Video Quality Assessment in-the-wild

Dissertation zur Erlangung des

akademischen Grades eines Doktors der

Ingenieurwissenschaften (Dr.-Ing.)

vorgelegt von

Franz Götz-Hahn (geb. Hahn)

an der

Universität Konstanz

Mathematisch-Naturwissenschaftliche Sektion

Informatik und Informationswissenschaften

Konstanz, 2021

Tag der mündlichen Prüfung: 07. Juli 2022

1. Referent: Prof. Dr. Dietmar Saupe

2. Referent: Prof. Dr. Patrick Le Callet

3. Referent: Prof. Dr. Bastian Goldlücke

In dedication to my father.

Eadem mutata resurgo.

– Jakob Bernoulli

Acknowledgement

This thesis would not have been possible without the support, collaboration, and help of many people.

First, I would like to sincerely thank my supervisor, Dietmar Saupe, for all the advice, encouragement, and inspiration. Dietmar was a very patient listener and always had suggestions to investigate a problem from a different perspective. It was a pleasure working with someone who had spent many years thinking deeply about a broad field of topics.

Thanks to Vlad Hosu, for participating in countless hours of discussions and explorations of topics closely and tangentially related to this thesis and giving useful advice on being mindful and not straying away too far from what is important. You taught me a lot of what I have learned. It was a pleasure brainstorming, working, and travelling together and I'm looking forward to more collaborations in the future.

A special thank you also goes out to my dear colleague Oliver Wiedemann, for being a steadfast, headstrong, and level-headed friend that helped me in innumerable ways and occasions. I'm extremely grateful to have met you and look forward to many more years of friendship to come.

Thank you Hanhe Lin, Hui Men, and Mohsen Jenadeleh, for making the group feel like a second home. Ingrid Baiker, for being the heart and soul of our group and always having an open ear and solutions for any troubles. Claudia Widmann, for being the grounding force holding the TRR together.

Many thanks to Juan Quintana Duque for his help structuring and correcting preliminary versions of this thesis.

Finally, I'd like to thank Bianka, for always reminding me that I needed to write my thesis, for supporting me through thick and thin, and for her unending love.

Abstract

This thesis investigates topics related to technical video quality assessment. At its core, it is comprised of three parts.

First, it describes the developments of two state-of-the-art datasets for video quality assessment. These datasets capture an unmatched breadth of distortions, qualities, and contents, as they are based on videos obtained from public video repositories. This makes them particularly useful for the prediction of video quality in-the-wild. The entire dataset construction process is explained, including the collection of the video sources, the annotation process, and thorough comparisons to related datasets.

Second, it presents an approach for automated feature extraction that is novel in the domain of video quality assessment. Based on deep neural networks pre-trained on another perceptual task, features useful for visual quality assessment are obtained by averaging the activation maps of kernels in many layers of the source network. It is shown that these features are not only useful for the evaluation of content similarity, but also as a basis for image and video quality and aesthetics prediction. Furthermore, it explores the possibilities of finding features that carry utility when trying to bridge different quality-related domains, such as technical quality and aesthetic, which is discussed under the term generalized visual quality assessment. Three models introduced in this thesis leverage the power of these extracted deep features to obtain state-of-the-art for in-the-wild video quality prediction.

Finally, it highlights an aspect of the field that has so far been ignored, namely the impact of the distribution of an annotation budget over a dataset. This investigation shows, for the first time, that modern video quality prediction models benefit from a large amount of data that is only approximately annotated for quality. Contrary to a popular belief in the community, the precision of aggregated quality annotations are shown not to be the driving factor in predictive power of modern models. Rather, a delicate balance between this precision and breadth of content diversity benefits deep learning models in particular.

Zusammenfassung

Diese Dissertation behandelt Themen, welche in einem Zusammenhang mit der technischer Qualitäts-Bewertung von Videos stehen. Im Kern besteht sie aus drei Teilen.

Zuerst werden zwei aktuelle Datensätze zur Bewertung technischer Qualität in Videos, sowie alle relevanten Charakteristiken ihrer Entwicklung beschrieben. Diese erfassen eine - bis dato - unübertroffene Fülle an Qualitäten, Inhalten und Artefakten, welche sich daraus ableitet, dass die darin enthaltenen Videos von öffentlichen Video-Portalen stammen. Aus diesem Umstand ergibt sich ein besonderer Nutzen für die Vorhersage der Qualität von zweckungebundenen Videos. Es wird der gesamte Herstellungsprozess der Datensätze beschrieben, was sowohl die Rohdaten-Akquise, den Annotationsprozess, als auch einen detaillierten Vergleich mit anderen existierenden Datensätzen beinhaltet.

Zweitens wird in dieser Dissertation ein Ansatz zur automatischen Merkmalsextraktion präsentiert, welcher innerhalb der Domäne der Videoqualitäts-Einschätzung neuartig ist. Er basiert auf für einen anderen Zweck vortrainierten neuronalen Netzwerken, deren gelernte Merkmale für die Einschätzung von visueller Qualität verschiedener Art nützlich sind. Die Merkmale werden extrahiert, indem die Aktivierungsmatrizen aller Neuronen des Netzwerks einzeln gemittelt werden, um jeweils ein durchschnittliches Aktivitätslevel der Neuronen zu bekommen. Weiter wird gezeigt, dass diese Merkmale für verschiedene perzeptuelle Probleme nützlich sind, wie etwa dem der visuellen Ähnlichkeit zweier Bilder, aber auch als Basis für die Vorhersage von Bild und Video Qualität und Ästhetik. Darüber hinaus werden Möglichkeiten erörtert, einzelne bestimmte Merkmale zu ermitteln, welche besonderen Nutzen für perzeptuelle Probleme haben. So wird gezeigt, dass mit nur wenigen Dutzend dieser automatisch gelernten Merkmale bereits eine hohe Performanz in mehreren verwandten Domänen erzielt werden kann. Darauf aufbauend werden drei konkrete Modelle eingeführt, die unter Zuhilfenahme dieses Ansatzes den Stand der Technik in mehreren Aspekten der Prognose von Video Qualität stellen.

Zuletzt werden anhand der genannten Datensätze neue Themenfelder aufgebracht. Insbesondere der Zusammenhang zwischen der Verteilung eines Annotationsbudgets auf einen Video-Datensatz und die damit zusammenhängende Präzision einer Vorhersage ungesehener Daten steht hier im Mittelpunkt. Die Untersuchungen zeigen zum ersten Mal, dass moderne Modelle zur Vorhersage von Video Qualität davon profitieren, wenn der Trainingsdatensatz vergrößert wird und die Exaktheit der dazugehörigen Annotationen lediglich grob geschätzt wird. Diese Beobachtung steht im starken Kontrast zu dem geläufigen Glaube innerhalb der Domäne, dass die Exaktheit der Annotationen des Trainingsdatensatzes eine primäre Rolle in der Vorhersagegenauigkeit von Modellen spielt.

Contents

List of Figures

List of Tables

1 Introduction

The first quarter of the 21st century has witnessed the rise of video to be the primary source of media entertainment for consumers with the average US citizen spending 38 hours per week watching some form of video content. In 2008 Cisco noted in their global consumer internet traffic report* that approximately 20% of all internet traffic was used for some form of internet video streaming. Nearly a decade later in 2017 video viewing was reported to take up 75% of internet traffic† and is projected to reach 85% by 2022‡. In the same vein, the number of businesses using video as a marketing tool has increased by 41% since 2016§, meaning that 86% of businesses in the US reported using video as a marketing tool in 2021. Additionally, 93% of marketers who use video say that it's an important part of their marketing strategy.

Consequently, it comes as no surprise that a significant body of literature within the digital signal processing domain has arisen which focuses on both assessing and improving the quality of digital media content, as well as ensuring faithful reproduction at the consumer's end. For this purpose, *quality of experience* (QoE) has been defined and adopted in Recommendation ITU-T P.10 [ITU99] as:

> "The degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations concerning the utility and/or enjoyment of the application or service in the light of the user's personality and current state."

It can be considered as the technology-centered sibling to the human-computer interaction subfield of user experience, with a stronger focus on the perception of quality, rather than the concept of experience.

Three so-called influence factors have been identified and defined within QoE [Rei+14]. *Human influence factors* consider low-level and high-level processing factors on the consumer's end, meaning characteristics such as visual and auditory acuity, or cognitive processes and expectations. *Context influence factors* group situational properties describing the consumer's environment, such as the social context the digital media is consumed in, or interruptions during the consumption that are external to the media delivery system. Finally, *System influence factors* combine aspects that are related to the content, media, transmission, and the playback device. Any kind of encoding-related degradations of a video or delays in the transmission of a streamed digital media item are examples of influence factors in this category. The contents of this thesis is largely situated within this last category of influence factors of QoE, as video quality refers to the quantification of low-level perceptual degradation of a video stimulus.

* Cisco Visual Networking Index (VNI): Forecast and Methodology, 2008-2013
† Cisco VNI Complete Forecast Update, 2017–2022
‡ Cisco Annual Internet Report (2018–2023)
§ Wyzowl: The State of Video Marketing 2021

1.1 Video Quality

Digital media items undergo a multitude of steps from initial recording to their presentation to a consumer. During all parts of the content delivery pipeline, starting with the acquisition, followed by compression and transmission, and finishing with the processing, reproduction, and display at the destination, alterations to the media item are necessary and inevitable. Some of these alterations have more noticeable or salient perceptual impacts, which are often of negative impact to the delight of the observer. Specifically, video quality assessment (VQA) means evaluating whether the quality of the digital media item is high or low relative to some quality scale, due to the presence or absence of distortions. Thus, it tries to answer the following question: Does the video look qualitatively good? Since a video is comprised of individual images that are played in quick succession, its quality is also highly dependant on the quality of these individual frames, which almost necessarily links VQA to the field of image quality assessment (IQA).

As an illustration of different quality levels, Figure 1.1 depicts low, medium, and high quality images, taken from the popular IQA dataset KonIQ-10k [Hos+20a]. The low quality image, showing a close-up of a chicken coop, is impacted by an overall blurriness and tone fuzziness, as well as obvious ringing artifacts (the cyan discoloring where white and red tones meet around the chickens' heads), which are an indication for quantized block-based discrete cosine transform coding. The group of bicycle enthusiasts in the medium quality image is slightly blurry, likely caused by the quantization of high-frequency components. Additionally, there is a lack of detail in the grass in the background of the image. Finally, the high quality image shows a detailed shot of an insect, where light ringing artifacts can be seen around the edges of the creature only at full resolution.

All of the artifacts described for the examples shown in Figure 1.1 are representative for still images. The playback of a video can hide or amplify these problems and video encoding methods can also introduce additional effects that can impact the perceived quality in both directions on the subjective quality scale. For example, subtle instances of staircasing noise, a special form of blockiness along diagonal or curved edges, can be masked during video playback, even though it is clearly visible in the video material's individual frames. Alternatively, frequent luminance or chrominance changes over time, known as flickering, can be a very salient temporal artifact in video playback, even though the individual video frames might be of pristine quality.

1.2 Video Quality Assessment

The only "true" approach of quantifying visual quality of digital media is through empirical studies involving human subjects. However, objective video quality assessment systems have been developed that have the goal of predicting subjective video quality accurately and automatically. They are a necessary step for a broad variety of applications, ranging from 1) the design and optimization of video acquisition, processing, compression, storage, distribution, reproduction, and display methods

Refer to Section 2.3.2 for a more comprehensive description of different image and video quality distortions.

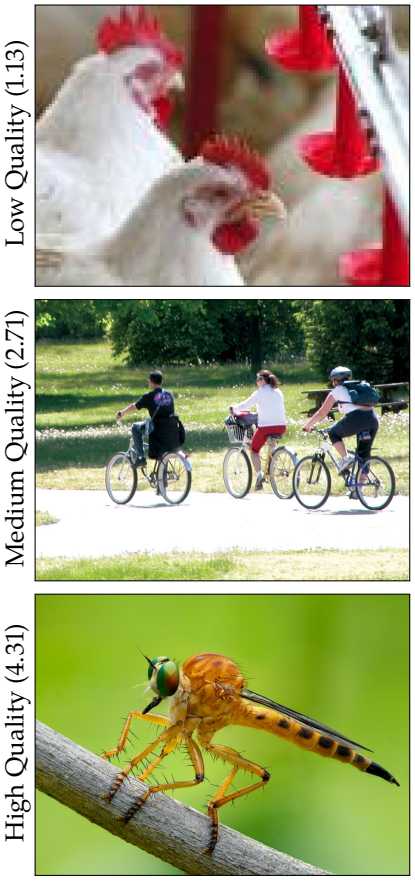


Figure 1.1: Image Quality Examples, ranging from low to high. Source: KonIQ-10k

and systems, to 2) comparing the quality of videos created using different parameters in the production pipeline, to 3) maintenance and resource allocation of visual communication systems.

Powerful and reliable models that automatically predict the quality of video items require subjective quality annotations for large video datasets. Conventionally, a standardized subjective test would require the recruitment of a diverse panel of volunteers to evaluate video stimuli in a laboratory environment. This would allow for careful control over the external factors impacting the accuracy and repeatability of the subjective study. In practice, however, human evaluation in a laboratory setting is both too time-consuming and expensive, especially considering the sheer amount of video data generated; for example, in 2017 it was estimated that every 60 seconds, 400 hours of video were uploaded to the online video site YouTube. In the past decade, crowdsourcing has risen as an alternative to lab studies to gather subjective data and has become the de facto standard for subjective video quality annotation. It trades the highly controlled environment of the laboratory with a more ecologically valid environment where an internet crowd evaluates the quality of multimedia files at their leisure. With the advent of this tool, it is now possible for the first time to obtain thousands of video quality annotations in a matter of a few days.

A secondary requirement for robust quality prediction is features that are informative of the quality scale. The best-performing models in the field of video quality are based on the features extracted from pre-trained convolutional neural networks (CNNs), which are made up of automatically learned kernels of hierarchical nature. In fact, deep features have effectively replaced traditional hand-crafted features for tasks related to the prediction of perceptual attributes of images and videos. This can be attributed to the circumstance that the complexity of features achieved in modern CNNs far exceeds that achievable by manual design. Through training processes that evaluate hundreds of thousands of images, CNNs can pick up on elaborate compositions of patterns that a human would be hard-pressed to discern, describe, or design a detector for.

1.3 Outline

The two main contributions of this thesis are (1) the design of KoNViD-1k and KonVid-150k, two novel datasets for VQA with a focus on ecological validity and the latter being the largest existing dataset within the field, and (2) the development of a set of objective VQA methods based on deep features that can train efficiently on datasets of the size of KonVid-150k, as well as generalize to videos in-the-wild at the example of legacy datasets.

Chapter 2 provides an in-depth perspective into three aspects. First, the conventional way of annotating the quality of digital media items in laboratory settings is described, along with a comprehensive summary of the most relevant ITU Recommendations. Second, we present how crowdsourcing relates to lab studies for subjective IQA and VQA, specifically, including the presentation of our research showcasing that

crowdsourcing-based IQA experiments can reproduce results obtained in lab studies with high levels of correlation. Additionally, we describe challenges that crowdsourcing campaigns face on a more general level and discuss relevant ITU-T Recommendations related to visual quality assessment in a crowdsourcing environment. Then, we provide a short introduction to the human visual system (HVS) alongside the most relevant image and video artifacts and their relation to both the HVS and the encoding process. Third, we conclude the chapter with a presentation of our research that takes advantage of knowledge about perceptual statistics of images in the encoding process to produce images of higher perceptual quality at the same bitrate as what is produced by conventional encoding systems.

Chapter 3 summarizes the most relevant image and video quality datasets and introduces our own efforts in furthering the field. Starting with a description of characteristics that are relevant in categorizing and understanding the purpose of quality assessment datasets in general we progress quasi-chronologically through legacy datasets that carry historical importance or are otherwise important in relation to our own contributions. We show the benefits and deficiencies of different considerations in the creation process and establish desirable attributes of an ideal video dataset to learn quality in-the-wild. Then, we describe considerations and steps taken for our own KoNViD-1k and KonVid-150k VQA datasets that attempt to more appropriately implement these desirable characteristics to establish modern benchmark datasets for VQA. Finally, we conclude the chapter with a quantitative evaluation of modern VQA datasets to investigate aspects like their ecological validity and annotation quality.

In Chapter 4 we first summarize historical approaches to provide features valuable in predicting I/VQA, as well as detecting artifacts in images and videos. Next, we describe learning-based techniques to construct features useful for perceptual tasks by summarizing related works that have implemented approaches that extract deep features from different sources and for the prediction of different perceptual attributes. Here, we describe our own procedure to extract deep features from pre-trained CNNs and what the impact of extracting features at different layers is on the predictive performance of simple models. Finally, we conclude the chapter with a presentation of our efforts in identifying specific deep features that are representative of a generalized perceptual quality, rather than just being useful for the prediction of a particular perceptual attribute. By using a simple deep feature selection strategy we can derive a subset of 10% of the original features that achieves the same performance as using the full set of deep features.

Chapter 5 is, then, primarily concerned with objective VQA. We first briefly summarize legacy approaches to the problem, followed by an in-depth overview of state-of-the-art methods. Here, we also show the dangers of using learning-based deep features suffer at the hand of our own research that uncovered a handful of published approaches that included different forms of data leakage, meaning that their published performances can not be compared to other works. Next, we introduce three VQA models based on deep features and evaluate them on the most relevant modern video datasets, both from in-the-wild and synthetic domains. We investigate the relationship between these domains in

inter-dataset and inter-domain evaluations and further show the power of our own dataset contributions we described in Chapter 3 compared to the state-of-the-art. Finally, we conclude the chapter with ablation studies that investigate the impact that different vote budget distribution strategies have on the predictive performance of our models, which raises doubts about the traditional rule of requiring several dozen individual quality annotations for each video.

Chapter 6, then, summarizes the core contributions of this thesis and articulates some of the open questions raised by this work.

1.4 Publications

During my time as a doctoral researcher, I authored and contributed to several articles in journals, conferences, and workshops that form the basis of this thesis:

[Sau+16] Dietmar Saupe, **Franz Hahn**, Vlad Hosu, Igor Zingman, Masud Rana, and Shujun Li. 'Crowd workers proven useful: A comparative study of subjective video quality assessment'. In: *QoMEX 2016: 8th International Conference on Quality of Multimedia Experience*. 2016

Contribution clarification. This paper is an extension to the MA thesis of Masud Rana, which was supervised by both me and Dietmar Saupe. It can be seen as the inauguration project of the crowdsourcing efforts of our group. My main contribution for this paper was supervising the implementation and execution of the crowdsourcing experiment, as well as the technical analysis of the resulting subjective video quality assessment annotations. The writing was spearheaded by Dietmar Saupe. Section 2.2 is based largely on this paper.

[Hos+16a] Vlad Hosu, **Franz Hahn**, Oliver Wiedemann, Sung-Hwan Jung, and Dietmar Saupe. 'Saliency-driven image coding improves overall perceived JPEG quality'. In: *2016 Picture Coding Symposium (PCS)*. IEEE. 2016, pp. 1–5

Contribution clarification. This paper is a collaborative effort between several authors, in particular Vlad Hosu, Oliver Wiedemann, and myself, who contributed approximately equally to the paper. At the suggestion of Dietmar Saupe, O. Wiedemann implemented a variation to the JPEG encoder that finds appropriate parameters to perform variable quantization of a target image based on an accompanying saliency map, such that the resulting compressed version meets a target bitrate requirement. For the encoding process, I performed some minor, although crucial, bugfixing of the encoding algorithm and contributed a saliency modeling approach that reduced the side information required for the decoder to be limited to 163 bits, which enabled the significant gains in perceptual quality at the same bitrate. Furthermore, I designed and executed the two crowdsourcing studies with assistance from V. Hosu, who also shared the majority of the

technical analysis and paper writing efforts with me. This paper is the basis for Section 2.4.

[Hos+17] Vlad Hosu, **Franz Hahn**, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 'The Konstanz natural video database (KoNViD-1k)'. In: *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2017, pp. 1–6

Contribution clarification. This paper is a collaborative effort between many authors, in particular Vlad Hosu and myself, who contributed equally to the paper. V. Hosu had the idea of using creative commons videos from flickr as a source for a new video quality assessment database. V. Hosu, Mohsen Jenadeleh, Hanhe Lin, Hui Men, and I jointly arranged the code base that evaluated a variety of attributes related to video quality, such as levels of blur, contrast, and colorfulness. The fair sampling strategy used to obtain the final set of 1,200 videos was authored by V. Hosu with the assistance of H. Lin. I led the crowdsourcing efforts, designing the interface and quality control mechanisms. The statistical analysis was a joint effort, and V. Hosu and I shared the majority of the paper writing efforts. Parts of this paper are used in Chapter 3, concretely in Section 3.3 and Section 3.5. Contents of this paper can also be found in the dissertation of M. Jenadeleh [Jen18].

[Göt+21] **Franz Götz-Hahn**, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 'KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild'. In: *arXiv preprint arXiv:1912.07966* (2021)

Contribution clarification. This paper can be understood as the big brother of [Hos+17], and is mainly a collaboration between Vlad Hosu and me, with Hanhe Lin and Dietmar Saupe providing key insights throughout the entire research and writing efforts. V. Hosu identified, that instead of the re-encoded videos that we used in our previous effort we could in many cases obtain original uploads from the authors of the videos and wrote a crawler to do so. I designed and executed several pilot studies that investigated the relation between playback duration and perceived quality, the impact of numbers of annotators on the mean opinion, as well as the impact of the level of quality control on the diligence of study participants. Based on my statistical analyses of these pilots I wrote our transcoding pipeline to obtain a large set of more ecologically valid videos. For the finalized dataset, I lead the development of the large-scale crowdsourcing study, based on our experiences from the pilot studies.

V. Hosu supplied a codebase for deep feature extraction for images that I adapted to work more coherently for videos. The design, implementation, and evaluation of the different models was also lead by myself, with the co-authors providing very useful input along the way. D. Saupe initiated the idea of comparing distributions of a fixed vote budget over different

sets of videos. The implementation and statistical analysis of this experiment, as well as the design of the other experiments carried out in the paper, were all done by myself. Moreover, the majority of the writing work was done by myself, with all co-authors giving valuable editorial input. Contents of this paper can be found in different Chapters of this thesis, namely in Chapter 3 in Section 3.4 and Section 3.5, as well as different parts of Chapter 4 and Chapter 5.

[GHS21] **Franz Götz-Hahn**, Vlad Hosu, and Dietmar Saupe. 'Critical analysis on the reproducibility of visual quality assessment using deep features'. In: *arXiv preprint arXiv:2009.05369* (2021)

Contribution clarification. This paper is a collaborative effort between Vlad Hosu, Dietmar Saupe, and myself. V. Hosu had observed in the evaluation of an initiative with another collaborator that some of the performance values published as a result of that effort were not reproducible. After a chain of publications from this same former collaborator that were evaluated on our work I started re-implementing these approaches and found my own implementations to also not match what was published. To understand the problems of my re-implementation I found several cases of data leakage that plagued these publications. The technical implementation and statistical evaluation were all done entirely by myself, and the paper was mainly written by me, with insightful editorial input by the co-authors, and V. Hosu supplying a section on his initial investigations that initiated these efforts. This paper is summarized in Section 5.1.2.

Furthermore, I published and contributed to additional publications which investigated tangentially related topics of interest, or inspired the needs and contributions of this thesis, but are not included therein. These publications are listed in the following:

[Hos+16b] Vlad Hosu, **Franz Hahn**, Igor Zingman, and Dietmar Saupe. 'Reported attention as a promising alternative to gaze in IQA tasks'. In: *PQS 2016: 5th ISCA/DEGA Workshop on Perceptual Quality of Systems*. 2016, pp. 117–121

[Spi+17] Marc Spicker, **Franz Hahn**, Thomas Lindemeier, Dietmar Saupe, and Oliver Deussen. 'Quantifying visual abstraction quality for stipple drawings'. In: *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering*. 2017, pp. 1–10

[Spi+19] Marc Spicker, **Franz Götz-Hahn**, Thomas Lindemeier, Dietmar Saupe, and Oliver Deussen. 'Quantifying visual abstraction quality for computer-generated illustrations'. In: *ACM Transactions on Applied Perception (TAP)* 16.1 (2019), pp. 1–20

2 Subjective Visual Quality

Subjective assessment of digital media items by human observers is a necessary benchmark for I/VQA. Aggregates of subjective quality annotations are used as ground truth data to train and test objective models that automatically predict quality scores. It is, therefore, important that the process by which this ground truth data is gathered produces precise and reliable answers.

In this chapter, we will describe the benefits and drawbacks of laboratory setups used in the annotation process of legacy datasets. Up until recent years, studies carried out in lab settings have marked the gold standard for gathering subjective human opinions of digital media items. We present crowdsourcing as the newly upcoming alternative for this laborious and expensive effort, as it presents the opportunity to annotate much larger datasets in a much shorter time, without necessarily sacrificing any quality in the gathered data. Experimental evaluation of our data shows that subjective annotations obtained via crowdsourcing correlate well with subjective data generated in a lab setting.

We describe the core functionalities of the HVS in regards to the information flow and processing steps of information sensed by the human eyes. Furthermore, we illustrate, characterize, and describe common classes of spatial and temporal artifacts that are ordinarily perceived as annoying and detrimental to image and video quality. The chapter concludes with a presentation of our research into leveraging information about the salient parts of images to encode images with spatially non-uniform levels of quality that tricks the HVS into perceiving an image of the same bitrate as higher quality than its conventionally encoded counterpart.

This chapter contains and extends material from the following publications. Please refer to Section 1.4 for the contribution clarification.

[Sau+16] Dietmar Saupe, **Franz Hahn**, Vlad Hosu, Igor Zingman, Masud Rana, and Shujun Li. 'Crowd workers proven useful: A comparative study of subjective video quality assessment'. In: *QoMEX 2016: 8th International Conference on Quality of Multimedia Experience*. 2016

[Hos+16a] Vlad Hosu, **Franz Hahn**, Oliver Wiedemann, Sung-Hwan Jung, and Dietmar Saupe. 'Saliency-driven image coding improves overall perceived JPEG quality'. In: *2016 Picture Coding Symposium (PCS)*. IEEE. 2016, pp. 1–5

2.1 Literature Review

2.1.1 Laboratory Setup

Laboratory environments are the established approach for the subjective quality evaluation of speech, audio, image, video, or multimedia, which has been known to produce reliable and accurate results. One major benefit of lab settings is the level of control over the testing environment, enabling a level of meticulousness concerning different influencing factors. These factors range from the hardware, i.e. the display hardware and its calibration, to the room settings, meaning the lighting and ambient illumination, and restrictions of the viewing behavior. The majority of these suggestions, restrictions, and factors have also been formalized in recommendations from the Radiocommunication, as well as the Telecommunication standardization sectors at the International Telecommunication Union, abbreviated as ITU-R and ITU-T, respectively. In the following, we will first summarize two important ITU Recommendations, being Rec. ITU-R BT.500 and Rec. ITU-T P.910, and highlight the most relevant suggestions within them.

Rec. ITU-R BT.500: Methodology for the subjective assessment of the quality of television pictures

This Recommendation [ITU19] is comprised of three parts that each govern different aspects of the subjective image quality assessment process.

Part 1 of ITU-R BT.500, called "Overview of subjective image assessment requirements", describes prerequisites for the environment, experimental hardware, material, and test subjects. It surmises recommended viewing distance and angle, as well as details regarding the display monitor, such as resolution and calibration settings. With respect to test material, it contains typical test material to address common assessment problems. It suggests using at least 15 observers of "(corrected-to-) normal visual acuity on the Snellen or Landolt chart, and for normal color vision using specially selected charts", such as Ishihara, for each item in the study.

A test session according to ITU-R BT.500 should not last more than half an hour, including the instructions. During instructions, participants are to be introduced to the assessment process and the grading scale. They are to be given examples of the types of degradations of visual impairments that they are presented with, which should not be displaying content that is used in the experiment. The beginning of the test should contain a handful of "dummy presentations" that help the participant get a

grounding for the types of stimuli he will be presented with, for which the subjective data is not recorded. The rest of the test items should be presented in random order, with the recommendation to repeat a subset of items if multiple sessions are carried out for the same participant.

This part of the recommendation also contains elaborate information on common methods of analysis. It recommends the mean opinion score (MOS)

$$\bar{u}_k = \frac{1}{N \cdot J \cdot R} \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{r=1}^{R} u_{ijkr}, \tag{2.1}$$

where u_{ijkr} describes the score of observer i under test condition j, for image k and repetition r, and N, J, and R are the total number of observers, test conditions, and repetitions, respectively. Here, test condition refers to variations to the source image. For example, when comparing different compression parameter settings of an encoding mechanism each setting applied to an image is considered a different test condition. Additionally, recommendations for the calculation of confidence intervals, screening of observers under different scenarios, and corrections for scores are provided.

Part 2 is called "Description of subjective image assessment methodologies" and provides recommended image assessment methodologies in the form of different stimulus presentation styles and accompanying scales. For the scope of this thesis, we will summarize the two relevant recommended methodologies, which are the single stimulus method using the five-grade adjectival categorical quality judgment scale, as well as the five-grade adjectival categorical judgments of the stimulus comparison method.

For single stimulus presentation methods, single items are displayed and rated by participants in sequence. The five-grade adjectival categorical judgment scale for quality designates five levels of quality with an appropriate adjective, as shown in Table 2.1. This scale is a scoring scale, meaning that a higher value is assigned to better quality items. It has been shown, however, that the absolute category rating scale, which is practically identical to the five-grade adjectival categorical judgment scale, has some drawbacks, such as participants having different interpretations of the categories, non-equidistance between the categorical adjectives, and others [Möl12]. In our single stimulus experiments, we used this type of scale, although the ordering was sometimes arranged horizontally, with "Bad" on the left side of the scale and "Excellent" on the right.

In stimulus comparison presentation methods the different stimuli are shown either in sequence or in parallel. For adjectival categorical judgments, Recommendation ITU-R BT.500 specifically suggests a seven-grade comparison scale as shown in Table 2.2, where participants provide relative judgments between the presented stimuli. In the experiments where we used the stimulus comparison method, we simplified this scale to a five-grade comparison scale and slightly adjusting the wording.

Table 2.1: An example for the five-grade adjectival categorical judgement scale for quality recommended to be used for single stimulus experiments.

Quality Scale	
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2.2: An example for the seven-grade adjectival categorical judgement scale for the comparison of two qualities recommended to be used for stimulus comparison experiments.

Comparison Scale	
−3	Much worse
−2	Worse
−1	Slightly worse
0	The same
1	Slightly better
2	Better
3	Much better

Part 3 of Recommendation ITU-R BT.500, called "Application specific subjective assessment methodologies for image quality", provides suggestions for application-specific subjective image quality evaluation experiments. It is similar to part 1, in that it provides recommendations for viewing conditions, test materials, and more, however considering the particular application context. We will omit further details for this part of the Recommendation since the applications covered within it are not relevant to this thesis.

Rec. ITU-T P.910: Subjective video quality assessment methods for multimedia applications

This Recommendation [ITU08] suggests methodologies for different aspects related to the subjective assessment of video quality. It discusses different considerations concerning the source material, such as the recording environment and system, as well as the characteristics of the content. Recommendation ITU-T P.910 suggests the use of the absolute category rating scale with or without hidden reference (ACR/-HR), or paired comparison (PC). The ACR/-HR scales are practically the same as the previously described five-grade adjectival categorical judgment scale for single stimulus rating methodologies (See Table 2.1) and suffer from the same problems. PC functions similarly to the stimulus comparison method described in Recommendation ITU-R BT.500, where both a sequential and simultaneous display method for stimulus pairs is described in detail.

Concerning the evaluation procedure, certain parameter settings for the viewing conditions are provided, such as a recommended viewing distance, as well as suggestions for luminance and illumination levels for various parts of the test setup. It recommends at least 15 participants, however, suggests that a range from 4 to 40 participants in a study is acceptable. Moreover, it is specifically stated that having more than 40 participants is only useful on rare occasions. Similar to Rec. ITU-R BT.500 typical screening for (corrected-to-) normal acuity and normal color vision is recommended. The instructions given to participants mirror those described for ITU-T BT.500 to a large degree.

With respect to content characteristics, Rec ITU-T P.910 introduces the spatial and temporal perceptual information measurements that are also relevant to this thesis.

Practical Considerations

Conventionally, experimenters would consider a single, fixed setup that would generally follow ITU recommendations, such that the only variable was the participants. Between different laboratories, however, these setups vary to large extents, and the descriptions and attention to detail of the experimental conditions differ widely. For example, for different image and video quality databases the viewing distance has been described as

- "comfortable seating distance" [Wan+04],
- "approximately 80 cm [...] controlled by a line hanging from the ceiling" [Vir+14],

- "comfortable for [the participant]" [Pon+15],
- and "$6 - 8H$ [...], where H is the height of the video" [De +09; De +10].

To illustrate it at the hand of a different example, some studies report very accurate lighting conditions measured at the monitors and in the surroundings - "The light hitting the monitors measured below 2 lx, and the ambient illumination from behind the monitors were 20 lx." [Vir+14] - while other studies report no such measurements at all.

The authors of the Tampere image database put it fittingly [Pon+15]:

> "in our opinion, visualization and analysis of image quality in slightly varying conditions provide reasonably good verification of quality metrics if these metrics are intended for visual quality assessment in practice in a priori unknown conditions".

Ultimately, digital media is consumed in environments that are unpredictable and likely do not conform to the strict boundaries that lab-based subjective annotations were obtained in. In this sense, it is arguable, whether the controlled lab environments are reflective of the way media is consumed in-the-wild.

There are two additional points of consideration to using a lab-based setup. First, performing a subjective experiment in a laboratory setting usually entails an expensive facility, which is commonly only set up once, restricting the pool of subjects. Most often, participants in these types of studies have been graduate and undergraduate students at the universities the experiments were carried out at. This can have the benefit of obtaining the opinions of young, technology-savvy people, which may have a more nuanced understanding of the visual quality of digital media. However, this potential bias in quality evaluations may also adversely affect any model trained on them, if its application is not primarily aimed at this particular demographic.

Secondly, carrying out an in-person experiment at a university lab is time-consuming and costly beyond the hardware setup. The facility can only be used by a single participant at any given time, and conventionally requires at least one overseeing researcher for the duration of the experiment. This heavily restricts the number of annotations that can be efficiently gathered and, thus, potentially limits the statistical relevance of the results. Typically, turnaround times for lab-based quality evaluation campaigns lie in the order of weeks, even though the datasets have thus far been relatively small with a maximum of a few hundred stimuli. Therefore, lab experiments are an insufficient tool, if the goal is to significantly increase visual quality assessment dataset sizes.

2.1.2 Crowdsourcing

Crowdsourcing has emerged as an affordable and fast alternative way of annotating multimedia datasets with subjective opinions. Although there is no clear consensus on the definition of the term crowdsourcing, a recently published white paper on the topic can be understood to be the most elaborate attempt so far to establish an agreed-upon interpretation. Here, crowdsourcing was defined as "an action by an initiator

who outsources tasks to a crowd of participants to achieve a certain goal" [Hoß+20], where *action* encompasses study design, data capturing, and analysis, and *participants* specifically refers to human workers that process the tasks. This interpretation would, in fact, incorporate lab setups as described above under the term crowdsourcing, however, in the context of this thesis the term will refer exclusively to the outsourcing of tasks to an online crowd.

The distribution of micro-tasks to remote study participants via internet crowdsourcing platforms opens the reach to a diverse group of subjects that can provide results in a parallelized way, drastically decreasing turnaround time. Nonetheless, crowdsourcing also introduces new challenges, such as the further need to incentivize participants, as well as the lack of human-guided instructions and a reduced level of control over the environment, resulting in potentially lower quality of annotation. We will address both of these issues separately in the following.

User participation is a prerequisite for the success of any study. A model for worker's motivation has been proposed for the crowdsourcing context specifically [KSV11], which differentiates between enjoyment and community based intrinsic motivation factors, as well as extrinsic motivation stemming from social factors and immediate or delayed payoff expectations. In their evaluation of survey data collected from crowd workers on the Amazon Mechanical Turk platform showed intrinsic motivation factors dominating extrinsic ones. Although the extrinsic motivation factor of payment was rated as most important by a majority of the participants in the survey the authors concluded that their direct way of asking for the importance of money as a motivational factor for workers was non-objective [KSV11]. Still, further studies have shown that payment positively impacts task uptake, that is the number of tasks completed per participant, as well as tolerance towards more complex tasks [MW09; Rog+11]. Various studies have also noted, however, that payment beyond what is perceived as sufficient by the worker is not correlated to output accuracy [Fin+10; Rog+11; CK13].

In order to maximize the quality of the results of a crowdsourcing experiment, quality assurance and reliability methods help to identify individual erroneous submissions, as well as workers who deliberately act maliciously. Without such measures, participants may obtain payment without performing the tasks as requested, subsequently requiring repetitions of the study and incurring additional costs. Malicious participants in crowd studies have been categorized into different groups of behavior [Gad+15]. Workers may be ineligible, meaning that they participate in the experiment despite not conforming with specifically stated requirements, such as not using a particular device. Other participants are primarily driven by earning money quickly which causes them to try to take advantage of a lack of validation procedures, for example by supplying ill-fitting responses. Alternatively, some crowd workers may attempt to deceive simple validation procedures by apparently abiding by the given rules, but still providing responses that are not useful for the experimenter. All of these behaviors need to be addressed.

Beyond this, technical challenges ought to be considered for crowdsourcing campaigns aimed at subjective video quality assessment in particular. Users of crowdsourcing platforms have an international background,

with different platforms being populated by different crowds. To distribute data, and video data in particular, in a time-efficient manner, such that no buffering or other transmission-related artifacts outside of the control of the experimenter occur, the crowdsourcing platform has to be tested properly and the data potentially adjusted.

To summarize, there are several additional challenges associated with this relatively young method of collecting subjective data via crowdsourcing. The parallelization of tasks and decoupling of human-guided instructions generates vastly more data in a much shorter time. Additionally, the viewing conditions are harder to control and data distribution is - in most cases - outside the realm of control for the experimenter, potentially leading to lower data quality.

Ideally, the data generated by a crowdsourcing campaign should be *valid*, *reliable*, as well as *representative*. This means that that there should be a level of "confidence that a given finding shows that it purports to show"[HM14], a "confidence that a given empirical finding can be reproduced" [HM14], and a sufficient "degree to which the data sample is representative for the assumed population" [Hoß+20]. Although these three concepts share similarities, they need to be addressed separately.

The reliability of crowdsourcing studies for video quality assessment has been studied in related works [Gar+14]. In this work, two different quality control mechanisms are compared. The a posteriori mechanism entails the carrying out of the study with a subsequent filtering step. This has the downside of requiring more answers, as it is unclear prior to running the study how many workers will be filtered out. Moreover, the implementation and evaluation of this additional step is an administrative overhead. An alternative mechanism called in memento quality control is suggested that evaluates a participant's reliability during the assessment session. This not only allows the identification of unreliable users but also opens up the potential for feedback to the worker. Compared to the a priori mechanism, the in memento mechanism doubled the reliability of the results, while also reducing the total campaign completion time by a factor of ten. Interestingly, the study also found that repetitive involvement of only those workers that had been deemed reliable in previous studies improved the quality of the results only to a certain extent and had the additional downside of exhaustion of the crowd.

In recent years, crowdsourcing has also been recognized by the ITU as a solution to challenges posed by laboratory tests. We will briefly review those ITU-T Recommendations that discuss crowdsourcing within their scope.

Rec. ITU-T P.912: Subjective video quality assessment methods for recognition tasks

This Recommendation [ITU16] concerns itself with the identification of humans, faces, objects, and alphanumerics in video test material. Although these aspects of video quality assessment are not relevant to the thesis, the 2016 version has, for the first time, included crowdsourcing as an alternative to tests performed in a laboratory environment.

Based on related work in the field [Hoß+14], it describes "possible ways to increase the accuracy of results obtained [in the crowd]" [ITU16]. For the experimental design, it suggests using easy tasks with short completion times, such that a worker is unlikely to be interrupted during the session. Further, it recommends less detailed descriptions with heavy use of pictures as illustrations to make it as simple as possible to understand.

The Recommendation also touches on quality assurance and reliability testing, identifying the problem of random clickers and unreliable workers. The former kind of participant, which e.g. provides random answers on radio button rating scales, can be identified by using repeated questions. For the detection of unreliable test subjects, it suggests using specific obvious questions, enabling a simple filtering mechanism.

Rec. ITU-T P.808: Subjective evaluation of speech quality with a crowdsourcing approach

This Recommendation [ITU18] provides advice specifically for carrying out crowdsourcing campaigns to evaluate speech quality. It is much more specific to problems related to crowdsourcing, as it is the core assessment process considered within this Recommendation.

For the test procedure, it suggests a split into three parts, where first a worker has to qualify by passing an initial test with an adequate accuracy on questions with pre-conceived acceptable answers, so-called gold standard questions. Then, participants cycle through training tasks that grant temporary access to a fixed number of rating tasks. Training tasks are essentially re-affirmation for the experimenter that the worker knows how to do the job while rating tasks contain the items that are to be subjectively annotated. The Recommendation contains example job designs for all parts of the procedure, which can easily be adapted to image and video quality assessment.

2.2 Lab vs. Crowd Comparison

If we believe laboratory environments to be the gold standard for subjective assessment campaigns, the question remains how well crowdsourcing approaches perform in relation to them. A few comparison studies in fields related to and within the image and video quality domain exist.

For the evaluation of privacy filters in video surveillance [KCE12] an online crowdsourcing campaign was compared to results obtained in an offline lab-based environment with more rigorous control. The dataset consisted of video sequences that had been altered by privacy protection filters, e.g. blurring or pixelating human beings shown within them. Participants were asked to assess whether they were able to discern particular features of the subjects. This online evaluation correlated well with the results obtained in a previous lab-based experiment, with a 0.87 Pearson linear correlation coefficient (PLCC). The authors concluded that the crowdsourcing-based results were more preferable due to a more diverse pool of participants and the crowd setup more closely resembling real-life surveillance scenarios.

Within the IQA domain, previous work [RFN11] has evaluated the performance of a crowd-based assessment of the LIVE-IQA [SSB06] dataset. The evaluation procedure mirrored the original setup of a single stimulus methodology with an ACR scale carried out in a lab. To increase the comparability between the two assessment approaches a comparable number of workers were employed in the online setting as had been used offline, resulting in a correlation of 0.98 PLCC. This study employed a relatively conservative user screening that compared the answers of participants to the MOS values of the rest of the crowd. If the correlation between a worker and the crowd was lower than 0.25 the data was discarded. Additionally, workers were incentivized to perform the task well by being able to receive bonus payments for having answers that correlated strongly with the global MOS.

Similarly, related work in the field of video quality of experience [Wu+13] obtained assessment scores in both a lab and on Amazon Mechanical Turk (AMT) that were "reasonably consistent". This study employed a PC methodology and applied a fairly simple user screening approach that introduced the transitivity satisfaction rate. If a user had answered each comparison of a triplet of videos $\{A, B, C\}$, and A had been considered qualitatively superior to B, while B had been deemed of higher quality than C, then A had to also have been scored higher than C for a certain fraction of triplets. Otherwise, the worker was rejected and not paid.

These studies show that even with relatively simple quality control mechanisms a crowdsourcing setup can achieve an annotation quality comparable to lab setups. Inspired by this, we report on our own crowdsourcing-based experiment aimed at reproducing MOS values from a previously conducted lab-based experiment. Concretely, we set out to answer the following question:

> Can an online crowdsourcing study reproduce the results obtained in a strictly controlled lab setting, despite potential constraints induced by the nature of crowdsourcing studies?

We asked crowd workers to compare cropped versions of pairs of videos at different qualities side-by-side. Although we had to reduce the video sequence resolutions to 480x400 to fit a broad variety of displays, we found a very strong correlation between the high-resolution lab-based MOS values and our crowdsourcing-based low-resolution paired comparison (PC) results. The following Sections will elaborate on the individual elements of the study.

2.2.1 Dataset

For our experiments, we chose the IRCCyN IVC 1080i video quality database [PPL08a] that contains 24 groups of video sequences in 1920x1080 resolution and i50 (interlaced, 50 fps) format of 9 to 12 seconds, each group including 8 video sequences with different levels of visual quality ranging from excellent to bad. The video sequences with different levels of quality were obtained by encoding 24 raw source video sequences using different bitrates. Since this study is meant as a proof of concept, we used only 10 of the 24 available sources and only 4 of 8 levels of quality, evenly spanning the available range of quality. The original

study provided the absolute category ratings using the hidden reference testing methodology of 24 observers.

In order to facilitate our crowdsourcing-based study, several video preprocessing steps were carried out, using the open-source cross-platform conversion software FFmpeg. First, the interlaced HDTV video sequences in raw YUV format and 1920x1080 resolution were deinterlaced. To allow for the simultaneous, side-by-side display of a pair of video sequences on a typical crowd workers screen size we center-cropped the video sequences to a format of 480x400 pixels. This choice was informed by statistics that showed a majority of crowd workers having a resolution of at least 1024 pixels horizontally. To reduce the bandwidth requirements for online video streaming, we re-encoded the video sequences to a lower bitrate. We required a minimum PSNR of 40 dB, which ensured that no significant compression artifacts were generated by this step, and subsequently obtained video sequences of 2 to 4 MB size, which is a manageable size for a crowdsourcing study. For each of the 10 video groups each with 4 quality levels (L1, smallest bitrate and lowest quality, to L4, largest bitrate and best quality) we composed a series of $\binom{4}{2} = 6$ merged pairs of video sequences for the paired comparison tasks. In each pair, one of the videos was randomly selected for display on the left and the other on the right side. Thus, we obtained a total of 60 video sequences each showing one pair of video sequences for comparison. We used CrowdFlower to conduct the experiment and the implementation of the tests was based on the QualityCrowd framework [Kei+12].

CrowdFlower later became FigureEight, and was then acquired by Appen (https://www.appen.com)

2.2.2 Crowdsourcing

As previously described, ITU-R Rec. BT-500 [ITU19] suggests using a seven-grade adjectival categorical judgment scale for paired comparison methodologies. We adopted a coarser five-grade adjectival categorical judgment scale as shown in Table 2.3, which is similar to the five-level Likert scale.

Table 2.3: Five-grade adjectival categorical judgment scale similar to a five-level Likert-scale for quality preference between two side-by-side video stimuli.

Value	Interpretation
−2	Left is much better
−1	Left is slightly better
0	No difference
1	Right is slightly better
2	Right is much better

In crowdsourcing-based experiments, one of the key problems is to ensure the reliability of the performance of the crowd workers. Typically, in CrowdFlower this is ensured by three systems. First, users of the CrowdFlower platform are rated according to their global track record on experiments that they participated in, and are grouped into qualification levels, accordingly. Workers with higher qualification levels have access to better-paid tasks and crowd employers may limit the access to their jobs by imposing a minimum level of qualification. We limited the job to the highest level of qualification, meaning an average 90% accuracy over all previous jobs. Secondly, crowd workers must pass a set of qualification questions, often called *quiz*, to begin the actual experiment. In order to pass this quiz, a minimum level of rating accuracy set by the experimenter has to be reached, which we set to 70%. Finally, test questions for which ground-truth answers are provided are placed on each page of an experiment. A page consists of several random work items alongside these hidden test questions. Workers whose success rate on test questions dropped below 70% were excluded from further participation in our study and their previous answers were replaced.

To assess the effect of the degree of control we performed two crowdsourcing studies, the second one with less restricting requirements regarding reliability control in the qualification test and the test questions during the actual work. In total the experiment had 626 participants, 589 (94.1%) of which passed the quiz, and 576 (92%) of which passed the quality assurance testing during the experiment. We obtained 50 ratings per stimulus, yielding an average of 5.2 judgments per participant.

2.2.3 Results

In order to compare the original lab-based MOS ratings in the range $[1, 5]$ with our five-level paired comparisons in the range $[-2, 2]$ one must either transform the paired comparisons to the MOS, i.e., one must reconstruct absolute quality levels from the relative comparisons of video qualities or, vice versa, transform the MOS values of the ACR ratings to pairwise ratings of differences in quality. The first approach of reconstructing absolute quality ratings from differences is well researched only for binary judgments ('right is better or worse than left'), additionally allowing for a tie, see e.g. [Dav88]. For simplicity, we, therefore, convert the lab-based MOS of the left (S_l) and right (S_r) video stimuli, to a differential MOS (DMOS) by linearly mapping the difference $S_r - S_l \in [-4, 4]$ of the MOS to $\frac{5-\epsilon}{8}(S_r - S_l) \in (-2.5, 2.5)$, so that by rounding to the nearest integer one would get a value in the range of our five-grade judgment scale $\{-2, -1, 0, 1, 2\}$.

The MOS values from the lab-based study [PPL08a] together with their standard deviations, both averaged over 10 stimuli in each quality level L1 to L4, are presented in Table 2.4. For each of the 60 video pairs in both crowdsourcing studies we collected 50 DMOS values from the crowd. Table 2.5 and Table 2.6 provide the 6 comparative scores averaged over the 10 video groups together with the corresponding computed DMOS values from the lab-based study from each of our two crowdsourcing studies, respectively. The individual comparisons are additionally visualized as scatter plots in Figure 2.1, showing a very strong correlation (> 0.966) between our crowdsourcing-based results and the lab-based ones.

These results show that the crowdsourcing-based DMOS values are strongly related to the MOS values obtained in a strictly controlled study in a lab. In fact, the lab-based study also had compared results between two closely related methodologies, namely ACR and SAMVIQ.

Table 2.4: Original lab-based MOS values averaged over 10 stimuli for each quality level.

Level	MOS ($\pm\sigma$)
L1	1.6 ($\pm$0.6)
L2	2.8 ($\pm$0.8)
L3	3.6 ($\pm$0.8)
L4	4.5 ($\pm$0.6)

Table 2.5: Grouped DMOS comparison of the first crowd-based study with lab-based results.

	Crowd I	[PPL08a]
L1–L2	0.9 ($\pm$0.4)	0.9 ($\pm$0.4)
L1–L3	1.1 ($\pm$0.5)	1.3 ($\pm$0.3)
L1–L4	1.5 ($\pm$0.3)	1.7 ($\pm$0.3)
L2–L3	0.3 ($\pm$0.5)	0.5 ($\pm$0.5)
L2–L4	0.9 ($\pm$0.4)	1.0 ($\pm$0.4)
L3–L4	0.6 ($\pm$0.3)	0.6 ($\pm$0.2)

Table 2.6: Grouped DMOS comparison of the second crowd-based study with lab-based results.

	Crowd II	[PPL08a]
L1–L2	0.8 ($\pm$0.3)	0.9 ($\pm$0.4)
L1–L3	1.0 ($\pm$0.5)	1.3 ($\pm$0.3)
L1–L4	1.4 ($\pm$0.3)	1.7 ($\pm$0.3)
L2–L3	0.4 ($\pm$0.4)	0.5 ($\pm$0.5)
L2–L4	0.8 ($\pm$0.4)	1.0 ($\pm$0.4)
L3–L4	0.5 ($\pm$0.2)	0.6 ($\pm$0.2)

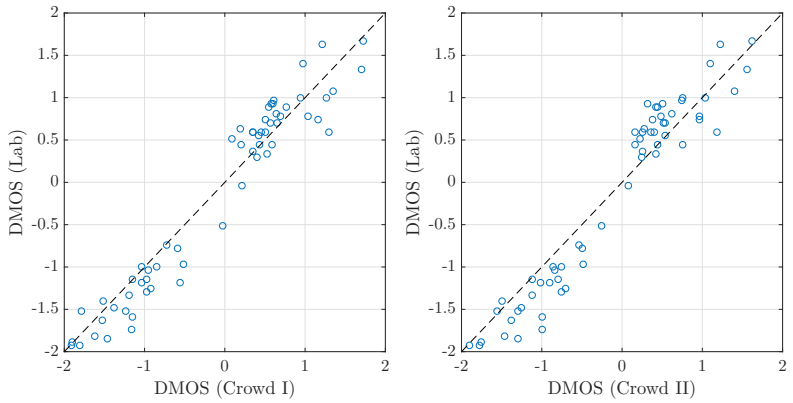


Figure 2.1: Scatter plots comparing three assessments of DMOS values for 60 paired comparisons of video quality. Left: Crowd study 1 (strict quality control) versus DMOS derived from lab-based MOS values. Right: Crowd study 2 (mild quality control) versus lab. The Pearson correlation coefficients are 0.9687 (left), 0.9661 (right).

The correlation coefficient between these scales was reported as 0.8993. Compared with our results of 0.9687 and 0.9661, one can see that DMOS estimates using crowdsourcing can be as precise as lab-based studies even though there was severe processing of the video sequences and the control of testing conditions and worker reliability are considered much weaker in crowdsourcing studies. Another result is that MOS values of the lab-based study were linearly correlated with DMOS values from the crowd workers. These findings hold for our particular case study, but a generalization to all video quality assessment scenarios is not straightforward.

2.3 Human Perception

In order to model human perception of quality, some approaches have taken a bottom-up strategy of simulating well-modeled functionalities of the human visual system (HVS), which they base their prediction on. At its core, the HVS is an information processing pipeline including the eyes as the information acquisition module, the lateral geniculate nucleus (LGN), a relay center that produces time correlated and spatially correlated outputs useful in three-dimensional representation of objects, and the visual cortex, which contains the core signal processing functionalities for visual information. In a sense, bottom-up approaches to objective IQA and, as a consequence, to VQA are directly connected with the characteristics of how the HVS functions. We will, therefore, briefly introduce the anatomy and early processing pipeline of the HVS, alongside an evaluation of some peculiarities of the system and their effect on quality perception in peripheral vision. As a visual aid, a simplified visualization of the HVS as shown in Figure 2.2 details the core modules of the HVS.

2.3.1 Human Visual System

Light hits the cornea and lens, where it is refracted, passing into the eye. Here, it is projected onto the retina, a membrane composed of several layers of neurons and cells, which is a structure typical for the different parts of the visual information processing pipeline. The first layer in the retina samples the retinal image projected onto it and then passes it to a layer comprised of rods and cones, which are two types of photoreceptor cells with different purposes. Cones activate according to different wavelengths (~420, ~540 and ~570nm), encoding color information in the process. Rods are more sensitive and, therefore, function better than cones at lower light, however they do not play in color perception, as all rods activate at approximately the same wavelength (~500nm). Moreover, rods and cones are distributed non-uniformly on the retina. In the fovea, which is the part of the retina at the center of the visual field along the visual axis, cones comprise the majority of photoreceptor cells and it is the place where cones are most densely populated. The further away from the fovea, the sparser populated the membrane is with cones, and the denser the distribution of rods becomes. An exception to this rule lies at the part of the retina where the optical nerve creates a blind spot with no rods and cones.

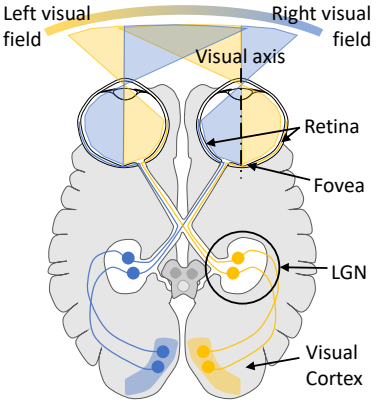


Figure 2.2: Schematic diagram of the early human visual system including the optical system of the eye, which connects the eyes with the visual cortex via the lateral geniculate nucleus (LGN). The LGN is an intermediate information processing and transmission structure.

The signal generated by the activation of rods and cones is passed through several layers of interconnecting neurons down to ganglions that finally transmit the signal to the LGN through the optical nerve for further processing. The LGN is also comprised of multiple layers of cells with different characteristics, such as whether cells in a layer process color information, or how fast the cells respond to stimulation. In the left and right LGN, the signals from the left and right visual fields of both eyes are merged before being relayed to the visual cortex.

Both hemispheres of the brain include a visual cortex that is responsible for the processing of information from the opposite hemisphere of the visual field that it is located in. It consists of several layers that are part of two primary pathways, called the ventral stream and the dorsal stream, which are functionally different. The ventral stream, comprised of a connection of layers V1 through V2 and V4 to the inferior temporal cortex, is associated with form recognition and object representation. Running from V1 through V2 into both V6 and V5, and further into the posterior parietal cortex is the dorsal stream, which is associated with motion, representation of object locations, as well as the coordination of eyes and arms when visual information is used to guide saccades or reaching. Each stream puts a different emphasis on the visual field, as the ventral stream emphasizes the central visual field, and the motion analysis concentrates on the periphery.

Overall, the hierarchy of the visual system can be divided into low-, intermediate-, and high-level visual processing functions. Here, low-level vision can be understood as processing of visual information that can be done without explicit knowledge that images come from surfaces in depth. Intermediate-level visual processing can be done without explicit knowledge of objects and their locations. Finally, high-level vision makes explicit inferences about objects and their action-dependent relations to the viewer.

2.3.2 Human Perception of Quality Artifacts

With the core functionality of the HVS discussed, it is important to reiterate that these descriptions are simplifications. The understanding about the exact mechanisms by which the different modules of the visual processing pipeline interact is still incomplete and undergoing active research. Still, we should highlight some HVS features and the resulting perception that relate to visual quality assessment. For example, the sensitivity of the HVS to contrast depends on the spatial and temporal frequencies of a visual stimulus, resulting in a human's perception of contrast not being entirely determined by visual content. Visual masking, which is the phenomenon where the perceptibility of a target feature in a stimulus is affected by the presence of another masking feature, is another peculiar feature of the HVS. For an observer, the target can appear with reduced contrast or intensity, the degree of which is also impacted by the spatial proximity of target and masker.

These and other features impact the evaluation of the visual quality of a stimulus. For example, the presence of some distortions, such as noise, may be less perceptible in areas of an image with strong textures. Alternatively, temporal discontinuities between two adjacent frames can

mask the presence of a variety of spatial distortions. At this point, it is useful to describe the most common types of visual artifacts found in images and videos. Broadly speaking, video quality artifacts can be categorized into spatial and temporal artifacts, where spatial artifacts are visible in individual frames, while temporal artifacts are perceived when a sequence of frames is displayed. Both classes can be further subdivided as shown in Figure 2.3 and we will summarize the different subclasses in the following. However, we do not claim this description to be exhaustive, and the inclined reader is referred to related work [SK98; YW98; Zen+14] that provides in-depth discussions of common distortion artifacts.

Refer to Section 4.1 for a description of different approaches to the detection of artifacts for use in objective quality models.

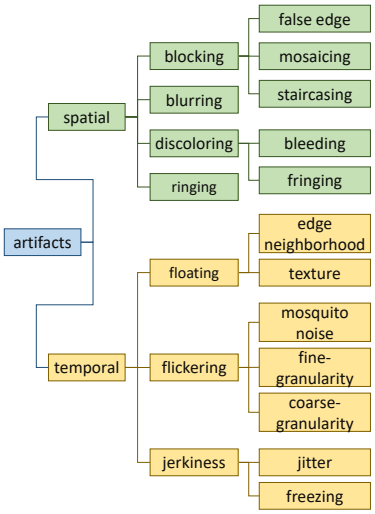


Figure 2.3: Categorization of video compression artifacts into subclasses of spatial and temporal domains.

Spatial Distortions

First, Blocking artifacts are caused by compression techniques that rely on block-based quantization, especially when limited by a low bitrate. Video frames are segmented into blocks, which are then quantized at different levels, without any consideration of inter-block correlations, resulting in discontinuities at block boundaries. Subcategories of artifacts of this type describe realizations of blockiness that are perceived particularly saliently. False edges are edges that appear in the encoded video stream that were not present in the original recording. The artifact is visible in still frames but results from temporal aspects of the encoding process, where motion compensation can cause a carry-over effect of blockiness from neighbouring frames both into the future and the past. Mosaic distortions appear in smooth regions of video frames that have non-uniform luminance levels, such as the sky at dusk or down or a uniformly colored wall in a room with localized lighting. Blocks in these smooth regions carry no high frequency components. They are, therefore, reconstructed using only the constant direct current (DC) component, which will vary slightly as luminance changes. Adjacent blocks with different DC values within these smooth regions will visibly stand out, as the color discontinuities at the block boundary are very salient. Finally, staircasing describes blockiness along (thin) diagonal lines with high levels of local contrast, which showcases a fundamental limitation of using rectangular blocks. The encoded video will show rectangular structures along the diagonal line that look like a staircase, hence the name.

Secondly, Blurring is caused at two stages of the encoding process. First, the quantization process favors low frequency components over high frequency components, resulting in the removal of texture and, thus, blurring. Second, to counteract blockiness modern encoders use de-blocking filters that target discontinuities at block boundaries by essentially applying spatial low-pass filters, further removing detail and causing blur.

Distortions of color are the third group of spatial artifacts and can be subdivided into two categories. First, color bleeding is caused by inconsistencies in luminance and chromatic channels resulting from encoding. Most often particular color channels are represented at lower resolution than luminance, which implies some form of interpolation when rendering the frames at full resolution. Additionally, the aforementioned artifacts of blur and blockiness are inconsistent across different color channels, resulting in irregular color spreading. Secondly, color fringing is a group of artifacts that result from hardware limitations. As sensors

of digital cameras decrease in size and pixel counts simultaneously increase, chromatic abberation can occur. Because the refractive index of optical glass varies with different wavelenghts not all light is brought into a single focus point on the image plane. Overall, this effect is rare in modern, high-quality, compound lenses, and is commonly only seen at the edges of the digital medium.

The final spatial artifact class we will mention here is ringing. Similarly to staircasing it appears along diagonal edges, while the perceptual effect is stronger with increasing levels of contrast at the edge. Due to the block-based nature of the compression, higher-frequency components in image blocks that overlap the edge in the image will affect the quantized reconstruction of the entire block even in the part of the block on the opposite side. The result are what is perceived as a ripple away from the edge that is parallel to the high contrast edge. If the area surrounding the edge is particularly smooth this effect is amplified, while high levels of texture can mask the ringing effect.

Temporal Distortions

Floating artifacts are a subclass of temporal distortions that are caused primarily by what is called the Skip coding mode of video encoders. As the name suggests, this mode indicates the skipping of encoding motion-comprensated predictions for a particular block, leaving the decoder to estimate motion vectors from neighbouring blocks. As a result, the skipped blocks will often appear to be moving differently than the observer might expect, often appearing as floating on a different plane than the surrounding background. This type of temporal artifact is further distinguished depending on whether the distortion appears in regions of an image that comprise texture or edges.

Frequent changes along the temporal dimension in the same spatial location of the luminance or chroma channels are called flickering. This type of temporal artifact is particularly noticeable and annoying for an observer. It can be subdivided into three distinct subclasses. Mosquito noise is the perception of temporal artifacts in a sequence of frames containing spatial artifacts such as ringing. As a sharp, high-contrast edge that is distorted by ringing moves in a sequence of frames the artifacts move with it, but with slight variations. This gives the observer the perception of small mosquito-like objects flying around the edge, giving this temporal artifact its name. Coarse-granularity flickering refers to sudden luminance changes in large parts of a frame caused by an independently encoded frame following the last of a chain of frames that were encoded using information from prior frames. If the number of these predicted frames is high, an increase in luminance over time may have been predicted incorrectly, and the next independent frame appears significantly and saliently brighter. Fine-granularity flickering, on the other hand, describes frequent flickering in small image regions that are the result of slightly different quantization in successive frames. Mosaic distortions in smooth regions of a video may vary slightly over the course of successive frames, giving the appearance of luminance or color changes at these locations.

Finally, jerkiness describes discontinuous motion of objects or scenes, which has a strong negative effect on perceived video quality [Pas+04; HG06]. The source of jerkiness can be transmission delays of the encoded bitstream to the decoder resulting in frozen frames, or unexpected frame skipping and frame loss which causes jitter.

Distortion Masking

Another interesting peculiarity of the HVS with respect to different distortions is non-uniform spatial resolution of the fixated part of the visual field, which is the result of the non-uniform distribution of cone photoreceptors on the retina. This can lead to masking of distortions in peripheral vision, which can be used in image compression techniques. Distortion masking in the peripheral visual field is the basis for region-of-interest (RoI) based image compression, which suggests to compress the background of an image more than the foreground for a better perceived quality at the same bitrates. Although the background will then be of significantly lower quality, the salient parts of the image will cause artifacts in the background to be masked simply by them not being in focus.

2.4 Saliency-driven Image Coding

In early studies on two-level RoI-based image coding a particular region selection strategy did not improve on the standard JPEG2000 overall [BS03]. Although slight improvements could be observed for very low bitrate encodings, the quality of the images were so degraded at that level that it does not warrant practical use. A more recent study on perceptual quality in images showed that image foreground regions are much more important than the background [Ale+13]. Additionally, some variable quantization techniques have already been shown to produce better results than standard JPEG encoding for special applications. For example, previous work adjusted the quantization scaling factors in such a way that text blocks were compressed at higher qualities than image blocks [KT00], resulting in an overall increase of perceptual quality. Another study used a measure of block activity and type to determine quantization [MT00]. Using 256 bytes of side information the resulting decoded image had significantly superior perceptual quality. However, none of these works evaluated their approaches in terms of perceptual improvement quantitatively, for example by performing subjective quality assessment user studies.

Nonetheless, the results of these reports suggest RoI-based image coding to be a promising approach to improving visual quality of digital media by exploiting characteristics of the HVS. Inspired by other related works [Ale+13], we consider saliency as a governing principle to the success of this approach. Our main research question is this:

> Using saliency as a basis for variable quantization, can we improve the perceptual quality when compared to standard JPEG?

The investigation comparing saliency-based coded images to standard JPEG encodings should consider multiple perceptual factors and evaluate the amount of perceived improvement in image quality at fine levels of quality differences. Several factors are to be considered, primarily driven by the need for a methodology of how RoIs are derived from saliency maps. The trivial way to do this would be to quantize the saliency map according to the block sizes. This, however, is problematic for the reason that both the encoder and decoder need access to this information, leading to a large overhead in the coded bitstream. One way to counteract this is to model the saliency map, which involves choices for the number of regions to consider, their span, and the granularity of levels of importance to consider. All of these choices affect the overhead of encoding variable quantization levels.

For simplicity we made a choice for the approach of variable quantization in JPEG Part 3 rather than for JPEG2000. One of the intended purposes of variable quantization is "the ability to use the masking properties of the human visual system more effectively, and thereby achieving greater compression rates for the same subjective quality." [ITU96] The technique offers a large number of quality levels rather than just two (for foreground and background) that we use for saliency-based coding. By applying a few simple transformations to a saliency map we can derive all the necessary parameters of the RoIs required by Rec. ITU-T T.84. The number of ROIs and their spans can be controlled by the standard deviation of a Gaussian filter that generates the saliency map from eye fixations. Their importance levels relate to the levels of saliency. The final (discrete) JPEG quality at 8×8 block level used for compression is a result of a quantization of this transformed saliency map. Parameterizing such a coding strategy allows to evaluate the impact of each factor on the perceived quality of reconstructed images. A suitable technique for performance evaluation shall be sensitive to small differences in perceived quality.

To evaluate the results of our image coding algorithm, we carried out a large-scale crowdsourced evaluation. Whereas many saliency-based coding approaches were evaluated using objective quality measures (e.g., [NCS06; Raj16; BS02; MM11; Gow+14]), only a few studies have employed subjective measures of preference, usually on small sized lab studies [BS03]. We involve a larger number of images with good variety for our crowdsourced studies. Having a diversity of content and choosing the images automatically reduces the level selection bias that an experimenter might create otherwise. Additionally we offer a way to go beyond opinion scores, aiming for something more tangible, by introducing the concept of perceived bitrate. This relates opinion scores to objective coding factors, in this case the bpp difference between two compared images.

2.4.1 Methodology

In contrast to previous work in RoI-based image coding we derive our JPEG quality levels from saliency maps, generated from many fixation points obtained via eye tracking. The variability of this approach make it a promising proposition.

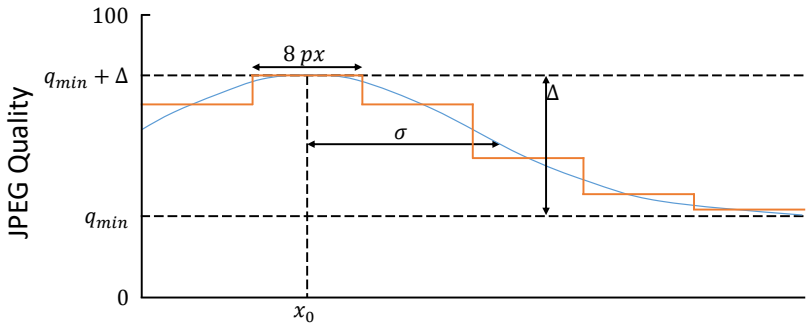


Figure 2.4: Illustration of the saliency-based variable quantization JPEG coding strategy. For simplicity we show the JPEG variable quantization quality factor $q_B[x, y_0]$ (the staircase curve) for just one scan line ($y = y_0$) of an image that has just one eye fixation at $[x_0, y_0]$, in the same scan line. The smooth curve is the graph of the linear function of the saliency $x \mapsto s[x, y_0] \cdot \Delta + q_{\min}$ (compare Eq. (2.2)). The JPEG quality is bounded to the interval $[q_{\min}, q_{\min} + \Delta]$ and $q_{\min}$ is adjusted to achieve the target bitrate.

The purpose of the saliency map of an image is to quantify the level of relevance at all pixels with respect to perceived subjective image quality. Our model for saliency maps is simple, taking into account a set of image (eye fixation) points $[x_i, y_i]$ with weights w_i, $i = 0, \ldots, k-1$. Then the saliency map is obtained by applying a Gaussian filter,

$$s[x, y] = \frac{1}{S} \sum_{i=0}^{k-1} w_i \delta[x - x_i, y - y_i] * g[x, y]$$

where S is a scaling value to normalize the maximal saliency on the image support to 1, $\delta[x, y]$ is the 2D unit impulse signal, and

$$g[x, y] = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

is the 2D Gaussian filter with standard deviation σ.

To reduce the (possibly large) set of eye fixation points to a small but still representative subset we used k-means clustering with $k = 8$ clusters $[x_i, y_i]$. The weights w_i were taken as the corresponding sizes of the k-means clusters.

In variable quantization JPEG a quality factor $0 \le q_B[x, y] \le 100$ controls the quantization for each 8×8-image block with upper left corner at $[x, y]$. The larger $q_B[x, y]$, the better is the reconstruction quality and, as a consequence, also the bitrate associated with the block. This factor is based on the average saliency per block,

$$s_B[x, y] = \frac{1}{64} \sum_{i=0}^{7} \sum_{j=0}^{7} s[x + i, y + j]$$

and set to

$$q_B[x, y] = \min(s_B[x, y] \cdot \Delta + q_{\min}, 100) \tag{2.2}$$

where $0 \le \Delta \le 100$ is a parameter denoting the quality difference between foreground ($s_B = 1$) and background ($s_B = 0$) blocks, and $0 \le q_{\min} \le 100$ is an offset equal to the JPEG quality of a background block. Figure 2.4 is a simplified visualization of the abovementioned parameters. Note that the bitrate of the coded image will be a monotonically increasing function of the base quality given by $q_{\min}$.

Thus, our approach to saliency-based variable JPEG coding has three parameters, σ, Δ, and $q_{\min}$. The parameter σ controls the size of the salient image region(s), Δ governs the quality difference between foreground and background blocks, and $q_{\min}$ is the base quality assigned

to background blocks. Note that standard JPEG coding, i.e., not using variable quantization, is given by the special case of $\Delta = 0$, in which the JPEG quality is equal to $q_{\min}$.

When comparing a coding strategy, parametrized by $(\sigma, \Delta, q_{\min})$, with a standard JPEG approach it is necessary to compare only images of the same bitrate which is a function of all three parameters. Given a JPEG coded image at a certain bitrate, we thus choose the base quality $q_{\min}$ for the variable quantization JPEG coded image such that the target bitrate is achieved as closely as possible. Computationally, we apply the bisection method for efficiency. This reduces the set of free parameters to just two of them, (σ, Δ).

For the reconstruction of an encoded image by variable quantization the decoder requires the block quantization factors $q_B[x, y]$. For this purpose JPEG prescribes a simple procedure similar to PCM. However, we found in experiments that for low bitrates the induced overhead is large and annihilates any gains that could be achieved using a saliency-based adaptive bitrate coding strategy. In our case, we can propose a more efficient coding scheme for this side information. We simply pass to the decoder the image saliency model itself together with the parameters $(\sigma, \Delta, q_{\min})$ such that the decoder can reconstruct all quality factors at the block level. For this purpose the x- and y-coordinates of the $k = 8$ fixation points are quantized to 8 bits, and their weights by only 3 bits. Thus, together with storage for σ (2 bits), Δ (2 bits) and $q_{\min}$ (7 bits), this side information amounts to only $8(8 + 8 + 3) + 2 + 2 + 7 = 163$ bits. Of course, the encoder must also use these same (quantized) saliency data and parameters.

2.4.2 Dataset

We constructed a dataset based on the MIT-1003 dataset [Jud+09] which contains 1003 images of natural indoor and outdoor scenes alongside the eye tracking data of 15 observers for a 3 second viewing duration. First, we estimated image blurriness by the ratio the average gradient magnitudes of the image at two different scales. Then we computed the SSIM [Wan+04] between the lowest and highest bitrate versions. We select the intersection between the 70% least blurry images and those 70% of images that had the lowest similarity between the lowest and highest desired bitrates. This selection resulted in 543 candidate images. For this subset we binned images according to the 8-bit entropy of their respective saliency maps to ensure a variety of test images. 11 images were randomly selected from each bin with rounded entropies of 3, 4, 5 and 6, which made up more than 95% of the dataset.

All 44 test images have a maximum dimension of 1024 pixels in height or width, with the other dimension ranging between 405 and 1024 pixels. The dataset consists of 41 RGB color images and 3 monochrome images. For the purpose of this study we considered 5 bitrates, 4 values for σ and 3 values for Δ, listed in Table 2.7. For each image source and each bitrate we compared a standard JPEG coded image with the $4 \cdot 3 = 12$ saliency-based adaptive bitrate coded variations. Thus, overall we had $44 \cdot 5 \cdot 12 = 2640$ paired comparisons.

Parameter	Number	Values	Unit
Bitrates	5	0.30, 0.36, 0.42, 0.50, 0.60	bpp
σ	4	5, 10, 15, 20	% of image width
Δ	3	15, 25, 35	JPEG quality factor

Table 2.7: Parameters and values considered for each source image in the crowdsourcing study.

2.4.3 Crowdsourcing Study

Subjective evaluation of the paired comparison was performed in two separate crowd experiments using the CrowdFlower platform. In the first experiment subjects were presented with paired comparisons consisting of standard JPEG images and different realizations of our proposed variable JPEG approach of the same bitrates. In the second experiment paired comparisons were comprised of standard JPEGs of 10 different bitrates. The original 5 bitrates used in the first experiment were augmented by 0.33, 0.39, 0.68, 0.82 and 1 to populate a broader range and giving better groundings in the low bitrate region. Workers were asked to denote which of the two shows more clear and sharp details present in the pictures on a 5-point Likert scale ranging from 1 to 5.

A CrowdFlower experiment is set up such that contributors can be tested. The questions are set by the experimenter for each task. Users are screened based on their performance on both the current task, and on their overall accuracy on the CrowdFlower platform. In both of our experiments we only allowed users with a cumulative accuracy rating of more than 70% to participate.

Workers were given very brief instructions, stating that they participate in a study of image encoding methods. They were asked to denote their answers quickly, so as to indicate their first impression of the comparison. No time constraint was imposed on the task, and each contributor was allowed to rate 500 pairs of images, or roughly 19% of the dataset.

Table 2.8: Five-grade adjectival score scale used to derive preference of the variable quantization-based test images over the corresponding standard JPEG version of the image.

Value	Test image preference
1.0	Strongly preferred
0.5	Slightly preferred
0	No preference
−0.5	Slightly disliked
−1.0	Strongly disliked

In order to ensure the quality of the results, we set test questions for both experiments. Test questions were comprised of paired comparisons between standard JPEG images encoded at different bitrates. Each test question had multiple allowed answers to capture the variability of the crowd. Contributors were presented a short qualification quiz comprised of test questions, the completion of which allowed them to perform the rest of the experiment. Throughout the experiment random test questions were presented. These were not explicitly marked as test questions such that contributors would pay close attention to each item. Whenever a contributor fell below 70% accuracy in answering test questions he was not allowed to continue. In this case all his previous ratings were discarded.

2.4.4 Performance Evaluation

In the first experiment each of the 2640 test images is compared to the corresponding standard JPEG image of the same bitrate. Figure 2.5 shows a selection of examples. For each judgement a score is set based on whether the test image is preferred over the standard JPEG version according

Figure 2.5: Standard JPEG vs. our variable coding strategy at the same bitrate. Some of the best parameter settings for the images in our collection with respect to DMOS. The salient parts of the image are considerably better quality than the less important and less noticeable background details. Refer to Table 2.10 for the parameters and study results for each image.

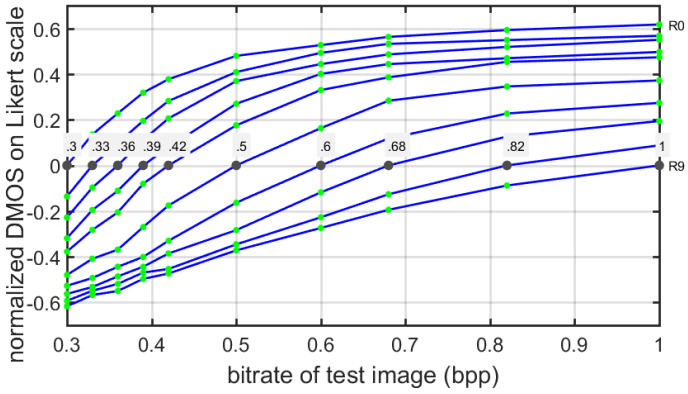


Figure 2.6: DMOS results (averaged over all 44 image sources) for standard JPEG images encoded at bitrate R (horizontal axis) with respect to reference bitrates $R_j = 0.3, \ldots, 1.0$, indicated at the DMOS $= 0$ line. These DMOS curves are used to estimate the bitrate advantage of our saliency-based adaptive bitrate JPEG encodings.

to the values shown in Table 2.8. The average score of all judgements is assigned to the test image and can be called a normalized DMOS. A positive DMOS indicates that on average the saliency-based coding yielded a reconstruction that the judges preferred over the corresponding standard JPEG image at the same bitrate.

To better assess the value of a given gain in DMOS we propose a quantification in terms of bitrate savings. For a given test image that was encoded to a bitrate R_j using one of the 12 coding strategies we ask for the (larger) bitrate $R^\star$ of a corresponding standard JPEG image (of the same source) that has the same perceptual image quality. Then the bitrate savings achieved by the saliency-based adaptive encoding is the difference $R^\star - R_j$.

In order to avoid the large cost for directly comparing all 2640 test images with a set of standard JPEG images, we estimate the bitrate savings as follows. The method is based on the data acquired from the second experiment in which we carried out paired comparisons for standard JPEG images with 10 different bitrates, $R_0, \ldots, R_9$. For all such comparisons of two images of different bitrates R_i, R_j, we computed the DMOS $D_{R_j}(R_i)$ of bitrate R_i with respect to the reference bitrate R_j, averaged over all 44 image sources. See Figure 2.6 for piecewise linear interpolations $D_{R_j}(R)$ for the 10 reference bitrates $R_0 = 0.3, \ldots, R_9 = 1.0$ bpp and $0.3 \le R \le 1.0$. Note that by design these functions $D_{R_j}(R)$ are monotonically increasing with the bitrate R.

Now, assume we are given a test image, adaptively encoded at bitrate R_j and with a DMOS advantage of $\epsilon > 0$ in comparison with the corresponding standard JPEG image at the same bitrate, R_j. Then it can be estimated that the corresponding standard JPEG encoded image at bitrate $R^\star = D_{R_j}^{-1}(\epsilon) > R_j$ has the same DMOS advantage. Thus, the difference $R^\star - R_j$ can be taken as the bitrate savings, and we will next report the relative bitrate savings $(R^\star - R_j)/R_j$, being more meaningful than the absolute values.

2.4.5 Results

In order to evaluate the performance of our approach we performed two crowdsourcing experiments. The first compared two versions of the same image encoded either using our variable quantization technique or standard JPEG technique at the same bitrate (VAR–STD). The second experiment compared different bitrates of the same image coded using standard JPEG (STD–STD).

A total of 453 workers from 53 countries participated in the first experiment, with 438 (96.69%) passing the qualification test, and 433 (95,58%) staying above the cumulative accuracy rating of 70%. The study ran for approximately 5 hours with a total cost of $106. The second experiment ran for roughly 10 hours, costing $133, and included a total of 950 workers from 72 countries. Here, 920 (96.84%) passed the qualification test, and 911 (95.89%) stayed above 70% test question accuracy throughout the course of the experiment. On average, each worker provided 184 and 108 answers for the two experiments, respectively. These stats are also reflected in Table 2.9.

We aggregated the results of each experiment by computing the normalized DMOS for each version of each image. Using the data from the second experiment (STD–STD) we computed the curves shown in Figure 2.6.

We aggregated the results of each experiment by computing the normalized DMOS for each bitrate version of each source image. The order of the images presented in each experiment is randomized. For establishing a clear relationship between images in the VAR–STD experiment, we reorder the images. In this new ordering the variable quantized JPEG is always on the right and the standard on the left. Thus, a positive DMOS score shows a preference for our approach, whereas a negative

	VAR–STD	STD–STD
no. of workers	453	950
ratio of workers that passed quiz	96.69%	96.84%
ratio of workers that passed work	95.58%	95.89%
avg. test question accuracy	95.73%	97.20%
max no. of answers per worker	500	500
avg. no. of trusted answers per worker	184	108
no. of answers per PC	40	50
no. of countries of origin	53	72

Table 2.9: Crowd statistics of the variable quantization JPEG experiments. Here, "STD–STD" refers to the anchoring experiment comparing two versions of the same image at different bitrates as obtained via standard JPEG which allows the estimation of bitrate savings, while "VAR–STD" is the experiment where our variable JPG approach is compared to standard JPEG compressed versions of the same image at the same bitrate.

	DMOS	Bitrate	Predicted bitrate	Bitrate savings (%)	Δ	σ(%)
musician	0.42	0.3	0.46	53	35	20
animal	0.32	0.3	0.39	30	15	10
flights	0.23	0.36	0.40	11	35	20
beach	0.37	0.42	0.51	21	35	20

Table 2.10: Parameter settings and results for the example images shown in Figure 2.5.

one implies a preference for the standard coding. We do this for the best parameter combinations (σ, Δ) for all bitrate versions of the variable quantization approach. This amounts to 220 image versions: 44 originals at 5 bitrates each. The results are shown in the histogram in Figure 2.7.

Using the DMOS data from the second experiment (STD–STD) we compute the curves shown in Figure 2.6. Relying on these trend lines we predict the relative bitrate savings for our VAR–STD experiment. The results are shown in Figure 2.7. We notice that in most cases our variable quantization approach shows an advantage over the standard JPEG encoding. On average we get 11% improvement in bitrate.

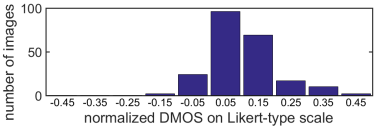


Figure 2.7: DMOS for the best parameters of our variable quantization approach in the VAR–STD experiment for all bitrate versions of each image. Positive values show the preference for our approach compared to the standard JPEG at the same bitrate. The average DMOS is 0.1 in favor of our method.

The results of our evaluation are promising. They show that the approach works well in some situations and it could be further improved in others. However, to apply it in practice we need to devise solutions for the limitations of our approach.

Our approach relies on a fixation map to prioritize the quality of some areas and decrease it in others. It is non-trivial to acquire eye-tracking data for new images on the fly. However, several saliency estimation methods have been proposed in the literature [Jia+15; KTB14; KAB15]. Incorporating a computational model for saliency prediction would make the approach practical.

We currently consider the ideal parameters for our method, based on the results of the crowdsourcing study. To apply the method, we would need a content-based strategy for choosing the best parameters. The performance of our approach is clearly dependent on image content. Images with well defined regions of interest, such as people, faces or objects perform very well, e.g., examples 1–3 in Figure 2.5. Nonetheless, images without a clear focus can be rated favorably as well, see the last image in Figure 2.5. Thus, analyzing the sizes of the foreground ROIs in the input images could help, and so could some form of content-based clustering.

Saliency driven coding is open for further improvement. Contrary to the findings of previous works [BS02; BS03], our user studies conclusively show that variable coding works across multiple bitrates. In some cases, our implementation prototype gives excellent results reaching over 50% relative bitrate savings, as shown in Figure 2.8. However, in other cases the best improvement is only marginal. Recent works have substantiated an even more pronounced gain in bitrate savings by using a related variable coding strategy for foveated video coding [Wie+20]. In the image domain, further research is required to improve our understanding of what makes images suitable for variable coding and how to best optimize the coding strategy.

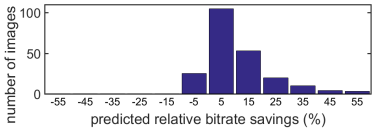


Figure 2.8: Predicted relative bitrate savings (%) for the best parameters with respect to DMOS in the VAR–STD experiment for all bitrate versions of each image. Positive percentage bins means our saliency-driven JPEG approach is better when compared to the standard JPEG at the same physical bitrate. This equates to an average of 11% bitrate savings.

3 Quality Assessment Datasets

To date the community surrounding quality assessment of digital media has developed many datasets that serve as a basis for research in the field. Most of them are designed with a specific purpose in mind, resulting in particular choices in the creation process that render them more or less useful for the problem of quality prediction of videos in-the-wild.

In this chapter we first identify important characteristics that distinguish datasets from each other and, moreover, heavily influence their applicability to different use-cases. We then summarize the state-of-the-art, describing the most relevant related works with the perspective of dataset characteristics in mind. Then, we present KoNViD-1k and KonVid-150k, the two datasets created as a solution to shortcomings by existing works specifically with respect to VQA in-the-wild. Both datasets are described in detail, discussing the video sources, additional information about the dataset creation process, and particularities of the annotation processes. The chapter concludes with an in-depth evaluation and quantitative comparison of the contributed datasets in relation to the most relevant related works. Here, we show that the proposed datasets set themselves apart from existing works in terms of diversity while adhering to established annotation quality standards.

This chapter contains and extends material from the following publications. Please refer to Section 1.4 for the contribution clarification.

[Hos+17] Vlad Hosu, **Franz Hahn**, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 'The Konstanz natural video database (KoNViD-1k)'. In: *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2017, pp. 1–6

[Göt+21] **Franz Götz-Hahn**, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 'KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild'. In: *arXiv preprint arXiv:1912.07966* (2021)

3.1 Dataset Characteristics

There are four key distinguishing characteristics that divide the field of image and video quality assessment datasets which are governed by decisions made by their creators. The first distinguishing factor that heavily influences the use of a dataset is the source of stimuli. The second characteristic is the way the videos are then potentially transcoded or otherwise altered, which distinguishes synthetic datasets from authentic ones. The subjective experiment is a third factor, where annotations obtained in a lab environment is differentiated from those obtained via crowdsourcing. Finally, the fourth distinguishing factor comes with the decision for the number of annotations per item in the dataset. We will cover the characteristics differentiating the wide variety of relevant related works separately.

3.1.1 Data Sources

Three types of approaches can be differentiated when looking at the way image and video material has been obtained for the purpose of creating image and video quality datasets.

The first is using self-recorded videos, with the most notable datasets employing this technique coming from the visual cognition research group at the University of Helsinki. In 2013 they released their Camera Image Database [Vir+14] (CID2013) that consists of photographs taken by the authors of the database using a variety of capturing devices. In total they used 79 devices, ranging from digital single-lens reflex cameras to mobile phones, to record 480 images of 8 scenes. For the video domain the same group introduced the Camera Video Database [Nuu+16] in 2014 (CVD2014), consisting of videos taken using 78 different capturing devices, again ranging from low-quality mobile phone cameras to high-quality digital single-lens reflex cameras. For this video dataset, the group recorded three out of a total of five scenes with each of the cameras in the pool, yielding a total of 234 videos. In a similar effort, the Laboratory for Image & Video Engineering (LIVE) at the University of Texas has previously used self-recording for their LIVE In the Wild Image Quality Challenge Database [GB15] (LIVE-itw) and LIVE-Qualcomm Mobile In-Capture Video Quality Database [Gha+17] (LIVE-Qualcomm). For the former, 1,162 unique pictures were captured by the group using many different cameras, with a majority being mobile phones. The latter comprises 208 videos capturing 54 scenes using 8 different mobile phone capturing devices. Here a particular focus was to ensure that certain in-capture distortions such as shakiness or underexposure were present in the videos. Evidently, the self-recording approach has the benefit of having full control over the recording equipment and recording settings, as well as knowing all the details of the processing pipeline. At the same time one is limited to the recording equipment at hand, and it cannot be expected to have multiple dozen different capturing devices available. Additionally, as is also evident in the few examples, it will usually lead to a very limited variety of scenes or contents.

The second approach for obtaining source material for databases aimed at image or video quality assessment is relying on third-parties, usually

by sampling from (multiple) existing datasets. In the image domain, the KODAK Photo CD dataset of 25 color images has been the source for the popular Tampere Image Database releases in 2008 [Pon+09] and 2013 [Pon+15] (TID2008, TID2013) from the Computational Imaging Group at Tampere University, as well as the LIVE Image Quality Assessment Database Release 2 [SSB06] from the University of Texas in 2006, the latter of which also included images from other sources. Similarly, the Multiply Distorted Image Database [SZL17] (MDID) released in 2015 sampled from the KODAK Photo CD dataset, as well as ImageNet [Den+09], the USC-SIPI Image Database, and others. In all of these cases the sampling databases were the starting point to derive variants of the source images that contained some (or multiple) forms of distortions. The video quality domain has seen less re-using of the same source material. Several datasets from the Images and Video-communcations (IVC) team at the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN) use videos supplied by the EURO1080 or Svergies Television AB television networks, such as the IVC 1080i dataset [PPL08b] consisting of high definition videos at 1080p. Others, like the Video Quality Expert Group High Definition Television dataset (VQEG-HDTV) and the IVP Subjective Quality Video Database [Zha+11] (IVP) from the Image and Video Processing group at the Chinese University of Hong Kong use their unique sources for videos. Adopting material from third-party sources reduces the front-loaded work in creating a dataset and allows a level of either specificity or diversity of the material that might otherwise not be attainable. This approach has mostly been used in conjunction with further processing of the image or video material, and most datasets in this category are therefore also confined to few unique scenes.

The KODAK Photo CD dataset and the USC-SIPI Image Database can be found at `https://www.math.purdue.edu/~lucier/PHOTO_CD/` and `http://sipi.usc.edu/database/`, respectively.

In the last five years a third way of sourcing images and videos for quality assessment datasets has been introduced. In 2017, the Multimedia Signal Processing Group at the University of Konstanz pioneered the use of internet videos with appropriate copyright sampled from the multimedia sharing platform Flickr as a source of videos for their Konstanz Natural Video Database [Hos+17] (KoNViD-1k), and subsequently used similar approaches for the creation of a variety of image and video quality datasets, such as KonIQ-10k [Hos+20b], consisting of 10,073 images sampled from the Yahoo Flickr Creative Commons 100 Million dataset [Tho+16] (YFCC100m), and the family of Konstanz Artificially Distorted Image quality Dataset [LHS19] (KADID-10k) and the Konstanz Artificially Distorted Image quality Set [LHS19] (KADIS-700k) in 2019 for the image domain, and KonVid-150k [Göt+21] in 2021 in the video domain. In the case of KADID-10k and KADIS-700k the popular international photo sharing platform Pixabay was used as a source, sampling 81 and 140,000 images from a set of 654,706 images for the two, respectively. From the over 150,000 videos within YFCC100m, KoNViD-1k contains 1,200 items that were sampled fairly, meaning that an attempt was made to select videos such that distributions of different quality indicators were as close to uniform as possible. Although the subsampling of YFCC100m for KoNViD-1k and KonIQ-10k is similar to re-using third-party datasets the difference lies in that YFCC100m is entirely online, rather than a centralized set of items. Finally, KonVid-150k is the group's newest video quality dataset with over 150,000 videos that was sampled directly from Flickr, which is also the source for YFCC100m. In 2020 the LIVE group also released the LIVE Video Quality Challenge

`https://www.flickr.com`

The described University of Konstanz video datasets do not constitute related work in the strictest sense, as they are part of the contributions of this thesis. However, for the purpose of explaining the different dataset characteristics we include them here.

`https://www.pixabay.com`

Database [SB18] (LIVE-VQC), by requesting videos from families, friends, and other peers of the team and obtaining a wide variety of over 1,000 videos. These were subsequently sampled into a set of 585 unique videos recorded by 101 different devices. This approach is technically neither sampling from internet videos, but also not self-recording in the form as we encountered before. It should be more understood as "in between" self-recording and using videos uploaded to video sharing platforms and we list it within the latter group, because the nature of the process seems to align more closely with the other datasets mentioned in this category.

3.1.2 Data Processing

There are two fundamentally diverging approaches to processing the obtained videos that influence the way an image or video quality assessment dataset will be used.

Classical datasets, as described above, include few source stimuli to begin with, which is not by happenstance. The traditional approach to curate a dataset for the purpose of quality assessment intends a processing step for the items that introduces particular distortions. Conventionally, a set of 4 to 25 distortions are applied to all source stimuli, where the introduced artifacts either correspond to those induced by elements in the processing pipeline or account for particular peculiarities of the HVS. Examples for the former are compression artifacts introduced by encoding an image or a video via JPEG or H.264, respectively, as well as introducing noise that resembles the type of noise that could be introduced by different exposure or ISO speed settings. Properties of the HVS that are accounted for might be its particular spatial frequency sensitivity by considering spatially correlated noise, or altering the color saturation to account for color sensitivity of the HVS. Datasets that take this approach are abundant, as it is the most prevalent method of creating quality assessment datasets. An inexhaustive list of example datasets in this category for the image quality domain are the TID2008, TID2013, and KADID-10k datasets, as they include the broadest sets of distortions within them, ranging from JPEG and JPEG2000 compression artifacts to various types of artifacts related to different types of noise, blur, and other degradations. Moreover, MDID deserves a special mention within this category, as images within this dataset have up to 4 different distortions applied to them at the same time. This approach is more realistic in the sense that images taken and shared by consumers tend to include artifacts from multiple degradation sources. In the video domain practically all datasets prior to 2017 fall into this category, with a lot of them focussing on H.264 encoding and transmission errors, such as the IVC datasets, as well as VQEG-HDTV, IVP, and others. Figure 3.1 and Figure 3.2 show examples of both frame and inter-frame artifacts, ranging from blockiness and blur to interlace and other temporal distortions.

Since 2017 a handful of datasets in both the image and video domain have been created that take a different approach, with the argument that the previous datasets suffer in ecological validity, as the degradations present in the processed stimuli are not authentic when compared to images and videos commonly consumed. Instead of introducing those types of distortion artifacts that an objective I/VQA model should be

Figure 3.1: Upscaled examples of video quality defects. Courtesy to the AGH Video Quality of Experience Team for the visualizations. https://qoe.agh.edu.pl/indicators/

able to accurately detect, these new datasets instead take items from image and video sharing platforms, so called images and videos in-the-wild. Additionally, they use strategies to maximize diversity along multiple quality dimensions, such as types of content, presence and level of distortions, as well as other quality-related indicators. KonIQ-10k, KoNViD-1k, and KonVid-150k, as well as LIVE-VQC are representative examples for the approach of ensuring ecological validity of videos contained within image and video datasets.

Figure 3.2: Upscaled examples of video quality defects. Courtesy to the AGH Video Quality of Experience Team for the visualizations. https://qoe.agh.edu.pl/indicators/

3.1.3 Subjective Annotation

Datasets in the image and video quality domain require subjective annotation of the items contained within them. The classical approach to subjective annotation is the collection of human opinion scores in fixed laboratory setups, where the images or videos are displayed on a handful of devices with fixed resolutions and strict adherence to viewing guidelines. Different variations of this legacy approach exist, where the subjects are sometimes experts in the domain of image and video quality, and sometimes naïve laypeople. Conventionally, participants in these lab studies would provide their opinions on the majority of the items of a dataset, if not on its entirety. Although this maximizes comparability of the different subjects and enables the anchoring of individuals within the trial, datasets annotated in this way are also restricted with respect to annotation diversity. Commonly, no more than 40 subjects participated in the subjective annotation of different image and video quality datasets. Up until 2016 this has been the norm, with TID2008 and TID2013 being a small exception in that they combined lab settings with strictly controlled online experiments and thereby reaching over 800 participants.

However, the increased availability and capacity of internet infrastructure has allowed recent related works to instead gather subjective human opinion scores of multimedia items in a crowdsourced setting. The LIVE group pinoneered crowdsourcing image quality annotations from over 8,100 individual subjects on AMT for their LIVE-itw dataset in 2016, which was also used to annotate LIVE-VQC (≥4,700 participants) and adopted for other datasets. For example, KonIQ-10k contains crowdsourced annotations from a total of over 1,400 workers on the Crowdflower platform, which was also used to annotate KoNViD-1k (≥600 participants), KADID-10k (≥2,200 participants), and KonVid-150k (≥1,200 participants). With the two TID datasets aside, crowdsourcing evidently reaches a more diverse pool of participants in subjective annotation experiments. Nonetheless, the online environment requires additional quality control mechanisms in order to ensure high quality results. Section 2.1.2 elaborates on this topic and Section 2.2 includes a comparison between lab and crowd experiments, showing how the latter can reproduce results from the former.

3.1.4 Vote Budget Distribution

The last factor impacting the characteristics and use of a quality assessment dataset is an emerging field of research. With a few exceptions, early works in lab environments ensured at least 25 ratings per stimulus

and distributed the budget (approximately) uniformly across all items, meaning that all participants rated most or all stimuli.

Recent works have increased the number of ratings per stimulus to above 200 with the goal of ensuring very high precision in the individual item annotations. However, given a fixed, affordable budget of annotations, one must consider the trade-off between the benefit of slightly more accurate quality scores for a small number of stimuli and the potential increase in generalizability when annotating more stimuli with fewer votes. For example, an 8-fold increase in numbers of ratings per stimulus when going from the generally accepted 25 to 200 ratings could just as well be invested in an 8-fold increase of numbers of stimuli, each rated 25 times. The increase of the precision of the experimental MOS suffers from diminishing returns as the number of raters increases. Since the precision gain per vote is highest at none or few ratings, careful considerations have to be made with respect to the distribution of annotation budgets across an unlabeled dataset. This is especially true in the wake of deep learning approaches outperforming classical methods in many computer vision tasks, as deep learning models are known to be robust to noisy labels [Rol+17] but also hungry for input data.

In 2018 a joint effort of researchers at the Department of Electrical Engineering and Computer Science at the University of California in Berkeley, OpenAI and Adobe showed that as few as 2 to 5 subjective annotations for paired comparisons of differently distorted image patches could enable machine learning models to accurately predict perceptual similarity of unseen patch pairs [Zha+18]. Inspired by this research and constrained by a fixed budget to annotate a large set of videos, KonVid-150k was the first video quality dataset designed to investigate the relationship of different vote budget distributions on the predictive performance of video quality assessment algorithms.

3.2 State of the Art

In the following we will list the most relevant state-of-the-art video quality assessment datasets. Here, we distinguish between two different classes of datasets based on some of the characteristics described in Section 3.1. *Synthetic* datasets describe those that create variants of a set of source stimuli by artificially distorting them, as they are not reflecting original authentic distortions. Conversely, *authentic* datasets contain images and videos that were obtained from an authentic environment and are unaltered, except for potentially re-encoding items at a visually near lossless level for an easier distribution in an online crowdsourcing study. Each of the discussed datasets is accompanied with a plot depicting the temporal and spatial information, as well as the color and contrast of the videos contained within them. For further information about the implementation of these measurements please refer to Section 3.3.1.

3.2.1 Synthetic Datasets

In the group of synthetic datasets we consider the most recent video quality assessment datasets as most relevant.

MCL-V

MCL-V was created with the specific purpose of covering scenes representative of common video applications including distortions caused by video up-scaling. 12 original videos were selected from various sources, based on the criterion that they should be professionally recorded and of 1920× 1080 resolution. They cover a variety of contents, ranging from animations, such as the "Big Buck Bunny" and "Fox Bird" scenes, to aerial videos, such as the "Old Town Cross" and "BQ Terrace" scenes, to videos of humans in different numbers and situations.

Each of the source videos was artificially distorted using two different approaches at four levels of degradation, each. The first distortion was introduced by encoding the source using x264 [MV06] with a specific reduced target bitrate that was subjectively selected to ensure distinguishability between the different variants. For this distortion the original scale was retained. For the video up-scaling distortion the sources were first down-sampled to a resolution of 1280×720 using the Lanczos algorithm, followed by x264 encoding in a similar as in the first case, and then up-sampling the re-encoded version using bilinear interpolation.

A total of 96 videos of 6 seconds were obtained, which is among the smallest dataset sizes in the VQA field. The nature and number of distortions is also narrow, rendering it a niche VQA dataset with respect to its utility for in-the-wild video quality prediction. Figure 3.3 shows the average spatial and temporal information for all recordings of the different scenes in the upper scatter plot, while the lower scatter plot compares the average color and contrast of the same videos The with different variations of the same source (plotted in the same color and marking) form small clusters, underlining the small coverage of different spatial and temporal information, as well as color and contrast. However, the latter is to be expected, given that the nature of the applied distortions conceptually do not alter color and contrast information significantly.

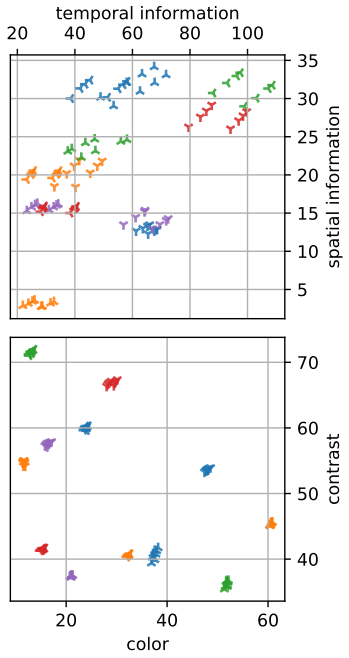


Figure 3.3: A selection of frame and video features for the different videos contained in MCL-V. The colors denote the individual scenes.

CVD 2014

CVD 2014 was created with the specific purpose of capturing the same scenes using a wide variety of cameras, as the name "Camera Video Dataset" suggests. The resulting videos should give a good representation of the types of degradations inherent to the capturing device.

Based on this assumption five scenes were selected that represent different spatial and temporal activities, such as the outdoor "City" scene containing more camera movement and objects in the scene moving at a higher rate, when compared to other scenes, while the indoor "Talking Head" scene has no camera movement and the single object in view barely moves throughout the video. Using 78 different cameras available to the authors, ranging from low-quality camera phones to high-quality digital single-lens reflex cameras. The video sequences of each of the five different scenes were captured one at a time using the different capturing devices, although not every scene was captured using every camera. This also means that the recordings of different cameras show slightly different contents for each scene.

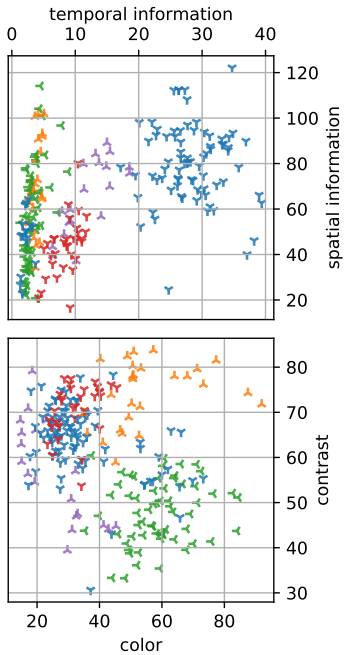


Figure 3.4: A selection of frame and video features for the different videos contained in CVD2014. The colors denote the individual scenes.

A total of 234 videos of 10-25 seconds were recorded, meaning that each camera was used to record three scenes, where each recording has its own unique mixture of in-capture distortions. In Figure 3.4 we can see a much broader distribution of spatial and temporal information, color, and contrast. Again, different variations of the same scene are plotted in the same color and marking, that form clusters that are less separable and of a larger size than in MCL-V. This is to be expected, as each stimulus in CVD2014 is a unique video rather than an alteration of a source video. However, the dataset only covers five unique scenes, which is the smallest number of unique scenes among all VQA datasets.

LIVE Qualcomm

LIVE Qualcomm was created with the purpose of balancing the level of diversity of in-capture video distortions caused by different mobile phone cameras with sufficient numbers of unique contents captured per recording device. In spirit it follows a similar approach to CVD2014 with a few key differences.

With 8 cameras forming the set of capturing devices, it considers roughly 10% of camera diversity compared to CVD2014. However, the dataset provides between 16 and 33 unique scenes for each camera, instead of the 5 in its Finnish cousin. Moreover, for the recording of an individual scene 4 different mobile phones were installed in a rig that allowed synchronized recording, which means that every unique content is only recorded using half of the cameras. Based on independent testing different arrangements of mobile phones were created that that considered the variety of expected quality and similarity in the field-of-view of the recording devices. Across multiple scenes each arrangement, then, yielded nearly-identical video material for the 4 cameras within it, with respect to scene content. The major differences in the video material between the cameras are caused by the hardware, ensuring the comparability of any pair of two cameras.

Each individual recording was carried out in such a way that the resulting videos would contain at least one of the distortions under consideration. These 6 distortions are noise and blockiness, color defects, over-exposure and under-exposure, degradations caused by the autofocus software, overall lack of sharpness, and camera shake during recording. 54 different scenes were recorded in this way, bringing the total video count to 208. Figure 3.5 depicts the overall broad distribution of temporal information and spatial information, as well as color and contrast for videos in the LIVE-Qualcomm dataset.

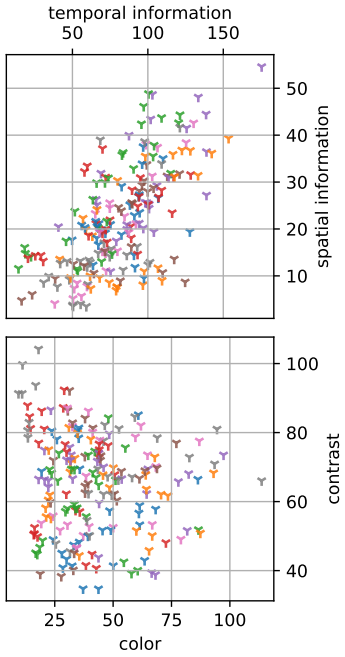


Figure 3.5: A selection of frame and video features for the different videos contained in LIVE-Qualcomm.

3.2.2 Authentic Datasets

LIVE-VQC

LIVE-VQC is the only VQA dataset that falls within our definition of authentic, meaning that the videos contained within were neither recorded for the purpose of VQA, nor post-processed to include particular distortions. It was created with the purpose of capturing the diversity of videos in-the-wild without any emphasize on particular distortions, contents, or recording devices. In doing so, the authors hoped to represent

the wide variety of authentic distortions that real world consumers of video material might encounter.

Although the videos were recorded by peers of the authors, their directors were not instructed that the videos would be used for the study of video quality. 585 videos of 10 second duration were sampled of over 1,000, by removing redundant, disturbing, and otherwise ill-fitting content, while seeking to preserve continuity. A total of 101 different recording devices were used in the process of capturing the video material, with the resulting material also spanning a wide variety of resolutions and including both landscape as well as portrait orientations. The material can broadly be grouped into 1080p, 720p, 404p portrait, and videos of other resolutions.

Figure 3.6 shows that the distribution of video characteristics for LIVE-VQC is broader than the synthetic cases. This is can be attributed to each video being a unique scene in combination with the variety of recording devices.

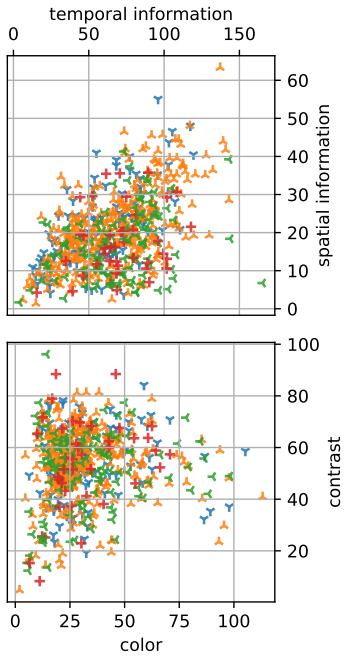


Figure 3.6: A selection of frame and video features for the different videos contained in LIVE-VQC.

3.3 KoNViD-1k

Summarizing the previous Section, current synthetic VQA datasets contain only a small number of video sequences with little content diversity, thus offering limited support for designing and evaluating no-reference VQA methods for videos in-the-wild effectively and fairly. Additionally, these databases were mostly designed to include only artificially distorted video sequences to simulate quality loss in compression, transmission, and other parts of the video processing and distribution pipeline. Some databases capture imagery with a variety of cameras to encompass authentic video acquisition distortions, however, with content restricted to a small number of physical scenes. LIVE-VQC as the only authentic VQA dataset covers a larger number of unique contents, when compared to its synthetic cousins. However, the videos contained within it were obtained from peers, friends, and family of the authors and it is, therefore, not clear how representative of videos in-the-wild the material is.

Previous work into the comparison of IQA and VQA datasets has proposed several quantitative criteria for source material, test conditions, and subjective ratings, applying them to 27 image and video databases [Win12]. Here, most collections were found to be unsatisfactory in terms of content range and uniformity of the described criteria. Only few databases showed good uniformity for the test conditions (i.e. image or video quality), but not over the whole quality range. Furthermore, the distortion variety was found lacking in most databases covering mainly compression and transmission, but not the many other types of natural distortions found “in the wild".

To overcome these limitations we introduce the Konstanz natural video database (KoNViD-1k), a large publicly available database of video sequences based on YFCC100m. KoNViD-1k is a first step towards an even larger database containing a diverse set of video content, enriched with meta data both from YFCC100m and additional video stream level information not in YFCC100m. In this Section we will report the filtering mechanisms and sampling procedures necessary to construct

high-quality VQA databases of this kind, focusing on their usefulness in a variety of applications. Basically, the dataset is introduced with the following goal in mind:

> Curation of a large set of videos sampled from public repositories that cover the full range of quality characteristics, while also being small enough to perform online crowd-based quality annotation.

Video authenticity and distortion diversity are the two focal points of this dataset. That is to say the recordings contained within it should be representative of those uploaded, shared, and consumed on video sharing platforms, and their distortions should be of an authentic nature and distributed as broadly and uniformly as possible. For this reason, the starting point for KoNViD-1k was the well-known public YFCC100m, comprised of 793,436 Creative Commons (CC) licensed videos. From this dataset we first selected videos based on practical requirements, such that the videos:

- Were still available for download at the time
- Played at more than 15 frames per second (FPS)
- Lasted longer than 8 seconds
- Did not have a "No Derivative Works" CC attribute
- Had a resolution of at least 960×540
- Were in landscape orientation

Applying these filtering steps left a subset of 144,889 videos, which were downloaded from the Flickr servers for further sampling. Due to Flickr's constraints of 30 seconds duration for uploaded videos, those running longer than this maximum duration had been cropped to fit this criteria.

3.3.1 Quality Indicators

In order to address the distortion diversity we rely on attributes such as the previously used temporal and spatial information, or assessment of color and contrast of the videos. These measures can serve as proxies to content diversity of video datasets [Win12]. Guided by previous research, we chose six video attributes and used them not only for the analysis of the resulting dataset, but for its creation as well. For each attribute we relied on the best-performing technique available in the literature (to the best of our knowledge). Since the computational complexity of some of the selected metrics is high relative to the amount of frames to be analysed, we created cropped and scaled versions of the videos to run them on. All attributes, except the one related to color, were computed on grayscale frames, and in some cases, only a subset of frames were processed.

Blur

Bluriness of individual frames was assessed by the cumulative probability of blur detection (CPBD) metric [NK11].

The method considers 64x64 blocks of the input image at a time, further processing those where more than 0.2% (≥820) of pixels are part of a horizontal edge. For those edge blocks the local contrast C is computed and used to derive the just-noticeable-blur (JNB) edge width $w_{\text{JNB}}(e_i)$ and edge width $w(e_i)$ of all pixels e_i belonging to an edge within the block. If $w(e_i) > w_{\text{JNB}}(e_i)$ the blur is considered to be detected at the edge, as the width of the edge is bigger than the JNB edge width. Intuitively, the metric measures whether blur at edges within the image will be detected by humans, based on the principle that the level of blur correlates with the width of an edge.

In order to apply the computationally complex CPBD metric to videos we averaged the CPBD values of a 1 fps version of the video.

Color

Color of individual frames was assessed by Hasler and Suesstrunk's colorfulness metric for natural images [HS03].

With the RGB channels of a frame as matrices R, G, and B, two matrices $\text{rg} = \text{R} - \text{G}$ and $\text{yb} = \frac{1}{2}(\text{R} + \text{G}) - \text{B}$ are computed. Then, the frame colorfulness is calculated as

$$\sqrt{\sigma_{\text{rg}}^2 + \sigma_{\text{yb}}^2} + \frac{3}{10}\sqrt{\mu_{\text{rg}}^2 + \mu_{\text{yb}}^2},$$

where $\sigma_{\cdot}^2$ and $\mu_{\cdot}$ denote the variance and mean of the values in their respective matrices. This metric has been shown to correlate to about 94% with subjective evaluations of colorfulness in psychophysical experiments for images depicting natural scenes.

For video level colorfulness, the metric's output is averaged over all frames of the input video.

Contrast

Contrast of individual frames was measured simply by the standard deviation of pixel grayscale intensities, which is a common approach [Pel90].

The average frame-level standard deviation then gave the video level contrast metric.

Spatial Perceptual Information

Spatial perceptual information (SI) is measured by applying a Sobel filter to each video frame F_n at time n to extract the gradient magnitude for each pixel. Then, the standard deviation over pixel dimensions is computed for each frame. In a final step, the maximum value among all frame-wise standard deviations is the SI for the video:

$$\text{SI} = \max_{\text{time}}\{\text{std}_{\text{space}}[\text{Sobel}(F_n)]\} \tag{3.1}$$

SI is an indicator of edge energy [Wol97] and is commonly understood to correlate with image complexity [YW13]. The clearer an edge the bigger

the spread in pixel values of the Sobel filter output and, thus, the larger the standard deviation.

Temporal Perceptual Information

Temporal perceptual information (TI) is computed using the motion difference frame $M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$, where $F_n(i,j)$ denotes the pixel values at (i,j) in frame F at time n. Similar to SI, for a given video the TI is obtained by combuting the standard deviation over pixel dimensions for each motion difference frame M_n and subsequently taking the maximum standard deviations across all n:

$$\mathrm{TI} = \max_{\mathrm{time}}\{\mathrm{std}_{\mathrm{space}}[M_n(i,j)]\} \tag{3.2}$$

Intuitively, the more objects move between adjacent frames, the higher differences in M_n and, therefore, the higher the standard deviations, resulting in higher values of the TI measure.

Video quality

Video quality was measured using the Natural Image Quality Evaluator [MSB13] (NIQE) for individual frames as a proxy to assess video level quality by computing the mean NIQE value of all frames. This method (VNIQE) does not require knowledge of distortion types nor quality ratings for training. The NIQE of a frame is simply the distance between certain ideal features and a particular frame's features. It can be interpreted as the degree of frame-level deviation from naturalness, according to the NIQE model.

3.3.2 Filtering

Based on these six attributes, videos were selected such that they are suitable for VQA. We removed videos depicting non-natural scenes like screen recordings or stop-motion sequences, as well as overly dark or bright videos. Most of these situations were encountered at extremes of the attribute values. For instance, very low TI videos were often found to be screen-text recordings due to little change between frames, while dark videos have low contrast and SI. Uniformly and brightly colored videos have a high level of colorfulness.

Therefore, we decided to remove videos that have extreme attribute values. The filtering thresholds were empirically chosen based on a qualitative inspection such that most of the filtered videos show obvious artificial content. The selected thresholds are available in Table 3.1. Upon further inspection we noticed that stop-motion videos do not reside on the extremes of a sole attribute, so we devised another approach to detect and remove them. Here, we relied on two observations; namely (1) stop-motion videos show periodical changes in TI, and (2) a high percentage of consecutive frames show no change at all. We used the difference of TI of consecutive frames to quantify both factors.

Table 3.1: Attribute thresholds for filtering outlier videos.

name	low	high
blur	0.05	0.88
color	4.37	123.00
contrast	7.51	97.48
SI	7.70	187.76
TI	3.07	56.81
VNIQE	3.58	23.08

With respect to the periodicity, we found local maxima of TI with an inter-peak distance greater than 1 second. By computing the distances d_i between consecutive local maxima together with their mean μ and variance σ^2, we extracted a measure of regularity as $R = \mu/\sigma^2$. If the variance σ^2 is low, then the regularity R is high. The mean of the distances is the average spacing between peaks. If the peaks are further apart the formula tolerates smaller changes in the peak timing, and R is higher. $R > 1/300$ was found to be a good indicator for detecting stop-motion videos. Furthermore, if more than 30% of consecutive frames showed no difference in TI, we assumed to be dealing with a non-natural video sequence.

These criteria further eliminated about 500 videos from our video collection, leading to a dataset of 124,865 videos, which we will call KoNViD-125k. The distributions of the normalized attribute values are displayed in Figure 3.7.

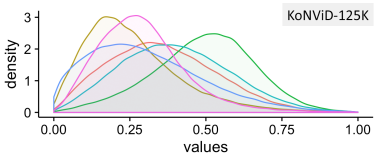


Figure 3.7: The distribution of values for the six quality-related attributes on KoNViD-125k.

3.3.3 Sampling

Initially our goal was to sample and annotate a set of 10,000 diverse videos from KoNViD-125k. As an intermediate step, we wanted to evaluate the impact of sampling procedures on the utility of VQA datasets for the prediction of quality of videos in-the-wild.

We devised a "fair-sampling" strategy with which we generated a subset of 10,000 videos. From the resulting set we took a random collection of 1,200 videos, which forms the KoNViD-1k dataset. We performed subjective studies to assess the visual quality of the videos in this dataset. As a consequence, we arrived at a better understanding of the diversity of the 10,000 fair-sampled videos, and the efficiency of our approach.

A "fair sampling" mechanism should produce a broader diversity of video properties than a random sampling mechanism. Our videos are represented as points in a 6-dimensional attribute space. We can think of the KoNViD-125k collection as a sample of $M = 124,865$ points of a multivariate and approximately normal distribution, for which random subsampling of 10,000 items would yield a subset with a similar normal distribution. The sampling procedure is engineered to give a more uniform 6-dimensional sample distribution.

Each attribute relates to a particular subjective property. Most videos having extreme values for one or more attributes show severe quality degradations. Random sampling will sample these "unusual" videos at a lower rate, because they occur less frequently. Nonetheless, they are as important as the more common videos with average attribute values. A preliminary qualitative inspection of several hundred videos (both randomly and fairly sampled) suggested that our strategy creates a balanced mix of videos. The dataset shaping approach described by Winkler [Win12] does not apply in our case, and thus we're forced to devise our own strategy that works on the joint distribution

With respect to the sampling procedure, the method of Vonikakis et al. [VSW16] can ensure a uniform distribution for each attribute independently. However, we are also interested in sampling videos with joint

distortions, having extreme values in several attributes simultaneously. Thus, we applied a different approach.

Note that our attributes are correlated as shown in Figure 3.8. For instance, contrast and spatial information (SI) have a 0.62 correlation, whereas VNIQE and SI have a negative correlation of −0.43. Thus, as a preprocessing step we applied a principal component analysis (PCA), that shows that 37.1%, 57.7%, 73.4%, 85.8%, and 95.2% of the variance of the data is explained by the 1st to 5th principal components, respectively. We decorrelated the attributes by taking five components, maintaining all but 5% of the signal energy.

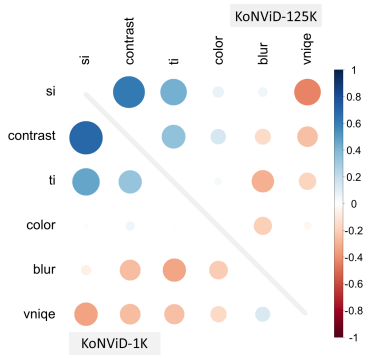


Figure 3.8: Correlations between attribute values. Correlations considering KoNViD-125k dataset are above the main diagonal, and for KoNViD-1k below. Larger and darker circles represent a stronger absolute correlation coefficient. Red hues encode negative correlations, whereas blues are positive.

One way to solve the subsampling problem is to design a sampling method that favors videos that are part of a low-density region in the attribute space as follows.

Let $\rho : \mathbb{R}^5 \rightarrow \mathbb{R}$ denote the probability density function of the 5-dimensional PCA attribute vectors v_i for natural videos, estimated from a video collection such as our KoNViD-125k. We used a k-NN method to estimate the density in the neighbourhood of a video in the PCA attribute space. For the i-th video attribute v_i we first found its k-th nearest neighbour $v_{k(i)}$ (we chose $k = 500$) and computed the distance $||v_i - v_{k(i)}||$. This distance is the smallest radius of a hypersphere around v_i that encompasses k videos. The density $\rho(v_i)$ was then taken inversely proportional to the volume $||v_i - v_{k(i)}||^n$, with $n = 5$, the number of dimensions.

The task then is to assign suitable sampling probabilities $p_i, i = 1, \ldots, M$ for the set of videos in KoNViD-125k having attribute vectors $v_i, i = 1, \ldots, M$. Then, 10,000 subsamples will be drawn with corresponding probabilities p_i and without replacement. A natural choice is to set the probabilities p_i proportional to the inverse of the attribute densities, $1/\rho(v_i)$, at the corresponding attribute vectors, v_i, for all $i = 1, \ldots, M$.

From the 10,000 fairly sampled videos, we randomly subsampled 1,200 to form the KoNViD-1k dataset. A Gaussian kernel density estimation along each dimension shows the distribution of the samples, which is reproduced in Figure 3.9. We can see that attribute values are more evenly spread in most dimensions, except for the temporal information. This might be caused by local correlations in the data that have not been removed by PCA.

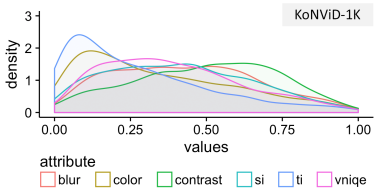


Figure 3.9: The distribution of values for the six quality-related attributes on KoNViD-1k.

3.3.4 Subjective Annotation

Quality control is a key component when designing an experiment for the crowd, so we considered how to set up our procedure carefully. Initially, each worker was instructed according to VQEG recommendations [ITU04], which were modified to fit our single stimulus presentation technique. In the instructions workers were informed about types of degradation (e.g., related to motion, color, brightness, and details) and about how they would be asked to evaluate the overall quality of each video. Next, examples of videos with “Good", “Fair" and “Bad" quality were displayed for anchoring and workers were instructed on the steps required to rate a video. Initially, a button was displayed below the video, which started playing the video muted. Once the video was finished

playing a rating scale was displayed and workers had to select one of five categories to proceed. Only if a worker had watched and rated all 10 videos on a page, could he proceed to the next.

In order to control for the quality of workers' performance, it is common to use gold standard questions, which is also a feature provided by CrowdFlower. Since there is no ground truth for our dataset, we have devised a plan to filter unreliable workers. We randomly sampled a subset of 100 videos from our pilot for an uncontrolled (no gold standard questions) crowdsourcing experiment of 50 ACR scores per video. From these, we computed the 95% confidence intervals of the MOS values to select those videos with a size of the confidence interval smaller than 0.5 on the ACR scale. The resulting 65 highly agreed upon videos and their MOS values were used as the ground truth for test questions for the evaluation of the entire data set.

We employed the same setup for KoNViD-1k as above, with the addition of the test questions for quality control. We considered the rounded MOS ±1 as eligible answers to these test questions. Workers that fell below 70% accuracy on test questions were removed from the experiment along with the data they had generated. Moreover, we only allowed CrowdFlower workers of Level 1 and above (more than 70% accuracy in all previous tasks) to participate in our experiment. Due to the fixed number of 65 test questions, workers could rate at most 550 videos in batches of 10 per page (10 questions for the quiz page, 55 batches with one test question each), with a maximum of 550 ratings per worker, based on the number of test questions.

3.4 KonVid-150k

With KoNViD-1k we created a fairly sampled, subjectively annotated video database, showing authentic distortions, consisting of diverse videos published by different users using various cameras with different shooting skills. It overcomes some drawbacks of previous works. Namely, fair sampling from the large dataset YFCC100m ensured a high diversity in content, and the distortions are authentic in the sense that they were not synthetically introduced, as is the case in artificially degraded datasets. Nonetheless, with a size of 1200 videos the overall diversity in content is still limited and Flickr's own encoding resulted in a noticeable degradation in quality and introduction of artifact compared to source material. Consequently, we set out to resolve these two key issues with the creation of KonVid-150k, an ecologically valid VQA dataset over 100 times larger than KoNViD-1k. Moreover, we created it with the goal of analyzing the impact of vote distributions on the performance of VQA models, which is an insufficiently investigates topic, as explained in Section 3.1.4. Basically, the dataset is introduced with the following goal in mind:

> Curation of an extremely large set of videos obtained from public repositories that reflect a wide range of quality and content characteristics, which is non-uniformly annotated to facilitate research into vote budget distribution, as well as the development of VQA algorithms at scale.

Figure 3.10: Comparison of the quality of the original (center) to the version Flickr provides (bottom) and our transcoded version (top).

For KonVid-150k our main objective was to create a video dataset that covers a wide variety of contents and quality-levels as commonly available on video sharing websites. Therefore, we took a similar approach to collect our data as was done for KoNViD-1k, with an additional step to improve the quality of the videos. In KoNViD-1k all collected videos had been transcoded by Flickr, to reduce their bandwidth requirements and standardizing them for playback. Consequently, noticeable degradation was introduced relative to the original uploads. Flickr allows the uploading of video files of most codec and container combinations, resolutions, and durations. However, they re-encode the uploaded videos to common resolutions such as HD, Full HD, strongly compressing them. Instead of using the versions provided by Flickr, we use the Flickr API, which allows access to metadata that links to the original, raw uploads. As these raw uploads are often very large and come in many different formats, they cannot directly be used for crowdsourcing. Therefore, we proceeded as follows. We downloaded authentic raw videos that had an aspect ratio of 16:9 and resolution higher than 960×540 pixels. Then we rescaled them to 960×540, if necessary, and extracted the middle five seconds. Finally, we re-encoded them using FFmpeg at a constant rate factor of 23, which balances visual quality and file size. The resulting files have an average size of 1.23 megabytes.

This resulted in fewer encoding artifacts while keeping the file size manageable for distribution in a crowdsourcing study with an average of 1.23 megabytes per video. Figure 3.10 is a visual comparison of the differences, showing a small crop of a frame of the originally uploaded video together with the two re-encodings offered by Flickr and our own version. Compression artifacts are clearly visible in the Flickr re-encoded version, whereas our re-encoding is very similar to the original.

For each video, we extracted meta-information that identifies the original encoding, including the codec and the bit-rate. Furthermore, we collected social-network attributes such as the number of views and likes and publication dates that indicate the popularity of videos. In total, this collection amounts to 153,841 videos. We believe that all the additional measures we have taken to refine our dataset significantly improved its ecological validity, and thus the performance of VQA methods trained on it in the future.

3.4.1 Subjective Annotation

We annotated all 153,841 videos for quality in a crowdsourced setting on Figure Eight. First, each participant was presented with instructions according to VQEG recommendations [ITU04], which were modified to our requirements. Here, participants were introduced to the task and provided with information about types of degradation, e.g., poor levels of detail, inconsistencies in color and brightness, or imperfections in motion. Next, we provided examples of videos of a variety of quality levels with a brief description of identifiable flaws and instructed the reader on the workflow of rating videos, which is illustrated in Figure 3.11. Finally, we informed participants about ongoing hidden test questions that were presented throughout the experiment, as well as the minimum resolution requirement that enabled them to continue participating in the experiment. This was checked before the playback of any video.

`http://www.figure-eight.com/` (now `https://appen.com/`)

Figure 3.11: Illustration of the crowdsourcing video playback workflow. A worker is first presented with a white box of 960x540 pixels. Upon clicking the box, the video plays in its place. Playback controls are disabled and hidden. Upon finishing, the video is hidden and replaced with a white box that informs the participant to rate the quality on the Absolute Category Rating (ACR) scale shown below. The rating scale is only shown upon completion of video playback.

During the actual annotation procedure, for each stimulus, workers were first presented with a white-box of the size of the video that also functioned as a play button. Then, the video was shown in its place with the playback controls hidden and deactivated. After playback finished, it was hidden, and the rating scale was revealed below it. This setup ensured that neither the first nor the last still frame of the video were influencing the worker's rating, and no preemptive rating could be performed before the entirety of the video had been seen. An option to replay the video was not provided so as to improve attentiveness and ensure that the obtained score is the intuitive response from the worker. Additionally, playback of any other video on the page was disabled until the currently playing video was finished, in order to better control viewing behavior and discourage unreliable or random answers.

According to Figure Eight's design concept, crowd workers submit batches of multiple ratings in so-called pages. Each page has a fixed batch size of rows, where each row conventionally represents a single item. Due to constraints on the number of rows allowed per study, we grouped 15 stimuli by random selection into each row, with a page size of ten rows per page, totaling to 150 videos per batch, respectively page. Moreover, the design concept intends a two-stage testing process, where workers are first presented with a quiz of test questions followed by subsequent pages where test questions are randomly inserted into the data acquisition process. Test questions are not distinguishable from conventional annotation items.

In our implementation, illustrated in Figure 3.12, we interspersed three test videos with twelve videos randomly sampled from the dataset in each row with test questions. The test videos were sampled from hand-picked set of videos, which in one part was made up of very high-quality videos obtained from Pixabay and in another of heavily degraded versions of them. Therefore, we defined the ground truth quality of each test video as either excellent or bad, respectively. We performed a confirmation study to ensure that the perceived quality of these videos was rated at the very top or bottom ends of the 5-point ACR scale.

`http://pixabay.com`

In the second stage, after the quiz, consisting of only test rows, workers annotated 150 videos in 10 rows per page. On each page, we included one further test row at a random position.

Participants had to retain at least 70% accuracy on test questions through-

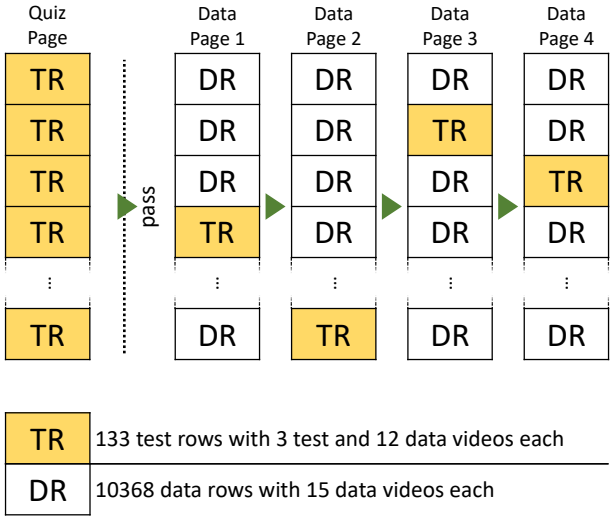


Figure 3.12: Simplified workflow diagram of the experiment. A worker is first presented with a quiz page of test rows (TR, in yellow) with three test videos and twelve data videos each. Upon passing the quiz with ≥ 70% accuracy they proceed to answer data pages with one test row per page. Data rows (DR, in white) contain 15 data videos. Data rows are annotated by five unique participants. Test rows can be answered once by each worker.

out the experiment. Data entered from workers that dropped below this threshold were removed from our study, and the corresponding videos were scheduled for re-annotation.

When running a study on Figure Eight, the experimenter decides the number of ratings per data row, as well as the pay per page. The latter was set such that with eight seconds per video, including five seconds for viewing and three seconds for making the decision, a worker would be paid USD 3 per hour. We had compiled 10,368 data rows of 15 data videos each. These data rows were presented to five workers each, yielding 155,520 annotated video clips. In some rare ($\leq$ 1%) cases users bypassed our restrictions by disabling javascript and were able to proceed without actually rating the videos. In that case the required 5 votes were not met, and we had to discard this video. Additionally, not all videos were readable by the Python libraries we used as feature extractors. Those videos were also removed. In total, 152,265 videos were retained as valid, forming our larger dataset, called KonVid-150k-A.

Each of the 10,368 data rows was presented to five workers. There were altogether 133 test rows for presentation to all crowd workers. However, each crowd worker could annotate any given test row at most once. Since 12 of the 15 videos in a test row were sampled from the set of data videos, we thus obtained far more than five ratings for each of these individual videos. In total, 1,596 data videos were used in the 133 test rows and were rated between 89 and 175 times, due to randomness in test question distribution. We separated 1,575 valid videos of this very extensively annotated set in a new dataset and call it KonVid-150k-B. As a random subset of the entirety of our videos selected from Flickr, it is ecologically valid and from the same domain as the other data videos. This dataset will be used as a test set for the evaluation of our models trained on KonVid-150k-A.

The choice for five individual ratings per data row was based on a small scale pilot study with a subset of 600 randomly sampled videos. For this subset we obtained two sets of 50 opinion scores for each video with a similar experimental setup as described above. We then evaluated the SRCC between a MOS comprised of a random sample of n votes from one set to the MOS of the other set. At 5 votes this SRCC reached 0.8,

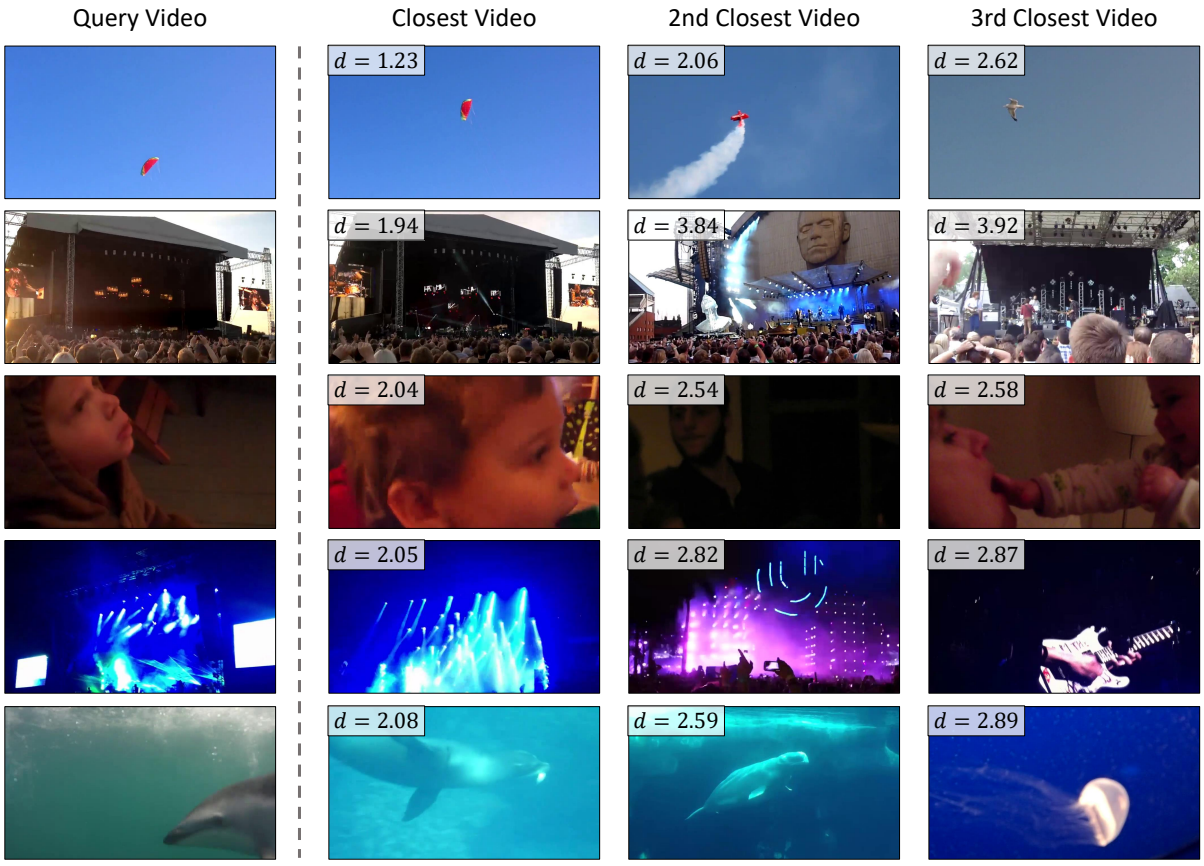


Figure 3.13: Still images from videos closest to the query video on the left as measured by the Euclidean distance d in the feature space of top-layer features from Inception-ResNet-v2. This shows the utility of activations of layers from pre-trained DCNNs for usage in a content similarity measure. Even though only the 1792 activations of the last layer were used, which are commonly understood to focus on semantic entities more so than low level structures, these features encode useful information.

which we considered to be a good threshold. For reference, the SRCC between the two independent samplings of 50 votes settled at 0.9. Further investigation of the feasibility of our choice of 5 ratings is contained in more detail in Section Section 5.3.

3.5 Dataset Comparison

3.5.1 Coverage

In order to evaluate the diversity of our datasets, which was a key objective in their creation, we will now demonstrate that they are not only the largest annotated VQA dataset in terms of video items, but also the most diverse in terms of content. First, we need a measure for content diversity. For this purpose, we extract the activations of the last fully-connected layer of an Inception-ResNet-v2 model pre-trained on ImageNet for each frame. To represent a given video, we average these activations over all frames to obtain a 1792-dimensional content feature. A similar approach has been used in the image quality domain before to create a subset of data that is diverse in content [Hos+20a].

Figure 3.13 is an illustration of the usefulness of these content features to

assess content similarity. Given a query video taken from KoNViD-1k on the left we compute the Euclidean distance in content feature space to all other videos in the dataset. On the right we show still frames from the three videos with smallest distance to the query. We can see that close proximity in content feature space seems to correspond to semantically similar video content. The images in the first row show flying objects in a blue sky, where the color of the object as well as the color of the sky seem to influence the distance in content feature space. In the second row we can see that crowds in front of a stage are located in close proximity in content feature space. Images in the third row show that videos containing heads, but especially babies are encoded similarly in the 1792-d content feature vectors. Light shows and underwater videos, as seen in the fourth and fifth rows, can also be retrieved by querying nearest neighbours of an appropriate video. It is to be noted that the closest videos for rows one, two and four are near duplicates. The recordings seem to be from different periods of time of the same scene.

Therefore, the extracted features are useful as an information retrieval tool, and we make use of it to quantify the degree by which a video dataset covers the content of competing datasets. For this purpose we represent a video dataset by its corresponding set of content feature vectors, $X = \{x_i \mid i = 1, ..., N\}$, where N is the number of videos in the dataset. We consider the Euclidean distance of a point x in feature space to a (finite) point set Y, $d(x, Y) = \min\{d(x, y) \mid y \in Y\}$. For two finite point sets $X = \{x_1, ..., x_n\}, Y = \{y_1, ..., y_m\}$ and any given distance $s \geq 0$, we define the fraction or ratio of the first dataset X, that is covered by the dataset Y at distance s as

$$C_{Y,s}(X) = \frac{|\{x \in X \mid d(x, Y) \leq s\}|}{|X|}$$

where $|A|$ denotes the cardinality of a set A. For example, if $X \subseteq Y$, then Y covers X perfectly at distance zero, i.e., $C_{Y,0}(X) = 1$. Or, if $C_{Y,1}(X) = 0.8$, then this means that the union of all balls of radius 1 centered at the points of the set Y contain 80% of the points in X. The function $s \mapsto C_{Y,s}(X)$ thus comprises the cumulative histogram of the individual distances $d(x, Y)$ for all $x \in X$.

When comparing the coverage two datasets with respect to each other, we check the corresponding cumulative histograms showing the coverage of one dataset by the other. The dataset with the topmost cumulative histogram then can be considered to be the dominant one that covers the competing one.

To compare the diversity of content for several given datasets $X_1, \ldots, X_K$, let us form their union $Z = X_1 \cup \cdots \cup, X_k$ and consider how well each dataset X_k covers all the others, i.e., the complement $X_k^c = Z \backslash X_k$. For this purpose we compute the cumulative histograms $C_{X_k,s}(X_k^c)$ for $k = 1, \ldots, K$. Figure 3.14 shows the result for the five datasets KonVid-150k, KoNViD-1k, VQC, Qualcomm, and CVD 2014. Here, KonVid-150k clearly has the best coverage of contents present in the other datasets, as it has the largest area under the curve.

To summarize the coverage of one dataset X by another, Y, by a single number rather than the curves of the cumulative histogram of distances,

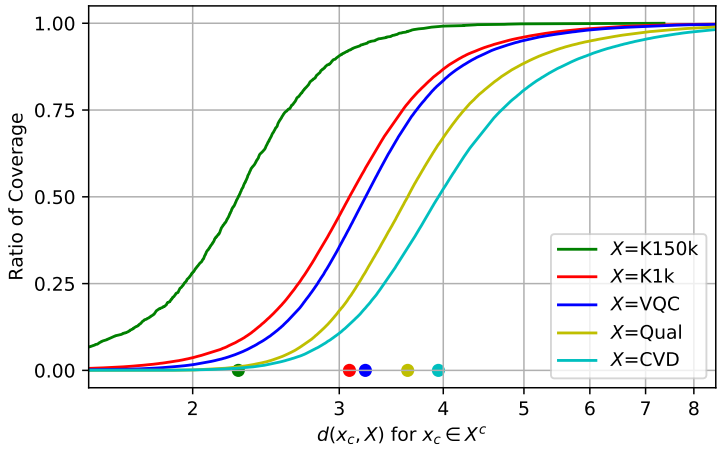


Figure 3.14: This plot shows how well a video dataset covers all others together. The curves are the empirical cumulative histograms of Euclidean distances $d(x_c, X)$ for all $x_c \in X^c$, where X^c is the complement to X, i.e., the union of the other datasets. The green, red, blue, yellow, and cyan lines refer to X being KonVid-150k, KoNViD-1k, VQC, Qualcomm, and CVD 2014, respectively. KonVid-150k covers the other datasets the best, as the green plot has the largest area under the curve and it has the smallest median distance of approximately 2.3 at coverage ratio 0.5. This means that for half of the videos in all other datasets, there is a similar video in KonVid-150k that has a distance in content feature space of at most 2.3.

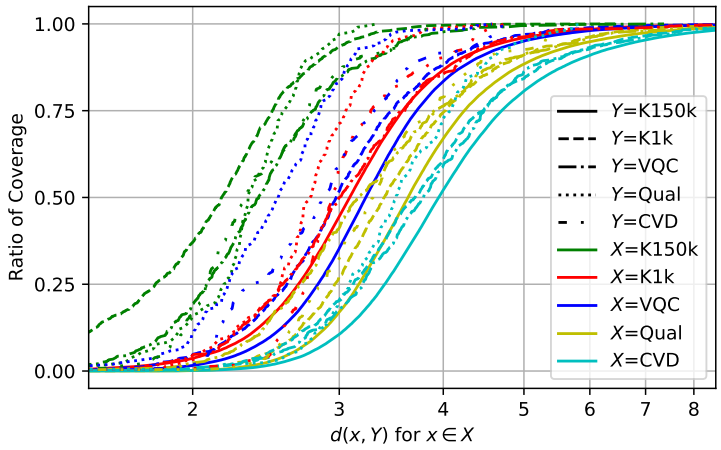


Figure 3.15: Pairwise comparison of content coverage. Empirical cumulative histograms of $d(x, Y)$ for all $x \in X$. The green, red, blue, yellow, and cyan line colors refer to the covering set Y and the different line styles refer to X being KonVid-150k, KoNViD-1k, CVD 2014, Qualcomm, and VQC, respectively. As expected from the previous figure, KonVid-150k covers the other datasets the best, indicated by the four green plots consistently falling to the left of their counterparts. The summarizing statistics, $d(X, Y)$ can be taken from the intersections of the graphs with the

we define the one-sided distance of X from Y as

$$d(X, Y) = f(d(x_1, Y), d(x_2, Y), \ldots, d(x_n, Y))$$

where f is a scalar, non-negative function. For example, if f is the maximum function, then $d(X, Y)$ is known as the one-sided Hausdorff distance. For our purpose, the median is better suited as it is less sensitive to outliers. The distance $d(X, Y)$ can be understood as a simplified indicator for the coverage of X by Y. These medians are shown in Figure 3.14 by the bullet dots at the coverage ratio of 0.5.

Figure 3.15 then shows $d(X, Y)$ for the competing dataset pairs individually. It can be seen that KonVid-150k covers the contents of competing datasets the best, as the green curves are strictly above the cumulative histograms for the other datasets. Moreover, the other datasets cover the content space of KonVid-150k the worst, as the solid lines depicting the coverage of KoNViD-1k, CVD 2014, Qualcomm, and VQC of KonVid-150k are generally to the right of the other three for the respective dataset.

These findings are an indication that our proposed dataset KonVid-150k is comprised of a large variety of contents with good coverage of the contents contained in existing works.

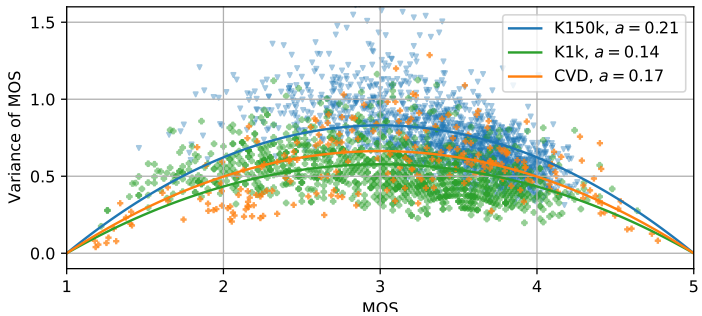


Figure 3.16: Comparison of the SOS hypothesis [HSE11] of KoNViD-1k, CVD2014, and KonVid-150k-B. The SOS parameter for the three datasets are $a = 0.14$, $a = 0.17$, and $a = 0.21$, respectively. For VQA the typical range is $a \in [0.11, 0.21]$, which shows that KonVid-150k can be considered a typical example in terms of annotation quality.

3.5.2 SOS hypothesis

Another common characteristic to compare the annotation quality of different studies is by evaluating the standard deviation of opinion scores (SOS) as a function of MOS. It follows the basic idea that in quality controlled experimental studies subjective opinions will vary only to a certain extent, as the experimental setup ensures similar test conditions. In the case of the 5-point scale we used in our experimental setup, the maximum SOS is expected near a MOS of 3, while the minimum will be attained near the extremes of the rating scale (i.e., 1 and 5). Therefore, computing the average SOS over all videos is not an unbiased indicator, as common datasets have differing distributions of MOS values. Instead, the variance σ^2 is modelled as a quadratic function of the MOS [HSE11], which in the case of a 5-point scale is described as:

$$\mathrm{SOS}(\mathrm{MOS})^2 = a(-x^2 + 6x - 5), \tag{3.3}$$

and the SOS parameter a is a better indicator of the variance of subjective opinions for any particular experimental study.

Moreover, the SOS parameter has been shown to correlate with task difficulty and can be used to characterise application categories [JP15]. For VQA the SOS parameter has been reported in the range $a \in [0.11, 0.21]$, with $a_{\text{KoNViD-1k}} = 0.14$ and $a_{\text{CVD2014}} = 0.17$. In the case of LIVE-Qualcomm and LIVE-VQC, no SOS parameter has been reported and the publicly available annotation data does not allow for such an analysis, as only the MOS values for videos in these specific datasets are available.

We computed and visualized the SOS parameter for KonVid-150k-B as well, see Figure 3.16. For the case of the larger KonVid-150k-A set, we have 5 ratings per stimulus which allows only for 21 different MOS values, and therefore we did not include it in the figure. Nonetheless, KonVid-150k-B is a good estimation of what can be expected in terms of annotation quality of KonVid-150k as a whole. The figure shows the comparison between KoNViD-1k, CVD2014, and KonVid-150k-B, where the latter has an SOS parameter of $a_{\text{KonVid-150k-B}} = 0.21$, which lies within the typical range for VQA experiments.

Considering the similarities between KoNViD-1k and KonVid-150k, the difference in a seems surprisingly large at first. However, some differences in the design choices of the subjective annotation process can be identified as potential causes for the larger SOS parameter for KonVid-150k.

Videos from KoNViD-1k and KonVid-150k are both sampled from Flickr.com. However, their compression settings are different. While the videos in KonViD-1k are heavily compressed, those in KonVid-150k are representative of the originals as uploaded by their respective authors. This means that KonVid-150k videos are more diverse in terms of distortion types, as heavy compression can have a strong masking effect. A wider variety of distortions is expected to cause a higher disagreement between raters, and thus a higher variance of their answers.

Moreover, the sources for the test videos in each dataset used during the crowdsourcing experiment are different. KoNViD-1k test videos were sampled from the same source and with ground truth annotations from a prior study, while the test videos in KonVid-150k are sampled from another source, and involve artificial distortions.

On the one hand, the choice of test videos for KoNViD-1k can cause workers to pay more attention, and agree better, however, at the cost of having more biased answers. First, the test and data videos are impossible to distinguish at a glance. This means that crowd-workers need to constantly pay attention to all work items, and not just to those that are easy to identify as test items. Second, the test videos have similar levels and types of distortions. There are no other items to anchor user opinions at the extreme of the quality scale. This means that the range of the quality scale may not be used well. The downside of this choice is that the accepted answers for the test videos are derived from a pilot study, and this can introduce a bias towards the opinions expressed in that study.

On the other hand, KonVid-150k uses pristine quality videos from a different source (pixabay.com), alongside highly degraded variants of the same videos. These videos are easier to distinguish from data videos. Consequently, workers are not forced to pay attention to all items the same, they can theoretically put more thought in answering test videos than they do for data videos. The tests in this case are more lenient, as they are selected to represent the extremes of the quality range (both highest and lower quality). However, they also serve as anchors for the quality scale, which are not available for KoNViD-1k. The approach is less biased, but can result in more disagreement between annotators, which in turn leads to a higher variance of the answers. It is preferable to have less bias rather than a higher agreement on the wrong ratings.

3.5.3 Annotation Quality

Figure 3.17 shows the convergence of the size of the 95% confidence interval on the MOS, averaged over all images, in relation to the number of workers. We compare this convergence from both KonVid-150k and KoNViD-1k with CVD2014, with a large overlap. With the goal of a mean size of the confidence interval of below 0.5 on the ACR-scale, we arrived at 50 answers per stimulus as an optimal number of judgments. This setup resulted in a total of 642 workers from 64 countries participating in our experiment. On average each video received 114 votes (including test questions) with a mean accuracy of 94% on the test questions. Figure 3.18 depicts a histogram of common user statistics in the KoNViD-1k crowdsourcing study. With regards to screen size, 94% of workers had a

resolution above 1024 × 600 pixels (width × height), which allowed full size display of the videos. It is to be noted that we did not enforce 1:1 pixel display. However, nearly two-thirds of workers did in fact view the videos unzoomed at 1:1, while 20% used a ratio of 0.9 to 0.75 and 14% displayed the videos at a zoom of up to 2 (including font scaling).

When comparing video quality ratings across different studies, commonly MOS, standard deviations and confidence intervals are considered. However, it has been shown that when comparing different rating scales, the design and discretization of rating scales have a strong influence on the standard deviations of the opinion scores (SOS). In [HSE11] a quadratic model is proposed for the dependence of the variance σ^2 on the MOS values. For the 5-point ACR scale the function $\sigma^2(\text{MOS}) = a(\text{MOS}-1)(5-\text{MOS})$ is fitted to empirical data, which yields the SOS parameter a. The parameter a quantifies the variance of the user ratings more appropriately than the average over all stimuli. Moreover, it characterises application categories and correlates with task difficulty [JP15]. For VQA, the SOS parameter a was reported to fall in the range $[0.11, 0.21]$ [HSE11].

We compared the standard deviations of the crowdsourced ratings of our KoNViD-1k to the CVD2014 (normalized from a 0-99 scale to a 5-point ACR scale) and IRCCyN-IVC-1080i datasets [Nuu+16; PPL08c], where the subjective scores were gathered in a lab setting, see Figure 3.16. With $a = 0.14$ our SOS hypothesis parameter is lower than those of the two compared datasets (0.17 for CVD2014 and 0.21 for IRCCyN-IVC-1080i, respectively). This suggests that our task was simpler than the compared lab studies and resulted, as is desirable, in lower variances of worker ratings with smaller confidence intervals for the MOS values.

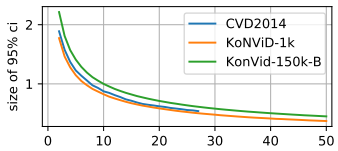


Figure 3.17: Convergence of the confidence intervals with respect to number of workers of our datasets (in orange/-green) compared to CVD2014 (blue). The confidence intervals for KoNViD-1k are slightly smaller than those of CVD2014, while KonVid-150k-B performs slightly worse by this measure.

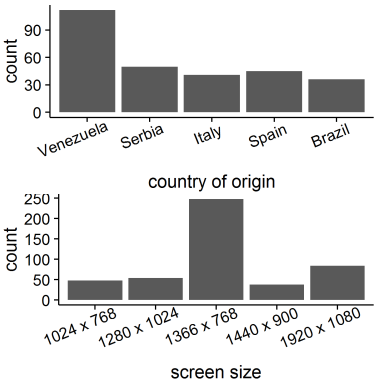
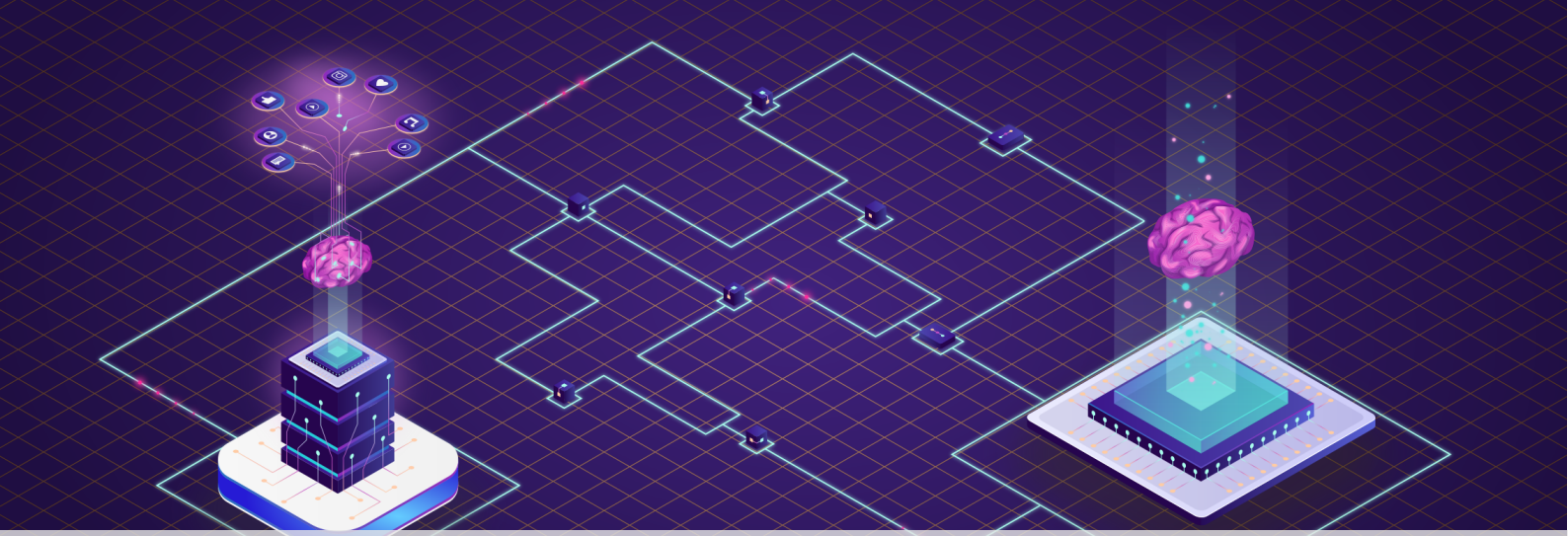


Figure 3.18: Crowdworker statistics on country of origin and screen resolution from the KoNViD-1k annotation. The five most frequent groups are shown, making up over 94% of workers.



4 Quality-Related Features

One of the main drivers of any machine learning task is effective representations of data that capture salient semantics relevant to the task itself. For the purpose of quality assessment measuring the presence of particular classes of distortions is a sensible approach. Conventionally, the design process for distortion detectors was manual work based largely on an in-depth understanding of both the human visual system and image and video compression techniques. Recently, however, it has been operationalized by designing systems that learn representations automatically. In the last thirty years multi-layered deep convolutional architectures have emerged as the strongest approach to learning generic visual features for the purpose of a variety of perceptual tasks, such as object classification, recognition, and localization in images in-the-wild. In this chapter we will show the use of these learned features for the purpose of IQA, VQA, and other perceptual tasks. We summarize both historical and state-of-the-art features used in practice and present new research into the field of feature selection for generalized perception of quality.

This chapter contains and extends material from the following publications. Please refer to Section 1.4 for the contribution clarification.

[Hos+17] Vlad Hosu, **Franz Hahn**, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 'The Konstanz natural video database (KoNViD-1k)'. In: *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2017, pp. 1–6

[Göt+21] **Franz Götz-Hahn**, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 'KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild'. In: *arXiv preprint arXiv:1912.07966* (2021)

4.1 Feature Detection

Compression of digital media is the primary, albeit not sole, source of quality degradations. In the following we will attempt to summarize and reference the key approaches to represent presence of and measures for distortions and quality-related features more comprehensively. Section 3.3.1 already described a variety of detectors for specific distortions, artifacts, or other image characteristics, which were also successfully used as proxies for assessment of quality of visual media. That list is not exhaustive, however, since the breadth of research into distortion measures is too large to summarize here in its entirety.

Please refer to Section 2.3.2 for a description of the most common distortions and their perceptual appearance.

4.1.1 Quality Artifacts

Video quality artifacts can be categorized into spatial and temporal artifacts, which both subdivide into further types of distortions, as shown in Figure 4.1. Spatial artifacts are visible in individual frames, while temporal artifacts are perceived when a sequence of frames is displayed.

Spatial Artifacts

Within the class of spatial artifacts there are four major subclasses to consider.

The measurement of these discontinuities at block boundaries is one way to estimate the level of blockiness [WY96]. When the block boundaries are unknown, for example when a video has been encoded and cropped multiple times, blockiness can be estimated using harmonic analysis [TG00].

Blind estimation of blur work off of a variety of different concepts, including the width and amplitude of edges [Dij+03], the detection of their spread [Mar+02; Ong+03], kurtosis [CO04; Zha+03], particular features of the frequency domain [KH96; MMZ99], or the local contrast of 2-D analytic filters [Win01]. Other work focuses on the estimation of the probability that blur is detected at a particular edge, called just noticeable blur [FK09; NK11].

As shown in Section 3.3.3, the colorfulness of an image is weakly correlated with image quality. Colorfulness by itself can been measured, for example, by considering the mean and variance of color distributions [HS03] as described in Section 3.3.1.

Algorithms to detect chromatic abberation have been proposed in related works [LK16] that are based on either the correlation between different color planes or measuring the level of corner displacement.

Blind ringing metrics are highly designed, complex algorithms that identify and extract line segments relevant for ringing detection and the estimation of whether ringing is perceived by involving a model based on the HVS [LKH09].

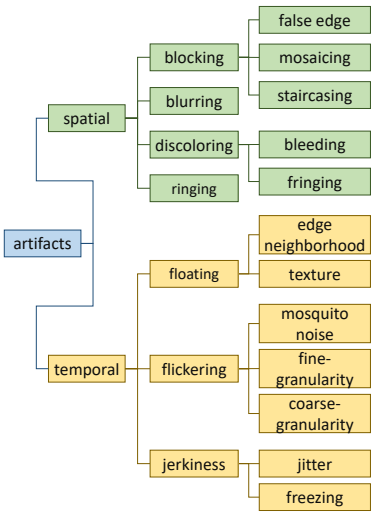


Figure 4.1: Categorization of video compression artifacts into subclasses of spatial and temporal domains.

Temporal Artifacts

Distortions manifesting during playback of a video can be subdivided into three classes.

Texture floating detection has been proposed that evaluates relative motion in mid-energy and mid-luminance regions of a video stream [Zen+14].

Different jerkiness metrics have been proposed, mostly based on inter-frame correlation [Mon+06] or inter-frame dissimilarity [PG06]. A more recent, significantly more complex and constructed metric also evaluates the temporal quality fluctuation [Bor10], as the same amount of frame loss within one sequence could lead to different levels of subjective temporal degradation when considering different temporal distributions of frame loss.

In summary, the metrics specifically targeted at detecting the above mentioned distortions have become increasingly complex and little research has evaluated their stability as video quality and encoder performances has improved over the past years. Tweaking these metrics to videos encoded using modern codecs is likely an involved process. Additionally, evaluating a large breadth of these features to capture a wide range of quality indicators is a time-consuming process. In recent related work a set of 75 quality-related features were considered as a basis for their VQA model [Kor19]. Their 46 temporal (consistency) features describe, for example, motion intensity, direction, uncertainty, and spread, as well as jerkiness and spatial activity. They also included spatially relevant features, such as blurriness and blockiness. An additional 30 so-called high-complexity features further evaluated various blockiness, noise, sharpness, contrast, colorfulness and brightness measures. These high-complexity features were not computed on all frames of the target video, rather on a significantly smaller subset of frames that were selected based on information from the temporal features computed prior. For five ten to twenty second long videos in HD ready resolution of 1280x720 pixels the computation of all 75 features took an average of 222.2 seconds. They also showed that the computational complexity of the evaluation of their chosen set of features was approximately linear with the resolution. Since video resolution is ever growing, and their choices of metrics used for features such as jerkiness or flickering are on the lower end of the computational complexity scale, it is easily understood that this approach of evaluating complex, hand-crafted feature detectors does not scale very well.

4.1.2 Deep Features

Designing detectors for particular image features by hand, evaluating them on an input, and then ultimately feeding the resulting feature vector into a model was a common and high-performing approach for many different perceptual tasks. An alternative to it was originally shown for the purpose of handwritten zip code recognition [LeC+89a], where the network architecture would instead take the input in its original form and learn the representation of useful features automatically. In the subsequent decades the approach was adopted in improved forms in practically all scientific fields related to perception and even completely

unrelated fields. At this point it is no exaggeration to say that the internal representations of useful features within deep neural networks (DNNs) have effectively replaced traditional hand-crafted features for tasks related to the prediction of perceptual attributes of images.

Modern DNNs are comprised of a few initial layers often referred to as the stem, followed by multiple layers of particular modules, where each module is a small network in itself, and finally a few layers at the top, called the head of the network. The modules commonly include a small number of parallel streams, each consisting of chains of a handful of convolutional and pooling layers, which are concatenated at the end. Oftentimes the internal structure of these modules also changes based on the location of the module within the network. Modules closer to the input of the network will often feature convolutional layers with smaller kernel sizes, while those closer to the head sometimes use larger kernels. This is based on the assumption that the distance of the layer to the input is related to the level of abstraction the features represent. The spatial concentration of higher level features can be expected to decrease, suggesting the use of more convolutional kernels of larger sizes.

In other words, the depth at which a layer is located is a basic measure for the level of complexity that the feature can represent. This has been validated by recent works [SVZ14; MV15; MOT15; DB16] which tried to visualize what a neural network considers to be a good representative image for a particular class it was trained to classify with some surprising results [NYC15]. More recently, these efforts were further elaborated on by generating images that maximize the activation of a specific kernel in a particular layer [Sch+20], allowing more introspection into what individual neurons in a network are trained to detect. To give an illustration of this, images that particular features in the lower, middle, and upper part of the Inception-v1 network activate highly on are shown in Figures 4.2, 4.3, and 4.4, respectively. Evidently, early layer features activate on particular alignments of edges, sometimes including specific color information on either side of the edge. Features in the middle part of a network activate to entities with simple semantic meanings such as circular objects on a green background, eyes, or even animal faces. Further up the network individual features might activate on very specific objects of increasing complexity that are frequenty present in the training data, such as poodles, bugs, or statues of clocks.

Given some dataset to train on, the choice of network will heavily influence what types of features it will generate. The ImageNet large scale visual recognition challenge [Rus+15] has brought about a wide variety of network architectures with new best performing models every year [KSH12; Sze+17; SVZ14; He+16; Xie+17; Cho17; Hut18]. All of these have different structures, with AlexNet [KSH12] only using 8 convolutional layers, while some ResNet [He+16] variations use up to 152 convolutions layers. As a result, the features that were learned in these two examples are also drastically different. Neurons in the upper layers of an AlexNet are quite specific and the visualization techniques used before result in human interpretable images. In contrast, neurons in the upper layers of a ResNet-101 are much less interpretable on a first glance and require more inquiry by, for example, extracting image patches that the neuron activates highly on, in order to make sense of what types of semantics the neuron captures. Furthermore, the hierarchical

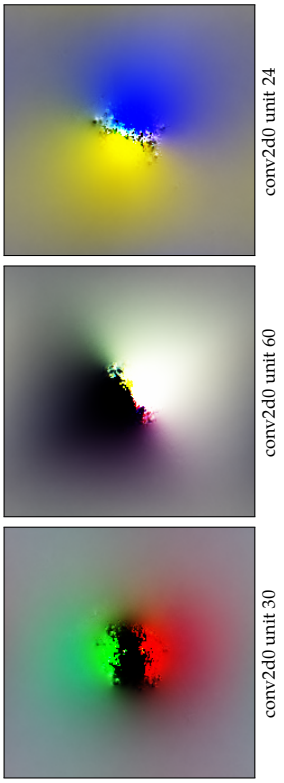


Figure 4.2: Visualization of features from an early layer in the Inception-v1 network. Courtesy to OpenAI Microscope for the visualization.

Figure 4.3: Visualization of features from a layer located in the middle of Inception-v1 network. Courtesy to OpenAI Microscope for the visualization.

structure, as well as the interconnectedness of the hierarchy, allows for near arbitrary complexity of the learned features. Constrastingly, the overparameterization and a substantial level of redundancy are widely accepted characteristics of these large deep neural networks, with studies showing that the performance of deep networks does not deteriorate even under heavy weight pruning [LeC+89b; Den+13; HMD15], weight quantization [HMD15; Zho+17], and huffman coding [HMD15].

A benefit of deep features stems from the fact that the output of a kernel in any given layer of a neural network is a feature activation map, which inherently also includes positional information. Changes in the response of these feature maps can, therefore, encode temporal information. For example, it is reasonable to assume that fine-granularity flickering, i.e. a slight back and forth translation of an edge between successive frames, would show a change in the spatial activation map of low-level edge detector features. Consequently, learning from frame-level features could allow to learn the effect of temporal degradations on video quality indirectly. However, the large size of pre-trained networks yields a much larger amount of features compared to the number of existing hand-crafted features. The first convolutional layer of an AlexNet alone has more kernels than the 75 hand-crafted features used in the work described at the end of the last Section. The broadness of the filterbank that is generated by DNNs transforms the task from simply learning a model based on them to identifying those features that are useful for the detection of spatial or temporal distortions.

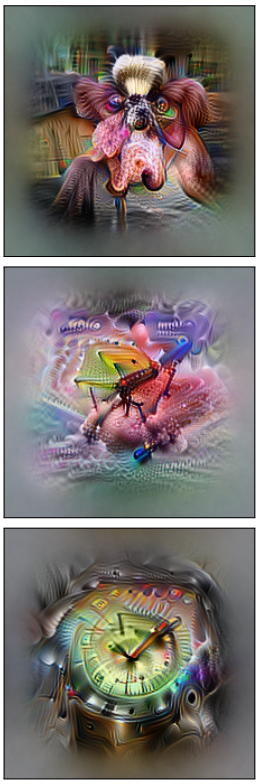


Figure 4.4: Visualization of features from an upper layer in the Inception-v1 network. Courtesy to OpenAI Microscope for the visualization.

4.1.3 Deep Feature Extraction

The extraction of deep features is a fairly recent approach, but has been applied for the purpose of visual quality prediction in several related works. TL-Xception [OAB18] was an initial work that utilized deep-features to predict image quality in a transfer learning setting. Using an Xception-net [Cho17] as a base-model, they added two 1×1 convolutional layers on top, followed by both a global average pooling layer and a global maximum pooling in parallel. The outputs of the pooling served as an input to a small fully connected head which was topped off with a 5-neuron output layer that represents the opinion score as a distribution. Using this approach, the authors achieved state-of-the-art performance.

The paper also references DeepRN [VSS18] as a better model, however the results of DeepRN for KonIQ-10k have since been shown to be incorrect [GHS21]

The BLINDER framework [Gao+18] improved upon the approach by using multiple layers of the base-model to extract deep features. They resized images to 224×224 and extracted a feature vector from each layer of a pre-trained VGG-net. Each of these features vectors was then fed into separate SVR heads and trained, such that the average layer-wise scores predict the quality of an image. BLINDER was evaluated on a variety of IQA datasets and reported an improvement of the state-of-the-art.

[HGS19] went a step further by utilizing deeper architectures to extract features, such as Inception-v3 and InceptionResNet-v2. Furthermore, features were aggregated from multiple levels and extracted from images at their original size. This retained detailed information that would have been lost by down-sizing the inputs. Moreover, it allowed linking information coming from early levels (image dependent) and general category-related information from the latter levels in the network. This

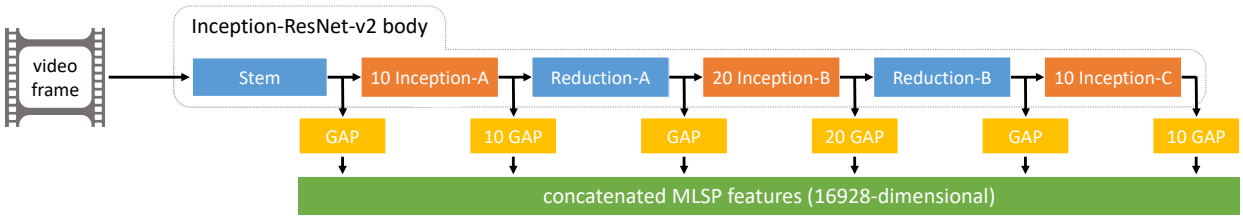


Figure 4.5: Extraction of multi-level spatially-pooled (MLSP) features from a video frame, using an InceptionResNet-v2 model pre-trained on ImageNet. The features encode quality-related information: earlier layers describe low-level image details, e.g. image sharpness or noise, and later layers function as object detectors or encode visual appearance information. Global Average Pooling (GAP) is applied to the activations resulting from the Stem, each Inception-module, as well as the Reduction-modules, and finally concatenated to form MLSP features. For more information regarding the individual blocks please refer to the original paper [Sze+17].

approach has since been further elaborated on with DeepFL [LHS20], which incorporated a supervised fine-tuning step prior to feature extraction to drastically improve state-of-the-art NR-IQA performance on the complex artificially degraded KADID-10k dataset.

These two efforts are the closest inspiration for our approach, namely we extracted narrow multi-level spatially-pooled (MLSP) features, but for individual frames of videos, as shown in Figure 4.5. In principle it is an expansion of what we described previously in Section 3.5.1 for the comparison of content diversity. The core difference lies in the extraction of features in all Inception modules of the network, rather than only the last layer in the network head. This approach of extracting activations from individual layers of a network can be applied to any popular architecture. Related work has shown that this approach works with an Inception-ResNet-v2 network as a feature extractor in the IQA domain [Hos+20a; LHS20]. For the extraction process we, therefore, passed individual video frames to an InceptionResNet-v2 network, pre-trained on ImageNet [Sze+17]. We then performed global average pooling on the activation maps of all kernels in the stem of the network, as well as on each of the 40 Inception-ResNet modules and the two reduction modules. Concatenating the results yielded our MLSP feature vector consisting of average activation levels for 16,928 kernels of the InceptionResNet-v2 network. These MLSP feature vectors were extracted for all frames of all videos.

In [Var20] a similar approach was used for the purpose of NR-VQA. The method extracted features for intra-frames, averaging them along the temporal domain to obtain a video-level feature vector. The final video quality prediction is done by an SVR. In our approach we go beyond this by considering both an average feature vector with our MLSP-VQA-FF architecture, as well as an LSTM model that takes a set of consecutive features of frames as input, leveraging temporal information of feature activations.

Figure 4.6 shows a visualization of parts of the MLSP feature vector for multiple consecutive frames.

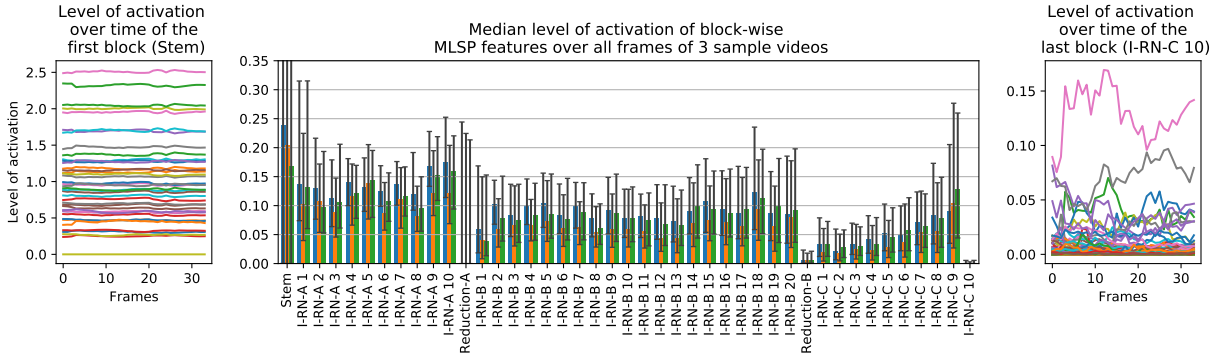


Figure 4.6: Visualization of the variation of activation levels of MLSP features over the course of KonVid-150k videos. In the center, the median level of activation for each of the 43 blocks from the Inception-ResNet-v2 network is displayed for 3 sample videos. The black whiskers indicate the 50% confidence interval on the level of activation. For the first block (Stem), the whiskers extend to 0.7. The left and right plots show the activation of 1/8th of the first and last blocks' features over time.

4.2 Feature Selection for Perceptual Tasks

Pre-trained networks used as a starting point for fine-tuning or a source for feature extraction are usually trained on a different task than what they are ultimately used for. Therefore, it is reasonable to assume that some subset of them are either irrelevant or redundant for the target prediction. Here, we describe a feature as irrelevant when it does not affect the prediction in any way, while redundant features do not add any new information. Both types of features will slow down the speed of the learning algorithm due to a higher dimensionality of the feature space. Additionally, redundant features may even add noise, consequently potentially reducing predictive performance of the trained model. To combat this, one can employ feature selection, meaning selecting a subset of features, such that sufficient information is preserved to retain a high level of performance. It has been shown that selecting appropriate features positively benefits learning performance in neural networks significantly [MM91; Bat94; Pir04].

Despite the large quantity of different network architectures that have been developed for the object classification task, the same source networks have been adopted for different tasks related to perceptual qualities. In particular, both for the prediction of image quality and image aesthetics Inception-style networks have been shown to be a solid source for representative features. This suggests that the features used for these tasks carry information useful for both tasks. Some related work even investigated the tasks jointly [TM18]. In this case, the authors not only trained the same type of model for both technical quality prediction and aesthetics assessment, individually, they also introduced cross-task testing, where the models trained for one task were tested on the other. However, the model optimization was only ever considering the tasks independently. In literature on measuring aesthetic quality it was revealed that technical quality is an important attribute to consider when predicting aesthetics, as a high level of technical quality was necessary for high aesthetics ratings [CL09]. Despite instructions stating to focus only on aesthetic quality, observers are influenced by the integrity or quality of an image in their decision making [CL09; RH12]. On the other hand, characteristics commonly attributed to aesthetics such as desireability of

the media content were found to be a strong influence on the perceived technical quality of videos [KS10], which likely translates to images as well. With the goal of solving perceptual quality assessment in more general terms, the intrinsic dimensionality of the problem's feature space should be considered.

We explore the potential of deep feature selection for the use case of perceptual attribute prediction. Specifically, we investigate the following problem:

> What is the minimal size of a subset of the full feature vector that still approximates the original performance on both an IQA task, as well as an aesthetics quality assessment (AQA) task, within a "small" epsilon.

Based on feature importance obtained from gradient boosting machine (GBM) models we derive a subset of 10% of the deep features that is competitive with full feature performance. Moreover, as the selection heuristic considers both the feature's importance for the IQA and the AQA tasks, it performs en par with an optimal subset of deep features for the individual tasks. Finally, the joint optimal subset surpasses state-of-the-art in cross-task testing.

4.2.1 Related Work

The field of image quality assessment matured over the years from approaches that exploit natural scene statistics [MMB12; Ye+12; Xu+16] to deep-learning based predictors. A conceptual predecessor of our approach is presented by Bianco et al. [Bia+18], wherein the authors experiment with extracting features from convolutional neural networks that were pre-trained for different tasks.

Aesthetics prediction is closely related to and intertwined with quality assessment, but focuses more on subjective appeal than on technical image characteristics. Recent works by Talebi and Milanfar introduced NIMA [TM18], an approach that investigates both aspects and retrains object recognition networks on one aesthetics and two quality datasets. Though the training was done separately, the authors were the first to conduct cross-task testing on one dataset after fine-tuning on the others in all directions, which suggested a weak correlation between the tasks.

Hosu et al. [HGS19] presented a landmark paper that introduced multi-layered spatially pooled features and cemented the Spearman's rank-order correlation coefficient as a metric over the prevalent binary classification accuracy in aesthetics prediction. The approach improved the SROCC on AVA from the previous state of the art of .64 to .67 using only a small network of three fully connected layers applied to the entire MLSP feature vectors. Furthermore, the authors reproduced the results produced by NIMA and made substantial points towards the claim that training on extracted features is more powerful than fine-tuning of neural networks. An application of this approach to an IQA task is portrayed in [Hos+20a], which also extends the paper that originally introduced the KonIQ-10k database [LHS18].

The line between technical image quality and aesthetics is ambiguous and the two concepts can even be in contrast to each other. Stylistic means, such as bokeh blur, can degrade the image quality, while they might increase the aesthetic appeal. A few studies investigated this distinction and its consequences for quality and aesthetics assessment. Cerosaletti and Loui [CL09] researched the effects of selected image characteristics such as main subject size, sharpness and composition on perceived technical and *artistic* properties. They presented visible clusters when plotting the first versus the second principal component of these characteristics that could be summarized through short descriptions of the respective image's characteristics. Their investigation was carried out separately for images containing humans and images not containing humans. They argued that initial filtering for sufficient technical quality first might be reasonable when aiming to identify aesthetically pleasing images.

Redi and Hyenderickx [RH12] aimed to establish a more encompassing definition of visual quality by subsuming technical aspects under "image integrity", and contrasting these with aesthetics. They ran a subjective study on the aesthetic appeal of a set of images that was once given in pristine quality and once impaired by visible technical distortions. Participants rated severely distorted images lower in comparison to less affected images, albeit being asked not to consider integrity in their judgement, which corroborates the presumed link between quality and aesthetics.

4.2.2 Feature Importance

Gradient boosting machines (GBMs) are a class of predictive machine learning models based on the assumption, that ensembles of weak learners can be used as a single strong learner. A weak learner is considered to be only slightly better at predicting some outcome than random chance. In GBMs the weak learners are decision trees with shallow depth, which are constructed sequentially according to an optimization problem using a procedure roughly similar to gradient descent. Extreme gradient boosting (XGBoost or XGB) [CG16] is an advanced implementation of GBMs that utilized various improvements that have been developed over the past two decades to boost GBM performance. "Feature importance" is calculated for a single XGB model by the amount that each node improves the performance measure, weighted by the number of observations the node is responsible for. The performance measure is the weighted MSE of improvement at all nodes of the tree divided by the numbers of observations.

By considering ensembles of XGB models we obtain an importance value for each feature per member of the ensemble. There are two main ways these are commonly aggregated. Either, one considers the number of times a feature appears among the top features for all models of the ensemble, or the sum of the individual normalized importance values over all ensemble members is used. We use the latter interpretation and also consider the joint feature importance as the product of the features importance values from the two tasks of IQA and AQA. Based on this approach Figure 4.7 depicts a histogram of feature importance values for all three cases. This shows that no individual feature has very high

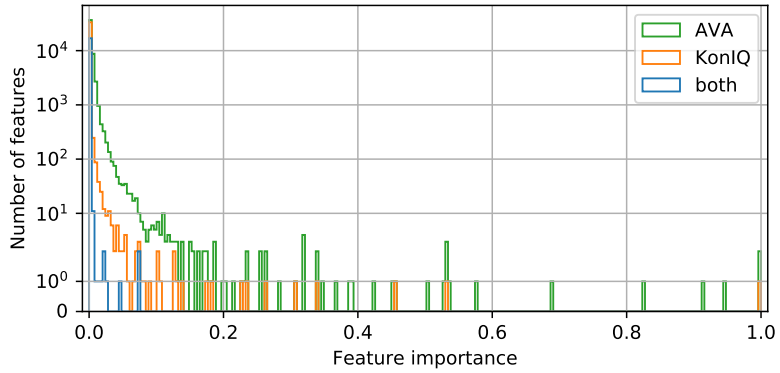


Figure 4.7: Histogram of feature importance for KonIQ, AVA, and their product ("both"). For both of the individual tasks only few individual features are considered important by the XGB models.

importance in both tasks, as the highest importance value in the product is below 0.1. Figure 4.8 shows a histogram of features according to the Inception modules within the Inception-ResNnet-v2 parent network for the top 1% (169) of features. It is apparent that particular blocks (I-RN-B 3 and Reduction-B) seem to be more related to the prediction of both perceptual attributes.

Features useful for IQA and AQA are not an exhaustive list of features useful for the whole domain of general perceptual quality tasks. However, these tasks have by far the biggest annotated datasets, and so we limit ourselves to these two. Further research has to be done to evaluate the utility of features found here for other tasks in the domain.

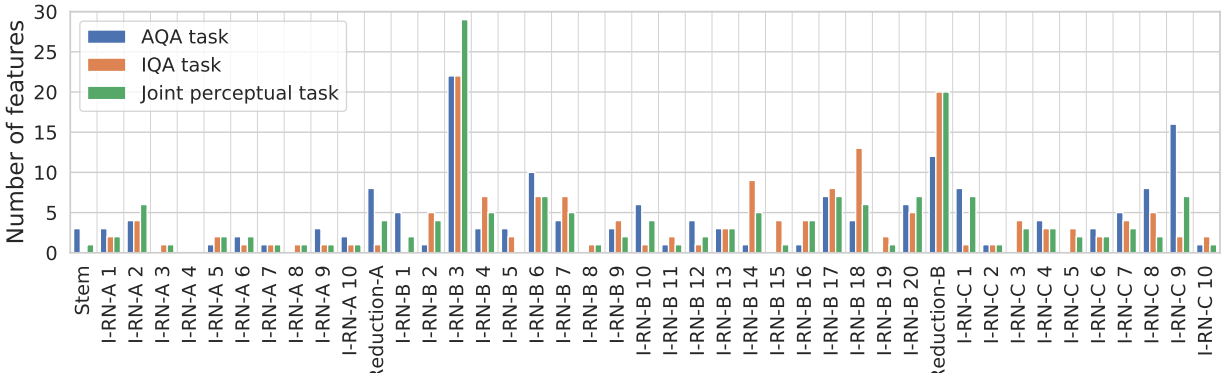


Figure 4.8: Histogram of top 1% features according to feature importance grouped by the Inception module they reside in for AQA and IQA tasks, as well as jointly.

4.3 Experiments

In this Section we focus on three parts. First, we validate the heuristic of selecting features based on the aggregated feature importance values obtained from the 100 GBM models trained on all features. Next, we compare the performance of IQA and AQA related features with the joint perceptual feature subset. Finally, we further evaluate the generalization performance of XGB models trained on perceptual features in cross-task tests.

4.3.1 Feature Selection Evaluation

Since we focus on the goal of finding features that solve both tasks, we present results based on the product of feature importance values. As a baseline we are comparing it to the random selection of the same number of features. We sampled 100 random feature combinations of $n \in \{1, 4, 16, 66, 264, 1058, 4232\}$ features and subsequently trained a GBM model on a unique 90/10 train/validation split. The resulting performances evaluated on the official test sets for KonIQ and AVA are visualized as estimated contour plots for the joint probability densities in Figure 4.9. Additionally, we plotted the top feature combinations in the same color, obtained by taking the same number of features with the highest importance values of the individual tasks, as well as the joint importance.

Any of the top feature combinations outperform random selections by a significant amount. Nonetheless, objectively evaluating the performance improvement is not trivial. A naïve way could be to take the euclidean

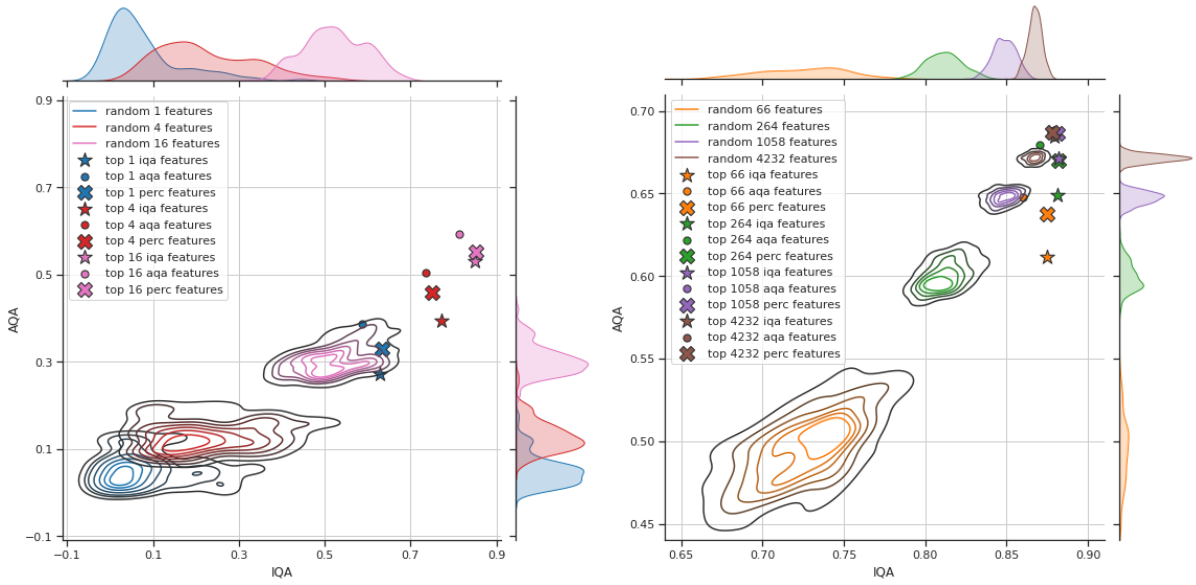


Figure 4.9: Performance of XGBoost models on KonIQ and AVA when trained on different numbers of features taken either randomly (density plots) or as the top features according to individual joint feature importance on both tasks ('x' markers). For the density plots we obtained 100 random feature combinations.

norm as

$$\sqrt{\rho_{\text{IQA}}^2 + \rho_{\text{AQA}}^2} \tag{4.1}$$

where ρ indicates the SROCC on the respective dataset, and comparing the difference between the mean performance of a randomly sampled feature set to that of the most important perceptual features. However, this difference does not represent a fair measure, as an improvement in this norm has to be evaluated within the context of the baseline performance. The significance of an improvement of both ρ_{IQA} and ρ_{AQA} by 0.05 by using the feature selection heuristic depends on the performance of a random feature combination. If a random selection of features performs at $\rho_{\text{IQA}} = \rho_{\text{AQA}} = 0.2$ then the improvement is likely to be considered insignificant. On the other hand, if the baseline performs at $\rho_{\text{IQA}} = \rho_{\text{AQA}} = 0.8$ then an improvement of 0.05 is much more impactful.

In order to better estimate how much better the selection of any feature combination performs compared to random sampling, we derive the norm reduction from the theoretical maximum performance of $\rho = 1$ for both tasks d as

$$d_n = 1 - \frac{\sqrt{2} - ||x_{\text{top}_n}||}{\sqrt{2} - ||x_{\text{rand}_n}||}. \tag{4.2}$$

Here, x_{top} and x_{rand} are vectors in performance space of the single top feature combination and randomly sampled feature combinations, respectively. Effectively, this measure represents the distance reduction to the theoretically maximum performance of $\rho_{\text{IQA}} = \rho_{\text{AQA}} = 1$. At $d = 0.5$ the performance of the top feature combination would be twice as close to the theoretical maximum correlation for both tasks as the average random sample. Table 4.1 contains the joint performances of both the sampling strategies, as well as the respective d for all n considered. The performance improvement achieved by training on perceptual features is most notable at smaller numbers of feature n, as is expected. When training on less than 0.1% of the features the perceptual feature selection strategy cuts the distance to the theoretical maximum performance in half.

Additionally, Figure 4.10 depicts a plot of d for all n, as well as logarithmic functions fit to all data (in blue) or all data points where $n > 1$ (in red). Since the probability distribution for $n = 1$ is heavily skewed according to the IQA performance, as can be seen in Figure 4.9 on the left, the average is not an accurate representation as an expected performance value. If we omit $n = 1$ in the function fit, as shown by the red fit, the coefficient of determination r^2 exceeds .98, indicating a good fit.

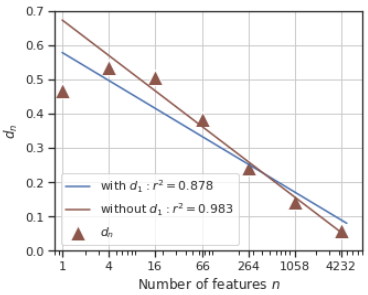


Figure 4.10: Fitting a logarithmic curve to the distance metric d with or without the value for single feature performance at $n = 1$.

Table 4.1: Prediction performance comparison on joint perceptual tasks of randomly sampled features and most perceptually relevant features. The "rand" and "perc" columns denote the joint perceptual performance in the form of Equation 4.1, while d is the norm reduction to the theoretically best performance as shown in Equation 4.2.

n	rand	perc	d
1	0.093	0.713	0.469
4	0.260	0.879	0.537
16	0.604	1.014	0.506
66	0.877	1.083	0.383
264	1.009	1.107	0.243
1058	1.067	1.117	0.144
4232	1.097	1.115	0.058

4.3.2 Comparison of Task Specificity

So far we have looked at feature combinations that were based on joint task importance. However, as shown in Figure 4.8, the different tasks favored particular features differently. The question arises, how the joint

	IQA			AQA		
n	aux	perc	nat	aux	perc	nat
1	0.574	0.612	0.612	0.244	0.244	0.328
2	0.588	0.632	0.628	0.273	0.329	0.386
4	0.672	0.677	0.732	0.361	0.408	0.442
8	0.735	0.751	0.772	0.394	0.458	0.503
16	0.773	0.805	0.822	0.471	0.506	0.564
33	0.813	0.851	0.850	0.531	0.552	0.592
66	0.845	0.869	0.864	0.583	0.603	0.624
132	0.860	0.875	0.875	0.612	0.638	0.648
264	0.867	0.880	0.879	0.636	0.657	0.668
529	0.871	0.882	0.881	0.649	0.670	0.679
1058	0.874	0.883	0.883	0.661	0.683	0.687
2116	0.878	0.881	0.882	0.671	0.686	0.688
4232	0.879	0.880	0.880	0.677	0.687	0.688
8464	0.878	0.878	0.880	0.684	0.687	0.687
16928		0.875			0.687	

Table 4.2: Performance of auxiliary ("aux"), perceptual ("perc"), and native ("nat") feature combinations for IQA and AQA.

best features compare to features useful for a particular task. Moreover, it is interesting how a set of features that was selected based on one task performs on the other task. To simplify the description, we introduce terminologies to refer to different ways of selecting features. First, we will refer to features that were selected based on a particular task and used in the same task as *native* features. In contrast, *auxiliary* features are those that were selected according to the performance on the opposite task. For example, the top 8 features selected based on the feature importance from an IQA task are called auxiliary when used to train a model to predict AQA task. Finally, features selected based on the joint performance on both task are referred to as *perceptual* features. It is important to note, that this is not the same as cross-testing. The models are trained and tested on the same task. In the case of the auxiliary features the other task is only used to select a subset of features that are used as a basis for training.

For each of the two target tasks we evaluated the top features for all three selection cases. Table 4.2 summarizes the results for each task, separately, indicating the native, auxiliary, and perceptual feature selection criteria as "nat", "aux", and "perc", respectively. With XGB models trained on the entire feature vector using the "optimized" parameter set described in Table 4.4, the theoretical maximum SROCC of 0.88 for IQA and 0.69 for AQA is achieved. For both tasks, 10% of the top perceptual features achieve comparable results. In the case of IQA, even less than 1% of the features ($n = 132$) achieve 0.88 SROCC. With smaller n we can observe the single most important perceptual feature already achieving a performance of 0.61 SROCC on IQA. On the AQA task, however, it only performs up to 0.24 SROCC, indicating that the AQA task is more complex based on Inception features.

4.3.3 Cross-Task Performance

To further evaluate the generalization performance of models, it has become common practice in the IQA domain to train on one dataset and test on another. Since the sources of data, as well as the subjective

parameter	initial	optimized
lr	0.1	0.05
estimators	200	1500
max depth	6	6
gamma	10	1
reg alpha	0.3	0.3
colsample tree	0.5	0.5
subsample	0.5	0.6
sampling method	gradient based	
objective	squared error	
eval metric	rmse	
early stopping	-	5

Table 4.4: Model parameters used for the initial estimation of per-task feature importance, as well as the set of parameters obtained via hyperparameter optimization that was used in the evaluation of those models trained on the subset of features selected.

annotation study design and annotators are different between the datasets, these cross-tests reveal more of the capabilities of the models to judge IQA in the wild. Both [TM18] and [HGS19] performed cross-task tests, meaning that they trained their models on the AQA task using AVA and tested it on IQA datasets. Indeed, as already pointed out in Section 4.2.1, some literature showed that low technical quality images would often be judged as less aesthetic than higher technical quality versions of the same content. This is one interpretation for the circumstance, that cross-task testing from AQA to IQA seems to work quite well (0.6 SROCC [HGS19]), while the opposite direction of training on an IQA dataset and testing on an AQA dataset gives much lower performance (0.20 SROCC [TM18]).

Given that our feature selection incorporates information from both an IQA and AQA task, the resulting features will best predict the intersection of the two. This intersection emphasizes those features that connect the domains. From the previous experiments we derived, that the top perceptual features perform better on IQA than on AQA. Table 4.3 summarizes cross-task testing performances for our XGB models trained on perceptual features. With $n = 264$ perceptual features trained on AVA and cross-tested on KonIQ we achieve state-of-the-art performance at 0.60 SROCC. Although the performance on AVA itself is not at its peak at 0.66, the model can leverage the close relationship between the two tasks to predict IQA fairly accurately, without ever having been trained on the task itself. Interestingly, the reverse cross test is at peak performance at $n = 264$, with an SROCC of 0.316. This is also state of the art for the KonIQ to AVA cross-test, which was previously set at 0.20 [TM18]. As the number of features is further increased, the performance drops again.

Table 4.3: Performance of perceptual feature combinations in AQA→IQA (A→I) and IQA→AQA (I→A) cross-task tests.

n	A → I	I → A
1	0.582	0.289
4	0.583	0.293
16	0.587	0.301
66	0.584	0.298
264	0.599	0.316
1058	0.627	0.303
4232	0.629	0.227

5 Video Quality Prediction

One of the challenging and fundamental problems in the field of image and video processing is the evaluation of the quality of digital media items. Accurately estimating image and video quality has widespread practical applications, such as the optimization and monitoring of visual information acquisition and transmission systems or efficient image and video compression, storage, and presentation. Although signal fidelity measures like the mean square error (MSE) or the peak signal-to-noise ratio (PSNR) are simple and well defined approaches to evaluating technical image quality, they often times do not correlate well with self-reported perceived visual quality of a human observer. This perceptual component to image and video quality is an additional important and highly complex factor that is not fully understood and difficult to model, especially when dealing with temporal components in videos. Consequently, subjective assessment of digital media items by human observers is still a common benchmark for image video quality assessment.

In this chapter we summarize the field of objective visual quality assessment, including a short description of the human visual system, as some methods derive their functionality from it.

This chapter contains and extends material from the following publications. Please refer to Section 1.4 for the contribution clarification.

[Göt+21] **Franz Götz-Hahn**, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 'KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild'. In: *arXiv preprint arXiv:1912.07966* (2021)

[GHS21] **Franz Götz-Hahn**, Vlad Hosu, and Dietmar Saupe. 'Critical analysis on the reproducibility of visual quality assessment using deep features'. In: *arXiv preprint arXiv:2009.05369* (2021)

5.1 Objective Visual Quality Assessment

The goal of objective visual quality assessment is the design of mathematical models that make a prediction of the quality of an image or video which correlates with the subjective quality evaluation of a human observer. Broadly, it can be categorized based on the availability of a reference image of pristine quality. Most of the classical objective quality measures proposed in the literature assumes that this reference image is fully available when evaluating a distorted versions of it. This class of metrics is called *full-reference* visual quality assessment, and resembles a relative similarity measure, since the output is only meaningful in relation to the reference. Although it has been the focus for research for the better half of the 20th century, this way of estimating the relative quality of an image pair is impractical in many real world applications, simply because the transmission of the reference in its entirety is too costly, as is the case in e.g. real-time video streaming applications. In this case, particular features of the reference can be extracted and transmitted in an ancillary channel, such that they can be incorporated in the quality estimation of the distorted version. This class of metrics is known as *reduced-reference* visual quality assessment.

However, in a lot of practical applications a reference image is not available at all. This task, called *no-reference* visual quality assessment, although rather easily handled by a human observer, poses a supreme challenge for objective methods. The model must evaluate the quality of a given item without additional information about what a pristine version of it might look like, or, indeed, what a good or bad quality version of any item might look like.

It is this problem that we are faced with when evaluating videos in-the-wild. In the worst case, the videos we aim to predict the quality of have been captured using an unknown device, encoded and re-encoded an unknown number of times using encoders and encoding parameters unknown to the method, and, therefore, have unknown artifacts present within them. Even in the best case, where the metadata contains information on the camera and parameters to the single encoder that was used, some unknown quality degradations remain.

For example, the quality of a video depicting a confetti shower will most likely be very poor, due to most conventional video encoders not being able to handle many small object moving relatively quickly in many different directions. Of course, we could specifically include an estimator that assesses the likelihood of the video containing a confetti shower, but this quickly turns into a sisyphean task, where we need to anticipate all deficiencies of the image and video processing pipeline and construct systems that detect them accurately. Nonetheless, certain systems that measure, for instance, blockiness or saturation of a still image, or shakiness of a video can be very useful in handling most commonly encountered scenarios that alter perceived video quality.

For this reason we will briefly summarize the most relevant application-specific no-reference IQA and VQA measures, as well as those that aim to predict general image and video quality.

5.1.1 State of the Art

Existing NR-VQA methods can be differentiated based on whether they are based solely on spatial image-level features or also explicitly account for temporal information. In general, however, all recently developed models are learning-based.

Image-based NR-VQA methods are mostly based on theories of human perception, with natural scene statistics (NSS) [Sri+03] being the predominant hypothesis used in several works, such as the naturalness image quality evaluator (NIQE) [MSB12], blind/referenceless image spatial quality evaluator (BRISQUE) [MMB12], feature-map-based referenceless image quality evaluation engine (FRIQUEE) [GB17] and high dynamic-range image gradient-based evaluator (HIGRADE) [Kun+17]. NSS hypothesizes that certain statistical distributions govern how the human visual system processes particular characteristics of natural images. Image quality can be derived by measuring the perturbations of these statistics. The approaches above have been extended to videos by evaluating them on a representative sample of frames and aggregating the features by averaging.

Approaches that consider temporal features, so-called general-purpose VQA methods, are less numerous and more particular in their approach. In [SBC14], the authors extended an image-based metric by incorporating time-frequency characteristics and temporal motion information of a given video using a motion coherence tensor that summarizes the predominant motion directions over local neighborhoods. The resulting approach, coined V-BLIINDS, has been the de facto standard that new NR-VQA methods are compared with.

Apart from V-BLIINDS, several other machine-learning-based models for NR-VQA have been proposed. Regrettably, most have only been evaluated on older datasets such as LIVE-VQA, making comparisons across multiple datasets difficult. Moreover, their codes are not publicly available, further exacerbating this issue. The three most notable examples are the following. V-CORNIA [Xu+14] is an unsupervised frame-base feature-learning approach that uses Support Vector Regression (SVR) to predict frame-level quality. Temporal pooling is then applied to obtain the final video quality. SACONVA [Li+16] extracts feature descriptors using a 3D shearlet transform of multiple frames of a video, which are then passed to a 1D CNN to extract spatio-temporal quality features. COME [WSZ18] separated the problem of extracting spatio-temporal quality features into two parts. By fine-tuning AlexNet on the CSIQ dataset, spatial quality features are extracted for each frame by both max pooling and computing the standard deviation of activations in the last layer. Additionally, temporal quality features are extracted as standard deviations of motion vectors in the video. Then, two SVR models are used in conjunction with a Bayes classifier to predict the quality score.

TLVQM [Kor19] and 3D-CNN+LSTM [YK19] are recently published approaches in blind VQA which claim state-of-the-art performance. The former is a hierarchical approach for feature extraction. It computes two types of features: low complexity features characterizing temporal aspects of the video for all video frames, and high complexity features representing spatial aspects. High complexity features relating to spatial activity,

exposure, or sharpness, are extracted from a small representative subset of frames. TLVQM achieves the best performance on LIVE-Qualcomm. 3D-CNN+LSTM is an end-to-end DNN approach, where 32 groups of 16 224×224 crops of frames are extracted from the original video and individually fed into a 3D-CNN architecture that outputs a scalar frame-group quality. This is then subsequently passed to an LSTM that predicts the overall video quality. This approach sets the state-of-the-art for KoNViD-1k, besting TLVQM slightly.

State-of-the-art for CVD2014 is achieved by VSFA [LJJ19], which is an approach that leverages feature extraction at the head of a ResNet-50 model for each frame of a video. For each video, all frame features are fed into a recurrent neural network, with the aim of modeling temporal dependencies in the frame-wise features. The approach was designed specifically for quality assessment of in-the-wild videos.

Finally, PVQ [Yin+21] is the most recent approach to blind VQA that marks state-of-the-art performance on the LIVE-VQC dataset. It combines frame-level feature extraction using PaQ-2-PiQ [Yin+20] with spatio-temporal feature extraction on patches of frame stacks using a 3D ResNet-18 [HKS17] pre-trained on the Kinetics dataset [Kay+17]. Both the frame-level features as well as the 3D features are pooled twice independently, before being fed into the InceptionTime [Faw+20] model that is used to predict the quality of a given video.

There has been a body of work by another author on NR-VQA [VS19; Var19; Var20]. However, there are concerns about the validity of the published performance values [GHS21]. Specifically, it has been shown that the performance values reported in both [VS19] and [Var19] were obtained with implementations containing some forms of data leakage. In both cases, the fine-tuning stage of the two-stage process embedded information about the test sets into the model used for feature extraction. Furthermore, in [GHS21] it was shown that fine-tuning prior to feature extraction had much less impact on the final performance than claimed. Since [Var20] is using a similar two-stage approach involving fine-tuning and feature extraction and there is a substantial improvement in performance from the non-fine-tuned to the fine-tuned implementation, we hold some reservations as to the validity of the reported performance values.

5.1.2 Data Leakage in Related Work

Modern algorithms that predict an image or video's visual quality depend on machine learning techniques such as support vector regression or deep neural networks, trained in a supervised fashion. For this purpose, benchmark datasets are commonly split three-fold into a training, validation, and test set. The training set is understood as "a set of examples used to fit the parameters of a classifier" [Rip07] or regressor. The validation set is used as a stopping criterion of the learning process and a means for selecting the optimal model hyperparameters, such as the number of layers or hidden units in a neural network layer. A model with its parameters fitted on the training set is tested on the validation set to estimate its generalization performance. While the validation set is chosen to produce an unbiased initial estimation of the generalization

performance, with every evaluation and change in the hyperparameters towards the optimal performance, more information about the data in the validation set is incorporated into the model.

Finally, the test set performance remains the only unbiased estimate. It is used solely for the performance evaluation of a fully trained model. This means, only once the model's configuration is finalized, the test set is employed to determine the model's generalization performance and, therefore, its real-world applicability. For this purpose, it is paramount that adjustments of a machine learning model are never based on its evaluation on the test set, as the test set's purpose is then lost by influencing the model's performance on itself.

To ensure the validity of this procedure data is commonly first sampled randomly without replacement according to some ratio for the test set, followed by the validation set in the same fashion. The remaining data is left for training. For small datasets, a typical split is 60/20/20% for training, validation, and testing, respectively, while larger datasets often use smaller validation and test set sizes. The performance of a quality predictor is then primarily measured by the Pearson linear correlation coefficient (PLCC) or the Spearman rank-order correlation coefficient (SROCC) of the predictions with the ground-truth qualities in the corresponding benchmark datasets.

Reproducibility and explainability are machine learning topics that gained increased traction in recent years. Various surveys have shown that a vast majority of papers do not make their code available [Hut18; GK18] and nearly half of them do not include pseudocode, either. Moreover, the simple inclusion of pseudocode does not guarantee reproducibility [Raf19].

Despite the efforts mentioned above to validate and test independently, data leakage is considered by many experts as one of the biggest problems in machine learning. It is a primary culprit for irreproducibility. Data leakage in machine learning relates to training a model on information that should only be available at test time. One of the simplest ways data leakage can occur is when the target itself is used as an input to the model. However, data leakage can manifest in machine learning in many subtle ways, such as being introduced in several different training procedure stages. In fact, one could argue that the definition of data leakage could be elaborated to mean any kind of information flow between data used in any part of the machine learning pipeline, such as information of the validation set being available at training time. This kind of data leakage should conceptually not improve the performance on an independent test set, but in nature it is a problem very similar to what is understood as classical data leakage. Therefore, it becomes increasingly difficult to spot data leakage when multiple processing steps are involved, or statistical information is extracted during pre-processing.

For example, if one considers the entirety of a dataset when normalizing it, the information would be leaked between the training, validation, and test sets. Afterward, performing cross-validation might unintentionally change the estimated performance on the validation and test sets, although the impact might be relatively small. Instead, one should first

normalize the training set, represent the transformation parameters independent of the data, and apply the same parameterized transformation to the validation and test sets before other learning algorithms are used.

Generally, state-of-the-art machine learning approaches for IQA and VQA are marked by small, incremental improvement. In contrast, five recent papers showed remarkable progress for deep learning models for IQA and VQA and certainly deserve special attention in the field. In this contribution, we provide a study on the validation and reproducibility of these existing findings. However, our results turn out to be negative in that the existing findings are found to be irreproducible. The problems with the questionable contributions stem from adequately training machine learning models to predict data and validating their expected performance correctly.

In the following, we share and discuss these five data leakage cases in the visual quality assessment domain. We summarize the papers that proposed similar image and video quality prediction approaches and describe the subtle ways data leakage caused overly optimistic results that do not hold under scrutiny and careful reimplementation. Specifically, we discuss three IQA publications in [Bia+18], [VSS18], and [Var20], as well as [VS19] and [Var19] from the VQA domain. We henceforth refer to the approaches described in these publications as DeepBIQ, DeepRN, MultiGAP-NRIQA, CNN-SVR, and CNN-LSTM, respectively, in order to increase readability.

The concept of the first four very similar approaches can be broken down into the three stages of fine-tuning, feature extraction, and quality prediction. Depending on the final predictor, an additional step of feature aggregation may be required. Figure 5.1 is a high-level representation containing the broader differences between the methods, ranging from different inputs to the network, to the features that are extracted from the network, as well as the way features are aggregated to serve as an input to the final regressor. In the following, we will first describe the three separate stages and outline the difference and then discuss the differences to [VSS18] separately. For clarity, whenever we reference frames, we are relating it to CNN-SVR and CNN-LSTM, whereas the term images refers to DeepBIQ, DeepRN and MultiGAP-NRIQA.

For the individual investigations of each pulication we refer the inclined reader to the full paper [GHS21], where we report the complete corrected results for the approaches, the reverse-engineering process, and reconstruct the mistakes that have likely resulted in the incorrect published performance numbers. The conclusions drawn from this analysis can be summarized as follows:

The DeepBIQ method does not yield the level of performance claimed in the paper [Bia+18]. On LIVE-in-the-wild, an IQA dataset, the SROCC on test sets are only 0.76 ± 0.02, instead of the 0.89 claimed, and on TID 2013, an artificially degraded IQA dataset, the model performs with 0.64 ± 0.05 SROCC, instead of the 0.96 claimed in the paper. The discrepancy between these results and the true values can be attributed to a case of data leakage, where the model's fine-tuning step illegitimately had access to data from the test set.

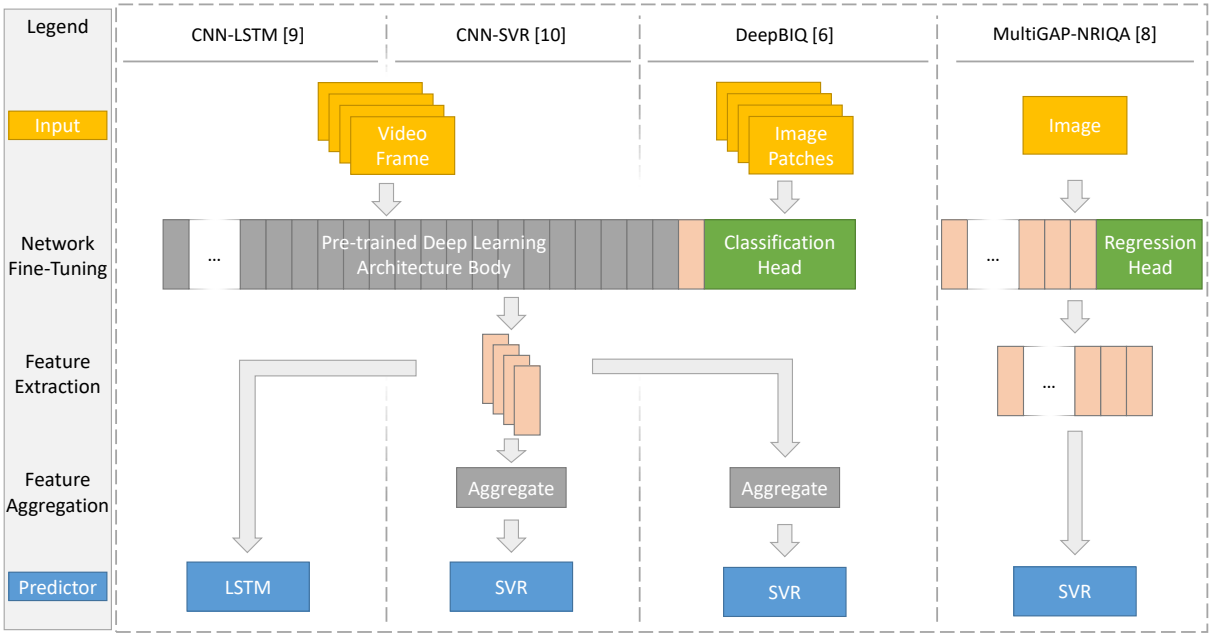


Figure 5.1: This is a high-level flowchart of the procedures employed in four of the five referenced visual quality papers. Video frames, image patches, or images are input into a pre-trained deep learning network with a classification or regression head replacement. The entire network is fined-tuned and then used as a feature extractor. The approaches differ by using the last layer or all layers as a feature source. The feature representations are then aggregated, where appropriate, and used to train the final quality predictor.

In both CNN-SVR and CNN-LSTM, which are very close adaptations of DeepBIQ for the field of VQA, a similar case of data leakage as in DeepBIQ occurs, despite contrasting claims. Furthermore, an additional case of data leakage occured, where the validation set used for fine-tuning was not properly separated from the training set. They, therefore, do not yield the performances as claimed. On KoNViD-1k, a large-scale VQA dataset, the SROCCs on test sets are only 0.67 ± 0.04 and 0.63 ± 0.05, respectively, instead of 0.85.

MultiGAP-NRIQA, an enhanced version of DeepBIQ suffers from a different kind of data leakage causing illegitimate performance values for two artificially degraded IQA datasets, KADID-10k, and TID2013. On KADID-10k, the SROCCs on test sets are only 0.81 ± 0.05, instead of 0.97 as claimed.

The published performance of DeepRN also cannot be reproduced. The introduction of the simple types of data leakage the author revealed in personal communication to have happened does not explain the published results.

5.2 Video Quality Assessment based on Multi-Level Spatially Pooled Features

Our proposed NR-VQA approach of extracting features from a pre-trained classification network and training DNN architectures on them has been designed to predict video quality in-the-wild. We evaluate the potential of the MLSP features when used for training the shallow feed-forward and recurrent networks by measuring their performance

on four widely used datasets (KoNViD-1k, LIVE-VQC, CVD2014, and LIVE-Qualcomm) and our newly established dataset KonVid-150k. We consider two basic scenarios, namely (1) intra-dataset, i.e. training and testing on the same dataset, and (2) inter-dataset, i.e., training (and validating) on our large dataset KonVid-150k and testing on another.

There are two fundamental limitations in these datasets that affect the performance of our approach. The first one relates to the video content, in the form of domain shifts between ImageNet and the videos in the datasets. The other one is due to the different types of subjective video quality ratings (labels) in the datasets, that may affect the cross-testing performance.

First, the features in the pre-trained network have been learnt from images in ImageNet. There are situations when the information in the MLSP features may not transfer well to video quality assessment. Namely, some artifacts are unique to video recordings; this is the case of temporal degradations such as camera shake, which does not apply to photos. Moreover, compression methods are different for videos in comparison to images. Thus, the individual frames may show encoding-specific artifacts that are not within the domain of artifacts present in ImageNet. Lastly, in-the-wild videos have different types and magnitudes of degradations compared to photos. For example, motion blur degradations can be more prevalent and of a higher magnitude in videos compared to photos. This could affect how well MLSP features from networks pretrained on ImageNet transfer to VQA.

The second limitation is concerned with the subjective video quality ratings to be predicted when cross-testing. Although there are similarities between the rating scales used in the subjective studies corresponding to each dataset, the ratings themselves may suffer from a presentation bias. For example, in the case of a dataset with highly similar scenes, but minuscule differences in degradation levels, as is the case for LIVE-Qualcomm and CVD2014, a human observer may become very sensitive to particular degradations. Conversely, video content becomes less critical for quality judgments. The attention of the human observer is diverted to parts in the video he might otherwise not have looked at, had he not seen the same or a very similar scene many times before. Whether the resulting subjective judgments can be regarded as fair quality values is arguable. A human observer would rarely watch a scene multiple times before rating the quality. This bias of subjective opinions may greatly influence how the quality predictions trained in one setting generalize to others. Similarly, quality scores obtained in a lab environment will be much more sensitive to differences in technical quality than a worker in a crowdsourcing experiment might be able to pick up. Therefore, it may be challenging to generalize from one experimental setup to another. While consumption of ecologically valid video content happens in a variety of environments and on a multitude of devices, it is arguable whether one experimental setup is superior.

Different learning-based regression models, such as Support Vector Regression (SVR) or Random Forest Regression (RFR), have been employed to predict subjective quality scores from frame features, with SVR yielding generally better results [Kor19]. However, most existing works only extract a few dozen to a few hundred features. Since SVR is sub-optimal

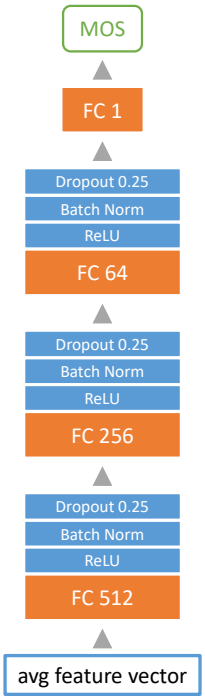


Figure 5.2: The MLSP-VQA-FF model, that relies on average frame MLSP features and a densely connected feed forward network.

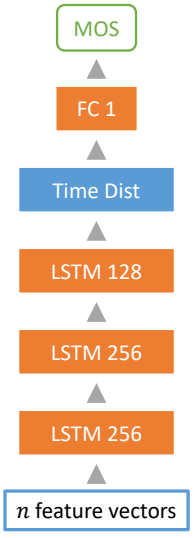


Figure 5.3: The MLSP-VQA-RN recurrent model, implementing a stacked long short-term memory network. This model takes corresponding frame features at each time step as an input to the network.

when applied to very large dimensional features like our MLSP feature, we instead train three small-capacity DNNs.

FF (Figure 5.2) A feed-forward DNN where the average feature vector is the input of three blocks of fully connected layers with ReLU activations, followed by batch normalization and dropout layers.

RN (Figure 5.3) A deep Long Short-Term Memory (LSTM) architecture, where each LSTM layer receives the feature vector or the hidden state of the lower LSTM layer as an input and outputs its hidden state. This stacking of layers allows for the simultaneous representation of input series at different time scales [HS13]. The bottom LSTM layer can be understood as a selective memory of past feature vectors. In contrast, each additional LSTM layer represents a selective memory of past hidden states of the previous layer.

HYB (Figure 5.4) A two-channel hybrid of both the FF and RN variants. The temporal channel is a copy of the RN model's architecture, while the second channel is a mirror of the FF network scaled up to match the number of kernels in the temporal branch in the last layer. The outputs of the two channels are concatenated and a small 32 kernel fully connected layer feeds into the last prediction layer.

Our tests showed that employing dropout of any kind within the recurrent networks, such as input/output dropout or recurrent dropout, resulted in reduced performance. We, therefore, do not employ any dropout in these architectures.

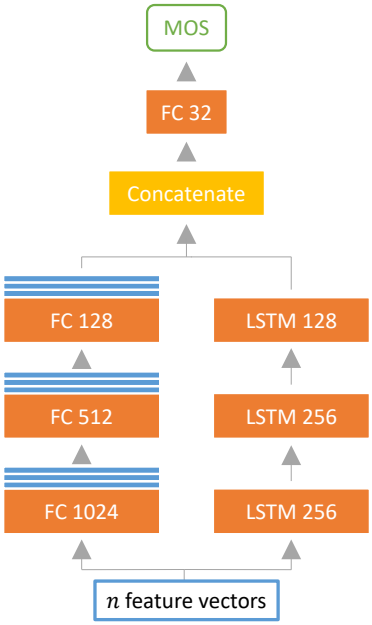


Figure 5.4: The hybrid MLSP-VQA-HYB dual channel model, that has a bigger variant of the FF network on the left and the recurrent part of the RN network on the right. Both channels output activations at each timestep and are merged along the feature dimension, before feeding into a small prediction head. This model takes corresponding frame features at each time step as an input to the network.

5.2.1 Comparison to Transfer Learning

The appropriation of features obtained by training for a particular task that is conceptually related to the target task is called transfer learning. The naïve way to perform transfer learning for tasks related to visual features with small sets of data is removing the head of a pre-trained base-model and replacing it with a small fully connected head. By freezing the layers in the base-model it's predictive power can be used to perform well on the new task. After training this new header, it is not uncommon to unfreeze all layers and fine-tuning the entire trained network with a low learning rate to improve predictive power even more. However, this approach has three important downsides.

First, the new task is trained based on the highest level features in the base-model. These features are particularly tuned to detecting high-level semantic features that are useful in the detection of objects present in the image. However, for tasks such as quality, low-level features with a small receptive field are arguably more important.

Secondly, for each forward and backward pass the entire base-model has to be present in memory, which contain many more weights than the header network that is being trained. Consequently, training is slowed down a lot.

Finally, the last fine-tuning step is prone to overfitting, as the high capacity of the base-model alone allows the network to memorize training data rather than extracting useful general features. Careful hyperparameter tuning is therefore required, to ensure this step is successful in improving performance.

The two-step strategy of feature extraction followed by training a regressor is similar to the classical way of performing transfer learning, but much faster. However, it is difficult to fairly assess the difference, as a lot of factors play a role. For example, when fine-tuning an Inception-net, the speed at which the videos are read from the hard-drive can become a bottle-neck, if a very powerful GPU is performing the training procedure. Our proposed approach with an Inception-ResNet-v2 as a feature extraction network has a benefit for this scenario. Since the input data for each frame is fixed at 16,928 floating point values, the requirements for hard-drive reading speed are not exacerbated when using datasets with larger resolution videos. In contrast, if the GPU used to perform the training is not as powerful, it itself can become a bottle-neck of the system. In this case, our proposed approach has the alternative benefit that the small network size allows for much larger batches and quicker forward and backward passes.

In order to quantify the difference, we compare different setups of transfer learning and fine-tuning to our proposed two-step MLSP feature-based training procedure on a machine that reads from an NVMe connected SSD and trains the networks using Tensorflow 2.4.1 on an NVIDIA A100 with 40GB of VRAM. To simplify the setup, we are evaluating only the MLSP-VQA-FF model on the pre-extracted first frames of KonVid-150k-B. One might argue that the first frame is not as representative of the opinion scores, but our aim is to investigate the differences in training speed, rather than an exhaustive performance evaluation. The transfer learning scenarios are all performed using an Inception-ResNet-v2 base-model with our FF model sitting on top for 40 epochs. However, we compare four slightly different scenarios:

Koncept The FF model takes the last layer of the base-model as an input, much like the Koncept model proposed in [Hos+20a]. The weights of the base-model are not frozen, so the entire model is fine-tuned over the course of the training. We employ two training stages, one with a learning rate of 1×10^{-3}, and the second with a learning rate of 1×10^{-5}.

IRNV2 Instead of fine-tuning the entire model throughout both stages, we freeze the layers of the Inception-ResNet-v2 base-model for the first stage, so as to avoid the large update steps caused by the random initialisation of the header network to destroy the useful features in it. For the second stage we unfreeze the weights in all layers.

IRNV2-MLSP As stated before, one downside of the above approaches lies in the circumstance that the header network relies only on the top level features as inputs. For the third comparison we concatenate the activation layers of all Inception-modules and feed that as an input to the header network. Here, we also freeze the base-model weights for the first stage, and unfreeze all weights for the second stage.

MLSP The final item in the comparison takes the MLSP features described above as an input. This means, the model is much smaller, as the base-model does not need to be loaded. However, the model can not leverage the spatial information about the activations to make it's prediction. No explicit weight freezing is performed in this scenario.

These different cases are compared in Figure 5.5. The green graph,

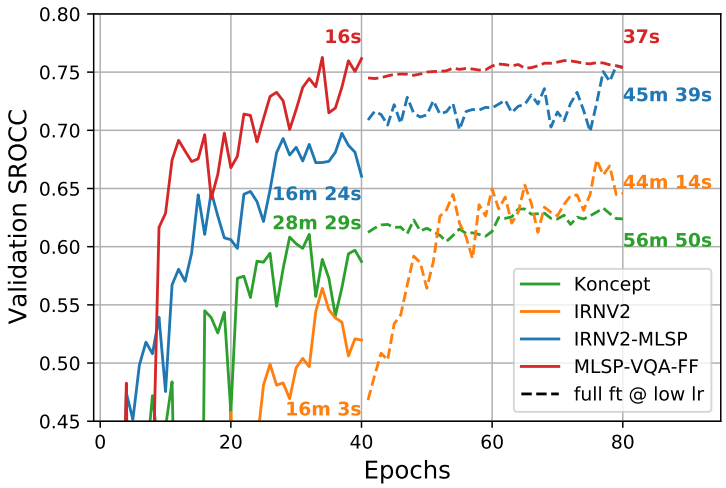


Figure 5.5: A visualization of the convergence of different transfer learning techniques along with information about the training times. The solid lines depict the first training stage of 40 epochs, where the IRNV2 (orange) and IRNV2-MLSP (blue) architectures have their weights frozen. The dashed lines represent the second training stage of 40 epochs where all models had their weights unfrozen. For the second stage we start from the best performing model according to validation loss from the previous stage. This is the reason for the discontinuities between the graphs. Koncept (green) and IRNV2 connect the last layer to the small header network, while IRNV2-MLSP concatenates all individual Inception-module outputs to feed into the head. Finally, MLSP-VQA-FF works off of extracted MLSP features, which for this scenario took 38 seconds.

corresponding to the Koncept model, takes the longest to train in total and achieves the worst validation performance at the end of the 80 epochs. The reason for the slow training in the first stage is that none of the weights are frozen and the backpropagation step therefore takes additional time. Both the orange IRNV2 and blue IRNV2-MLSP models train faster by approximately 22%, as the weights are frozen in the first stage. However, they differ in that the inclusion of all Inception-modules in the concatenation layer for the latter increases performance significantly. Finally, the red graph, representing the MLSP-VQA-FF model trained on extracted MLSP features achieves the best performance while beating the IRNV2-MLSP model in terms of speed by factor 74. Moreover, peak performance is achieved much earlier, as the second training stage is not required, raising the speed-up to factor 171.

However, feature extraction has to be performed once as well, which for the first frames of KonVid-150k-B took 38 seconds. Including this time in the comparison still renders the MLSP-VQA-FF model faster by factor 36, when considering both training stages. This factor is dependant on input resolutions, however with videos increasing in resolution the speed-up will only change in favor of the MLSP-based model, as its training speed will not change, while the training speed of the fine-tuning approach is inversely correlated with input resolution. This shows the power of using pre-extracted MLSP features.

Furthermore, we have observed the success of fine-tuning an Inception-style network in this manner is very sensitive to hyperparameters, while training the small FF network on MLSP features is fairly robust.

Table 5.1 gives an overview of some hyperparameter settings used in the training of our MLSP-based models for the compared datasets. Mean square error (MSE) was used as a loss function for a duration of 250 epochs, stopping early if the validation loss did not improve in the most recent 25 epochs at an initial learning rate of 10^{-4}. By default, the MLSP-VQA-FF model was trained with a learning rate of 10^{-2}, and both the MLSP-VQA-RN and the MLSP-VQA-HYB models were trained with a learning rate of 10^{-4}.

Type	MLSP-VQA-FF frames	bs	lr	MLSP-VQA-RN/-HYB frames	bs	lr
KoNViD-1k	all	128	10^{-2}	180	128	10^{-4}
LIVE-Qualcomm	all	8	10^{-3}	150	8	10^{-4}
CVD2014	all	8	10^{-3}	140	8	10^{-4}
LIVE-VQC	all	8	10^{-3}	150	8	10^{-4}
Proposed	all	128	10^{-2}	180	128	10^{-4}

Table 5.1: Training settings and parameters for the MLSP-VQA models.

5.2.2 Evaluation

We first evaluate the performance of the proposed model on four existing video datasets. KoNViD-1k and LIVE-VQC both pose the unique challenge that they are in-the-wild video datasets, containing authentic distortions that are common to videos hosted on Flickr. LIVE-Qualcomm contains self-recorded scenes of different mobile phone cameras that were aimed at inducing common distortions. CVD2014 differs from the previous two, in that it is a dataset with artificially introduced acquisition-time distortions. It also contains only five unique scenes depicting people. Finally, LIVE-VQC was a collaborative effort of friends and family of the LIVE research group that were asked to submit video files of a variety of contents to capture diversity in capturing equipment and distortions.

We are comparing our proposed DNN models against published results for other methods that have been thoroughly evaluated on these datasets using SVR and RFR. Detailed information regarding the experimental evaluation and results of the classical methods can be found in [Kor19].

We adopt a similar testing protocol by training 100 different random splits with 60% of the data used for training, 20% used for validation, and 20% for testing in each split. Table 5.2 summarizes the SRCC with respect to the ground-truth for the predictions of the classical methods (taken from [Kor19]) alongside our DNN-based approach. It is to be noted that the random splits we used are different from the ones used to evaluate the classical methods in [Kor19]. For brevity, we are only reporting the results for classical methods obtained using SVR, although four individual results are slightly improved using RFR.

The FF network outperforms the existing works on KoNViD-1k, improving state-of-the-art SRCC from 0.80 to 0.82, while the RN and HYB models remain competitive with an SRCC of 0.78 and 0.79, respectively. This shows that the proposed approaches are performing close to state-of-the-art on authentic videos with some encoding degradations. Since the feature extraction network is trained on images with natural image distortions, some of the extracted features are likely indicative of these distortions, which are not unlike the video encoding artifacts introduced by Flickr.

Existing methods had not been evaluated exhaustively on LIVE-VQC at the time of writing. Our recurrent networks achieve 0.70 (RN) and 0.69 (HYB) SRCC, while the FF model performs at 0.72 SRCC, rendering it competitive with state-of-the-art for the dataset. Recently, a new publication on arXiv disusses a new approach called RAPIQUE that achieves an SRCC of 0.76 on LIVE-VQC [Tu+21]. However, this work has not yet been peer reviewed. One of the difficulties inherent to VQC with respect to

our models is the circumstance, that it is comprised of videos of various resolutions and aspect ratios. An evaluation of the performance of the models with respect to the video resolutions can be found in the top part of Figure 5.6. Since 1080p, 720p, and 404p in portrait orientation are the predominant resolutions with 110, 316, and 119 videos, respectively, we grouped the other resolutions into the *other* category. We can see that both the FF and RN models perform worse on the 1080p and 720p videos, whereas the HYB model performs better on the higher resolution videos.

In the case of LIVE-Qualcomm our best performance of 0.75 SRCC of the hybrid model is surpassed only by TLVQM with 0.78. Since the dataset is comprised of videos containing six different distortion types, we also evaluated the performance of the models according to each degradation, as depicted in the middle plot of Figure 5.6. Here, we show the deviation of the RMSE of each model for each distortion type from the average performance in percent. Little deviation between all three models is observed for both Exposure and Stabilization type distortions. However, for Artifacts and Color the RN model deviates from the other two drastically, performing worse on the former and better on the latter. Videos in the focus degradation class show auto-focus related distortions where parts of the video are intermittently blurry or sharp over time and are overall the biggest challenge for our recurrent models, that both perform over 20% worse on them than average. Finally, the Sharpness distortion is best predicted by the recurrent networks, with the hybrid model outperforming the pure LSTM network.

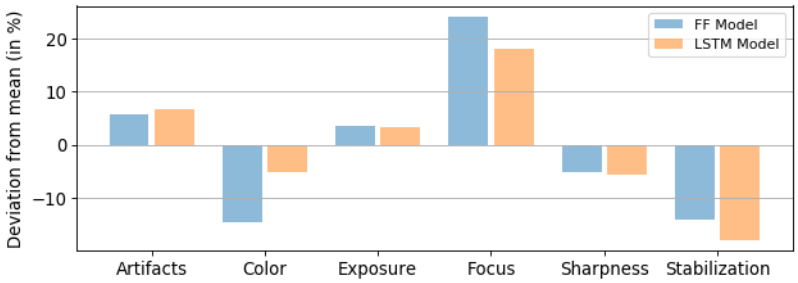


Figure 5.6: Percent deviation of the mean RMSE of the proposed models on each of the six degradation types present in LIVE-Qualcomm (top), each of the six test scenarios in CVD2014 (middle), and the different resolutions in LIVE-VQC (bottom).

On CVD2014, our proposed models with SRCCs of 0.77, 0.75, and 0.79 for the FF, RN and HYB models, respectively, are outperformed by both FRIQUEE and TLVQM at 0.82 and 0.83 SRCC. CVD2014 is a dataset of videos of two different resolutions, with artificially introduced capturing distortions and only five unique scenes of humans and human faces. The magnitude of the artifacts is at a level that is not commonly seen in videos in-the-wild, and the types of defects are also not within the domain of distortions present in ImageNet. Therefore, this is the most challenging dataset for our approach and, consequently, the relative performance of our approach is worse. CVD2014 is split into six subsets with partially overlapping scenes but distinct capturing cameras. The bottom part of Figure 5.6 shows the relative deviation of the RMSE from the mean performance for each of these test setups. The first two setups include videos at 640×480 pixels resolution, which are generally rated with a lower MOS than videos in the other test setups, which could both be an important factor in our models' increased performance here. Although all setups include scenes 2 and 3, scene 1 is only included in test setups 1

Table 5.2: Results of different NR-VQA metrics on different authentic VQA datasets

		in-the-wild		synthetic	
	Name	KoNViD-1k SRCC (±σ)	LIVE-VQC SRCC (±σ)	LIVE-Qualcomm SRCC (±σ)	CVD2014 SRCC (±σ)
SVR	NIQE (1 fps)	0.34 (±0.05)	0.56 (±–.—)	0.46 (±0.13)	0.58 (±0.10)
	BRISQUE (1 fps)	0.56 (±0.05)[1]	0.61 (±–.—)	0.55 (±0.10)	0.63 (±0.10)[1]
	CORNIA (1 fps)	0.51 (±0.04)	–.— (±–.—)	0.56 (±0.09)	0.68 (±0.09)
	V-BLIINDS	0.65 (±0.04)[1]	0.72 (±–.—)	0.60 (±0.10)	0.70 (±0.09)[1]
	HIGRADE (1 fps)	0.73 (±0.03)	–.— (±–.—)	0.68 (±0.08)	0.74 (±0.06)
	FRIQUEE (1 fps)	0.74 (±0.03)	–.— (±–.—)	0.74 (±0.07)	0.82 (±0.05)
	TLVQM	0.78 (±0.02)	–.— (±–.—)	**0.78 (±0.07)**	**0.83 (±0.04)**
DNN	3D-CNN + LSTM[2]	0.80 (±–.—)	–.— (±–.—)	0.69 (±–.—)	–.— (±–.—)
	MLSP-VQA-FF	**0.82 (±0.02)**	**0.72 (±0.06)**	0.71 (±0.08)	0.77 (±0.06)
	MLSP-VQA-RN	0.78 (±0.02)	0.70 (±0.06)	0.72 (±0.07)	0.75 (±0.06)
	MLSP-VQA-HYB	0.79 (±0.02)	0.69 (±0.07)	0.75 (±0.04)	0.79 (±0.05)

[1] Performance improves when using random forest regression.
[2] The authors did not supply any standard deviations for the performance measures, and did not evaluate the method on CVD2014.

and 2, scene 4 is only included in test setups 3 and 4, and scene 5 is solely included in test setups 5 and 6. Since the features we use are tuned to identify content, as we showed in Section 3.5, inclusion or exclusion of particular scenes can have an impact on the performance of our method. Moreover, since each test setup contains videos taken from different cameras than the rest, it is possible that the in-capture distortions caused by particular cameras in any individual test setup may be closer to the types of distortions present in ImageNet.

We now consider the performance evaluation when training and testing on our new dataset, KonVid-150k-B of 1,596 videos, each with at least 89 ratings comprising the quality score. We separate these tests from the previous ones because, in this case, we have the option to train the networks on the additional 150k videos in KonVid-150k-A that stem from the same domain. From the previous experiments, it is evident that TLVQM is the best performing classical metric on the similar domain, given by KoNViD-1k, by a large margin. Therefore, we compare our MLSP-VQA models only against TLVQM and the standard V-BLIINDS.

Table 5.3 summarizes the performance results. Compared to the performance on KoNViD-1k, V-BLIINDS (row 1) improves slightly, while TLVQM (row 2) performs significantly worse. Since the main difference between KoNViD-1k and this dataset is the reduced re-encoding degradations, it appears as though the classical methods over-emphasize their prediction on these artifacts. The third through fifth row list the performance of our models, which outperform both classical methods, beating TLVQM's 0.71 SRCC with 0.83 (FF), 0.78 (RN) and 0.75 (HYB) when trained and tested on the B variant exclusively.

Finally, the last three rows show the results from training on the large dataset, KonVid-150k-A, with 150k videos. For these last three evaluations a random subset of 50% of KonVid-150k-B was used for validation during training. The remaining part of KonVid-150k-B was used for testing. We note an additional substantial performance increase for our networks. The FF model's performance increases from 0.81 SRCC to 0.83, while the

Table 5.3: Results of NR-VQA metrics on KonVid-150k-B. The bottom three rows describe the performance when training on the entirety of KonVid-150k-A, using half of KonVid-150k-B as a validation set, and the other as a test set.

	Name	PLCC ($\pm\sigma$)	SRCC ($\pm\sigma$)	RMSE ($\pm\sigma$)
SVR	V-BLIINDS (SVR)	0.68 ($\pm$0.04)	0.68 ($\pm$0.04)	0.27 ($\pm$0.02)
	TLVQM (SVR)	0.68 ($\pm$0.12)	0.71 ($\pm$0.04)	0.26 ($\pm$0.04)
DNN	MLSP-VQA-FF	0.83 ($\pm$0.02)	0.81 ($\pm$0.02)	0.26 ($\pm$0.01)
	MLSP-VQA-RN	0.80 ($\pm$0.02)	0.78 ($\pm$0.02)	0.29 ($\pm$0.01)
	MLSP-VQA-HYB	0.76 ($\pm$0.04)	0.75 ($\pm$0.04)	0.32 ($\pm$0.03)
	MLSP-VQA-FF (Full)	0.86 ($\pm$0.01)	0.83 ($\pm$0.01)	0.19 ($\pm$0.01)
	MLSP-VQA-RN (Full)	0.83 ($\pm$0.01)	0.81 ($\pm$0.01)	0.21 ($\pm$0.01)
	MLSP-VQA-HYB (Full)	0.83 ($\pm$0.01)	0.81 ($\pm$0.01)	0.21 ($\pm$0.01)

RN model improves from 0.78 SRCC to 0.81. The largest performance gain can be observed for the HYB network, as it improves from 0.75 SRCC to 0.81 SRCC as well. This demonstrates, for the first time, the enormous potential gains that can be achieved by vast training datasets for VQA. Although KonVid-150k-A only has MOS scores comprised of five individual votes, by training on them and validating on the target dataset we drastically improve performance. It is to be noted as well that the test sets in this scenario are larger than when training and testing solely on KonVid-150k-B. This renders the test performance to be even more representative. However, the change in variance of the resulting correlation coefficients can not directly be attributed to the increase in training dataset size. The difference likely arises from the fact that the models trained using KonVid-150k-A have the same training data, and are therefore more likely to learn similar features. Nonetheless, this effect should be investigated further.

5.2.3 Inter-Dataset Performance

Considering the diversity in content and distortions in KonVid-150k we highlight the power of KonVid-150k in combination with our MLSP-VQA models in inter-dataset testing scenarios. At the time of writing, LIVE-VQC has not been considered in any performance evaluations across datasets. The previously best reported cross-test performances between the other three legacy datasets are three different combinations of NR-VQA methods and training datasets. Specifically, TLVQM trained on CVD2014 performs best on KoNViD-1k cross-testing with 0.54 SRCC. V-BLIINDS trained on KoNViD-1k is the best combination for cross-testing on LIVE-Qualcomm with 0.49 SRCC. Finally, FRIQUEE trained on KoNViD-1k performs best when cross-testing on CVD2014 with 0.62 SRCC. It is apparent from these results that no single NR-VQA and dataset combination generally outperforms in inter-dataset testing scenarios.

These results are taken from [Kor18].

We evaluate the performance of our models when cross-testing on other datasets, trained on KonVid-150k-A and validated and tested on each 50% of KonVid-150k-B. The average SRCC performances of 10 models are reported in Table 5.4. For ease of comparison we also include the best within-dataset performance in the first row, as well as the previous best cross-dataset test performances as taken from [Kor18] in the second row of the table. Although the performances between our different models do not vary much, the results reveal some interesting findings.

Table 5.4: Inter-dataset test performance of our three models averaged over 10 splits trained on the entirety of KonVid-150k-A. The different splits only affect the validation and test sets, as all videos of KonVid-150k-A are used for training.

	in-the-wild		synthetic	
	KoNViD-1k SRCC ($\pm\sigma$)	LIVE-VQC SRCC ($\pm\sigma$)	LIVE-Qualcomm SRCC ($\pm\sigma$)	CVD2014 SRCC ($\pm\sigma$)
Intra-dataset best	0.82 (±0.02)	0.72 (±0.06)	0.78 (±0.07)	0.83 (±0.04)
Prev. inter-dataset best[Kor18]	0.54 (±–.—)	–.— (±–.—)	0.49(±–.—)	**0.62 (±–.—)**
MLSP-VQA-FF	**0.83 (±0.01)**	**0.75 (±0.01)**	**0.64 (±0.01)**	0.55 (±0.02)
MLSP-VQA-RN	0.80 (±0.01)	0.71 (±0.01)	0.61 (±0.03)	0.52 (±0.02)
MLSP-VQA-HYB	0.79 (±0.01)	0.71 (±0.01)	0.62 (±0.03)	0.52 (±0.02)

- The cross-dataset test performance of the FF model on KoNViD-1k of 0.83 SRCC is higher than all other within-dataset test performances and especially any cross-test setups. This again underlines the potential power of data, even if it is annotated with lower precision. Although KonVid-150k does not have the Flickr video encoding artifacts present, it can predict the distorted videos of KoNViD-1k better than training on videos taken from the same dataset.
- Our models trained on KonVid-150k and cross-tested on LIVE-VQC achieve state-of-the art performance and even surpass the best within-dataset performance in the case of the FF model with 0.75 SRCC.
- On LIVE-Qualcomm the cross-dataset test performances of all our models are slightly better than V-BLIINDS (0.60), when it is trained and tested on LIVE-Qualcomm. Since V-BLIINDS has been the de facto baseline method, this is a remarkable result. Additionally, for a cross-dataset test our proposed KonVid-150k dataset shows the best generalization to LIVE-Qualcomm, improving the previous best 0.49 SRCC to 0.64.
- Next, our models struggle with CVD2014, as none of them beat even the most dated classical models trained and tested on CVD2014 itself. This may be in part due to the nature of the degradations induced in the creation of the dataset, which are not native to the videos present in KonVid-150k. Moreover, the domain shift between KonVid-150k and CVD2014 seems to be larger than to the other datasets, as the previous best cross-dataset performance is also not achieved.

The cross-test performance drops notably when testing on synthetic video datasets. This has already been observed in the IQA domain [LHS20], where training and testing on the same domain resulted in much higher performance than when the source and target domains were different. The types of distortions in individual frames of videos from two different domains result in different characteristics of the activations of Inception-net features, resulting in reduced performance.

5.3 Fixed Vote Budget Distribution

As described in Section 3.1.4, the choice of the number of ratings per video is a distinguishing, yet so far unexplored factor in the design of

VQA datasets in the context of optimizing model training performance. In order to study the effect of varying the number of ratings per video, we trained a large set of corresponding models in two experiments. In the first one, we increased the number of ratings to reduce the level of noise in the training set. In the second one, we additionally introduced the natural constraint of a vote budget, limiting the total number of ratings to a constant.

It is common to use an equal number of votes for each stimulus so that the MOS of the training, validation, and test sets have the same reliability, respectively, the same level of noise. Deep learning is known to be robust to label noise [Rol+17], however, this has been only studied when the same amount of noise is present for all items in all parts of the dataset (train/test/validation). Thus, the first question we investigate is:

- *What impact do different noise levels in the training and validation sets have on test set prediction performance?*

More precisely, we are interested to know the change in prediction performance when fewer votes are used for training and validating deep learning models, compared to the number of votes used for test items.

In order to answer this question, we randomly sampled $v = 1, 2, 4, 7, 14, 26,$ and 50 votes five times for each video within KonVid-150k-B and computed the corresponding MOS values (7×5 MOS per video). We then trained our MLSP-VQA-FF model by varying both training set, and validation set MOS vote counts while keeping the test set MOS vote count at 50. For each pair of training and validation MOS, we considered twenty random splits with 60% of the data for training, 20% for validation, and 20% for testing, with the above mentioned five versions of the MOS each. Therefore, we trained $5 \times 20 \times 7 \times 7 = 4900$ models in total.

The graph in Figure 5.7 depicts the mean SRCC between the models' predictions and the ground truth MOS of the test sets. Each line in this graph represents a different number of votes comprising the validation MOS, whereas the x-axis indicates the number of votes comprising the training MOS. Note that the x-axis is scaled logarithmically for better visualization. There are three key observations concerning the prediction performance:

- The prediction performance improves as the number of votes comprising the training MOS increases, regardless of the number of votes used for validation.
- The performance improvements scale approximately logarithmically with the number of votes comprising the training MOS.
- The test set performance varies less due to changes in the number of votes used for validation than it does due to the number of votes for items in the training set.

The fact that performance improves with lower training label noise is not surprising. Nonetheless, the gentler slope for the performance curves beyond four votes comprising the training MOS is an indicator that the common policy to gather 25 votes for all stimuli in a dataset may be sub-optimal, due to diminishing returns. In fact, at approximately five votes (1/10th of the analysed budget) the model achieves roughly 92% of the peak performance, suggesting it to be a good trade-off between precision and cost.

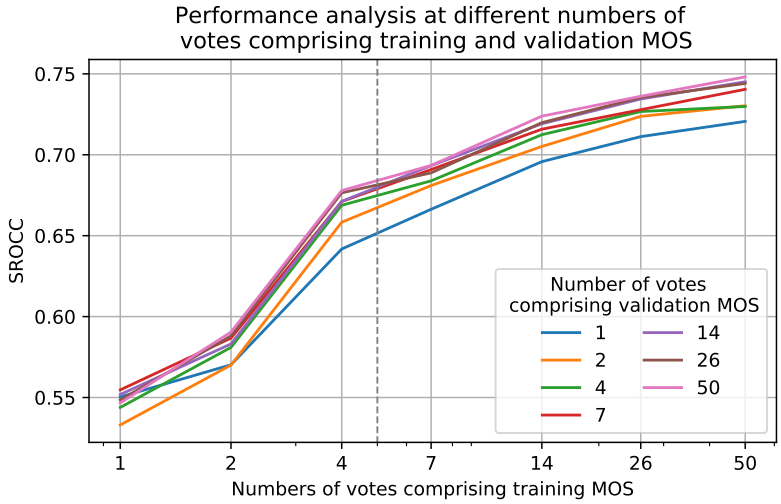


Figure 5.7: This plot summarizes the evaluation of MLSP-VQA-FF models trained on KonVid-150k-B using different numbers of votes comprising the training or validation MOS, indicated by the x axis and the color of the graphs, respectively. The y-axis shows the average of 20 models' SRCC between the predicted MOS values on the test set and the ground truth data, which is comprised of 50 votes.

The comparison between data splits in this experiment is not balanced, because the data points in the graphs of Figure 5.7 correspond to different vote budgets, ranging from 1 rating per video in one instance on the left up to 50 per video on the right. The annotation of datasets in the lab and also in the crowd usually is constrained by a budget in terms of total hours of testing or overall cost of crowdsourcing. This translates to a maximum number of votes that can be attained for a given dataset. Therefore, the second question we investigate is:

- *Given a fixed vote budget, how does the allocation of votes on the training set affect test performance?*

In other words, is it better to collect more votes for fewer stimuli, or less votes for more videos?

In order to answer this question, we first divided KonVid-150k-B into five disjoint test sets (each with 20% of all videos) and sampled the same number of videos from the remaining set of KonVid-150k-B for validation. We then considered three levels of precision at 100, 5, and 1 votes comprising the MOS of videos used in training, as well as six vote budgets of 100,000, 25,000, 10,000, 2,500, and 1,000 votes. We built the training sets accordingly, sampling from the remaining videos in KonVid-150k-B first, and then adding in videos from KonVid-150k-A, if needed, such that the smaller sets are proper subsets of the larger variants. For the vote budget of 100,000 votes we consequently created three training sets of 1,000, 20,000, and 100,000 videos at training MOS precision levels of 100, 5 and 1 vote(s), respectively. It is to be noted that the overlap between the different samples of the same sets increases as the set size increases, as the whole KonVid-150k-B set is only comprised of ≈150,000 videos, which in turn has an effect on the standard deviation of the predictions.

We trained both MLSP-VQA-FF and MLSP-VQA-RN on the five different splits for all three vote budget distributions and reported the results in Table 5.5. We give the average SRCC, PLCC, and RMSE between the models' predicted scores and the MOS computed by using all available votes. There are are few key takeaways from these results:

Set	PLCC	SRCC	RMSE
1000@100	0.76 (±0.03)	0.73 (±0.04)	0.24 (±0.01)
20000@5	0.76 (±0.02)	0.74 (±0.03)	0.24 (±0.01)
100000@1	0.77 (±0.02)	0.74 (±0.03)	0.24 (±0.01)
250@100	0.75 (±0.01)	0.70 (±0.01)	0.26 (±0.01)
5000@5	0.77 (±0.02)	0.72 (±0.02)	0.25 (±0.01)
25000@1	0.76 (±0.02)	0.72 (±0.02)	0.25 (±0.01)
100@100	0.68 (±0.03)	0.62 (±0.05)	0.28 (±0.01)
2000@5	0.68 (±0.02)	0.64 (±0.03)	0.28 (±0.02)
10000@1	0.69 (±0.06)	0.66 (±0.05)	0.28 (±0.01)
25@100	0.56 (±0.08)	0.51 (±0.07)	0.32 (±0.02)
500@5	0.59 (±0.04)	0.54 (±0.07)	0.34 (±0.02)
2500@1	0.57 (±0.04)	0.52 (±0.05)	0.36 (±0.04)
10@100	0.46 (±0.07)	0.41 (±0.09)	0.34 (±0.02)
200@5	0.55 (±0.05)	0.50 (±0.07)	0.34 (±0.02)
1000@1	0.46 (±0.12)	0.44 (±0.10)	0.45 (±0.05)

Table 5.5: Performance of our FF model at a fixed vote budget of 100,000, 25,000, 10,000, 2,500, and 1,000 votes.

- As one would suspect, the performance drops as the total vote budget decreases.
- Surprisingly, however, the performance appears to be stable across the different distribution strategies for budgets of more than 1,000 votes.
- For smaller vote budgets a middle ground choice between MOS precision and numbers of videos seems to be favorable, as indicated by the 5 vote MOS distribution strategy outperforming the more and less precise extreme strategies. This suggests that for very small vote budgets in particular the focus should be on fewer than the commonly suggested 30 rating MOS recommendations that are found in literature.

6 Conclusions

With KoNViD-1k and KonVid-150k this thesis introduced two large-scale in-the-wild datasets for VQA in-the-wild. These datasets take a novel approach to VQA and KonVid-150k in particular exceeds the competition in several aspects. Not only is it two orders of magnitude larger than previous published datasets, it is, more importantly, authentic both in terms of variety of content types and distortions, but also due to the compression settings of the videos. The original video files were retrieved as the original version that was uploaded by users from Flickr, without the default re-encoding that is generally applied by any video sharing platform to reduce playback bandwidth costs. Additionally, the re-encoding process considered a balance between quality and size constraints for crowdsourcing, which was ensured by encoding the raw video files at a quality higher than what is common for these platforms while not exceeding a maximum file size.

The crowdsourcing study, with which subjective quality annotations were obtained, was designed carefully to enable investigations into the effects of different levels of label-noise and how a fixed vote budget affects model performance. Concretely, it enabled ablation studies into the effect of different vote budget distribution strategies, meaning that the number of annotated videos was adjusted according to the desired MOS precision. Under a sizeable but fixed budget, the number of votes allocated to each video was found not to be an important factor in the final model performance when using feature-based approaches.

Furthermore, we introduced three novel state-of-the-art no-reference VQA methods for videos in-the-wild. The proposed learning approach, called MLSP-VQA, outperforms the best existing VQA methods trained end-to-end on several datasets and is substantially faster to train without sacrificing any predictive power. This is achieved by working off of deep features pre-extracted from prominent off-the-shelf networks trained for object detection. By global average pooling the activation maps of all kernels in the Inception modules of an InceptionResNet-v2 network trained on ImageNet, a wide variety of features are extracted, ranging from detections of oriented edges to more abstract features related to higher order semantics of objects. These features are input to the architectures, which, in the case of the RN and HYB variants, make use of

the temporal sequence of the frame features. In terms of training speed, the MLSP-VQA-FF model was faster to train than na ive end-to-end transfer learning approaches by factor 36, while also achieving the overall best performance on the introduced datasets.

Finally, this thesis extended the horizon of how different parts of the visual quality assessment domain are related and should be considered jointly. It was shown that a particular set of deep features are more closely related to visual tasks than others. This was investigated by comparing the performances of models trained on features that were determined to be good for either an image quality prediction or image aesthetics prediction task, with those of models trained using a balanced set of features that are useful for both tasks. The results showed no statistically significant difference between the models trained on native or perceptual features, but a significant difference between models trained on perceptual features and auxiliary features. Moreover, models trained on these perceptual features were shown to be achieving state of the art performance in cross-domain testing scenarios.

Similarly, for the VQA domain, by training on the entirety of KonVid-150k the inter-dataset test performance on KoNViD-1k and LIVE-Qualcomm was improved and is competitive in an inter-dataset setup on LIVE-VQC. Inter-dataset performance tests can be understood to be have a small domain-shift and an important measure for the generalizability of models, as well as the specificity of individual datasets. The obtained inter-dataset performance on KoNViD-1k even outperformed the intra-dataset performance, where a model was trained and tested on the dataset itself.

To summarize, it seems clear that one way or another, the existing practices within the field of visual quality assessment have been focused on a narrow set of problems. Important subjects related to scalability of objective models, feature engineering and selection, as well as thorough evaluation of models need to be approached with a broader perspective. However, it remains to be seen which of the above contributions will be most useful. I hope that this thesis will contribute to our understanding of the strengths and weaknesses of these different approaches, and towards the use of more diverse and powerful predictors and features.

Bibliography

[ITU99] P ITU-T RECOMMENDATION. 'Subjective video quality assessment methods for multimedia applications'. In: *International telecommunication union* (1999) (cited on page 1).

[Rei+14] Ulrich Reiter et al. 'Factors influencing quality of experience'. In: *Quality of experience*. Springer, 2014, pp. 55–72 (cited on page 1).

[Hos+20a] Vlad Hosu et al. 'KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment'. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4041–4056 (cited on pages 2, 49, 60, 62, 78).

[Sau+16] Dietmar Saupe et al. 'Crowd workers proven useful: A comparative study of subjective video quality assessment'. In: *QoMEX 2016: 8th International Conference on Quality of Multimedia Experience*. 2016 (cited on pages 5, 8).

[Hos+16a] Vlad Hosu et al. 'Saliency-driven image coding improves overall perceived JPEG quality'. In: *2016 Picture Coding Symposium (PCS)*. IEEE. 2016, pp. 1–5 (cited on pages 5, 9).

[Hos+17] Vlad Hosu et al. 'The Konstanz natural video database (KoNViD-1k)'. In: *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2017, pp. 1–6 (cited on pages 6, 31, 33, 55).

[Jen18] Mohsen Jenadeleh. 'Blind Image and Video Quality Assessment'. PhD thesis. 2018 (cited on page 6).

[Göt+21] Franz Götz-Hahn et al. 'KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild'. In: *arXiv preprint arXiv:1912.07966* (2021) (cited on pages 6, 31, 33, 55, 69).

[GHS21] Franz Götz-Hahn, Vlad Hosu, and Dietmar Saupe. 'Critical analysis on the reproducibility of visual quality assessment using deep features'. In: *arXiv preprint arXiv:2009.05369* (2021) (cited on pages 7, 59, 69, 72, 74).

[Hos+16b] Vlad Hosu et al. 'Reported attention as a promising alternative to gaze in IQA tasks'. In: *PQS 2016: 5th ISCA/DEGA Workshop on Perceptual Quality of Systems*. 2016, pp. 117–121 (cited on page 7).

[Spi+17] Marc Spicker et al. 'Quantifying visual abstraction quality for stipple drawings'. In: *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering*. 2017, pp. 1–10 (cited on page 7).

[Spi+19] Marc Spicker et al. 'Quantifying visual abstraction quality for computer-generated illustrations'. In: *ACM Transactions on Applied Perception (TAP)* 16.1 (2019), pp. 1–20 (cited on page 7).

[ITU19] ITU. 'Methodology for the subjective assessment of the quality of television pictures'. In: *ITU-R Recommendation BT.500-14* (2019). Available online, at `https://www.itu.int/rec/R-REC-BT.500` (cited on pages 9, 17).

[Möl12] Sebastian Möller. *Assessment and prediction of speech quality in telecommunications*. Springer Science & Business Media, 2012 (cited on page 10).

[ITU08] ITU. 'Subjective video quality assessment methods for multimedia applications'. In: *ITU-T Recommendation P.910* (2008). Available online, at `https://www.itu.int/rec/T-REC-P.910/en` (cited on page 11).

[Wan+04] Zhou Wang et al. 'Image quality assessment: from error visibility to structural similarity'. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612 (cited on pages 11, 26).

[Vir+14] Toni Virtanen et al. 'CID2013: A database for evaluating no-reference image quality assessment algorithms'. In: *IEEE Transactions on Image Processing* 24.1 (2014), pp. 390–402 (cited on pages 11, 12, 32).

[Pon+15] Nikolay Ponomarenko et al. 'Image database TID2013: Peculiarities, results and perspectives'. In: *Signal processing: Image communication* 30 (2015), pp. 57–77 (cited on pages 12, 33).

[De +09] Francesca De Simone et al. 'Subjective assessment of H. 264/AVC video sequences transmitted over a noisy channel'. In: *2009 International Workshop on Quality of Multimedia Experience*. IEEE. 2009, pp. 204–209 (cited on page 12).

[De +10] Francesca De Simone et al. 'A H. 264/AVC video database for the evaluation of quality metrics'. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2010, pp. 2430–2433 (cited on page 12).

[Hoß+20] Tobias Hoßfeld et al. 'White Paper on Crowdsourced Network and QoE Measurements–Definitions, Use Cases and Challenges'. In: *arXiv preprint arXiv:2006.16896* (2020) (cited on pages 13, 14).

[KSV11] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 'More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.' In: *Amcis*. Vol. 11. 2011. Detroit, Michigan, USA. 2011, pp. 1–11 (cited on page 13).

[MW09] Winter Mason and Duncan J Watts. 'Financial incentives and the" performance of crowds"'. In: *Proceedings of the ACM SIGKDD workshop on human computation*. 2009, pp. 77–85 (cited on page 13).

[Rog+11] Jakob Rogstadius et al. 'An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets'. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 1. 2011 (cited on page 13).

[Fin+10] Tim Finin et al. 'Annotating named entities in twitter data with crowdsourcing'. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 2010, pp. 80–88 (cited on page 13).

[CK13] Dana Chandler and Adam Kapelner. 'Breaking monotony with meaning: Motivation in crowdsourcing markets'. In: *Journal of Economic Behavior & Organization* 90 (2013), pp. 123–133 (cited on page 13).

[Gad+15] Ujwal Gadiraju et al. 'Understanding malicious behavior in crowdsourcing platforms: The case of online surveys'. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015, pp. 1631–1640 (cited on page 13).

[HM14] S Alexander Haslam and Craig McGarty. *Research methods and statistics in psychology*. Sage, 2014 (cited on page 14).

[Gar+14] Bruno Gardlo et al. 'Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing'. In: *2014 IEEE International Conference on Communications (ICC)*. IEEE. 2014, pp. 1070–1075 (cited on page 14).

[ITU16] ITU. 'Subjective video quality assessment methods for recognition tasks'. In: *ITU-T Recommendation P.912* (2016). Available online, at https://www.itu.int/rec/T-REC-P.912/en (cited on pages 14, 15).

[Hoß+14] Tobias Hoßfeld et al. 'Best Practices and Recommendations for Crowdsourced QoE-Lessons learned from the Qualinet Task Force "Crowdsourcing"'. In: (2014) (cited on page 15).

[ITU18] ITU. 'Subjective evaluation of speech quality with a crowdsourcing approach'. In: *ITU-T Recommendation P.808* (2018). Available online, at https://www.itu.int/rec/T-REC-P.808/en (cited on page 15).

[KCE12] Pavel Korshunov, Shuting Cai, and Touradj Ebrahimi. 'Crowdsourcing approach for evaluation of privacy filters in video surveillance'. In: *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*. 2012, pp. 35–40 (cited on page 15).

[RFN11] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. 'Crowdsourcing subjective image quality evaluation'. In: *2011 18th IEEE International Conference on Image Processing*. IEEE. 2011, pp. 3097–3100 (cited on page 16).

[SSB06] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 'A statistical evaluation of recent full reference image quality assessment algorithms'. In: *IEEE Transactions on image processing* 15.11 (2006), pp. 3440–3451 (cited on pages 16, 33).

[Wu+13] Chen-Chi Wu et al. 'Crowdsourcing multimedia QoE evaluation: A trusted framework'. In: *IEEE transactions on multimedia* 15.5 (2013), pp. 1121–1137 (cited on page 16).

[PPL08a] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet. 'Suitable methodology in subjective video quality assessment: a resolution dependent paradigm'. In: *International Workshop on Image Media Quality and its Applications, IMQA2008*. 2008, p. 6 (cited on pages 16, 18).

[Kei+12] Christian Keimel et al. 'Qualitycrowd—a framework for crowd-based quality evaluation'. In: *2012 Picture coding symposium*. IEEE. 2012, pp. 245–248 (cited on page 17).

[Dav88] Herbert Aron David. *The method of paired comparisons*. Charles Griffin & Co. Ldt, 1988 (cited on page 18).

[SK98] Mei-Yin Shen and C-C Jay Kuo. 'Review of postprocessing techniques for compression artifact removal'. In: *Journal of visual communication and image representation* 9.1 (1998), pp. 2–14 (cited on page 21).

[YW98] Michael Yuen and Hong Ren Wu. 'A survey of hybrid MC/DPCM/DCT video coding distortions'. In: *Signal processing* 70.3 (1998), pp. 247–278 (cited on page 21).

[Zen+14] Kai Zeng et al. 'Characterizing perceptual artifacts in compressed video streams'. In: *Human Vision and Electronic Imaging XIX*. Vol. 9014. International Society for Optics and Photonics. 2014, 90140Q (cited on pages 21, 57).

[Pas+04] Ricardo R Pastrana-Vidal et al. 'Sporadic frame dropping impact on quality perception'. In: *Human Vision and Electronic Imaging IX*. Vol. 5292. International Society for Optics and Photonics. 2004, pp. 182–193 (cited on page 23).

[HG06] Quan Huynh-Thu and Mohammed Ghanbari. 'Impact of jitter and jerkiness on perceived video quality'. In: *Proc. Workshop on Video Processing and Quality Metrics*. 2006 (cited on page 23).

[BS03] Andrew P. Bradley and F. W M Stentiford. 'Visual attention for region of interest coding in JPEG 2000'. In: *Journal of Visual Communication and Image Representation* 14.3 (2003), pp. 232–250 (cited on pages 23, 24, 30).

[Ale+13] Hani Alers et al. 'Studying the effect of optimizing image quality in salient regions at the expense of background content'. In: *Journal of Electronic Imaging* 22.4 (2013) (cited on page 23).

[KT00] Konstantinos Konstantinides and Daniel Tretter. 'A JPEG variable quantization method for compound documents'. In: *IEEE Transactions on Image Processing* 9.7 (2000), pp. 1282–1287 (cited on page 23).

[MT00] Nasir D Memon and Daniel R Tretter. 'Method for variable quantization in JPEG for improved perceptual quality'. In: *Image and Video Communications and Processing 2000*. Vol. 3974. International Society for Optics and Photonics. 2000, pp. 24–34 (cited on page 23).

[ITU96] ITU. *Digital compression and coding of continuous-tone still images: Extensions*. Available online, at https://www.itu.int/rec/T-REC-T.84/en. 1996 (cited on page 24).

[NCS06] Anthony Nguyen, Vinod Chandran, and Sridha Sridharan. 'Gaze tracking for region of interest coding in JPEG 2000'. In: *Signal Processing: Image Communication* 21.5 (2006), pp. 359–377 (cited on page 24).

[Raj16] Omprakash S Rajankar. 'Effect of Single and Multiple ROI Coding on JPEG2000 Performance'. In: April (2016), pp. 29–38 (cited on page 24).

[BS02] Andrew P. Bradley and Fred W. M. Stentiford. 'JPEG 2000 and Region of Interest Coding'. In: *Digital Image Computing Techniques and Applications* January (2002), pp. 303–308 (cited on pages 24, 30).

[MM11] Usama S. Mohammed and Abd-Elhafiez. Walaa M. 'New Approaches for DCT-Based Image Compression Using Region of Interest Scheme'. In: 5.1 (2011), pp. 29–43 (cited on page 24).

[Gow+14] G Gowripushpa et al. 'Implementation of ROI Based Baseline Sequential Adaptive Quantization'. In: *International Journal of Emerging Technology and Advanced Engineering* 4.2 (2014), pp. 361–367 (cited on page 24).

[Jud+09] Tilke Judd et al. 'Learning to predict where humans look'. In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, pp. 2106–2113 (cited on page 26).

[Jia+15] Ming Jiang et al. 'SALICON: Saliency in context'. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2015, pp. 1072–1080 (cited on page 30).

[KTB14] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. 'Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet'. In: *arXiv preprint arXiv:1411.1045* (2014) (cited on page 30).

[KAB15] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. 'Deepfix: A fully convolutional neural network for predicting human eye fixations'. In: *arXiv preprint arXiv:1510.02927* (2015) (cited on page 30).

[Wie+20] Oliver Wiedemann et al. 'Foveated video coding for real-time streaming applications'. In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2020, pp. 1–6 (cited on page 30).

[Nuu+16] Mikko Nuutinen et al. 'CVD2014—A database for evaluating no-reference video quality assessment algorithms'. In: *IEEE Transactions on Image Processing* 25.7 (2016), pp. 3073–3086 (cited on pages 32, 54).

[GB15] Deepti Ghadiyaram and Alan C Bovik. 'Massive online crowdsourced study of subjective and objective picture quality'. In: *IEEE Transactions on Image Processing* 25.1 (2015), pp. 372–387 (cited on page 32).

[Gha+17] Deepti Ghadiyaram et al. 'In-capture mobile video distortions: A study of subjective behavior and objective algorithms'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.9 (2017), pp. 2061–2077 (cited on page 32).

[Pon+09] Nikolay Ponomarenko et al. 'TID2008-a database for evaluation of full-reference visual quality assessment metrics'. In: *Advances of Modern Radioelectronics* 10.4 (2009), pp. 30–45 (cited on page 33).

[SZL17] Wen Sun, Fei Zhou, and Qingmin Liao. 'MDID: A multiply distorted image database for image quality assessment'. In: *Pattern Recognition* 61 (2017), pp. 153–168 (cited on page 33).

[Den+09] Jia Deng et al. 'Imagenet: A large-scale hierarchical image database'. In: (2009), pp. 248–255 (cited on page 33).

[PPL08b] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet. 'Suitable methodology in subjective video quality assessment: a resolution dependent paradigm'. In: (2008), p. 6 (cited on page 33).

[Zha+11] Fan Zhang et al. *IVP Subjective Quality Video Database*. The Chinese University of Hong Kong. 2011. URL: %5Curl%7Bhttp://ivp.ee.cuhk.edu.hk/research/database/subjective/%7D (cited on page 33).

[Hos+20b] Vlad Hosu et al. 'KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment'. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4041–4056 (cited on page 33).

[Tho+16] Bart Thomee et al. 'YFCC100M: The new data in multimedia research'. In: *Communications of the ACM* 59.2 (2016), pp. 64–73 (cited on page 33).

[LHS19] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 'Kadid-10k: A large-scale artificially distorted iqa database'. In: (2019), pp. 1–3 (cited on page 33).

[SB18] Zeina Sinno and Alan Conrad Bovik. 'Large-scale study of perceptual video quality'. In: *IEEE Transactions on Image Processing* 28.2 (2018), pp. 612–627 (cited on page 34).

[Rol+17] David Rolnick et al. 'Deep learning is robust to massive label noise'. In: *arXiv preprint arXiv:1705.10694* (2017) (cited on pages 36, 85).

[Zha+18] Richard Zhang et al. 'The unreasonable effectiveness of deep features as a perceptual metric'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 586–595 (cited on page 36).

[MV06] Loren Merritt and Rahul Vanam. 'x264: A high performance H. 264/AVC encoder'. In: *online] http://neuron2. net/library/avc/overview_x264_v8_5. pdf* (2006) (cited on page 37).

[Win12] Stefan Winkler. 'Analysis of public image and video databases for quality assessment'. In: *IEEE Journal of Selected Topics in Signal Processing* 6.6 (2012), pp. 616–625 (cited on pages 39, 40, 43).

[NK11] Niranjan D Narvekar and Lina J Karam. 'A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)'. In: *IEEE Transactions on Image Processing* 20.9 (2011), pp. 2678–2683 (cited on pages 40, 56).

[HS03] David Hasler and Sabine E. Suesstrunk. 'Measuring colorfulness in natural images'. In: *Electronic Imaging 2003*. International Society for Optics and Photonics. 2003, pp. 87–95 (cited on pages 41, 56).

[Pel90] Eli Peli. 'Contrast in complex images'. In: *Journal of the Optical Society of America, A* 7.10 (1990), pp. 2032–2040 (cited on page 41).

[Wol97] Stephen Wolf. 'Measuring the end-to-end performance of digital video systems'. In: *IEEE Transactions on Broadcasting* 43.3 (1997), pp. 320–328 (cited on page 41).

[YW13] Honghai Yu and Stefan Winkler. 'Image complexity and spatial information'. In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE. 2013, pp. 12–17 (cited on page 41).

[MSB13] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 'Making a "completely blind" image quality analyzer'. In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212 (cited on page 42).

[VSW16] Vassilios Vonikakis, Ramanathan Subramanian, and Stefan Winkler. 'Shaping datasets: Optimal data selection for specific target distributions across dimensions'. In: *Proceedings of 2016 International Conference on Image Processing*. IEEE, 2016, pp. 3753–3757 (cited on page 43).

[ITU04] ITU. 'Objective perceptual assessment of video quality: Full reference television'. In: *ITU-T Tutorial* (2004). Available online, at `https://www.itu.int/itudoc/itu-t/manuels/87170.html` (cited on pages 44, 46).

[HSE11] T. Hoßfeld, R. Schatz, and S. Egger. 'SOS: The MOS is not enough!' In: *Third International Workshop on Quality of Multimedia Experience*. 2011, pp. 131–136 (cited on pages 52, 54).

[JP15] Lucjan Janowski and Margaret Pinson. 'The accuracy of subjects in a quality experiment: A theoretical subject model'. In: *IEEE Transactions on Multimedia* 17.12 (2015), pp. 2210–2224 (cited on pages 52, 54).

[PPL08c] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet. 'Suitable methodology in subjective video quality assessment: A resolution dependent paradigm'. In: *Proceedings of 2008 International Workshop on Image Media Quality and its Applications (IMQA 2008)*. 2008 (cited on page 54).

[WY96] HR Wu and M Yuen. 'A generalized block-edge impairment metric (GBIM) for video coding'. In: *IEEE Signal Processing Letters* 4.11 (1996) (cited on page 56).

[TG00] KT Tan and Mohammed Ghanbari. 'Blockiness detection for MPEG2-coded video'. In: *IEEE Signal Processing Letters* 7.8 (2000), pp. 213–215 (cited on page 56).

[Dij+03] Judith Dijk et al. 'A new sharpness measure based on Gaussian lines and edges'. In: *International Conference on Computer Analysis of Images and Patterns*. Springer. 2003, pp. 149–156 (cited on page 56).

[Mar+02] Pina Marziliano et al. 'A no-reference perceptual blur metric'. In: *Proceedings. International conference on image processing*. Vol. 3. IEEE. 2002, pp. III–III (cited on page 56).

[Ong+03] EePing Ong et al. 'No-reference JPEG-2000 image quality metric'. In: *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*. Vol. 1. IEEE. 2003, pp. I–545 (cited on page 56).

[CO04] Jorge Caviedes and Franco Oberti. 'A new sharpness metric based on local kurtosis, edge and energy information'. In: *Signal Processing: Image Communication* 19.2 (2004), pp. 147–161 (cited on page 56).

[Zha+03] Nien F Zhang et al. 'A kurtosis-based statistical measure for two-dimensional processes and its applications to image sharpness'. In: (2003) (cited on page 56).

[KH96] Deepa Kundur and Dimitrios Hatzinakos. 'Blind image deconvolution'. In: *IEEE signal processing magazine* 13.3 (1996), pp. 43–64 (cited on page 56).

[MMZ99] Xavier Marichal, Wei-Ying Ma, and HongJiang Zhang. 'Blur determination in the compressed domain using DCT information'. In: *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*. Vol. 2. IEEE. 1999, pp. 386–390 (cited on page 56).

[Win01] Stefan Winkler. 'Visual fidelity and perceived quality: Toward comprehensive metrics'. In: *Human Vision and Electronic Imaging VI*. Vol. 4299. International Society for Optics and Photonics. 2001, pp. 114–125 (cited on page 56).

[FK09] Rony Ferzli and Lina J Karam. 'A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)'. In: *IEEE transactions on image processing* 18.4 (2009), pp. 717–728 (cited on page 56).

[LK16] Anupama B Lamb and Madhuri Khambete. 'Perceived no reference image quality measurement for chromatic aberration'. In: *Journal of Electronic Imaging* 25.2 (2016), p. 023004 (cited on page 56).

[LKH09] Hantao Liu, Nick Klomp, and Ingrid Heynderickx. 'A no-reference metric for perceived ringing artifacts in images'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 20.4 (2009), pp. 529–539 (cited on page 56).

[Mon+06] Marco Montenovo et al. 'Objective quality evaluation of video services'. In: *Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. 2006 (cited on page 57).

[PG06] Ricardo R Pastrana-Vidal and Jean-Charles Gicquel. 'Automatic quality assessment of video fluidity impairments using a no-reference metric'. In: *Proc. of 2nd Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*. 2006 (cited on page 57).

[Bor10] Silvio Borer. 'A model of jerkiness for temporal impairments in video transmission'. In: *2010 second international workshop on quality of multimedia experience (QoMEX)*. IEEE. 2010, pp. 218–223 (cited on page 57).

[Kor19] Jari Korhonen. 'Two-level approach for no-reference consumer video quality assessment'. In: *IEEE Transactions on Image Processing* 28.12 (2019), pp. 5923–5938 (cited on pages 57, 71, 76, 80).

[LeC+89a] Yann LeCun et al. 'Backpropagation applied to handwritten zip code recognition'. In: *Neural computation* 1.4 (1989), pp. 541–551 (cited on page 57).

[SVZ14] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 'Deep inside convolutional networks: Visualising image classification models and saliency maps'. In: *In Workshop at International Conference on Learning Representations*. Citeseer. 2014 (cited on page 58).

[MV15] Aravindh Mahendran and Andrea Vedaldi. 'Understanding deep image representations by inverting them'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5188–5196 (cited on page 58).

[MOT15] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. *Inceptionism: Going Deeper into Neural Networks*. 2015. URL: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (visited on 08/05/2021) (cited on page 58).

[DB16] Alexey Dosovitskiy and Thomas Brox. 'Inverting visual representations with convolutional networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4829–4837 (cited on page 58).

[NYC15] Anh Nguyen, Jason Yosinski, and Jeff Clune. 'Deep neural networks are easily fooled: High confidence predictions for unrecognizable images'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 427–436 (cited on page 58).

[Sch+20] Ludwig Schubert et al. *OpenAI Microscope*. 2020. URL: https://openai.com/blog/microscope/ (visited on 08/05/2021) (cited on page 58).

[Rus+15] Olga Russakovsky et al. 'ImageNet Large Scale Visual Recognition Challenge'. In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y (cited on page 58).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 'Imagenet classification with deep convolutional neural networks'. In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105 (cited on page 58).

[Sze+17] Christian Szegedy et al. 'Inception-v4, inception-resnet and the impact of residual connections on learning'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017 (cited on pages 58, 60).

[He+16] Kaiming He et al. 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cited on page 58).

[Xie+17] Saining Xie et al. 'Aggregated residual transformations for deep neural networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492–1500 (cited on page 58).

[Cho17] François Chollet. 'Xception: Deep learning with depthwise separable convolutions'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258 (cited on pages 58, 59).

[Hut18] Matthew Hutson. *Artificial intelligence faces reproducibility crisis*. 2018 (cited on pages 58, 73).

[LeC+89b] Yann LeCun et al. 'Optimal brain damage.' In: *NIPs*. Vol. 2. Citeseer. 1989, pp. 598–605 (cited on page 59).

[Den+13] Misha Denil et al. 'Predicting parameters in deep learning'. In: *arXiv preprint arXiv:1306.0543* (2013) (cited on page 59).

[HMD15] Song Han, Huizi Mao, and William J Dally. 'Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding'. In: *arXiv preprint arXiv:1510.00149* (2015) (cited on page 59).

[Zho+17] Aojun Zhou et al. 'Incremental network quantization: Towards lossless cnns with low-precision weights'. In: *arXiv preprint arXiv:1702.03044* (2017) (cited on page 59).

[OAB18] Hatef Otroshi-Shahreza, Arash Amini, and Hamid Behroozi. 'No-reference image quality assessment using transfer learning'. In: *2018 9th International Symposium on Telecommunications (IST)*. IEEE. 2018, pp. 637–640 (cited on page 59).

[VSS18] Domonkos Varga, Dietmar Saupe, and Tamás Szirányi. 'DeepRN: A content preserving deep architecture for blind image quality assessment'. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2018, pp. 1–6 (cited on pages 59, 74).

[Gao+18] Fei Gao et al. 'Blind image quality prediction by exploiting multi-level deep representations'. In: *Pattern Recognition* 81 (2018), pp. 432–442 (cited on page 59).

[HGS19] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. 'Effective aesthetics prediction with multi-level spatially pooled features'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9375–9383 (cited on pages 59, 62, 68).

[LHS20] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 'DeepFL-IQA: Weak supervision for deep IQA feature learning'. In: *arXiv preprint arXiv:2001.08113* (2020) (cited on pages 60, 84).

[Var20] Domonkos Varga. 'Multi-pooled Inception Features for No-reference Video Quality Assessment.' In: *VISIGRAPP (4: VISAPP)*. 2020, pp. 338–347 (cited on pages 60, 72, 74).

[MM91] Heidar A Malki and Alireza Moghaddamjoo. 'Using the Karhunen-Loe've transformation in the back-propagation training algorithm'. In: *IEEE Transactions on Neural Networks* 2.1 (1991), pp. 162–165 (cited on page 61).

[Bat94] Roberto Battiti. 'Using mutual information for selecting features in supervised neural net learning'. In: *IEEE Transactions on neural networks* 5.4 (1994), pp. 537–550 (cited on page 61).

[Pir04] Selwyn Piramuthu. 'Evaluating feature selection methods for learning in data mining applications'. In: *European journal of operational research* 156.2 (2004), pp. 483–494 (cited on page 61).

[TM18] Hossein Talebi and Peyman Milanfar. 'NIMA: Neural image assessment'. In: *IEEE Transactions on Image Processing* 27.8 (2018), pp. 3998–4011 (cited on pages 61, 62, 68).

[CL09] Cathleen Daniels Cerosaletti and Alexander C Loui. 'Measuring the perceived aesthetic quality of photographic images'. In: *2009 International Workshop on Quality of Multimedia Experience*. IEEE. 2009, pp. 47–52 (cited on pages 61, 63).

[RH12] Judith A Redi and Ingrid Heynderickx. 'Image integrity and aesthetics: towards a more encompassing definition of visual quality'. In: *Human Vision and Electronic Imaging XVII*. Vol. 8291. International Society for Optics and Photonics. 2012, p. 829115 (cited on pages 61, 63).

[KS10] Philip Kortum and Marc Sullivan. 'The effect of content desirability on subjective video quality ratings'. In: *Human factors* 52.1 (2010), pp. 105–118 (cited on page 62).

[MMB12] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 'No-reference image quality assessment in the spatial domain'. In: *IEEE Transactions on image processing* 21.12 (2012), pp. 4695–4708 (cited on pages 62, 71).

[Ye+12] Peng Ye et al. 'Unsupervised feature learning framework for no-reference image quality assessment'. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 1098–1105 (cited on page 62).

[Xu+16] Jingtao Xu et al. 'Blind image quality assessment based on high order statistics aggregation'. In: *IEEE Transactions on Image Processing* 25.9 (2016), pp. 4444–4457 (cited on page 62).

[Bia+18] Simone Bianco et al. 'On the use of deep learning for blind image quality assessment'. In: *Signal, Image and Video Processing* 12.2 (2018), pp. 355–362 (cited on pages 62, 74).

[LHS18] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 'KonIQ-10K: Towards an ecologically valid and large-scale IQA database'. In: *arXiv preprint arXiv:1803.08489* (2018) (cited on page 62).

[CG16] Tianqi Chen and Carlos Guestrin. 'Xgboost: A scalable tree boosting system'. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794 (cited on page 63).

[Sri+03] Anuj Srivastava et al. 'On advances in statistical modeling of natural images'. In: *Journal of mathematical imaging and vision* 18.1 (2003), pp. 17–33 (cited on page 71).

[MSB12] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 'Making a "completely blind" image quality analyzer'. In: *IEEE Signal processing letters* 20.3 (2012), pp. 209–212 (cited on page 71).

[GB17] Deepti Ghadiyaram and Alan C Bovik. 'Perceptual quality prediction on authentically distorted images using a bag of features approach'. In: *Journal of vision* 17.1 (2017), pp. 32–32 (cited on page 71).

[Kun+17] Debarati Kundu et al. 'No-reference quality assessment of tone-mapped HDR pictures'. In: *IEEE Transactions on Image Processing* 26.6 (2017), pp. 2957–2971 (cited on page 71).

[SBC14] Michele A Saad, Alan C Bovik, and Christophe Charrier. 'Blind prediction of natural video quality'. In: *IEEE Transactions on Image Processing* 23.3 (2014), pp. 1352–1365 (cited on page 71).

[Xu+14] Jingtao Xu et al. 'No-reference video quality assessment via feature learning'. In: *2014 IEEE international conference on image processing (ICIP)*. IEEE. 2014, pp. 491–495 (cited on page 71).

[Li+16] Yuming Li et al. 'No-reference image quality assessment with deep convolutional neural networks'. In: *2016 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE. 2016, pp. 685–689 (cited on page 71).

[WSZ18] Chunfeng Wang, Li Su, and Weigang Zhang. 'COME for no-reference video quality assessment'. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2018, pp. 232–237 (cited on page 71).

[YK19] Junyong You and Jari Korhonen. 'Deep neural networks for no-reference video quality assessment'. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 2349–2353 (cited on page 71).

[LJJ19] Dingquan Li, Tingting Jiang, and Ming Jiang. 'Quality assessment of in-the-wild videos'. In: *Proceedings of the 27th ACM International Conference on Multimedia*. 2019, pp. 2351–2359 (cited on page 72).

[Yin+21] Zhenqiang Ying et al. 'Patch-VQ:'Patching Up'the Video Quality Problem'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14019–14029 (cited on page 72).

[Yin+20] Zhenqiang Ying et al. 'From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3575–3585 (cited on page 72).

[HKS17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 'Learning spatio-temporal features with 3d residual networks for action recognition'. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 3154–3160 (cited on page 72).

[Kay+17] Will Kay et al. 'The kinetics human action video dataset'. In: *arXiv preprint arXiv:1705.06950* (2017) (cited on page 72).

[Faw+20] Hassan Ismail Fawaz et al. 'Inceptiontime: Finding alexnet for time series classification'. In: *Data Mining and Knowledge Discovery* 34.6 (2020), pp. 1936–1962 (cited on page 72).

[VS19] Domonkos Varga and Tamás Szirányi. 'No-reference video quality assessment via pretrained CNN and LSTM networks'. In: *Signal, Image and Video Processing* 13.8 (2019), pp. 1569–1576 (cited on pages 72, 74).

[Var19] Domonkos Varga. 'No-reference video quality assessment based on the temporal pooling of deep features'. In: *Neural Processing Letters* 50.3 (2019), pp. 2595–2608 (cited on pages 72, 74).

[Rip07] Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007 (cited on page 72).

[GK18] Odd Erik Gundersen and Sigbjørn Kjensmo. 'State of the art: Reproducibility in artificial intelligence'. In: *Thirty-second AAAI conference on artificial intelligence*. 2018 (cited on page 73).

[Raf19] Edward Raff. 'A step toward quantifying independently reproducible machine learning research'. In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 5485–5495 (cited on page 73).

[HS13] Michiel Hermans and Benjamin Schrauwen. 'Training and analysing deep recurrent neural networks'. In: *Advances in neural information processing systems* 26 (2013), pp. 190–198 (cited on page 77).

[Tu+21] Zhengzhong Tu et al. 'RAPIQUE: Rapid and Accurate Video Quality Prediction of User Generated Content'. In: *IEEE Open Journal of Signal Processing* (2021) (cited on page 80).

[Kor18] Jari Korhonen. 'Learning-based prediction of packet loss artifact visibility in networked video'. In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2018, pp. 1–6 (cited on pages 83, 84).