Research proposal: Just noticeable difference (JND) and satisfied user ratio (SUR) prediction for compressed video

Jingwen Zhu, Patrick Le Callet

To cite this version:
Jingwen Zhu, Patrick Le Callet. Research proposal: Just noticeable difference (JND) and satisfied user ratio (SUR) prediction for compressed video. The 13th ACM Multimedia Systems Conference MMSys ’22, Jun 2022, Athlone, Ireland. pp.393-397, 10.1145/3524273.3533933 . hal-03788614

HAL Id: hal-03788614
https://hal.science/hal-03788614
Submitted on 26 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Research proposal: Just Noticeable Difference (JND) and Satisfied User Ratio (SUR) prediction for compressed video

Jingwen Zhu
jingwen.zhu@etu.univ-nantes.fr
Nantes Université
Ecole Centrale Nantes
CNRS, LS2N, UMR 6004
Nantes, France

Patrick Le Callet
patrick.lecallet@univ-nantes.fr
Nantes Université
Ecole Centrale Nantes
CNRS, LS2N, UMR 6004
Nantes, France

ABSTRACT
This paper contains the research proposal of Jingwen ZHU that was presented at the MMSys 2022 doctoral symposium. Just noticeable difference (JND) is the minimum amount of distortion from which human eyes can perceive difference between the original stimuli and distorted stimuli. With the rapid raise of multimedia demand, it is crucial to apply JND into the visual communication systems to use the least resources (e.g., bandwidth and storage) without damaging the Quality of Experience (QoE) of end-users. In this thesis, we focus on the JND prediction for compressed video to guide the choice of optimal encoding parameters for video streaming service. In this paper, we analyse the limitations of the current JND prediction models and present five main research questions to address these challenges.

CCS CONCEPTS
• Applied computing; • Human-centered computing;

KEYWORDS
Just noticeable difference (JND), Satisfied User Ratio (SUR), video quality assessment

1 INTRODUCTION
Human Visual System (HVS) cannot perceive small distortions, Just Noticeable Difference (JND) threshold is the minimum amount by which stimulus intensity must be changed to produce a noticeable variation for HVS. Nowadays, with the increasing multimedia demand such as video streaming, JND plays an important role to reduce the resources (e.g., bandwidth, storage) consumption without decreasing the Quality of Experience (QoE) for end-users. In addition, JND has been widely employed in many other vision applications, including digital image/video processing, visual signal restoration/enhancement, and watermarking. To find the JND video in a series of videos with different distortion levels for a given anchor/reference, video with every distortion level needs to be compared with the anchor. Wang et al. [24] proposed a binary search method to conduct the JND search which can help to reduce the number of comparison during the JND search. Lin et al. [9] proposed two methods to improve the efficiency of binary search: slider and keystroke. However, compared to ordinary subjective tests such as Absolute Category Rating (ACR) or Degradation Category Rating (DCR) [6], subjective JND test is much more time consuming. As subjective test is time/money consuming, it is important to develop JND prediction models.

JND depends on 3 factors: (1) display setting, e.g., viewing distance [17], monitor profiling, etc.; (2) subjects; (3) image/video contents [11]. In this thesis, we focus on how different video contents impact the location of JND in the context of video compression. Specifically, when the anchor is the no-compression version, e.g., sources(SRCs) video, the JND is called 1st JND, and if we take the 1st JND as the new anchor, the next JND is then the 2nd JND, etc. In practice, it is far enough to measure the 1st, 2nd and 3nd JND, because the video quality worse than the 3nd JND is too poor in current internet video streaming applications [24].

The prediction of JND has been investigated in both images [2, 10, 13, 27, 28] and videos [23, 25, 29, 30]. However, most of the current JND prediction methods are application/codec specific. With the rapid development of multimedia compression standards, it is not practical to collect large-scale dataset for every new codec. Therefore it is important to develop codec agnostic JND prediction methods or a general model which could be adapted easily for new codecs. Moreover, current JND prediction models are not practical, because they take the SRC and its corresponding Processed Video Sequences (PVSs) with many different encoding parameters as input. Therefore before applying these models, SRC need to be compressed several times, which is computation and time consuming. In this thesis, we present five main research questions to target the above mentioned issues in the video-wise JND prediction for compressed video. We plan to address them using existing publicly available JND datasets and machine learning methods.

2 RELATED WORK
The JND prediction models in literature can be divided into: (1) Pixel-wise JND models; (2) Sub-band domain JND models; (3) picture-wise JND model and (4) Video-wise JND model. Pixel-wise JND models...
predict either a JND map, which concretely contains per pixel JND in a given image [27, 28], or a probability map that contains for each pixel the probability that an average observer will notice a difference between the reference image and the distorted image [16]. Sub-band domain JND models calculate the JND threshold for each sub-band coefficient in transform domain (e.g., Discrete cosine transform (DCT)) [15, 21, 26]. These models are "generic" models which are application-free (e.g., JPEG). However, even if the pixel-wise JND prediction is accurate enough, it is still challenging to mimic HVS for a global JND threshold for images, especially for small spatial or temporal shifts. Moreover, pixel-wise JND prediction models cannot be used directly to guide the image compression. The sub-band domain models consider each sub-band blocks separately, therefore the cross sub-band blocks information is not taken into consideration.

For a better understanding of the global JND for pictures, several picture-wise JND datasets have been collected from subjective JND test under different image compression methods and for different image types [3, 7, 14, 19]. For a given content, the JND annotation can vary for different subjects. Satisfied User Ratio (SUR) was introduced in [24] to measure the disagreement between subjects. SUR curve is the Complementary Cumulative Distribution Function (CCDF). For a given distortion level \( D \), the corresponding value \( p \) of SUR on the SUR curve is the ratio of subjects who cannot perceive any difference between the reference video and the distorted video with distortion level \( D \). By fixing a threshold \( p \% \) of SUR, the corresponding distortion level is defined as \( p\%\text{SUR} \) or \( p\%\text{JND} \). The most common used threshold is 75%. Based on these picture-wise JND datasets, some deep-learning base picture-wise JND prediction models have been proposed. Liu et al. [13] first transferred the multi-class classification problem to a binary classification (lossy/lossless) problem and used a sliding window search strategy to predict JND points. The binary classifier is a lossy/lossless predictor based on convolutional neural networks (CNN), which is trained and evaluated on the MCL-JCI [7] dataset. The sliding window search strategy is designed to find the first JND point using the output of the classifier. SUR-Net [2] is based on a Siamese CNN to predict the SUR value for different distortion levels of the image. The discrete points obtained by the predictor are used to fit the SUR curve by least squares assuming that the JND annotations of a group observers follow normal distribution. The first JND point is derived by the fitted SUR curve with 75% threshold. Transfer learning and data augmentation are used to avoid over-fitting caused by the small-scale training dataset MCL-JCI [7]. Based on SUR-Net, Lin et al. [10] proposed SUR-FeatNet to predict the SUR curve and JND. Maximum likelihood estimation and Anderson-Darling test are used to obtain a more accurate probability distribution model for the JND distribution for different observers. The ground-truths are generated by fitting the SUR curve in order to train the network which is based on Inception-V3 [20]. There are two stages to train the network. During the first stage, it is trained to predict the fixed full-reference (FR)-IQA score of the input distorted image versus the input reference image. While during the second stage, the SUR value will be predicted instead. The least-squares will be used to fit the final SUR curve using the discrete SUR points obtained in the second stage.

Video-wise JND is less investigated compared to picture-wise JND. Two video-wise JND datasets have been collected in [22, 24] under H.264 video compression. Wang et al. [23] proposed a learning-based model to predict SUR curve by using the Support Vector Regression (SVR) under the assumption that the individual JND points of different subjects follow a normal distribution. This SUR predictor is trained on VideoSet [24] using Quality Degradation Features using VMAF [8] and Masking Features [4, 5]. The 75%SUR (75%JND) can be derived from the predicted SUR curve. Similar method is used in [25] to predict the 2nd and 3rd JND points using 3 different settings in which the reference inputs of the predictor are different. Instead of predicting encoding parameter QP of H.264 as JND profile, Zhang et al. [29] proposed a novel perceptual model to predict SUR versus bitrate, which is more widely used in practice. Three kinds of features, Masking features, re-compression features, and basic attribute features, are extracted from SRC and several PVSs. Gaussian Processes Regression (GPR) is applied to predict SUR. Zhang et al. [30] proposed a deep-learning based SUR prediction model that extracts spatial and temporal features via CNNs and a weighted average pooling to predict SUR value of a given PVS.

However, these picture-wise and video-wise JND prediction models are not application-free, they are limited to a certain image/video codec. For instance, the video-wise JND prediction models proposed in [24, 30] are learning-base models, they are both trained on VideoSet [24] which is compressed by H.264, and the JND proxy is the quantization parameter (QP) value. However with the rapid development of video compression techniques, it is neither practical nor efficient to collect a large-scale datasets each time for a new codec and to train a new model which is compatible with the new codec. Moreover, these JND prediction models usually need a large number of PVSs, e.g., compressed videos with QP from 1 to 51 in [23, 30], to predict the SUR curve. However, for a video streaming service provider, it is very resource-consuming to first compress the SRC with every possible encoding parameters and then predict the JND from them. Furthermore, the learning-based JND prediction models are not only sensitive to codecs, but also to the video resolutions [30], i.e., cross resolution decreases the prediction accuracy.

3 RESEARCH QUESTIONS

To address the previous issues in related work, we plan to investigate the following research questions (RQs):

RQ1: How to design JND subjective test methodology to collect JND dataset more efficiently?

The current large-scale JND datasets VideoSet [24] include 220 5-second SRCs in 4 encoding resolutions. Each SRC is encoded with H.264 codec with QP from 1 to 51. More than 30 subjects participated into the viewer group for each SRC. However, it is well known that JND search is extremely time consuming, we try to find a new subjective test methodology to collect JND more efficiently. This research question could be divided into several sub-questions: (i) how to select contents to cover a wide-range of characteristics? (ii) how to define the candidate PVS list instead of all possible PVS from
**RQ1:** We plan to collect a new large-scale JND datasets for compressed video under H.265 for HD, UHD and HDR contents. The proxy of the JND is set to be Constant Rate Factor (CRF). We use the content selection features in [12] and add compression-base features: including the number of I, P, and B frames with respective average sizes, and the clip bitrate as features to select contents press. JND searches are not limited in one resolution, SRCs are used as anchor to make JND search in both encoding resolutions in order to find the best encoding parameters. 20 observers who have either normal or corrected-to-normal visual acuity were included in the experiment, and A 55-inch calibrated “UHD Grundig Finearts 55 FLX 9492 SL” was employed as display screen. The viewing distance is set to 1.5H for UHD videos and 3H for HD as recommended in ITU-R BT.2022 [1], where H is the height of the screened video. Binary search [24] is used to find JND for each content. This new JND datasets gives us opportunities to address the cross-resolution issues in RQ3 and develop codec agnostic prediction model in RQ5 along with VideoSet [24] under H.264 compression.

**RQ2** We proposed a pipeline for the prediction of SUR and JND based only on SRC as illustrated in Fig.1. There are two main steps: (i) mathematical modeling of the SUR curve to obtain the parameters as ground truth that describe the SUR curve; (ii) prediction of the parameters and SUR accordingly from the features of SRC. We
In this thesis, we focus on five mains research questions about the JND prediction for compressed video to guide the choice of encoding parameters for a high quality of experience with limited storage and internet bandwidth. Each research question is based on the existing problems in related work, and we try to address these questions on the benefit of on-hand VQA methods and machine learning methods. We present briefly the methodologies and conclusions of ongoing and future work.

ACKNOWLEDGMENTS

This work is supervised by Patrick Le Callet, full professor of Nantes University. I would like to thank him for the guidance and support in this thesis. The authors also thank Dr. Suiyi Ling and Dr. Sriram Sethuraman for the helpful discussions during the thesis. The financial support of this thesis is Amazon Prime Video.