4

Quality Assessment in Video Surveillance

Mikołaj Leszczuk, Piotr Romaniak and Lucjan Janowski
AGH University of Science and Technology
Poland

1. Introduction

Anyone who has experienced artifacts or freezing play while watching a film or a live sporting event on TV is familiar with the frustration accompanying sudden quality degradation at a key moment. Notwithstanding, video services with blurred images may have far more severe consequences for video surveillance practitioners. Therefore, the Quality of Experience (QoE) concept for video content used for entertainment differs considerably from the quality of video used for recognition tasks. This is because in the latter case subjective user satisfaction depends only or almost only on the possibility of achieving a given functionality (event detection, object recognition). Additionally, the quality of video used by a human observer for recognitions tasks is considerably different from objective video quality used in computer processing (Computer Vision).

So called, task-based videos require a special framework appropriate to the video's function — i.e. its use for recognition tasks rather than entertainment. Once the framework is in place, methods should be developed to measure the usefulness of the reduced video quality rather than its entertainment value. The precisely computed usefulness can be used to optimize not only the video quality but the whole surveillance system. It is especially important since surveillance systems often aggregates large number of cameras which streams has to be saved for possible future investigation. For example in Chicago at least 10,000 surveillance cameras are connected to a common storing system (ACLU, 2011).

To develop accurate objective measurements (models) for video quality, subjective experiments must be performed. For this purpose, the ITU-T[1] P.910 Recommendation "Subjective video quality assessment methods for multimedia applications" (ITU-T, 1999) addresses the methodology for performing subjective tests in a rigorous manner.

In this chapter the methodology for performing subjective tests is presented. It is shown that subjective experiments can be very helpful nevertheless they have to be correctly prepared and analyzed. For illustration of the problem, license plates recognition analysis is shown in detail.

2. Related work

Some subjective recognition metrics, described below, have been proposed over the past decade. They usually combine aspects of Quality of Recognition (QoR) and QoE. These metrics have been not focused on practitioners as subjects, but rather on naïve participants.

[1] International Telecommunication Union — Telecommunication Standardization Sector

The metrics are not context specific, and they do not apply video surveillance-oriented standardized discrimination levels.

One of the metrics being definitively worth mention is Ghinea's Quality of Perception (QoP) (Ghinea & Chen, 2008; Ghinea & Thomas, 1998). Anyway, the QoP metric does not entirely fit video surveillance needs. It targets mainly video deterioration caused by frame rate (fps), whereas fps not necessarily affects the quality of CCTV and the required bandwidth (Janowski & Romaniak, 2010). The metric has been established for rather low, legacy resolutions, and tested on rather small groups of subjects (10 instead of standardized 24 valid, correlating subjects). Furthermore, a video recognition quality metric for a clear objective of video surveillance context requires tests in fully controlled environment (ITU-T, 2000), with standardized discrimination levels (avoiding ambiguous questions) and with minimized impact of subliminal cues (ITU-T, 2008).

Another metric being worth mention is QoP's offshoot, Strohmeier's Open Profiling of Quality (OPQ) (Strohmeier et al., 2010). This metric puts more stress on video quality than on recognition/discrimination levels. Its application context, being focused on 3D, is also different than video surveillance which requires rather 2D. Like the previous metric, this one also does not apply standardized discrimination levels, allowing subjects to use their own vocabulary. The approach is qualitative rather than quantitative, whereas the latter is preferred by public safety practitioners for e.g. public procurement. The OPQ model is somehow content/subject-oriented, while for video surveillance more generalized metric framework is needed.

OPQ partly utilizes free sorting, as used in (Duplaga et al., 2008) but also applied in the method called Interpretation Based Quality (IBQ) (Nyman et al., 2006; Radun et al., 2008), adapted from (Faye et al., 2004; Picard et al., 2003). Unfortunately, these approaches allow mapping relational, rather than absolute, quality.

Extensive work has been carried out in recent years in the area of consumer video quality, mainly driven by two working groups: VQiPS (Video Quality in Public Safety) (VQiPS, 2011) and VQEG (Video Quality Experts Group) (VQEG, n.d.).

The VQiPS Working Group, established in 2009 and supported by the U.S. Department of Homeland Security's Office for Interoperability and Compatibility, has been developing a user guide for public safety video applications. The goal of the guide is to provide potential public safety video customers with links to research and specifications that best fit their particular application, as such research and specifications become available. The process of developing the guide will have the desired secondary effect of identifying areas in which adequate research has not yet been conducted, so that such gaps may be filled. A challenge for this particular work is ensuring that it is understandable to customers within public safety, who may have little knowledge of video technology (Leszczuk, Stange & Ford, 2011).

In July 2010, Volume 1.0 of the framework document "Defining Video Quality Requirements: A Guide for Public Safety" was released (VQiPS, 2010). This document provides qualitative guidance, such as explaining the role of various components of a video system and their potential impact on the resultant video quality. The information in this document as well as quantitative guidance have started to become available at the VQiPS Website in June 2011 (VQiPS, 2011).

The approach taken by VQiPS is to remain application agnostic. Instead of attempting to individually address each of the many public safety video applications, the guide is based

on commonalities between them. Most importantly, as mentioned above, each application consists of some type of recognition task. The ability to achieve a recognition task is impacted by many parameters, and five of them have been selected as being of particular importance. As defined in: (Ford & Stange, 2010), they are:

1. **Usage time-frame.** Specifies whether the video will need to be analyzed in real-time or will be recorded for later analysis.
2. **Discrimination level.** Specifies how fine a level of detail is sought from the video.
3. **Target size.** Specifies whether the anticipated region of interest in the video occupies a relatively small or large percentage of the frame.
4. **Lighting level.** Specifies the anticipated lighting level of the scene.
5. **Level of motion.** Specifies the anticipated level of motion in the scene.

These parameters form what are referred to as generalized use classes, or GUCs. Figure 1 is a representation of the GUC determination process.

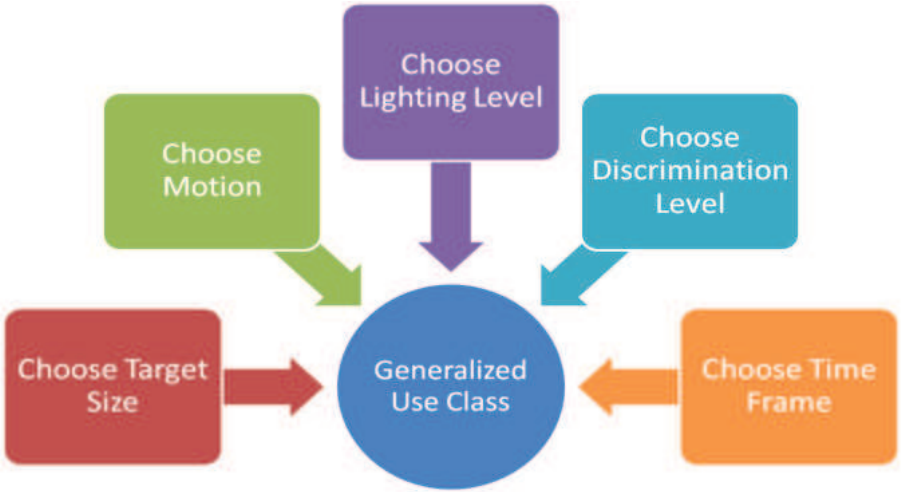


Fig. 1. Classification of video into generalized use classes as proposed by VQiPS (source: (Ford & Stange, 2010)).

The VQiPS user guide is intended to help the end users determine how their application fits within these parameters. The research and specifications provided to users is also to be framed within those parameters. The end user is thus led to define their application within the five parameters and will in turn be led to specifications and other information most appropriate for their needs (Leszczuk, Stange & Ford, 2011).

Recently, a new VQEG project, Quality Assessment for Recognition Tasks (QART), was created for task-based video quality research. QART will address the problems of a lack of quality standards for video monitoring. The aims of QART are the same as the other VQEG projects — to advance the field of quality assessment for task-based video through collaboration in the development of test methods (including possible enhancements of the ITU-T Recommendations), performance specifications and standards for task-based video, and predictive models based on network and other relevant parameters.

3. How it influences the subjective experiment?

The task-based QoE is substantially different from the traditional QoE by different manners. Firstly, as long as a user is able to perform the task we do not have to care if he/she is happy with the overall quality or not up to the level when watching such quality result in fast fatigue of the viewer. Therefore, question about the overall quality does not make much sense. It obviously changes the subjective quality tests significantly.

The second difference between QoE and QoR tests are the source sequences. Let us assume that the task is to recognize if a person on the screen is carrying a gun. In this case more than one source sequence is needed since some alternatives have to be provided. Such alternative sequences have to be very carefully prepared since they should differ from the "gun" sequence by only this one detail. It means that lighting, clouds or any objects at the camera view have to be exactly the same.

The third difference is subjective experiment preparation. In the most traditional QoE experiment a set of parameters of HRC (Hypothetical Reference Circuit) is chosen to produce so called PVS (Processed Video Stream) i.e. a sequence presented to the subjects. A single SRC (Source Reference Circuit) distorted by n different HRCs result in generating n PVSes. In the QoE all those PVSes are shown to a subject so the impact of HRCs can be analyzed. In case of the QoR such methodology is difficult to use. For example in case of plate recognition if a subject recognizes the plates once, he/she can remember them making the next recognition questionable.

Issues of quality measurements for task-based video are partially addressed in the ITU-T P.912 Recommendation "Subjective video quality assessment methods for recognition tasks" (ITU-T, 2008). This Recommendation introduces basic definitions, methods of testing and ways of conducting psycho-physical experiments (e.g. Multiple Choice Method, Single Answer Method, and Timed Task Method), as well as the distinction between Real-Time- and Viewer-Controlled Viewing scenarios. While these concepts have been introduced specifically for task-based video applications in ITU-T P.912, more research is necessary to validate the methods and refine the data analysis methods. In this chapter we present detailed description for which task-based experiments which methodology can be used.

4. What kinds of objective metrics are needed?

In the traditional assessment tests video quality is defined as a satisfaction level of end users, thus QoE. This definition can be generalized over different applications in the area of entertainment. The sources of potential quality degradation are located in different parts of the end-to-end video delivery chain. The first group of distortions can be introduced at the time of image acquisition. The most common problems are noise, lack of focus or improper exposure. Other distortions appear as a result of video compression and processing. Problems can also arise when scaling video sequences in the quality, temporal and spatial domains, as well as, for example, when introducing digital watermarks. Then, for transmission over the network, there may be some artifacts caused by packet loss. At the end of the transmission chain, problems may relate to the equipment used to present video sequences.

In the task-based scenario QoE can be substituted with QoR. The definition of QoR will change along with the specific task requirements and cannot be universally defined. Additionally, subjective answers can be strictly classified as correct or incorrect (i.e. a ground truth is available, e.g. a license plate to be recognized). This is in contradiction to the traditional quality assessment case where there is no ground truth regarding quality.

Because of the above reasons there are additional requirements related to quality metrics. These requirements reflect a specific recognition task but also the viewing scenario. The Real-Time viewing scenario is more similar to the traditional quality assessment tests, although even here additional parameter such as relative target size has to be taken into account. In the case of the Viewer-Controlled viewing scenario additional quality parameters are related to a single shot quality. This is especially important for monitoring objects with a significant velocity. Sharpness of a single video frame (referred to as motion blur) may be a crucial parameter determining the ability to perform a recognition task.

There is one another quality parameter inherent for both viewing scenarios, i.e. source quality of a target. It reflects the ability to perform a given recognition task under the perfect conditions (when additional quality degradation factors do not exist). An example of two similar targets having completely different source quality is two pictures containing car license plate, one taken in a car dealer showroom and one during an off-road race. The second plate may be not only soiled but also blurred due to high velocity of the car. In such a case the license plate source quality is much lower for the second picture what affects significantly recognition ability.

All the additional factors have to be taken into account while assessing QoR for the task-based scenario. The definition of QoR changes between different recognition tasks and requires implementation of dedicated quality metrics.

In the rest of this chapter, as we have already mentioned at the end of Section 1, we would like to review the development of techniques for assessing video surveillance quality. In particular, we introduce a typical usage of task-based video: surveillance video for accurate license plate recognition. Furthermore, we also present the field of task-based video quality assessment from subjective psycho-physical experiments to objective quality models. Example test results and models are provided alongside the descriptions.

5. Case study — License plate recognition test-plan

This section contains a description of the car plate recognition experiment. The purpose of this section is to illustrate an example of a task-based experiment. The presented experiment design phase reveals differences between the traditional QoE assessment and the task-based quality assessment tests. In the following sections, issues concerning the validation of testers and the development of objective metrics are presented.

5.1 Design of the experiment

The purpose of the tests was to analyze the people's ability to recognize car license plates on video material recorded using a CCTV camera and compressed with the H.264/AVC codec. In order to perform the analysis, we carried out a subjective experiment.

The intended outcome of this experiment was to gather the results of the human recognition capabilities. Non-expert testers rated video sequences influenced by different compression parameters. We recorded the video sequences used in the test at a parking lot using a CCTV camera. We adjusted the video compression parameters in order to cover the recognition ability threshold. We selected ITU's ACR (Absolute Category Rating, described in ITU-T P.910 (ITU-T, 1999)) as the applied subjective test methodology.

The recognition task was threefold: 1) type-in the license plate (number), 2) select a car color, and 3) select a car make. We allowed subjects to control playback and enter full screen mode.

We performed the experiment using diverse display equipment in order to be eventually able to analyze the influence of display resolution on the recognition results.

We decided that each tester would score 32 video sequences. The idea was to show each source (SRC) sequence processed under different conditions (HRC) only once and then add two more sequences in order to find out whether testers would remember the license plates already viewed. We screened the n-th tester with two randomly selected sequences and 30 SRCs processed under the following HRCs:

$$\mathrm{HRC} = mod(n - 2 + \mathrm{SRC}, 30) + 1 \tag{1}$$

The tests were conducted using a Web-based interface connected to a database. We gathered

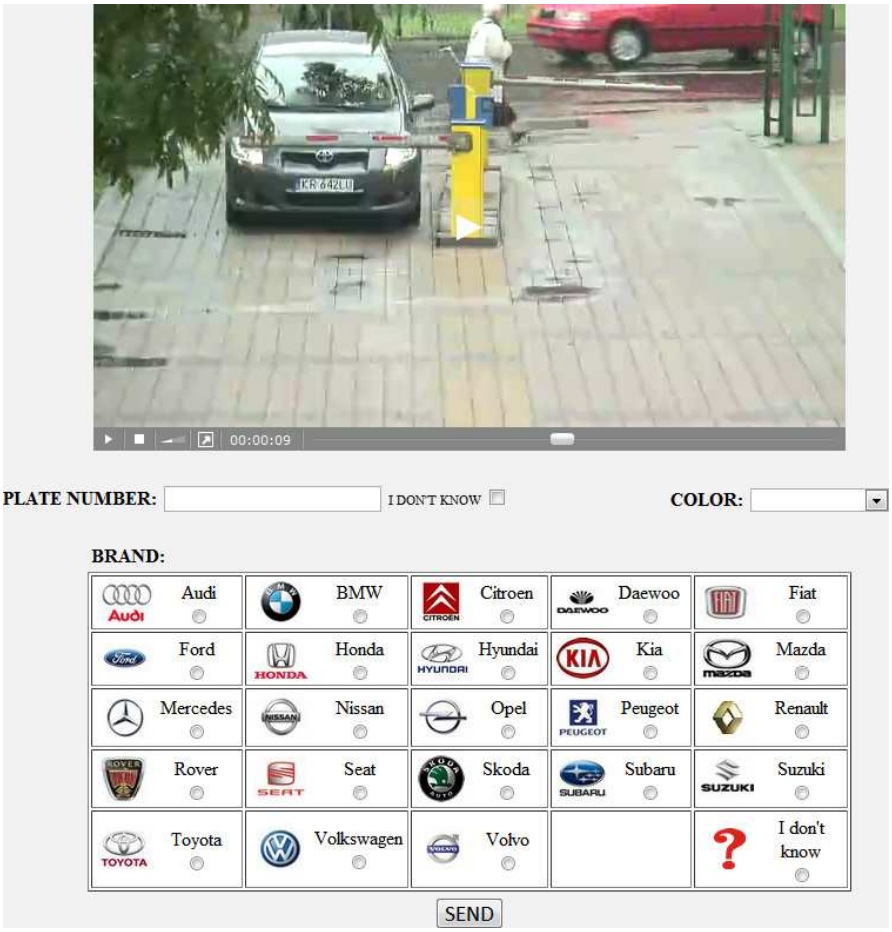


Fig. 2. Test interface.

both information about the video samples and the answers received from the subjects in the database. The interface is presented in Figure 2 (Leszczuk, Janowski, Romaniak, Glowacz & Mirek, 2011).

5.2 Source video sequences

We collected source video sequences at AGH University of Science and Technology in Krakow by filming a car parking lot during high traffic volume. In this scenario, we located the camera 50 meters from the parking lot entrance in order to simulate typical video recordings. Using ten-fold optical zoom, we obtained 6m×3.5m field of view. We placed the camera statically without changing the zoom throughout the recording time, which reduced global movement to a minimum.

We conducted acquisition of video sequences using a 2 mega-pixel camera with a CMOS sensor. We stored the recorded material on an SDHC memory card inside the camera.

We analyzed all the video content collected in the camera and we cut it into 20 second shots including cars entering or leaving the car park. The license plate was visible for a minimum 17 seconds in each sequence. The parameters of each source sequence are as follows:

- resolution: 1280×720 pixels (720p)
- frame rate: 25 frames/s
- average bit-rate: 5.6 — 10.0 Mbit/s (depending on the local motion amount)
- video compression: H.264/AVC in Matroska Multimedia Container (MKV)

We asked the owners of the vehicles filmed for their written consent, which allowed the use of the video content for testing and publication purposes.

5.3 Processed video sequences

If picture quality is not acceptable, the question naturally arises of how it happens. As we have already mentioned at the beginning of Section 5.2, the sources of potential problems are located in different parts of the end-to-end video delivery chain. The first group of distortions (1) can be introduced at the time of image acquisition. The most common problems are noise, lack of focus or improper exposure. Other distortions (2) appear as a result of further compression and processing. Problems can also arise when scaling video sequences in the quality, temporal and spatial domains, as well as, for example, the introduction of digital watermarks. Then (3), for transmission over the network, there may be some artifacts caused by packet loss. At the end of the transmission chain (4), problems may relate to the equipment used to present video sequences.

Considering this, we encoded all source video sequences (SRC) with a fixed quantization parameter QP using the H.264/AVC video codec, x264 implementation. Prior to encoding, we applied some modifications involving resolution change and crop in order to obtain diverse aspect ratios between car plates and video size (see Figure 3 for details related to processing). We modified each SRC into 6 versions and we encoded each version with 5 different quantization parameters (QP). We selected three sets of QPs: 1) {43, 45, 47, 49, 51}, 2) {37, 39, 41, 43, 45}, and 3) {33, 35, 37, 39, 41}. We adjusted selected QP values to different video processing paths in order to cover the license plate recognition ability threshold. We have kept frame rates intact as, due to inter-frame coding, their deterioration does not necessarily result in bit-rates savings (Janowski & Romaniak, 2010). Furthermore, we have not considered network streaming artifacts as we believed that in numerous cases they are related to excessive bit-streams, which we had already addressed by different QPs. Reliable video streaming solution should adjust video bit-stream according to the available network resources and prevent from packet loss. As a result, we obtained 30 different HRC.

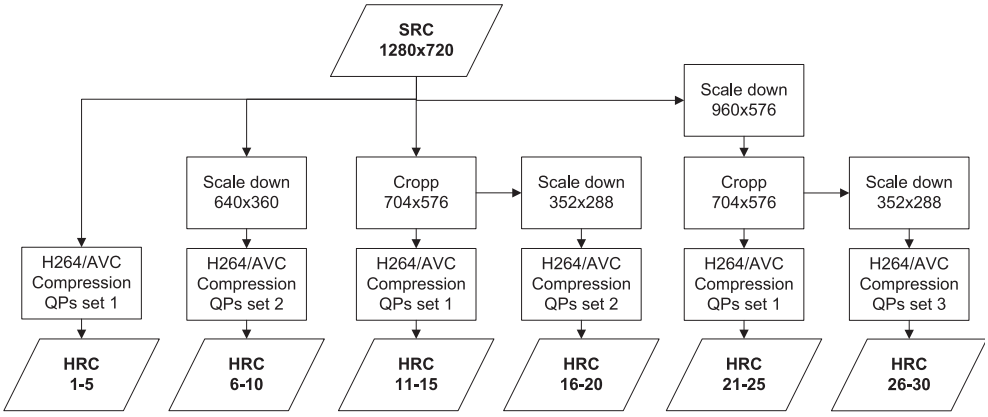


Fig. 3. Generation of HRCs.

Based on the above parameters, it is easy to determine that the whole test set consists of 900 sequences (each SRC 1-30 encoded into each HRC 1-30).

6. Testers' validation

One of the problems with each subjective experiments is reliability of the subjects, or, more precisely, of each individual subject. If a subject proves to be unreliable, any conclusions based on his/her answers may be misleading or simply useless. Therefore the issue of detecting and eliminating unreliable subjects is important in all types of subjective experiments.

Eliminating subjects needs to be based on strict rules, otherwise there is a risk that a subject is eliminated simply because his/her answers do not fit the theory being tested. In other words, any subjective criteria need to be eliminated. The correct methodology should be objective and allow for each subject's individual preferences.

On the other hand, it is necessary to detect subjects who do not take the experiment seriously, i.e. they answer randomly and do not care about giving correct and precise answers. There may also be subjects for whom a given test is too difficult (for example video sequences appear too fast).

The most popular way to validate subjects is correlation. It is a simple and intuitively correct method. We compute correlation between individual subject scores and the scores for all other subjects. We used this method in VQEG in (VQE, 2010). The problems entailed in are: (1) setting the subject elimination threshold (2) eliminating subjects no single answers, and (3) all subjects need to carry out the same tasks numerous times. For the first problem, an experienced scientist can specify correct threshold fitting the problem which he/she is analyzing. The second problem is more difficult to deal with. We know that even for reliable subjects some of their answers are likely to be incorrect. This may be a simple consequence of being tired or distracted for a short period of time. Correlation methodology cannot help in dealing with this problem. The third problem is not important for quality based subjective experiments, since the same sequences are scored in any case (e.g. the same source sequence encoded using different compression parameters). Nevertheless, in task-based subjective experiments the same source sequence should not be shown many times, because the correct

answer for a particular task could be remembered. For this reason different pool of sequences is shown to different subjects (e.g. each compression level for a given source sequences needs to be shown to a different subject).

A more formal way toward validation of subjects is the Rasch theory (Boone et al., 2010). It defines the difficult level for each particular question (e.g. single video sequence from a test set), or whether a subject is more or less critical in general. Based on this information it is possible to detect answers that not only do not fit the average, but also individual subjects' behavior. Formally the probability of giving correct answer is estimated by equation (Baker, 1985)

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

where β_n is ability of nth person to make a task and δ_i is the ith task difficulty.

Estimating both the task difficulty and subject ability make it possible to predict the correct answer probability. Such probability can be compared with the real task result.

In order to estimate β_n and δ_i values the same tasks have to be run by all the subjects which is a disadvantage of the Rasch theory, similarly to the correlation-based method. Moreover, the more subjects involved in the test, the higher the accuracy of the method. An excellent example of this methodology in use is national high school exams, where the Rasch theory helps in detecting the differences between different boards marking the pupils' tests (Boone et al., 2010). In subjective experiments, there are always limited numbers of answers per question. This means that the Rasch theory can still be used, although the results need to be checked carefully. Tasks-based experiments are a worst-case scenario. In this case each subject carries out a task a very limited number of times in order to ensure that the task result (for example license plate recognition) is based purely on the particular distorted video and is not remembered by the subject. This makes the Rasch theory difficult to use.

In order to solve this problem we propose two custom metrics for subject validation. They both work for partially ordered test sets (Insall & Weisstein, 2011), i.e. those for which certain subsets can be ordered by task difficulty. Additionally we assume that answers can be classified as correct or incorrect (i.e. a ground truth is available, e.g. a license plate to be recognized). Note that due to the second assumption these metrics cannot be used for quality assessment tasks, since we cannot say that one answer is better than another (as we have mentioned before, there is no ground truth regarding quality).

6.1 Logistic metric

Assuming that the test set is partially ordered can be interpreted in a numeric way: if a subject fails to recognize a license plate, and for n sequences with higher or equal QP the license plate was recognized 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.

Computing such coefficients for different sequence results in the total subject quality (Sq_i) given by

$$Sq_i = \sum_{j \in S_i} ssq_{i,j} \tag{3}$$

where S_i is set of all sequences carried out by ith subject, and $ssq_{i,j}$ is the subject sequence quality for sequence j, which is given by

$$ssq_{i,j} = \begin{cases} 0 \text{ if } r(i,j) = 1 \\ n \text{ if } r(i,j) = 0 \end{cases} \tag{4}$$

where

$$n = \sum_{k \in A_j} \sum_{m \in B} r(m,k) \tag{5}$$

where $r(i,j)$ is 1 if ith subject recognized the jth sequence and 0 otherwise, B is set of all subjects, and A_j is a set of all not easier sequences as defined above. In the case of this experiment A_j is a set of sequences with the same resolution and view although with a higher or equal QP than the jth sequence.

We computed Sq for each subject; the results are presented in Figure 4.

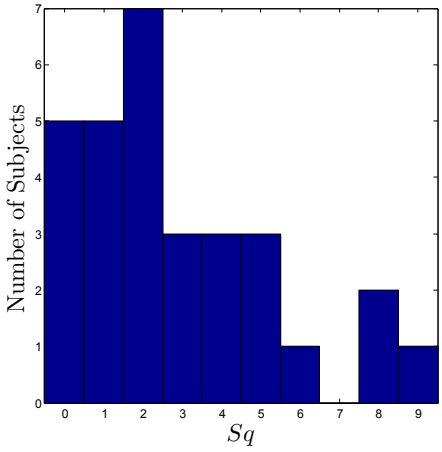


Fig. 4. Histogram of subject quality Sq obtained for all 30 subjects.

The histogram shows that the value of 6 was exceeded for just three subjects, denoted with IDs 18, 40 and 48. Such subjects should be removed from the test.

Sq_i metric assumes that a task can be done correctly or incorrectly. In case of recognition missing one character or all characters the metric returns the same value. In case of license plate recognition and many other tasks the level of error can be defined. The next proposed metric takes into consideration to incorrectness level of the answer.

6.2 Levenshtein distance

Levenshtein distance is the number of edits required to obtain one string from another. Subject quality based on Levenshtein distance Sql is given by

$$Sql_i = \sum_{j=1}^{30} ssql_{i,j} \tag{6}$$

where $ssql_{i,j}$ is subject quality metric based on Levenshtein distance obtained for subject i and sequence j given by

$$ssql_{i,j} = \sum_{k \in A_j} \sum_{m \in B} \begin{cases} 0 & \text{if } l(i,j) \leq l(m,k) \\ l(i,j) - l(m,k) & \text{if } l(i,j) > l(m,k) \end{cases} \quad (7)$$

where $l(i,j)$ is the Levenshtein distance between the correct answer and the subject i answer for the jth sequence, B is set of all subjects, and A_j is the set of all sequences for which the task is not easier defined at the beginning of the section.

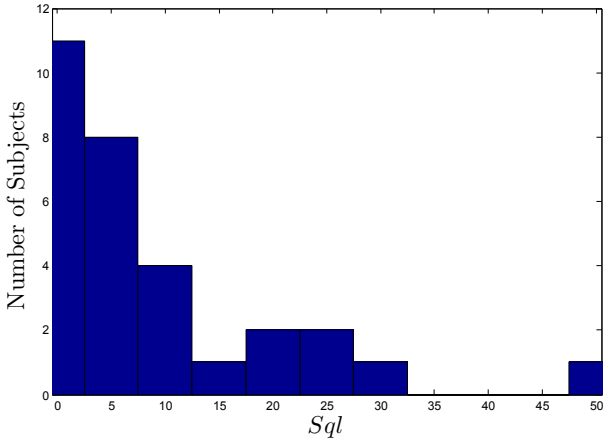


Fig. 5. Histogram of subject quality Sql obtained for all 30 subjects.

Figure 5 shows the histogram obtained for Sql. It is significantly different than the previous one obtained for Sq. It can be assumed that an Sql higher than 10 or 15 indicates a potentially irrelevant subject. One subject obtained significantly higher Sql value than the others (50). More detailed investigation of this case revealed that the subject provided additional text for one answer. After correction the corrected value for this subject is 25, which is still very high.

7. Modeling approaches

In the area of entertainment video, a great deal of research has been carried out on the parameters of the contents that are the most effective for perceptual quality. These parameters form a framework in which predictors can be created such that objective measurements can be developed through the use of subjective testing (Takahashi et al., 2010).

Analysis of the traditional QoE subjective experiment data is focused on the mean subject answer modeling. In addition subject reliability is controlled by the correlation test. Nevertheless, in case of the task-based QoE (QoR) it is impossible or very difficult to use such methodology. Therefore, modeling QoR subjective data calls for new methodology which is presented in this section.

The first step of the subjective experiment data analysis is subject evaluation which is presented in the previous section. The next step of the data analysis is finding the probability

of doing a particular task correctly. Again it is different from traditional QoE since the model has to predict probability not the mean value. It calls to use more general models like Generalized Linear Model (GLZ) (Agresti, 2002.).

The last open problems are explanatory variables i.e. the metrics which can correlate well with the probability of the correct task execution. Assessment principles for task-based video quality are a relatively new field. Solutions developed so far have been limited mainly to optimizing network Quality of Service (QoS) parameters. Alternatively, classical quality models such as the Peak Signal-to-Noise Ratio (PSNR) Eskicioglu & Fisher (1995) or Structural Similarity (SSIM) (Wang et al., 2004) have been applied, although they are not well suited to the task. The chapter presents an innovative, alternative approach, based on modeling detection threshold probabilities.

The testers who participated in this study provided a total of 960 answers. Each answer could be interpreted as the number of per-character errors, i.e. 0 errors meaning correct recognition. The average probability of a license plate being identified correctly was **54.8%** (**526/960**), and **64.1%** recognitions had no more than one error. **72%** of all characters was recognized.

7.1 Answers analysis

The goal of this analysis is to find the detection probability as a function of a certain parameter(s) i.e. the explanatory variables. The most obvious choice for the explanatory variable is bit-rate, which has two useful properties. The first property is a monotonically increasing amount of information, because higher bit-rates indicate that more information is being sent. The second advantage is that if a model predicts the needed bit-rate for a particular detection probability, it can be used to optimize the network utilization.

Moreover, if the network link has limited bandwidth the detection probability as a function of a bit-rate computes the detection probability, what can be the key information which could be crucial for a practitioner to decide whether the system is sufficient or not.

The Detection Probability (DP) model should predict the DP i.e. the probability of obtaining 1 (correct recognition). In such cases, the correct model is logit (Agresti, 2002.). The simplest logit model is given by the following equation:

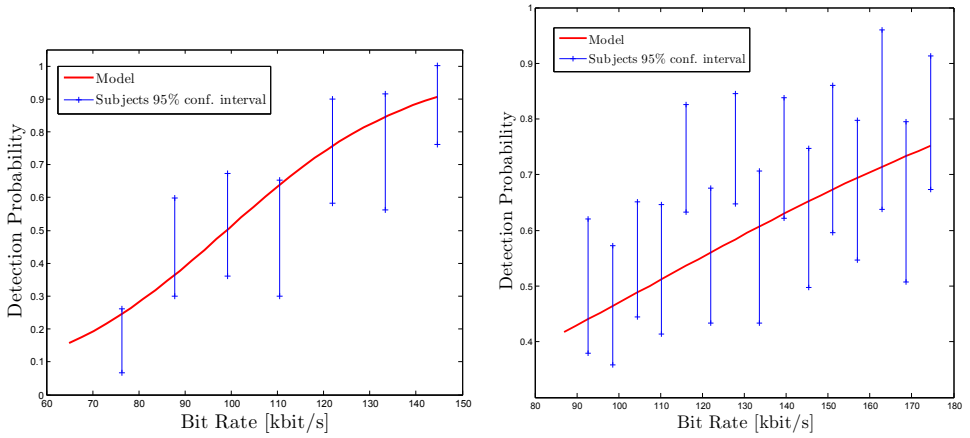
$$p_d = \frac{1}{1 + \exp(a_0 + a_1 x)} \tag{8}$$

where x is an explanatory variable, a_0 and a_1 are the model parameters, and p_d is detection probability.

The logit model can be more complicated; we can add more explanatory variables, which may be either categorical or numerical. Nevertheless, the first model tested was the simplest one.

Building a detection probability model for all of the data is difficult, and so we considered a simpler case based on the HRCs groups (see section 5.3). Each five HRCs (1-5, 6-10, etc.) can be used to estimate the threshold for a particular HRCs group. For example, in Figure 6(a) we show an example of the model and the results obtained for HRCs 20 to 25.

The obtained model crosses all the confidence intervals for the observed bit-rates. The saturation levels on both sides of the plot are clearly visible. Such a model could successfully be used to investigate detection probability.



(a) For HRC 20 to 25. (b) For all HRCs.

Fig. 6. Example of the logit model and the obtained detection probabilities.

We present an extension of the sequences analyzed to all HRCs results in the model drawn in Figure 6(b).

The result obtained is less precise. Some of the points are strongly scattered (see results for bit-rate 110 to 130 kbit/s). Moreover, comparing the models presented in Figure 6(a) and Figure 6(b) different conclusions can be drawn. For example, 150 kbit/s results in around a 90% detection probability for HRCs 20 to 25 and less than 70% for all HRCs. It is therefore evident that the bit-rate itself cannot be used as the only explanatory variable. The question then is, what other explanatory variables can be used.

In Figure 8(a) we show DP obtained for SRCs. The SRCs had a strong impact on the DP. We would like to stress that there is one SRC (number 26) which was not detected even once (see Figure 7(a)). The non-zero confidence interval comes from the corrected confidence interval computation explained in (Agresti & Coull, 1998). In contrast, SRC number 27 was almost always detected, i.e. even for very low bit-rates (see Figure 7(b)). A detailed investigation shows that the most important factors (in order of importance) are:

1. the contrast of the plate characters,
2. the characters, as some of them are more likely to be confused than others, as well as
3. the illumination, if part of the plate is illuminated by a strong light.

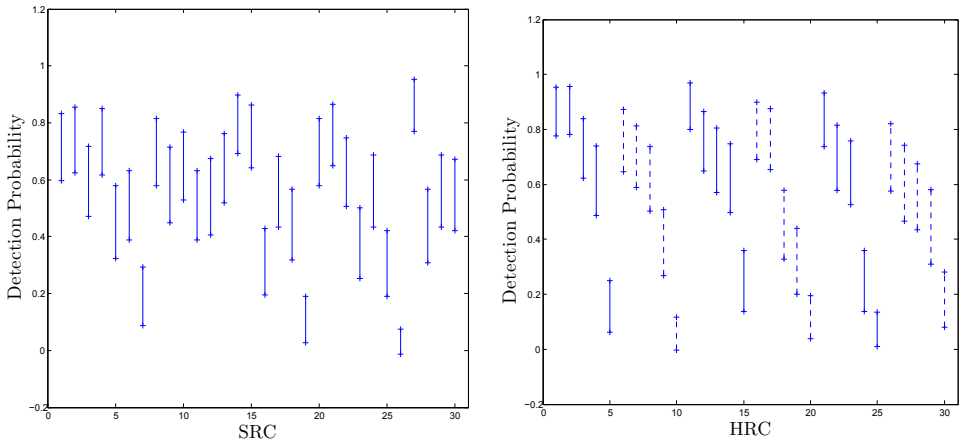
A better DP model has to include these factors. On the other hand, these factors cannot be fully controlled by the monitoring system, and therefore these parameters help to understand what kind of problems might influence DP in a working system. Factors which can be controlled are described by different HRCs. In Figure 8(b) we show the DP obtained for different HRCs.

For each HRC, we used all SRCs, and therefore any differences observed in HRCs should be SRC independent. HRC behavior is more stable because detection probability decreases for higher QP values. One interesting effect is the clear threshold in the DP. For all HRCs groups two consecutive HRCs for which the DPs are strongly different can be found. For example, HRC 4 and 5, HRC 17 and 18, and HRC 23 and 24. Another effect is that even for the same QP the detection probability obtained can be very different (for example HRC 4 and 24).

(a) One SRC (number 26) which was not detected even once.

(b) SRC number 27 was almost always detected, i.e. even for very low bit-rates.

Fig. 7. The SRCs had a strong impact on the DP.



(a) For different SRCs with 90% confidence intervals.

(b) For different HRCs. The solid lines correspond to QP from 43 to 51, and the dashed lines correspond to QP from 37 to 45.

Fig. 8. The detection probabilities obtained.

Different HRCs groups have different factors which can strongly influence the DP. The most important factors are differences in spatial and temporal activities and plate character size. We cropped and/or re-sized the same scene (SRC) resulting in a different output video sequence which had different spatial and temporal characteristics.

In order to build a precise DP model, differences resulting from SRCs and HRCs analysis have to be considered. In this experiment we found factors which influence the DP, but we observed an insufficient number of different values for these factors to build a correct model. Therefore, the lesson learned from this experiment is highly important and will help us to design better and more precise experiments in the future.

7.2 Alternative way of modeling perceptual video quality

For further analysis we assumed that the threshold detection parameter to be analyzed is the probability of plate recognition with no more than one error. For detailed results, please refer to Figure 9.

It was possible to fit a polynomial function in order to model quality (expressed as detection threshold probability) of the license plate recognition task. This is an alternative, innovative approach. The achieved R^2 is 0.86 (see Figure 9). According to the model, one may expect 100% correct recognition for bit-rates of minimum around 360 kbit/s and higher. Obviously accuracy of recognition depends on many external conditions and also size of image details. Therefore 100% can be expected only if other conditions are ideal.

Unfortunately, due to relatively high diversity of subjective answers, no better fitting was achievable in either case. However, a slight improvement is likely to be possible by using other curves.

Summarizing presentation of the results for the quality modeling case, we would like to note that a common method of presenting results can be used for any other modeling case. This

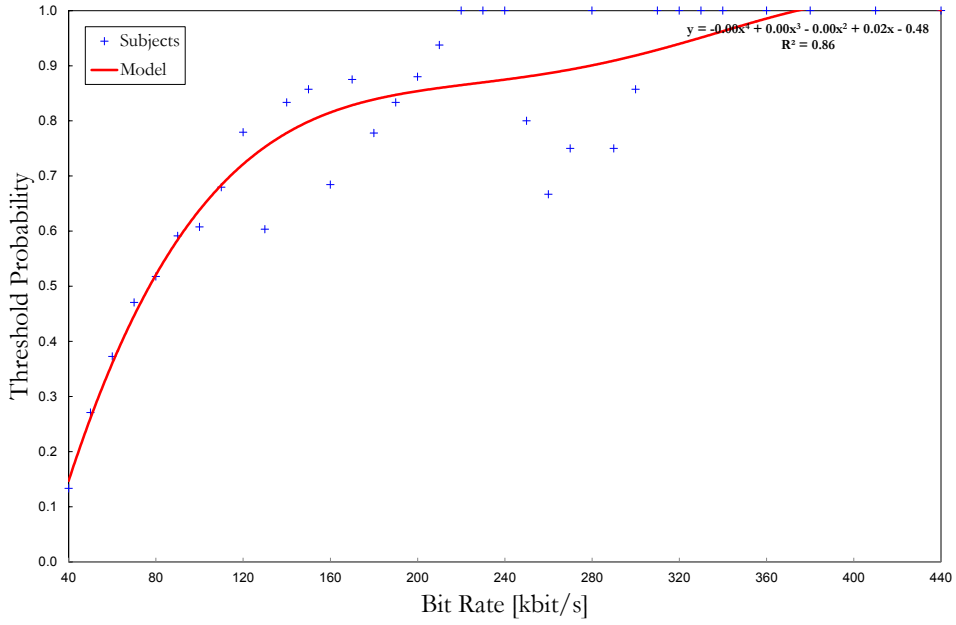


Fig. 9. Example of the obtained detection probability and model of the license plate recognition task.

is possible through the application of appropriate transformations, allowing the fitting of diverse recognition tasks into a single quality framework.

8. Future work

The methodologies outlined in the chapter are just a single contribution to the overall framework of quality standards for task-based video. It is necessary to define requirements starting from the camera, through the broadcast, and until after the presentation. These requirements will depend on scenario recognition.

So far, the practical value of contribution is limited. It refers to limited scenarios. The presented approach is just a beginning of more advanced interesting framework of objective quality assessment, described below (Leszczuk, 2011).

Further steps have been planned in standardization in assessing task-based video quality with relation to QART initiative. Stakeholders list has been initially agreed and action points have been agreed. The plans include: quantifying VQiPS' GUCs, extending test methods (standardization of test methods and experimental designs of ITU-T P.912 Recommendation), measuring camera quality, investigating H.264 encoders, investigating video acuity as well as checking results' stability. The final outcome should be to share ideas on conducting joint experiments and publishing joint papers/VQEG reports, and finally, to submit joint standardization contributions. Plans/next steps for standardizing test methods and experimental designs include verification of issues like: subliminal cues, Computer-Generated Imagery (CGI) source video sequences, automated eye charts as well as subjects' proficiency. The agreed tasks include verifying requirements, refining methods/designs and, finally, making subjective experiments both more accurate and feasible.

9. Acknowledgments

This work was supported by the European Commission under the Grant INDECT No. FP7-218086.

10. References

ACLU (2011). Chicago's video surveillance cameras, *Technical report*, ACLU of Illinois.

Agresti, A. (2002.). *Categorical Data Analysis*, 2nd edn, Wiley.

Agresti, A. & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions, *The American Statistician* 52(2): 119–126. ISSN: 0003-1305.

Baker, F. B. (1985). *The Basics of Item Response Theory*, Heinemann.

Boone, W. J., Townsend, J. S. & Staver, J. (2010). 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* n/a: 1–23.

Duplaga, M., Leszczuk, M., Papir, Z. & Przelaskowski, A. (2008). Evaluation of quality retaining diagnostic credibility for surgery video recordings, *Proceedings of the 10th international conference on Visual Information Systems: Web-Based Visual Information Search and Management*, VISUAL '08, Springer-Verlag, Berlin, Heidelberg, pp. 227–230.

Eskicioglu, A. M. & Fisher, P. S. (1995). Image quality measures and their performance, *Communications, IEEE Transactions on* 43(12): 2959–2965.
URL: `http://dx.doi.org/10.1109/26.477498`

Faye, P., Br?maud, D., Daubin, M. D., Courcoux, P., Giboreau, A. & Nicod, H. (2004). Perceptive free sorting and verbalisation tasks with naive subjects: an alternative to descriptive mappings, *Food Quality and Preference* 15(7-8): 781 – 791. Fifth Rose Marie Pangborn Sensory Science Symposium. URL: *http://www.sciencedirect.com/science/article/pii/S0950329304000540*

Ford, C. & Stange, I. (2010). A framework for generalising public safety video applications to determine quality requirements, *in* A. Dziech & A. Czyzewski (eds), *Multimedia Communications, Services and Security*, AGH University of Science and Technology Kraków.

Ghinea, G. & Chen, S. Y. (2008). Measuring quality of perception in distributed multimedia: Verbalizers vs. imagers, *Computers in Human Behavior* 24(4): 1317–1329.

Ghinea, G. & Thomas, J. P. (1998). Qos impact on user perception and understanding of multimedia video clips, *Proceedings of the sixth ACM international conference on Multimedia*, MULTIMEDIA '98, ACM, New York, NY, USA, pp. 49–54.
URL: *http://doi.acm.org/10.1145/290747.290754*

Insall, M. & Weisstein, E. W. (2011). *Partially Ordered Set*, MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/PartiallyOrderedSet.html.

ITU-T (1999). Recommendation 910: Subjective video quality assessment methods for multimedia applications, ITU-T Rec. P.910. URL: *http://www.itu.int/*

ITU-T (2000). Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures, ITU-R Rec. BT.500. URL: *http://www.itu.int/*

ITU-T (2008). Recommendation 912: Subjective video quality assessment methods for recognition tasks, ITU-T Rec. P.912. URL: *http://www.itu.int/*

Janowski, L. & Romaniak, P. (2010). Qoe as a function of frame rate and resolution changes, *in* S. Zeadally, E. Cerqueira, M. Curado & M. Leszczuk (eds), *Future Multimedia Networking*, Vol. 6157 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 34–45.
URL: *http://dx.doi.org/10.1007/978-3-642-13789-1_4*

Leszczuk, M. (2011). Assessing task-based video quality — a journey from subjective psycho-physical experiments to objective quality models, *in* A. Dziech & A. Czyzewski (eds), *Multimedia Communications, Services and Security*, Vol. 149 of *Communications in Computer and Information Science*, Springer Berlin Heidelberg, pp. 91–99. URL: *http://dx.doi.org/10.1007/978-3-642-21512-4_11*

Leszczuk, M. I., Stange, I. & Ford, C. (2011). Determining image quality requirements for recognition tasks in generalized public safety video applications: Definitions, testing, standardization, and current trends, *Broadband Multimedia Systems and Broadcasting (BMSB), 2011 IEEE International Symposium on*, pp. 1 –5.

Leszczuk, M., Janowski, L., Romaniak, P., Glowacz, A. & Mirek, R. (2011). Quality assessment for a licence plate recognition task based on a video streamed in limited networking conditions, *in* A. Dziech & A. Czyzewski (eds), *Multimedia Communications, Services and Security*, Vol. 149 of *Communications in Computer and Information Science*, Springer Berlin Heidelberg, pp. 10–18.
URL: *http://dx.doi.org/10.1007/978-3-642-21512-4_2*

Nyman, G., Radun, J., Leisti, T., Oja, J., Ojanen, H., Olives, J. L., Vuori, T. & Hakkinen, J. (2006). What do users really perceive — probing the subjective image quality experience, *Proceedings of the SPIE International Symposium on Electronic Imaging 2006: Imaging Quality and System Performance III, Vol. 6059*, pp. 1–7.

Picard, D., Dacremont, C., Valentin, D. & Giboreau, A. (2003). Perceptual dimensions of tactile textures, *Acta Psychologica* 114(2): 165 – 184. URL: http://www.sciencedirect.com/science/article/pii/S0001691803000751

Radun, J., Leisti, T., Häkkinen, J., Ojanen, H., Olives, J.-L., Vuori, T. & Nyman, G. (2008). Content and quality: Interpretation-based estimation of image quality, *ACM Trans. Appl. Percept.* 4: 2:1–2:15.
URL: http://doi.acm.org/10.1145/1278760.1278762

Strohmeier, D., Jumisko-Pyykkö, S. & Kunze, K. (2010). Open profiling of quality: a mixed method approach to understanding multimodal quality perception, *Adv. MultiMedia* 2010: 3:1–3:17. URL: http://dx.doi.org/10.1155/2010/658980

Takahashi, A., Schmidmer, C., Lee, C., Speranza, F., Okamoto, J., Brunnström, K., Janowski, L., Barkowsky, M., Pinson, M., Staelens, Nicolas Huynh Thu, Q., Green, R., Bitto, R., Renaud, R., Borer, S., Kawano, T., Baroncini, V. & Dhondt, Y. (2010). Report on the validation of video quality models for high definition video content, *Technical report*, Video Quality Experts Group.

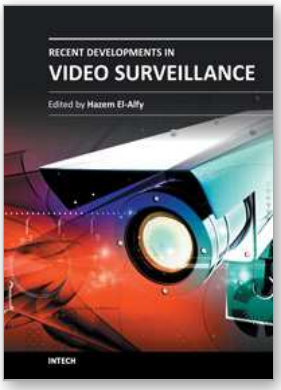
VQE (2010). *Report on the validation of video quality models for high definition video content*, version 2.0 edn. http://www.vqeg.org/.

VQEG (n.d.). The video quality experts group. URL: http://www.vqeg.org/

VQiPS (2010). Defining video quality requirements: A guide for public safety, volume 1.0, *Technical report*, U.S. Department of Homeland Security's Office for Interoperability and Compatibility. URL: http://goo.gl/TJOdU

VQiPS (2011). Video quality tests for object recognition applications.
URL: http://www.safecomprogram.gov/SAFECOM/library/technology/1627_additionalstatement.htm

Wang, Z., Lu, L. & Bovik, A. C. (2004). Video quality assessment based on structural distortion measurement, *Signal Processing: Image Communication* 19(2): 121–132.
URL: http://dx.doi.org/10.1016/S0923-5965(03)00076-6%20



Recent Developments in Video Surveillance

Edited by Dr. Hazem El-Alfy

ISBN 978-953-51-0468-1
Hard cover, 122 pages
Publisher InTech
Published online 04, April, 2012
Published in print edition April, 2012

With surveillance cameras installed everywhere and continuously streaming thousands of hours of video, how can that huge amount of data be analyzed or even be useful? Is it possible to search those countless hours of videos for subjects or events of interest? Shouldn't the presence of a car stopped at a railroad crossing trigger an alarm system to prevent a potential accident? In the chapters selected for this book, experts in video surveillance provide answers to these questions and other interesting problems, skillfully blending research experience with practical real life applications. Academic researchers will find a reliable compilation of relevant literature in addition to pointers to current advances in the field. Industry practitioners will find useful hints about state-of-the-art applications. The book also provides directions for open problems where further advances can be pursued.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mikołaj Leszczuk, Piotr Romaniak and Lucjan Janowski (2012). Quality Assessment in Video Surveillance, Recent Developments in Video Surveillance, Dr. Hazem El-Alfy (Ed.), ISBN: 978-953-51-0468-1, InTech, Available from: http://www.intechopen.com/books/recent-developments-in-video-surveillance/quality-assessment-in-video-surveillance

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821