# HOW MUCH BANDWIDTH DOES SURVEILLANCE SYSTEM REQUIRE?

*Zengmin Xu*[1,2,3], *Ruimin Hu*[1,2], *Jun Chen*[1,2], *Hongyang Li*[1,2], *Huafeng Chen*[1,2]

[1]National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan, 430072, China
[2]Collaborative Innovation Center of Geospatial Technology, Wuhan, 430072, China
[3]School of Mathematics and Computational Science, Guilin University of Electronic Technology, Guilin, 541004, China

## ABSTRACT

One of the main challenges in surveillance systems lies in the massive amount of video involved in providing potential key content with sufficient resolution. This paper shows that there exists a sweet spot, which we term critical video quality that can be used to reduce bitrate of video transmission without significantly affecting the accuracy of the surveillance tasks. We present a new city surveillance dataset which was divided into three types of scenarios, and we analyze subjective data collected via human subjective testing for object identification. These data are then used to create objective measurements (models) to drive video compression ratio based on the detection probability. The main idea is to find out the lowest bitrate of video transmission while maximizes the probability of detecting objects which are carried or abandoned. Experiment results shown that our generalized models can predict acceptable video quality for object identification in rational ways.

***Index Terms***— Task-based video, object identification, surveillance dataset, compression ratio.

## 1. INTRODUCTION

In 1999, the ITU-T1 P.910 Recommendation [1] introduces the methodology for performing subjective tests in a rigorous manner. Then, in order to solve the problem of quality measurements for task-based video, the ITU-T P.912 Recommendation [2] are proposed in 2008. However, this Recommendation only addresses basic definitions, methods of testing and ways of conducting psycho-physical experiments (e.g. Multiple Choice Method, Single Answer Method, and Timed Task Method). It points out that the traditional Quality of Experience (QoE) methods like absolute category rating is no longer suited to recognition task. Additionally, objective video quality used in Computer Vision (CV) is unfit for recognition tasks. As surveillance systems often streams and stores huge video records, it is very important to develop objective models for video quality assessment.

This research was supported by the National High Technology Research and Development Program of China (2013AA014602), the Internet of Things Development Funding Project of Ministry of industry in 2013(25), the Technology Research Program of Ministry of Public Security (2014JSYJA016), the Major Science and Technology Innovation Plan of Hubei Province (2013AAA020), the Guangdong-Hongkong Key Domain Breakthrough Project of China (2012A090200007), the Nature Science Foundation of Hubei Province (2014CFB712), and the National Nature Science Foundation of China (61367002, U1404618).

Many subjective recognition methods have been proposed over the past decade, but these methods are not context specific, and they do not apply video surveillance-oriented standardized discrimination levels. One of the methods being worth mention is Ghinea's Quality of Perception (QoP) [3,4] and QoP's offshoot—Strohmeier's Open Profiling of Quality (OPQ) [5]. Anyway, these methods do not entirely fit video surveillance needs. The QoP puts stress on video deterioration caused by frame rate (fps), whereas fps not necessarily affects the quality of CCTV and the required bandwidth [6]. The OPQ targets mainly video quality, but discrimination levels, it is more qualitative rather than quantitative.

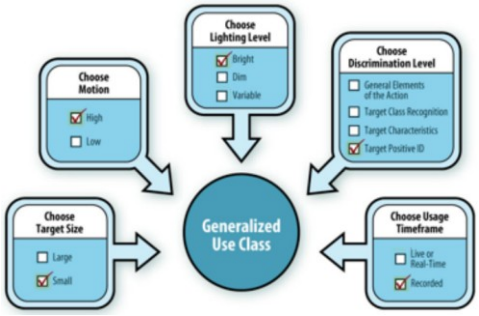



**Fig.1**: A representation of city surveillance application's GUC.

Another method being worth mention is Video Quality in Public Safety (VQiPS) working group [7]. The VQiPS user guide is intended to help the end users determine how their application fits within these parameters [8], these parameters form are referred to the Generalized Use Classes (GUC). Fig.1 is a representation of the GUC determination process. They have performed many subjective experiments to remedy this lack of video quality standards and measurements [9-13]. Other extensive works allow mapping relational rather than absolute quality [14-18], while for video surveillance more generalized framework is needed. Recently, Quality Assessment for Recognition Tasks (QART), driven by Video Quality Experts Group (VQEG) [9], was created for task-based video quality research. QART will address the problems of a lack of quality standards for video monitoring.

The main contribution of this work can be categorized into two folds: First, a new city surveillance dataset and its experimental design are proposed. Second, present a rational compression ratio for trading off video recognition and transmission of the surveillance system. This work provides an important guidance for future quality assessment of surveillance video. **Notice that the proposed dataset also can be used to estimate the object detection accuracy of CV algorithms.**

## 2. WHY SURVEILLANCE EXPERIMENT IS NEEDED?

Surveillance systems often connect a large number of cameras. For example a common storing system in Chicago aggregate at least 10,000 surveillance cameras [19]. We observe that surveillance tasks present an opportunity for a trade-off between the accuracy of the tasks and the bitrate of videos. The precisely computed models can be used to evaluate video quality and optimize the surveillance system. To our knowledge, **this is the first work on objective quality models for complex city surveillance environment**. The surveillance experiment is needed because of following reasons:

For one thing, CV technologies cannot handle all kinds of object detection because of its unsolved drawbacks. Although large-scale surveillance systems often rely on CV algorithms to automate surveillance tasks, in our experiment, **the sizes of objects occupied less than 0.3 percent of the total pixels in the video frame**. Even the state-of-art CV algorithms still struggle to increase the accuracy and speed of object detection in relatively simple conditions [20,21], let alone multi-targets in complex scenario involving occlusion, low lighting level, high motion, small target size or view angle variation. Therefore, CV's drawbacks enforce the development of Quality of Recognition (QoR) experiments. Human Vision cannot be substituted with CV in identifying areas where adequate research has not yet been conducted.

For another, the definition of QoR changes among different recognition tasks, and requires implementation of dedicated quality methods. For example, bronchoscopic diagnosis [22] and license plate recognition [23] aim to recognize specific similar objects in relatively similar scenes. While some identification tasks put stress on abnormal objects rare seen in various different scenes [2,10-12]. Moreover, there are some quality parameters influencing the objective methods, such as source quality of a target. Target velocity and sharpness of video frame referred to motion blur may be the crucial parameters determining the human recognition ability. More important, city surveillance system consists of a large number of cameras, how much bandwidth dose surveillance system requires for HD video transmission? So we should find other parameters like resolution and bitrate of video beyond Ford's framework [8].

## 3. OBJECT IDENTIFICATION EXPERIMENT

This section contains a description of the object identification experiment. The presented designing phase of the experiment reveals differences between the traditional QoE assessment and the task-based quality assessment tests.

### 3.1. Proposed Dataset and Experimental Design

In the latter case methodology of subjective tests is not suited to the task-based video quality assessment. Task-based videos require special methods of testing for different purposes. ITU-T P.912 Recommendation introduces basic definitions and ways of conducting experiments. Thus, a subjective experiment is carried out.

#### *3.1.1. Targets and Scenario Groups*

Given the practical applications in urban areas, we selected interphone, hammer, knife, beer bottle, plastic cup and brick as objects for identification. In order to find out the acceptable perceived quality of object identification, another 14 objects which have different material and similar shapes (e.g. gun, cell phone, axe, screwdriver, steel tube, crabstick, tin, mug, steel cup, flashlight, umbrella, wallet, book and packing box), are given into multiple-choice answers for confusion. In addition, there is "not mentioned above" option inside.

In order to perform the analysis, a 4.2 mega-pixel camera with a CMOS sensor was located 4 meters high to simulate the surveillance application. We chose three types of scenarios for testing human identification ability. The first two types of scenarios consisted of three parts: (1) a person walking with a handheld item, (2) some stationary objects in the middle of the screen, (3) an acuity chart on the wall 20 meters from camera location, or a car parking in the scene. The other type is presented including occlusions. Each type of scenarios has 6 different SRCs (Source Reference Circuit). Example frames cropped from scenes are shown in Fig.2.

(a)close range scene (b)wide range scene (c)wide&complex scene

**Fig.2**: Example frames cropped from our surveillance dataset.

As is shown in Fig.2(a), a person holding knife is walking across the field of view (FOV), while a brick is put on the ground and an acuity chart on the wall. Fig.2(b) states another cropped frame that a person is walking through the scene with plastic cup. Notice that there is a beer bottle at the left bottom of the picture, and a car parking nearby. Fig.2 (c) demonstrates that volunteers are walking around the street, sometimes motorbikes and cars go through. The identification task becomes harder due to occlusions. The distances of walking volunteers in scenarios were designed at 16, 20 and 40 meters far from camera respectively. Objects on the ground with different poses were 14 meters away. This is because these distances are common in realistic surveillance situations.

#### *3.1.2. Testers*

This test simulates recorded video by allowing viewers significant control over how and when video sequences are displayed. The video sequences represent a variety of target sizes, motion, and scene conditions. We invited 48 testers to complete the QoR test. Unlike the experimental procedure in those recognition works [2,23], **we do not screen the *n*-th tester with randomly selected sequences**, this is because our test only contained 9 selected objects. Therefore, in order to verify the identification effect of HRCs (Hypothetical Reference Circuit), we screened each tester 9 HRCs from 9 different SRCs only once. The *i*-th HRC with different parameters was shown to the *n*-th tester in the following way:

$$HRC(i) = mod\left(n + SRC(i),\ 9\right) + 1 \qquad (1)$$

A single SRC distorted by *n* different HRCs generate *n* PVSs (Processed Video Stream). Each PVS included at least three questions: (1) select a handheld item, (2) select some stationary objects on the ground, and (3) select objects' colors. In addition, there was a question only shown in the first PVS: (4) recognize the transformation of "E" in acuity chart or license plate number.

**Fig.3**: An example of SRC video sequence.

The subjective experiment began with two video sequences for training testers to family with scenarios and operation. Then, selected video sequences by equation (1) were screened to testers. Testers should complete the test with the most matches what they saw, while the procedures that video played counts and complete time were recorded. Testers were allowed to stop or playback the video as many times as wish, full screen mode was also permitted. We also provided another 12 AVI video files including handheld items 12 meters away, which were not applied in our subjective experiment but can be used to estimate the object detection accuracy of CV algorithms. **The proposed dataset can be downloaded on the following website: http://pan.baidu.com/s/1pJA1mBp.**

### 3.2. Video Sequences Processing

We gathered 18 SRC video sequences around University in the daytime. Original source sequences were filmed in HD video format with resolution of 1920×1280 pixels and a frame rate of 25 fps. The test clips were impaired using H.264 compression and down-converted to 1280×720 pixels. Using twelve-fold optical zoom, three different sizes of FOV were obtained. 10m×13m FOV was obtained in the close range scene, 15m×25m FOV was obtained in the wide range scene, and 20m×50m FOV was obtained in wide & complex scene approximately. The camera was placed statically without changing optical zoom. Each video was cut into 7 second shots.

Because the Detection Probability (DP) mainly correlates with the object size and size of image detail, it is possible to use lossy compression video which has not caused distortion visible to police for investigation, a rational solution for scaling compression is to operate the codec Quantisation Parameter (QP), while the frame rate is kept intact as their deterioration does not conduct to bitrates savings [6]. Thus, before encoding, a QP sets selected to cover the identification ability threshold, had applied to SRC modification. The QP set was:{28,30,32,34,36,38,40,42,44}. Each SRC had been encoded with H.264 by the QP set, resulting in 9 HRCs. The video processing is demonstrated in detail in Fig.4. After compression, the whole dataset contained 162 PVSs.

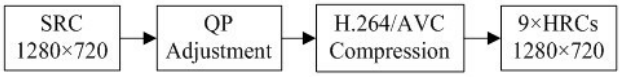



**Fig.4**: Generation of HRCs.

### 3.3. Testers' validation

One of the problems with subjective experiment is the reliability of the subjects. If a subject proves to be unreliable, any conclusions based on his/her answers may be questionable. Therefore it is necessary to detect subjects who do not take experiment seriously. The formal way toward validation of subjects is the Rasch theory [24]. The probability of giving reliable answer is estimated by equation:

$$P(X_{in}=1)=\frac{1}{1+\exp(\beta_n-\delta_i)} \qquad (2)$$

where $\beta_n$ is ability of $n$-th person to make a task and $\delta_i$ is the $i$-th task difficulty. In order to estimate $\beta_n$ and $\delta_i$ values, we combined two custom metrics with Rasch theory.

The first one is Logistic metric. If a subject fails to identify an object for $n$ sequences with higher or equal QP, while the same object was identified correctly by other subjects, the subject's inaccuracy level is increased by $n$. Higher $n$ values may indicate a better chance that the subject is irrelevant and did not pay attention to the recognition task. The next metric should be taken into account is Levenshtein distance, which can be used to estimate the incorrectness level of the answer. Finally, we found 5 subjects undesirable.

## 4. HOW MUCH BANDWIDTH DO WE REQUIRE?

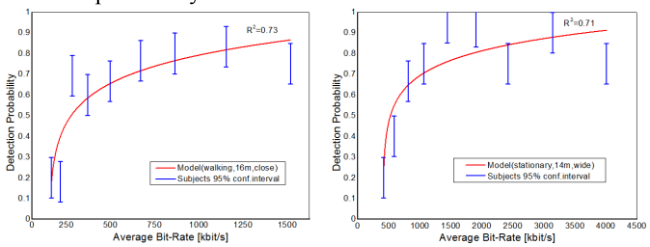
How to trade-off the high recognition ability related to HD video quality and the constraint resources (e.g. transmission bandwidth and storage space)? This section illustrates the procedure of data analysis and quality modeling appropriate to city surveillance. It is different from traditional QoE since the model has to predict probability, not the mean value [25]. It calls to use more general models like Generalized Nonlinear Model [26].

### 4.1. Data Analysis and Quality Model

The goal of this analysis is to find surveillance video quality measurements as a function of certain parameters, i.e. the explanatory variables. The most obvious choice for the explanatory variable is bitrate, which has two useful properties. The first property is a monotonically increasing amount of information, because higher bitrates indicate that more information is being sent. The second advantage is that if a model predicts the needed bitrate for a particular DP, it can be used to optimize the surveillance system.

Each answer of this experiment could be interpreted as two distances identification. Because the video quality is affected by compression with QP Adjustment seriously, **we perceived all QP statistical results of the same distances as the average detection probability**. The average probability of stationary objects (14 meters) being identified correctly was 0.632, and 0.679 recognitions have no more than three errors. The average probability of handheld items, which were 16, 20 and 40 meters far away being recognized correctly, was 0.684, 0.553 and 0.278 respectively. The recognitions of acuity chart and license plate were just used to verify the assessment methods.

The quality model should predict the detection probability of obtaining 1 (correct identification). In such cases, we chose logarithmic for modeling. The first model tested was the simplest one. The ideal obtained model should cross all the confidence intervals for the observed bitrates. Such a model could be used to predict detection probability.




(a) quality model for handheld items objects in specific scenario (b) quality model for stationary in specific scenario

**Fig.5**: Example of the logarithmic modeling and the obtained detection probabilities.

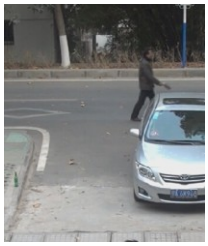
Nevertheless, the result obtained was less precise. Fig.5 shows that the logarithmic modeling approach is not good enough. Some of the points are strongly scattered (see results for bitrate 250 to 300 kbit/s). Hence it is evident that the logarithmic cannot be used. Then the question is, what other functions can be used.

We would like to stress that the SRCs had a strong impact on the DP. In Fig.6 two SRCs were shown for comparison, there is one SRC which was often not detected (see Fig.6(a)). In contrast, another SRC in the same type of scenarios was almost always detected, i.e. even for low bitrates (see Fig.6(b)). This identification problem exists in both stationary objects and handheld item. A detailed investigation shows that the most important factors are:

1. The contrast of the target characters,
2. The characters size, as some of them are more likely to be confused than others or the position of objects put on the ground,
3. The velocity, whether the target is moving fast.

These parameters help to understand what kind of problems might influence DP. The most important factors are differences in spatial and temporal activities and target character size. In this experiment we found factors which influence the DP, but we observed an insufficient number of different values to build a correct model. Therefore, this experiment will help us to design better and more precise experiments in the future.

(a) one SRC which was often not detected.



(b) one SRC was always detected, i.e. even for low bitrates.

**Fig.6**: The SRCs had a strong impact on the DP.

## 4.2. Optimization of perceptual video quality Modeling

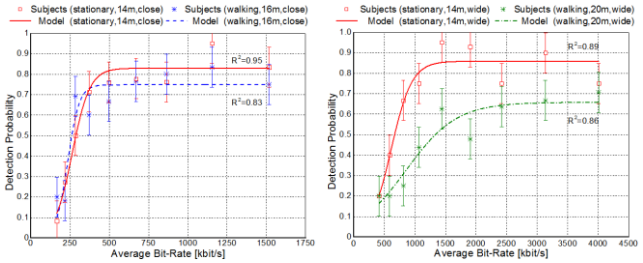
Unlike the license plate recognition task [7,8], **we do not use the threshold detection parameter as video quality**, this is because the detection probability (expressed as video quality) fit our needs. Obviously the accuracy of identification depends on many external conditions and also size of image details. When the optical zoom of a camera does not change, the target size may be too small to distinguish. Therefore, **100% identification cannot be expected, even if many conditions are ideal**. More precise results could be achieved through optimization using a logistic function in Fig.7.

### *4.2.1. Simple scenarios with logistic Modeling*

In our experiments the logistic function is more suitable than logarithmic. Its sigmoid growth curve for population P is widely used in binary response modeling. Ordinary regression deals with a function of the following equation:

$$p_d = \frac{c}{1+\exp(a_0 + a_1 x)} \tag{3}$$

Since each HRC had different bitrates in scenes, **we calculate the average bitrate of PVSs in the same type of scenario and the same QP of HRCs**. They are thus used as the x-axis's points. The average bitrate of PVSs in close range scenarios ranged from 165 to 1517 kbit/s. Fig.7(a) presents an example of the quality models. When bitrate was higher than 500 kbit/s, the DP of stationary objects achieved 0.82, and could not benefit from bitrate increasement. For wide range scenario involving more motions occupied higher bitrate, the average bitrate of PVSs ranged from 419 to 4009 kbit/s. In Fig.7(b) DP of stationary objects gained 0.04 improvement, when bitrate was higher than 1500 kbit/s.




(a) close range scenarios (b) wide range scenarios

**Fig.7**: The obtained detection probability in different scenarios.

Building a detection probability model for all data is difficult. Fortunately, Fig.7 shows both obtained models crosses all the confidence intervals for observed bitrates. The achieved $R^2$ are better than the obtained value via logarithmic model. Such a model could successfully be used to detection probability. Furthermore, Fig.7 shows that the video can be compressed to a lower bitrate as the threshold value for object identification. Therefore, the logistic function is suited to model quality assessment for surveillance task.

### *4.2.2. Combined scenarios with Generalized Modeling*

However, due to a relatively high diversity of scenarios, the bitrates are strongly different. Since the type of scenarios is different, it is evident that the bitrate itself cannot be used as the only explanatory variable. The question is, what other explanatory variables can be used. These two examples given above cannot be combined to produce a more generalized model for recognition task. Nevertheless, the results can be combined if their bitrates (Compressed Data Rates) are first normalized using the Compression Ratio parameter:

$$Compression\ Ratio = \frac{Compressed\ Data\ Rate}{Uncompressed\ Data\ Rate} \tag{4}$$

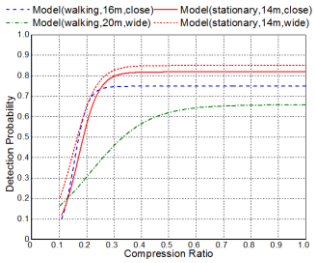



**Fig.8**: Generalized models in combined scenarios.

In Fig.8 we show that the regression curves maintain the same characteristics as Fig.7. We observe that HRCs with QP=28 had the same probability as the SRCs, so we used QP=28 to compress HRCs' bitrate as the uncompressed data rate, and normalize the compression ratios by other HRCs' data rate. Then the generalized model in combined scenarios was generated. The Compression Ratio threshold for stationary items can be observed at around 0.3. Similarly to the other scenarios, the threshold probability is visible.

## 5. CONCLUSION AND FUTURE WORK

The purpose of this paper is to advance the field of quality assessment for task-based video. We outlined a dataset, presented predictive models based on transmission bandwidth and other relevant parameters. Further steps have been planned on extending nightly dataset, researching the mapping function of FOV, investigating H.265 encoders, and studying action recognition of CV algorithms.

## 6. REFERENCES


[1] Telecommunication standardization sector of ITU, "Recommendation 910: Subjective video quality assessment methods for multimedia applications,"*ITU-T Rec.P.910*, pp.1-42, 1999.

[2] Telecommunication standardization sector of ITU, "Recommendation 912: Subjective video quality assessment methods for recognition tasks," *ITU-T Rec. P.912*, pp. 1-16, 2008.

[3] Ghinea Gheorghita, Chen Sherry Y., "Measuring quality of perception in distributed multimedia: Verbalizers vs. imagers," *Computers in Human Behavior*, ELSVIER, vol. 24, pp. 1317-1329, 2008.

[4] Ghinea Gheorghita, Thomas J. P., "Qos impact on user perception and understanding of multimedia video clips," in *sixth ACM international conference on Multimedia*, ACM, pp. 49-54, 1998.

[5] Strohmeier D., Jumisko-Pyykkö S. and Kunze K., "Open profiling of quality: a mixed method approach to understanding multimodal quality perception," *Advances in MultiMedia*, ACM, vol. 2010, 3:1-3:17, 2010.

[6] Janowski Lucjan, Romaniak Piotr, "QoE as a Function of Frame Rate and Resolution Changes," *Future Multimedia Networking*, Springer, vol. 6751, pp. 34-45, 2010.

[7] VQiPS working group. "Defining video quality requirements: A guide for public safety," *Technical report*, U.S. Department of Homeland Security, vol. 1, pp. 1-47, 2010.

[8] Ford Carolyn, Stange Irena, "A framework for generalising public safety video applications to determine quality requirements," in *IEEE Conference on Multimedia Communications, Services & Security*, 2010.

[9] K.Brunnstrom, D.Hands, F.Speranza and A.Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics," in *IEEE Signal Processing Society*, IEEE, vol. 26, pp.96-101, 2009.

[10] PSCR, "Task-Based Tactical and Surveillance Video Quality Tests," *Technical Report*, U.S. Department of Homeland Security, pp. 1-34, 2010.

[11] PSCR, "Recorded - Video Quality Tests for Object Recognition Tasks," *Technical Report*, U.S. Department of Homeland Security, pp. 1-46, 2011.

[12] PSCR, "Assessing Video Quality for Public Safety Applications Using Visual Acuity," *Technical Report*, U.S. Department of Homeland Security, pp. 1-48, 2013.

[13] PSCR, "Video quality tests for object recognition applications," Technical Report, U.S. Department of Homeland Security, pp. 1-24, 2010.

[14] Picard Delphine, Dacremont Catherine, Valentin Dominique and Giboreau Agnès, "Perceptual dimensions of tactile textures", *Acta Psychologica*, ELSVIER, vol. 114, pp. 165–184, 2003.

[15] Faye Pauline, Bremaud Damien, Daubin Mathieu Durand, Courcoux Philippe, Giboreau Agnès and Nicod Huguette, "Perceptive free sorting and verbalisation tasks with naive subjects: an alternative to descriptive mappings," *Food Quality and Preference*, ELSVIER, vol. 15, pp. 781–791, 2004.

[16] Nyman Göte, Radun Jenni, Leisti Tuomas, Oja Joni, Ojanen Hannu, Olives Jean-Luc, Vuori Tero and Hakkinen Jukka, "What do users really perceive - probing the subjective image quality experience," in *SPIE International Symposium on Electronic Imaging: Imaging Quality and System Performance III*, SPIE, vol. 6059, pp. 1-7, 2006.

[17] Radun Jenni, Leisti Tuomas, Häkkinen Jukka, Ojanen Hannu, Olives Jean-Luc, Vuori Tero and Nyman Göte, "Content and quality: Interpretation-based estimation of image quality," *ACM Transaction on Applied Perception*, ACM, vol. 4, pp. 2:1-2:15, 2008.

[18] Duplaga Mariusz, Leszczuk Mikołaj, Papir Zdzisław and Przelaskowski Artur, "Evaluation of quality retaining diagnostic credibility for surgery video recordings," in *10th international conference on Visual Information Systems: Web-Based Visual Information Search and Management*, Springer, vol. 5188, pp. 227–230, 2008.

[19] Adam Schwartz, "Chicago's Video Surveillance Cameras: A Pervasive and Poorly Regulated Threat to Our Privacy," *Northwestern Journal of Technology and Intellectual Property*, Northwestern University School of Law, vol. 11, pp. 47-60, 2011.

[20] Pavel Korshunov, Wei Tsang Ooi, "Video Quality for Face Detection, Recognition, and Tracking," *ACM Transactions on Multimedia Computing, Communications and Applications*, ACM, vol. 7, pp. 14:1-14:22, 2011.

[21] Cheng Ming-ming, Zhang Ziming, Lin Wen-Yan, Torr Philip, "BING: Binarized Normed Gradients for Objectness Estimation at 300 fps," in *Computer Vision and Pattern Recognition, 2014 IEEE Conference on*, IEEE, pp. 3286-3293, 2014.

[22] Leszczuk Mikołaj I, Duplaga Mariusz, "Algorithm for Video Summarization of Bronchoscopy Procedures," *BioMedical Engineering OnLine*, vol.10, pp.110:1-110:17, 2011.

[23] Leszczuk Mikołaj I, "Optimising task-based video quality: A journey from subjective psychophysical experiments to objective optimisation," *Multimedia Tools and Applications*, Springer, vol. 68, pp. 41-58, 2014.

[24] Boone William J., Townsend J. Scott and Staver John, "Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data," *Science Education*, vol. 95, pp. 258–280, 2010.

[25] VQEG, "Report on the validation of video quality models for high definition video content," *Technical report*, VQEG, 2010.

[26] Alan Agresti, "Categorical Data Analysis, 3rd Edition," *Wiley*, 2012.