may be participants who are unmotivated, incapable of judging the construct accurately, or whose judgments may be unreliable (e.g., they are making judgments based on factors that are not directly connected with the video quality, such as environment parameters and/or equipment capabilities.

In this paper, we address the issue of how participants should be screened in order to ensure that subjective ratings of services, such as online video, are reliable. The analyses reported below were carried out on the results of an experiment described in [1], where the results were analyzed in terms of MOS for TQ. For Telecom Operators, TQ of video is the only data element which can be monitored and managed through probes located in delivering network, so our screening analyses will focus on ratings of TQ.

In the current work, we focus on how to measure the reliability of subjective ratings of TQ for videos that have impairments and failures. A novel application of machine learning is presented and we assess its effectiveness in differentiating groups of participants according to their reliability.

The rest of this paper is organized as follows: In Section II, the related work is reviewed. The experimental design and methodology is described in Section III. Section IV gives a description of the screening methodology followed by a discussion on the results in Section V. Our conclusions are in Section VI.

II. Related Work

In recent years, a number of methods have been proposed for QoE assessemt of multimedia content. The methods can be either subjective [2]–[4] or objective [5]. Participants are required to indicate their opinion in an evaluation process. Their opinion can be used with the Absolute Category Rating (ACR) of the MOS approach, which is widely used in quality assessment studies [6]. In every proposed QoE assessment methodology, the user is the key while the proper screening of the subjects and outlier detection is the main challenge [7], [8], especially in lab experiments [9].

Crowdsourcing, has emerged as a cheaper and quicker alternative to traditional laboratory-based QoE evaluation for video streaming services [10], [11]. Although crowdsourcing makes it possible to reach a large group of people to perform video quality testing, challenges such as validity of results and the trustworthiness of participants still remain unresolved. A cheat detection mechanism was proposed in [12]. The authors propose a QoE framework for crowdsourcing where they also detected participants' problematic inputs through pairwise comparison.

In our previous work, we focused on the design of QoE experiments [1], [13], [14] and the characterization of different Impairments and Failures in a Session Oriented OTT video

[15]. In this work, we focus on the outlier detection mechanism. We propose a simple yet efficient detection approach that affects the TQ aspect of the experiment.

III. Experiment Design and Methodology

The following notational conventions will be used in this paper. Video that fails to start is denoted as NA (Non-Accessibility), and video that fails to play to the end is denoted as NR (Non-Retainability).

A. Experimental Setup

As described in [1], we selected a base set of 30 original unimpaired videos and added different levels of impairments and failures to these videos to create a video library that was used in the experiment. Since video freezing represents the dominant impairment in Hypertext Transfer Protocol (HTTP) based streaming video such as YouTube, our impaired video consisted of the 30 original videos with 1 to 4 freezing events of duration 10 seconds each as shown in Table I.

Videos that failed to play to the end (Retainability Failure, a.k.a NR) consisted of the 30 original videos with premature ending at the 20 second or 40 second mark of the video. The fail-to-start (Accessibility Failure, a.k.a NA) videos consisted of videos with a failure-to-play message displayed either immediately, or as a message displayed 10 seconds after the video was due to start playing.

Ratings were collected immediate after each video was viewed. The rating of interest for this paper was "your overall evaluation of the technical quality in the video is". Responses were on the five-point rating scale (1-bad, 2-poor, 3-fair, 4-good, 5-excellent).

The (spatial) resolution of each video was 512×288 pixels. The frame-rate of each video was 30 frames-per-second. The videos consisted of clips of different lengths that varied between 56 seconds to 123 seconds. Of the 30 video clips, 22 were movie trailers of short duration (teaser-trailers) and 8 were short movies. The total length of the experiment was between 1.5 and 2 hours.

B. Experimental Procedure

The experiment was divided into two Sessions, with assessment of 10 videos in Session 1 and 20 videos in Session 2.

- In Session 1, a subject viewed a mix of videos that either had no disruptions, or that had only Integrity impairments (i.e., video ran to the end, and exhibited only Freezing impairments).
- In Session 2, Accessibility and Retainability failures videos were added to the mix of undisrupted and Integrity-impaired videos.

At the beginning of the procedure, all participants filled out an online pre-questionnaire consisting of demographic questions, followed by questions related to video viewing habits and self-assessed levels of patience and tolerance to frustration. The participants then viewed 30 videos spread over the two Sessions, answering questions about acceptability, technical quality, content quality, and overall experience after viewing each video. In order to avoid subject fatigue, the video sessions were separated by a 10-minute break.

TABLE I
Video Disruptions Used in the Experiment

Impairment Set	No. of Videos	Description of Impairment
I_0	30	Unimpaired (Pristine)
I_1	30	Single temporary interruption of $10s$ duration happening at $40s$ (time after video playback start)
I_2	30	Two $10s$ (temporary) interruptions happening at $20s$ and $40s$ respectively
I_3	30	Three $10s$ (temporary) interruptions happening at $10s$, $20s$, $30s$ and $40s$ respectively
I_4	30	Four $10s$ (temporary) interruptions happening at $10s$, $20s$ and $40s$ respectively
NR_1	30	A permanent interruption happening at $20s$
NR_2	30	A permanent interruption happening at $40s$
NA_1	1	Video never starts to play. Video player displays failure-to-play message immediately
NA_2	1	Video never starts to play. Video player displays failure-to-play message after $10s$

Eight disrupted versions (having impairments and/or failures) were prepared for each of the 30 undisrupted videos. These eight versions corresponded to the Impairments and Failures, as listed in Table I. I_0 represented the case of a video that had no Impairments or Failures. Each participant viewed a version of each of the original 30 videos exactly once, so that no repeated content was seen by a particular participant.

There were 20 participants in the experiment. Each participant rated 30 video instances, for a total of 600 video instances rated in the experiment. Table II shows the number of times, per participant, Impairments and Failures appeared in Session 1 and Session 2.

IV. Screening Methodology

A. Labeled Data Creation

A number of methods for screening unreliable participants were reviewed in [16]. Cluster analysis can be used to segment users, with some clusters potentially representing unreliable users [17]. However, there is no guarantee that obtained clusters will differentiate users on the basis of reliability/

TABLE II
Video Disruptions Used in the Experiment

Disruption	Session 1 Freq.	Session 2 Freq.
I_0	3	6
I_1	2	2
I_2	2	2
I_3	2	1
I_4	1	1
NR_1	0	2
NR_2	0	2
NA_1	0	2
NA_2	0	2

TABLE III
DATA REPRESENTING BEHAVIOUR OF PROTOTYPICAL UNRELIABLE PARTICIPANTS.

Participant ID	**RC1**	**RC2**	**RC3**	**IR1**	**IR2**	**IR3**
Behavior Type	Random Clicker	Random Clicker	Random Clicker	Inconsistent Ratings	Inconsistent Ratings	Inconsistent Ratings
Data Creation Method	Creating new data set	Creating new data set	Creating new data set	Implanting to existing data	Implanting to existing data	Implanting to existing data
Data Creation Method Description	Each rating is created by random number generator in the range of scale.	2 rating values on the 5-point scale are selected by random value generator and these two values are randomly assigned to reach rating.	The range of rating values is less than 2 on the 5-point scale, the location of the range is chosen by random number generator. Each rating is then created randomly within the range.	Randomly choose data from existing participant, replace ratings for I0 with 1 or 2 on the 5-point scale randomly.	Randomly choose data from existing participant, replace ratings for NA with 4 or 5 on the 5-point scale randomly.	Randomly choose data from existing data, replace ratings for I0 with 1 or 2 and replace ratings for NA with 4 or 5 on the 5-point scale randomly.

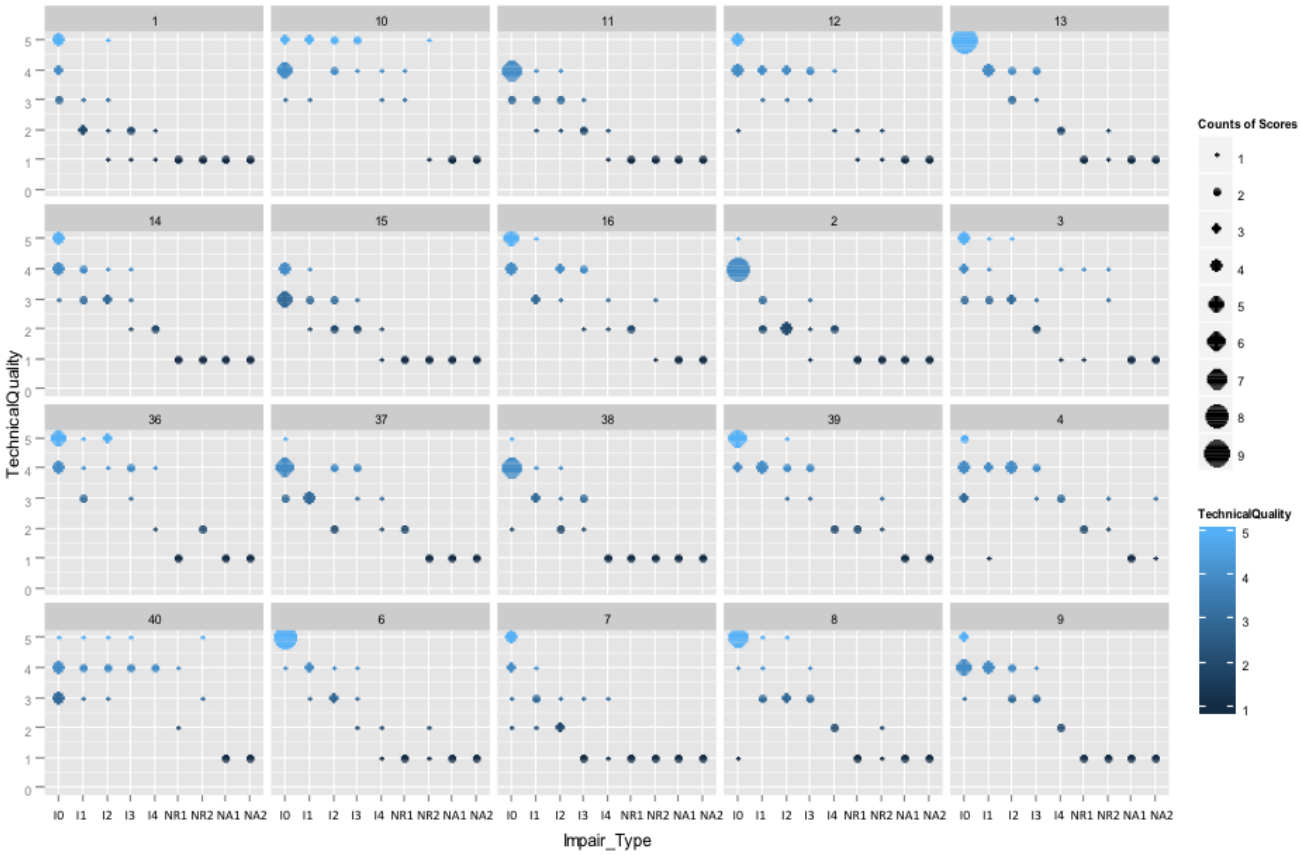

Fig. 1. TQ Ratings across the Disruption Types.

unreliability. For instance, differences in the use of the rating scale (e.g., predominantly higher values for some participants and lower values for other participants) may differentiate clusters in a way that is uninformative concerning the issue of participant reliability.

In this paper, we consider a novel clustering approach for assessing the reliability of ratings of TQ with a clustering approach. In this approach, artificial data, with prototypical patterns of unreliability, is added to the data set so that the patterns in that artificial data influence the clustering process (ideally it would focus clustering on the distinction between reliable and unreliable groups of participants). If the method is effective (in the work reported below we used K-means analysis [18]), then clusters containing the artificially defined unreliable data should also contain real participants who show unreliability in their assessments.

We constructed a set of features similar to [19], which were designed to identify behaviours that are indicative of unreliability, such as Random Clicking (RC) or Inconsistent Rating (IR). We assumed random clicking if ratings had

random values, or if less than half of the full range of the response scale was used. Evidence for inconsistent rating included:

- ratings for the best condition that were less than the midpoint of the scale,
- ratings for the worst condition that were greater than the midpoint of the scale, and
- rating reversals that contradicted an expected directionality in the ratings (e.g., a higher rating was given for a worse condition than for a better condition).

Six new artificial participants were created to represent six patterns of unreliable behaviour drawn from the list provided above. The details of how each data set was created are provided in Table III.

B. Clustering Variables

Three features were chosen for use as clustering variables:

1) Mean TQ score per condition (i.e, disruptions shown in Tables I and II) for each participant;
2) Difference between mean TQ score per condition for each participant and mean TQ score for the group;
3) Correlation between order of TQ ratings from each participant and the expected order of TQ ratings.

The rationale for using the correlation is that there is a natural tendency whereby the more freezings there are in a video, the lower the mean TQ ratings. Thus, for the Impairment data used here (where each instance of freezing is 10 seconds long) the order of TQ ratings should be $I_0 > I_1 > I_2 > I_3 > I_4$. Thus, there should be a correlation between order of TQ ratings from each participant and the expected order of TQ ratings. Note that the correlation is calculated using the Impairment data only.

V. Results and Analysis

Scatter plots for the 20 experimental participants are shown in Figure 1 with TQ rating shown on the y-axis and disruption type on the x-axis of each scatter plot.

It can be seen that there is quite a bit of variation between participants, particularly with respect to sensitivity to different number of freezings in the impairments. It can also be seen that almost all participants rate the TQ of Impairments higher than the TQ of Failures.

K-means clustering analysis was carried out and the two-cluster solution was used to differentiate the characteristics of reliable and unreliable participants. This procedure was applied to the composite data set containing 26 participants (20 real participants and 6 artificially generated participants), with the three clustering variables as described in the previous section.

Table IV lists the result of two-cluster screening of 26 participants, including both original and artificially-unreliable participants, using different combination of feature sets. X's in the cells indicate participants who were placed in the "unreliable" cluster.

Different columns in the Table represent different combinations of the three clustering variables. For instance, the first column shows the results of clustering with all three features (TQ mean, TQ mean difference, and TQ order correlation). It can be seen that the results for the M+MD+OC, M+OC, and MD+OC columns are identical. Thus, when at least two of the features are used, with one of them being OC, the participants identified as unreliable converge to a group that covers all the artificially-unreliable participants, plus a small group of the original participants, i.e., A04, A10, and A40.

As a comparison, Table V shows the results of 2-cluster screening of the original 20 participants using different combination of feature sets. Screening without artificially-unreliable participants produced a larger "unreliable" group. In addition to A04, A10 and A40, A12 and A36 are also identified as

TABLE IV
VIDEO DISRUPTIONS USED IN THE EXPERIMENT

Outlier Identification	**Outlier Feature Sets** M = TQ Mean, MD = TQ Mean Difference, OC = TQ Order Correlation Coefficient						
Participant ID	M + MD + OC	M + MD	M + OC	MD + OC	M	MD	OC
A01							
A02							
A03		x				x	
A04	x	x	x	x		x	x
A05							
A06							
A07							
A08							
A09							
A10	x	x	x	x	x	x	x
A11							
A12		x				x	
A13							
A14							
A15							
A16		x				x	
A36		x				x	
A37							
A38							
A39		x				x	
A40	x	x	x	x	x	x	x
IR1	x	x	x	x		x	x
IR2	x	x	x	x	x	x	x
IR3	x	x	x	x	x	x	x
RC1	x	x	x	x	x	x	x
RC2	x	x	x	x	x	x	x
RC3	x	x	x	x	x	x	

TABLE V
VIDEO DISRUPTIONS USED IN THE EXPERIMENT

Outlier Identification	**Outlier Feature Sets** M = TQ Mean, MD = TQ Mean Difference, OC = TQ Order Correlation Coefficient						
Participant ID	M + MD + OC	M + MD	M + OC	MD + OC	M	MD	OC
A01							
A02							
A03	x	x	x	x	x	x	
A04	x	x	x	x	x	x	x
A05						x	
A06							
A07							
A08	x			x		x	
A09	x			x		x	
A10	x	x	x	x	x	x	x
A11							
A12	x	x	x	x	x	x	x
A13	x	x		x	x	x	
A14	x			x		x	
A15							
A16	x	x	x	x	x	x	
A36	x	x	x	x	x	x	x
A37	x			x		x	x
A38							
A39	x	x	x	x	x	x	
A40	x	x	x	x	x	x	x

unreliable participants by all possible combination of feature sets. Visual inspection of the scatter plots for those participants suggests that their evaluations also differ from those of other participants and should be labeled as unreliable.

Examination of other "unreliable" participants in Table IV such as A03 and A13 also suggested that while they have some abnormal data points they are not necessarily unreliable in a broad sense.

Figure 2 shows a bar chart of TQ ratings across disruption types. In each pair of bars, the left bar shows the results without A03, A10 and A40, while the right bar shows the results for all 20 original participants. While the effect of removing these three unreliable participants is relatively modest, there is greater difference between I_0 and the other Impairments for the "reliable" group, and lower ratings for I_3 and I_4, and for the retainability Failures.

VI. Conclusions

Overall, traditional screening methods tend to screen in terms of one dimension of the unreliability space, but it may be preferable to use a screening that considers various types of unreliability. In this paper, we showed that seeding cluster analysis with idealized data that represents different types of reliability leads to a focused group of unreliable participants being identified, providing that an appropriate set of clustering variables are used in the analysis. In the example considered here, three of the 20 participants were identified as unreliable and the removal of their data led to differences in TQ ratings across video disruptions that were interpreted as beneficial. Correlation with the expected ordering of impairments, and differences between individual mean TQ scores (per disruption) and group mean TQ scores were found to be effective as cluster variables for the reliability analysis. Based on the present results it is proposed that seeding appropriately constructed cluster analyses with idealized artificial data (representing specific patterns of unreliability) should be considered as a useful technique for screening out unreliable participants in studies involving the measurement of QoE and related constructs.

VII. Acknowledgments

This research was supported by a grant from TELUS and a matching grant from NSERC/CRD. The authors thank Alberto Leon Garcia for his comments and Weiwei Li for her help with the experiments.

References

[1] W. Li, H. U. Rehman, D. Kaya, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Video quality of experience in the presence of accessibility and retainability failures," in *Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on*, Aug. 2014, pp. 1–7.

[2] K. T. Chen, C. C. Tu, and W. C. Xiao, "Oneclick: A framework for measuring network quality of experience," in *INFOCOM 2009, IEEE*, April 2009, pp. 702–710.

[3] ITU-T, "Methods for subjective determination of transmission quality," *ITU-R Recommendation P.800*, 1996.

[4] ——, "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, 2008.

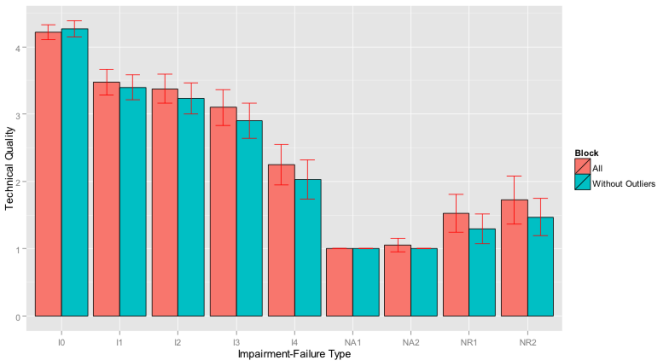

Fig. 2. TQ Ratings Across Disruption Types with and without Outliers.

[5] P. Brooks and B. Hestnes, "User measures of quality of experience: why being objective and quantitative is important," *IEEE Network*, vol. 24, no. 2, pp. 8–13, March 2010.

[6] T. Hossfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," in *Multimedia (ISM), 2011 IEEE International Symposium on*, Dec. 2011, pp. 494–499.

[7] J. Zhang and N. Ansari, "On assuring end-to-end qoe in next generation networks: challenges and a possible solution," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 185–191, July 2011.

[8] K. Mitra, A. Zaslavsky, and C. Åhlund, "Context-Aware QoE Modelling, Measurement, and Prediction in Mobile Computing Systems," *IEEE Transactions on Mobile Computing*, vol. 14, no. 5, pp. 920–936, May 2015.

[9] K. De Moor, I. Ketyko, W. Joseph, T. Deryckere, L. De Marez, L. Martens, and G. Verleye, "Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting," *Mobile Networks and Applications*, vol. 15, no. 3, pp. 378–391, 2010.

[10] L. Anegekuh, L. Sun, and E. Ifeachor, "A screening methodology for crowdsourcing video qoe evaluation," in *Global Communications Conference (GLOBECOM), 2014 IEEE*, Dec. 2014, pp. 1152–1157.

[11] B. Gardlo, M. Ries, and T. Hossfeld, "Impact of screening technique on crowdsourcing qoe assessments," in *Radioelektronika (RADIOELEKTRONIKA), 2012 22nd International Conference*, April 2012, pp. 1–4.

[12] C. C. Wu, K. T. Chen, Y. C. Chang, and C. L. Lei, "Crowdsourcing multimedia qoe evaluation: A trusted framework," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1121–1137, Aug. 2013.

[13] P. Spachos, W. Li, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Acceptability and Quality of Experience in over the top video," in *Communication Workshop (ICCW), 2015 IEEE International Conference on*, June 2015, pp. 1693–1698.

[14] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Impact of technical and content quality on overall experience of ott video," in *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Jan 2016, pp. 930–935.

[15] A. Leon-Garcia and L. Zucherman, "Generalizing MOS to assess technical quality for end-to-end Telecom session," in *Globecom Workshops (GC Wkshps), 2014*, Dec. 2014, pp. 681–687.

[16] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.

[17] M. Chignell, W. Li, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Enhancing reliability through screening and segmentation: An online video subjective quality of experience case study," *Procedia Computer Science*, vol. 69, pp. 55 – 65, 2015, the 7th International Conference on Advances in Information Technology. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915031701

[18] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979. [Online]. Available: http://dx.doi.org/10.2307/2346830

[19] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Norwell, MA, USA: Kluwer Academic Publishers, 1998.