Perceptual representations of structural and geometric information in images: bio-inspired and machine learning approaches Application to visual quality assessment of immersive media

Suiyi Ling

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Suiyi LING

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Informatique et applications
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Date du soutenance le 29 octobre 2018

Perceptual representations of structural and geometric information in images: bio-inspired and machine learning approaches

Application to visual quality assessment of immersive media

JURY

Présidente :	**M^me^ Luce MORIN**, Professeur des universités, INSA Rennes
Rapporteurs :	**M. Frederic DUFAUX**, Directeur de Recherche, CNRS-L2S Centrale Supélec **M. Dragan KUKOLJ**, Professor of Computer Engineering, University of Novi Sad
Examinateurs :	**M. Vincent COURBOULAY** , Maitre de conférences titulaire de l'HDR, Université de la Rochelle **M^me^ Nathalie GUYADER** , Maitre de conférences, Université Grenoble Alpes (UGA)
Directeur de thèse :	**M. Patrick LE CALLET**, Polytech Nantes, Université de Nantes

Dedication

To my parents, friends, and family

To my professor

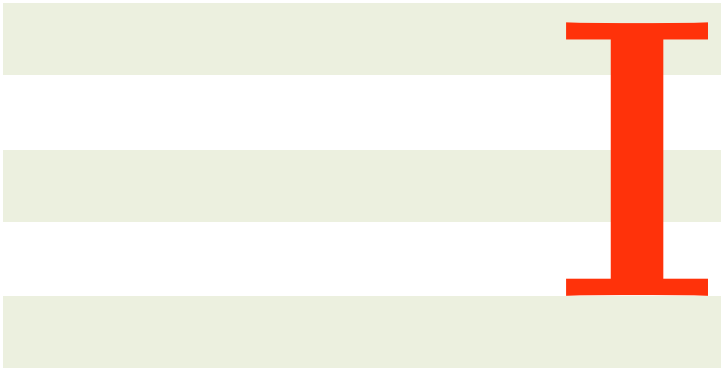
“I wish if you joined my happiness in the harvest day. I am optimistic, a lot of happiness come”

Acknowledgment

This work is supported by the Marie Skłodowska-Curie under the PROVISION (PeRceptually Optimised VIdeo CompresSION) project bearing Grant Number 608231 and Call Identifier: FP7-PEOPLE-2013-ITN and by UltraHD-4U project.

Special thanks to everyone helped me through my thesis. Thanks to my professor.

Contents



Hierarchy and Contents

1

Introduction

1.1 Quality Assessment of Immersive Multimedia

With the rise of more advanced 3D displays, head-mounted displays and other advanced equipment, Immersive Media applications such as Free-viewpoint TV (FTV), 3DTV, and Virtual Reality (VR) has become a hot topic for media ecosystems. Immersive media development requires usage of computer vision/image processing techniques that are likely subject to affect structures of images/videos. This happens in scenarios such as Free-viewpoint TV and Virtual Reality (where omnidirectional contents are presented):

— **Free-viewpoint TV and Quality of Synthesized Views:** FTV [1] aims to make possible for users to freely switching the viewpoints as they do in the real world. Super Multi-View (SMV) and Free Navigation (FN) applications are the two dominant applications that maybe qualified as FTV. Even though there are commonalities between them, they are usually optimizing different goals: SMV targets at compressing all the views more efficiently, while FN focuses more on developing better view synthesis so one can sample more coarsely the number of views (e.g camera arrangements, larger baselines ...). For Super Multi-View applications, densely arranged views are preferred to be compressed and transmitted without synthesizing virtual views and offer glasses-free 3D experience to viewers. High efficient compression mechanism for hundreds of views and smooth transition between adjacent views are the critical factors in this scenario. For Free Navigation, only a limited set of input views are expected to available and transmitted among all possible viewing angle that end user could select. As presented contents are synthesized using Depth-Image-Based Rendering technology (DIBR) [2, 3], in addition to compression and smooth transition between views, reliable synthesis algorithms that are robust to sparser camera arrangements and larger baselines are critical factors with respect to the rendered quality. DIBR based algorithms have the tendency to introduce local non-uniform structure-related distortions. In extreme cases, entire viewing experience of one Free Viewpoint Video (FVV) can be ruined by only one severely distorted region in one synthesized view [4]. As most of existing image/video estimators have been tuned and designed to

handle other type of distortions (traditional uniform compression distortions including blocking artifact, blurriness, ...), they are mostly not suitable for FTV systems. New image/video quality assessment tools that can deal with these structure related distortions are required for this scenario.

— **360° Image/Video in Virtual Reality:** 90% of the existing VR contents are in the form of panoramic images/videos [5]. Virtual Reality/360° images/videos offer to users immersive and interactive visual experiences notably supported by head-mounted or other new equipment [6]. Apart from all the common compression-related issues of 2D/flat images/videos, there are more unique distortions that are brought along with the delivery chain (from production to rendering). For instance, inappropriate mapping/projections from one layout to another may introduce geometric distortions. More importantly, to obtain 360° panoramic contents, stitching algorithms are commonly used for combining images/videos taken by a set of cameras/micro-cameras/fish-eye cameras. Depending on the stitching algorithm or the cameras calibration, very localized structure-related artifacts may be observed. As for the FVT systems scenario, usual image/video quality estimator are likely not sensitive to such distortions.

More generally, the way to assess the impact of structure-related distortions with respect to specific usage of images/videos is also relevant for the following cases:

— **utility assessment of image/video:** utility assessment is to evaluate the usefulness of one image/video in a certain task with respect to a reference. As demonstrated by Rouse and al., Image quality assessment is not a proxy for utility assessment. In most utility assessment tasks, as long as the structure of the image/video is still recognizable for observers, the image/video is considered as usable. However, in some other cases, an image may not be useful if its textures are severely degraded (e.g. in material recognition, textures are important for material identification). Due to the goal of the task itself, image/video may suffer from different levels of structures/textures related degradation in different systems. The definition of 'utility' varies when the goal of the task changes. As thus, how to quantify and leverage the amount of degradation on different information within images/videos, so that the 'utility' of the images/videos is evaluated properly according to the task, is important and challenging.

— **quality assessment of synthesized texture image:** visual texture synthesis is to infer a generating process from a texture sample. It allows then producing arbitrarily many new samples of that texture. This is an important technique in nowadays immersive multimedia system and widely used in many domains and applications. For example, it can be used in the DIBR process for inpainting the dis-occluded regions. As summarized in [7], applications of texture synthesis include image/video restoration [8], image/video generation [9], image/video compression [10], multimedia image processing [11], texture perception and description [12], texture segmentation, recognition [13, 14], and synthesis [15]. Qualifying a synthesized texture algorithm from quality performance point of view is important for all these applications. What makes the quality assessment of synthesized texture difficult relies on the fact that local structure of the synthesized texture may be very different from reference texture while still being perceived as equivalent by human observers. This is especially the case as long as some main properties of the texture patterns are preserved. Therefore, in synthesized texture image quality assessment, it is important and challenging to extract and quantify these texture attributes that convey the perceptually relevant information by taking structure into account.

1.2 Representing the Structure-Related Distortion for Quality Assessment: a Bio Inspired Approach

Intuitively, the best way to quantify the impact of those local, non-uniform, and structure-related distortions on visual quality is to adopt representation inspired by human visual system (HVS). The process of human analyzing a visual scene has been characterized by the presence of regions in the extrastriate cortex that are selectively responsive to scenes [16, 17]. These regions have often been interpreted as representing high-level properties of scenes (e.g., category) and they also exhibit substantial sensitivity to low-level (e.g., edge and texture) and mid-level (e.g., spatial layout) properties. Scene vision involves both foveal and peripheral information, and the representations of multiple features extracted from multiple levels (low, mid, high-level) of the hierarchy (e.g., gist and navigability). A recent bio-vision study [18] proposes a hierarchical framework of visual perception, which comprises a series of discrete stages that successively produce increasingly higher level representations. This framework is illustrated in Figure 1.1. It adopts well-adopted principles including the three levels of representation. **Structure-related distortions could be possibly detected and quantified at any levels.**

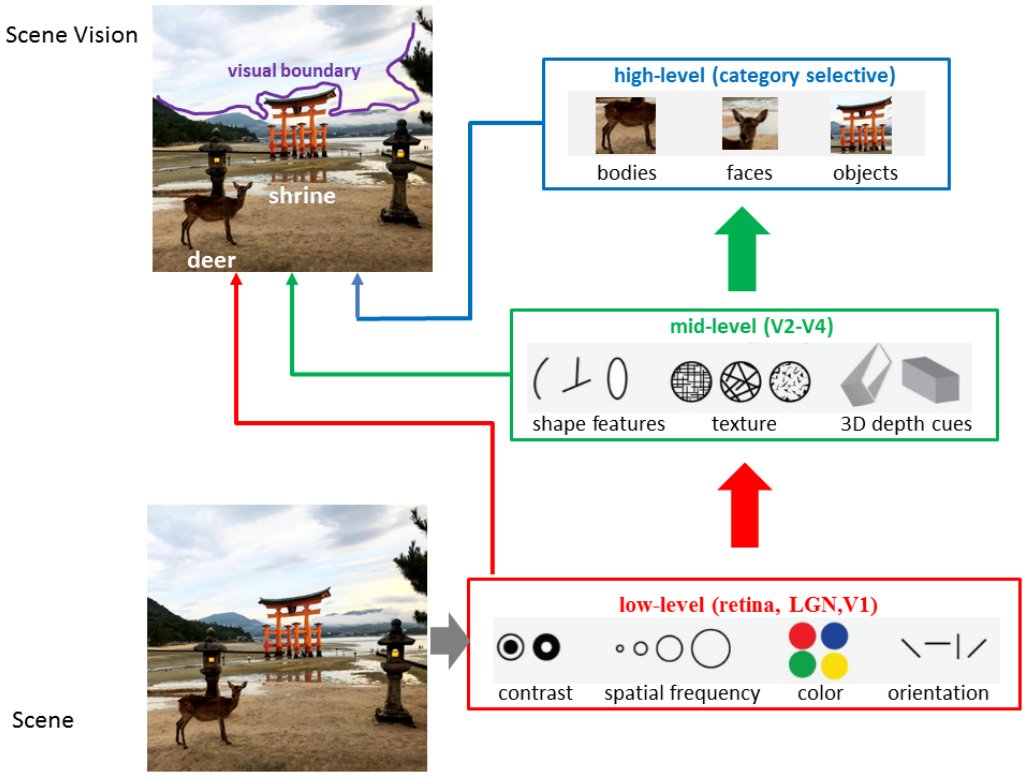


Figure 1.1 – The hierarchical framework of visual perception. Visual percept is formed based on successive extraction and representations of low-, mid- and high-level features. [16, 17]

1.2.1 Low-level of Visual Scenes in Human Brain

Low-level vision is thought to involve the representation of elementary features, such as local edge/contour, color, luminance, contrast, and texture. Such process is typically linked to the flow of information to primary visual cortex (V1) via the retinogeniculate and geniculostriate pathways [19], which translate light intensity at the retina into an orientated edge representation by means of small receptive fields (RFs) tiling the entire visual

field [20]. Neurons in the striate cortex (V1) encode nothing about the meaning of a scene, but they do provide a great deal of information about the image features within it.

In this dissertation, low-level representations of images/videos are defined as local features and descriptors that represent local image/video basic information (e.g., orientations of local edges, granularity of local textures).

1.2.2 Mid-level Representations of Visual Scenes in Human Brain

Mid-level vision is thought to involve intermediate visual patterns. The processes of 'mid-level' vision presumably provide the bridge between these 'low-level' representations of edges, colors, and textures and the 'high-level' semantic representations of objects, actions, and scenes. The immediate stages beyond V1, V2–V4 are often considered to encompass mid-level vision [18]. Overall, these areas are assumed to produce and convey representations of conjunctions of elementary features and properties such as surfaces, higher order image statistics, disparities, and intermediate shape features [21–23]. Recent studies have linked fMRI responses from these areas to representations of locally pooled low-level representations in computational models [24–26]. Furthermore, it is believed that the visual system is very efficient in encoding stimulus properties by utilizing available regularities in the inputs. These 'encoded' information can be considered as mid-level representation obtained based on low-level representations.

In this dissertation, mid-level representations of images/videos are defined as intermediate 'pattern-based encoded feature', where the patterns are learned by summarizing regularity/characteristics/properties of local low-level information (e.g., category of contour).

1.2.3 High-level Representations of Visual Scenes in Human Brain

High-level vision is thought to involve abstraction of visual input. It is well known that at some point there is a semantic, 'high-level' representation of the visual scene because human can describe verbally the contents that we are viewing and their meaning to us [18]. One of the most striking findings in visual neuroscience is that multiple distinct brain regions exhibit selective and highly reliable responses to stimuli from particular semantic categories [27]. Here, semantic is usually related to a certain given task, and categories are defined according to the task (e.g., categories would be different objects in object recognition). More specifically, high-level vision is considered to reflect the abstraction of the visual input into categorical or semantic representations that enable classification, identification with respect to the task. Another point about higher-level representation has been pointed out in [28], neural code in the higher-level cortex can be sparse code, where each element stands for meaningful characteristics of the world as sparsity is considered as one of the essential principles to sensory representation [29].

In this dissertation, higher-level representations of images/videos are defined as 'task-related abstraction', which learns a set of meaningful abstract patterns reflecting the characteristics of the task (e.g., distortion type in quality assessment). These representations are not direct liked to the semantics of the images/videos, but have better representation capability. Therefore, it is defined as 'higher-level representations' instead of 'high-level' representation.

1.3 Building and Learning Representation: White vs. Black Box Approaches

From visual quality assessment prospective, immersive media technologies are providing new challenges mostly related to structure information. To provide effective perceptual quality metric (e.g. in agreement with perceptual quality judgment of human observers), one can adopt bio inspired approaches to quantify the effect of distortions any levels. It implies then to investigate how to represent structure-related distortions. Nevertheless, the higher the representation level, the more difficult it is to derive pure parametric models in a white box manner where internal structures and functions of models are fully tractable and explainable. With the rapid development of machine learning techniques, advanced models have been proposed and employed in different domains. Compared to 'white box' approach, these 'black box' learning based models are more representative. Nevertheless, in many cases, 'black box' methods, e.g., deep learning models, are capable of achieving greater performance than human being [30] especially for high level tasks. These 'white/black box' methodologies are of potential to be adopted for representing images/videos concerning the representative mechanism especially for high level representation. In this thesis, we propose to explore both white and black approaches, mostly depending on the representation level, leading possibly to grey/hybrid approaches (e.g. combining white and black approaches for different stages of the models/representations).

Overall, this thesis aims at designing new effective representation of structural information at any perceptual levels from visual quality prediction application. In other words, effectiveness of representation will be only assessed based on the capacity to predict perceptual quality. While this investigation mostly targets immersive media applications, we also propose to study the characteristics of structure-related distortions in texture synthesis and utility assessment applications. Towards this end, we first review the type of structure-related distortion that occur in new visual media and identify the limitations of existing visual quality predictor to cope with these artefacts. As effectiveness of the representation will be assessed from perceptual quality point of view, it is also important to review the existing relevant visual quality datasets as the perfomance evaluation methodologies.

2

Structure-Related Distortions in Visual Media

2.1 Introduction

In this chapter, distortions that occur in recent images/videos applications are introduced and illustrated. Examples of distortions that challenge common objective quality measures are given along the following sections. As most of the efforts of this thesis are dedicated to FTV scenario, deeper details are provided on the distortions related to this scenario.

2.2 Distortions within Synthesized Views in Free-viewpoint TV

In FTV application, most of the structure-related distortions are introduced by the Depth-Image-Based Rendering technology (DIBR). To understand what types of structure-related distortions are introduced by DIBR and why they are difficult to handle, one should have an overall understanding of the DIBR-based framework for synthesized views generation. Different types of distortions come after certain processes within the DIBR framework. One of the most commonly used DIBR based view synthesized scheme can be summarized as a two-views framework similar to the one proposed by MPEG-FTV [31]. The diagram of this framework is represented in Figure 2.1. Different distortions introduced in different stages during the DIBR process are summarized below along with a brief description of the commonly used DIBR process:

(1) During the pre-processing procedure, camera parameters of both reference and synthesized views are utilized to obtain the projection matrix. This projection matrix is important as it is used to project coordinates in reference views to the ones in virtual views. If inaccurate parameters (i.e., P_L, P_V and P_R shown in the first part of the block diagram in Figure 2.1) are taken, or any mistakes occur during transformation operations, structure related distortions would appear in the next processes.

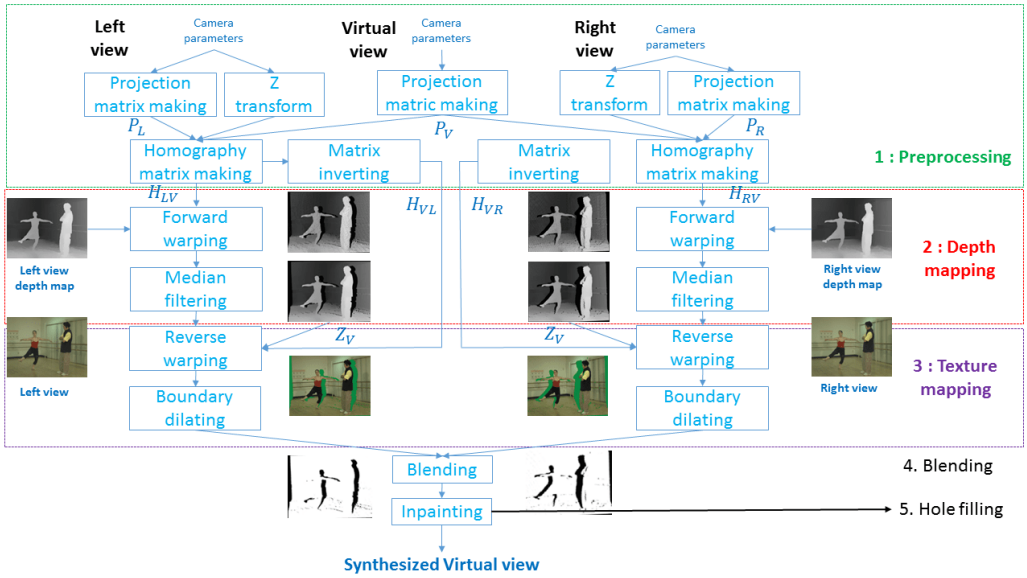


Figure 2.1 – Diagram of DIBR algorithm [31]: It consists of five main parts: (1) Preprocessing, (2) Depth mapping, (3) Texture mapping, (4) Blending, (5) Hole filling. Different distortions are introduced by these processes.

(2) During depth mapping/virtual depth map generation, both the left and the right reference depth maps are warped to generate the corresponding virtual depth map by doing forward warping (the 3D warping from the reference views to the virtual ones) with the relative transform matrix. To get rid of the small holes introduced due to the noise existing in the reference depth maps or the aliasing effects, median/low pass filters are commonly used. After using low-pass filters to remove noise in the depth maps, translation or change in the size of a region within the image, namely **'object shift'** or **'deformation of object shapes'** may be introduced. Furthermore, since the depth maps are needed for forward warping, any types of depth map related errors (including depth estimation, quantization errors and even inaccurate camera parameters obtained from the previous steps) may cause **'local geometric distortion'**.

(3) During texture mapping, the texture of the virtual view is synthesized by reverse warping, which is to map the texture from the reference view pixel-wise to the virtual one with the virtual depth maps. In this texture mapping process, regions that can be seen in the right/left views but occluded in the virtual views are remained as dark holes and commonly defined as the dis-occlusion/dis-occluded regions. Apart from the big holes (big dis-occluded regions) caused by the occlusions, **'small holes'** could also be introduced by the 'round-off error' [32] (if pixel coordinates are not mapped to an integer value at the virtual viewpoint, they would be usually either interpolated or rounded to their nearest integer positions. This type of incorrect coordinate mapping is name the 'round-off error').

(4) During the blending process, the left and right synthesized textures are then blended to recover the dis-occluded regions by borrowing information from the two reference views, which is defined as 'occlusion handling'. Geometric distortions can be amplified due to improper blending in this process.

(5) During the dis-occluded regions filling (hole filling) process, inpainting methods are commonly employed to fill up holes that cannot be handled in the previous steps. In this filling procedure, **'blurry regions'** may be introduced. When it comes to complex texture regions, where inpainting algorithms may fail to fill up the missing holes, incorrect rendering of texture regions may also occur (e.g., **'ghosting artifact'**). These blurry or poorly inpainted regions are more visible since they are always located along transitions regions between

foreground and background [33] and sometimes even degrade the structure.

In summary, the DIBR procedure may introduce the following special spatial and temporal distortions:

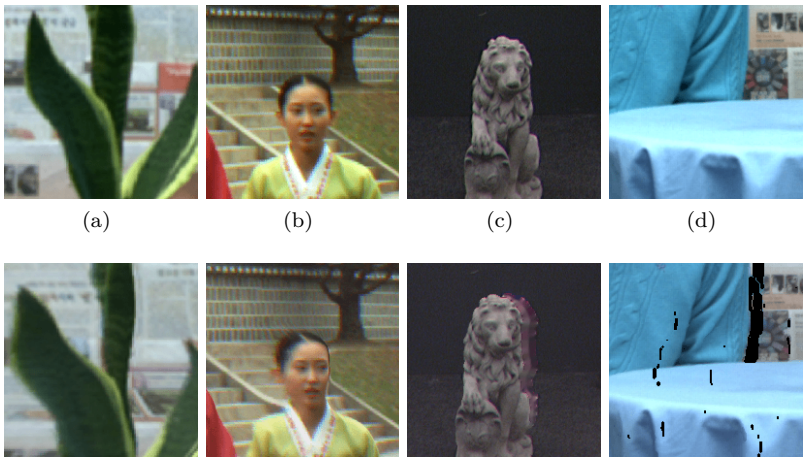


Figure 2.2 – Examples of special distortions introduced by DIBR algorithms in FTV system. Reference images are in the first row, while synthesized ones are in the second row.

Spatial Distortions in FTV Scenario:

(1) **Local non-uniform geometric distortions:** unlike traditional global uniform artifact, e.g., blocking artifact, the dominant spatial distortions of synthesized views are the local non-uniform geometric distortions around dis-occluded regions. These distortions are normally located around the boundaries of objects within 'Regions of Interest'. Although they are not distributed continuously throughout the image, they are less acceptable than the uniformly distributed distortions [34].

(2) **Structure deformations:** inappropriate 3D warping may modify/deform the shape of the objects as shown in 2.2 (f). This type of shape deformation is more annoying as it makes important objects appear unnatural.

(3) **Global shifting:** DIBR based algorithms may introduce global continuous shifting of objects as shown in 2.2 (e). Observers are normally more tolerant to this type of distortions than local severe ones [34]. However, this type of distortions is over-penalized by point to point metric like PSNR as (the acceptable shifted regions are considered errors by PSNR).

(4) **Dis-Occluded Regions/Dark Holes:** small and medium size of dis-occluded regions may be introduced as shown in 2.2 (h). Although these dark/black holes rarely exist in virtual views synthesized with more advanced view synthesis algorithms (e.g., the view synthesis reference software provided by the MPEG community), it could still appear in extreme situations (e.g., huge baseline distance).

(5) **Blurriness/ghosting artifacts:** inpainting algorithms that are used to inpaint the dis-occluded regions may introduce blurriness and unsmooth transition along objects' boundaries as shown in Figure 2.2 (g). If the dis-occluded regions are not well inpainted, it may cause changes of structures as well (e.g., strange blurred contours). When geometric distortions and blurriness are introduced at the same location, it could be more noticeable (or even become the dominant distortion) for human observers.

Temporal Distortions in FTV Scenario:

There are mainly two types of temporal structure-related distortions that are introduced by DIBR based algorithms:

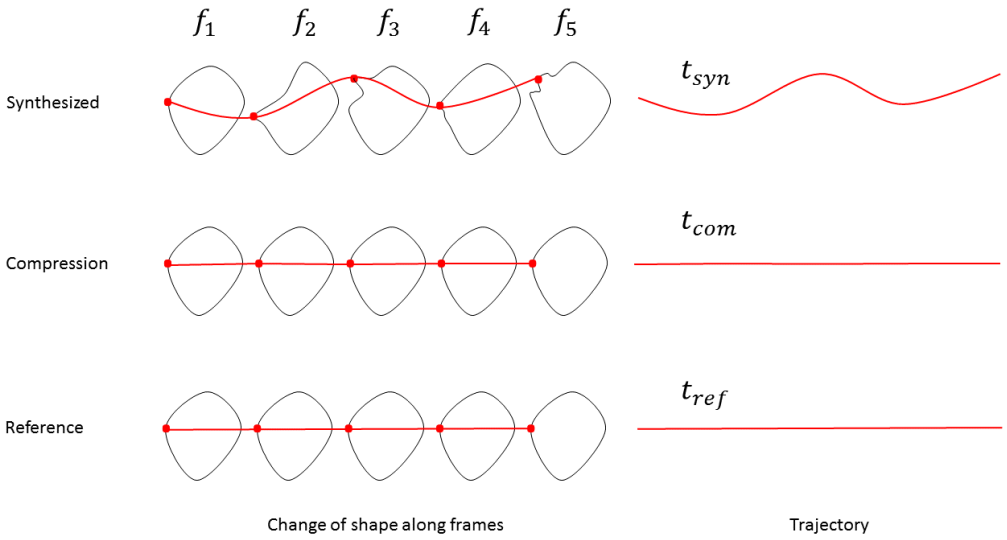


Figure 2.3 – Example explaining specific temporal trajectory deformations caused by spatial geometric distortions. t_{syn}: trajectory of one object's key point in synthesized video; t_{com}: trajectory of one object's key point in video contain traditional compression artifacts; t_{ref}: trajectory of one object's key point in reference video.

(1) **Temporal structure-related distortions at one viewpoint location/ temporal structure-related distortions within one viewpoint:** in the case of viewing one view at one individual view position/location, spatial geometric distortions introduced by DIBR processes may lead to temporal structure inconsistency. This type of temporal structure related distortion within one viewpoint can be reflected by the differences among the three trajectories as shown in Figure 2.3. As shown in the figure, shapes of objects in one frame could be deformed significantly compared to the ones in next frames when playing a synthesized view at one viewpoint location. In this situation, temporal flickering in the form of fluctuation of moving objects' boundaries may be observed within videos at a particular viewpoint location.

(2) **Temporal structure related distortions among viewpoints:** considering the scenario of navigating among different viewpoints, local structure related distortions (e.g., geometric distortions or inpainting related distortions) may introduce structure inconsistency from one viewpoint to another. Compared to temporal structure-related distortions at one viewpoint location, this type of distortions is observed due to view switch and could be more disturbing. Furthermore, the larger the baseline distance is used for view synthesis, the more abrupt/inconsecutive the structure changes would be when the users switch from one viewpoint to another. This rough transition between different viewpoints could be considered as temporal flickering. For example, Figure 2.3 shows the change of the shape of a static object along five frames across time in one multi-view content (i.e., the change of the object shapes from the first frame f_1 to the fifth one). Different degrees of local structure-related distortions could be introduced differently among different viewpoints with different contents. For example, to encode a set of multi-views sequences using 3D/MV HEVC [35], one rate-point for the entire set of multi-view plus depth format sequences is usually selected. Different viewpoints may contain significantly different contents, but the arrangement of compression budget for texture/depth map of different viewpoints are normally not assigned differently according to the contents. Structure information of different viewpoints may be degraded differently due to different degrees of distortions on depth map/texture of the sequence. As a result, unsmooth structures transition among viewpoints could be observed.

2.3 Distortions within Stitched Panoramic Images in Virtual Reality

The growing popularity of virtual reality (VR), augmented reality (AR) and mixed reality (MR) applications necessitates the generation of good-quality 360-degree panoramic images from multiple viewpoint images captured by different cameras on the same rig. However, stitching individual viewpoint images into one high-quality and coherent panoramic image is technically challenging, which could introduce disturbing structure-related distortions:

Ghosting and structure inconsistency are the two most common visual artifacts produced by modern image-stitching tools due to geometric misalignment and improper photometric correction [36]. Ghosting artifacts usually appear in the form of transparent objects, while structure inconsistency usually appears in the form of non-continuous contour transition along object shapes. Both of them are locally located in the 360-degree content, and in most of the case affect the perceived quality.

Overlap of different distortions: ghosting artifact, structure inconsistency, and even blurriness may be introduced together at the same location. This kind of phenomenon can be considered a 'masking effect' of different distortions. In most of the cases, this effect may amplify the annoyance level of the distortions.

Examples of local non-uniform structure-related disruptions are depicted in Figure 2.4. These artifacts cannot be easily captured by traditional image quality metrics, yet they are more visually disturbing than conventional distortion types like compression artifacts. Because severely distorted local structure regions are less acceptable for observers and greatly affect the quality of the entire image [34].

(a) (b) (c) (d)

Figure 2.4 – Examples of structure-related distortions introduced by stitching algorithms in VR System.

2.4 Distortions within Images in Utility Assessment

Different qualities can have equal or similar utilities in the task of utility assessment. In different tasks, the amount of structure/texture information needed to be maintained for an image/video to be useful is different.

Severely degraded structure: because of the goal of the task itself, images/videos with severe and extreme distortions on texture and structure could be kept as long as they are still useful for the system in related utility tasks. For example, in cases like object recognition, degraded image/video that contains only the shapes of the objects are still useful even though other detail structure/texture information is severely degraded. This is because human observers are capable of recognizing objects based on only the main structure of the image/video.

Different utility tasks, different structure/texture related distortions: distortion type differs de-

pending on the exact utility task and system. For example, in the case of predicting utility of synthesized views, the primary structure-related distortion could be synthesized-related geometric distortions. However, in the case of predicting the utility of images taken by surveillance cameras in adverse weather conditions, the main degradation comes from the camera systems. In this situation, the primary structure-related distortions could be blurry structures caused by lens distortion, which are different from the previous example.

Examples of images that are used in utility assessment are shown in Figure 2.5. These images are from the CU-Nantes database [37]. Description of the database is given in section 4.2.3. By checking the three degraded figures (b)-(d), it is not easy for someone to recognize the parrots in Figure 2.5 (c) or (d). However, we can still point out the two parrots in (b), since the structures of parrots in this sub-figure are roughly maintained. For Figure 2.5 (b), it would be useful if the task were object recognition but would be useless if the task were to tell the material of the cage. In utility assessment, how the utility should be evaluated should refer to the exact task. In most of the cases, utility assessment does not equal to quality assessment and should be designed differently according to the exact utility task.

(a) (b) (c) (d)

Figure 2.5 – Examples of different levels image distortions in task of utility assessment.

2.5 Distortions within Synthesized Texture Image in Nowadays Multimedia Applications

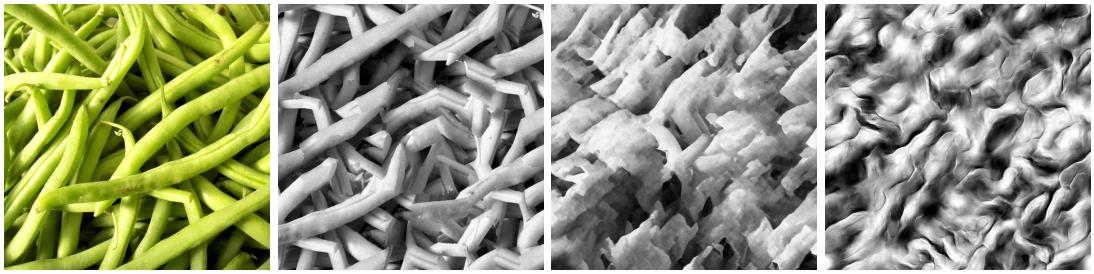
Distortions in synthesized texture images are typical. Synthesized texture images may include misalignment, blur, tiling, and loss in the periodicity of the primitives, depending on the texture synthesis algorithm. Generally, poor synthesis algorithm could result in altering of statistical texture properties and even image structure with respect to the original reference texture. According to a subjective test conducted in [38], the most detrimental distortion within synthesized texture images are the lack of structural details. Other pronounced artifacts include misalignment of the texture patterns, blurriness, and tiling introduced in the texture patterns.

Difficulties of quantifying texture-synthesis related distortions: The process of automatically assessing the perceived visual quality of synthesized texture images is an ill-posed because of the fact that

- Texture synthesis algorithms may modify the size of the original image.
- Global shifting of the image are commonly introduced. The synthesized textures are not required to have pixel-wise correspondences with the original texture but can still appear perceptually equivalent.
- In some cases, even the structures of the images have been modified, the quality of the synthesized texture images could still be high.

Examples of synthesized texture image are presented in Figure 2.6. Images in this figure are from the

SynTEX database [39–41] described in section 4.2.4. It can be seen from Figure 2.6 (c) and (d) that the structure of the 'green beans' has been destroyed compared to the one of Figure 2.6 (a). To better evaluate synthesized texture images, metrics should not only be able to quantify texture related distortions, but also the structure-related ones.



(a) (b) (c) (d)

Figure 2.6 – Examples of different levels of image distortions in task of synthesis texture images' quality assessment.

2.6 Conclusion

In this chapter, different types of structure-related distortions that appear in different applications are introduced. It is evident that those distortions are entirely different from the common distortions like blocking artifact, blur or quantization artifacts. Therefore, a particular model that is designed exactly for capturing these types of artifacts is needed. In the next chapter, limitations of commonly used metrics and existing metrics that are designed for the corresponding applications are described in detail.

3

Limitations of Existing Image/Video Quality/Utility Assessment Metrics

3.1 Introduction

In this chapter, examples of failure cases of commonly used image/video quality assessment metrics in different applications are first given. Afterwards, this chapter introduces special images/video quality models that has been designed for evaluating 1) the quality of synthesized images in FTV scenario; 2) the quality of synthesized videos in FTV scenario; 3) the quality of stitched panorama images in VR scenario; 4) the quality of synthesized texture images; and 5) the utility of images are introduced separately in the following sections. Meanwhile, limitations of those existing metrics are pointed out.

3.2 Commonly Used Image and Video Quality Assessment Metrics Fail to Quantify Structure-Related Distortions

In order to assess the quality of images/videos, where structure-related distortions (introduced in chapter 2) exist for different use cases or applications, it is intuitive to try existing commonly used image or video metrics first. As expected, the performances of these metrics are not acceptable as they fail to handle those novel structure-related distortions described in chapter 2. Following are some failure cases of using existing commonly used metrics for the quality/utility evaluation of images/videos in different applications. Their performance, in terms of Pearson Correlation Coefficient (PCC), Spearman's rank order Correlation Coefficient (SCC) or Root Mean Squared Error (RMSE), in different applications are reported (introductions of these performance evaluation methodologies are given in section 4.3).

- **In FTV Use Case**
  - **Spatial Structure-Related Distortions**

Spatial structure-related distortions like geometric distortions and object shifting are challenging for commonly used metrics. IVC-Images dataset consists of frames extracted from sequences synthesized with different DIBR synthesized algorithms. Images in this dataset thus contain mainly spatial non-uniform local geometric distortions, details of this database is given in section 4.2.1.1. The performances of commonly used image quality metrics on this dataset are shown in Table 3.1 as reported in [42]. According to Table 3.1, commonly used full reference image quality metrics, including peak signal to noise ratio (PSNR) [43], structural similarity index (SSIM) [44], multi-scales SSIM (MS-SSIM) [45], information content weighted PSNR (IW-PSNR) [46], and information content weighted SSIM (IW-SSIM) [46], show poor performance on this database. Moreover, natural scene statistic (NSS) based no reference metrics, i.e., natural image quality evaluator (NIQE) [47] and blind image integrity notator using DCT statistics (Bliinds) [48] perform poorly on this database too. Thus, it is necessary to develop a new reference metric to quantify the effect of degradations of structures spatially on perceived quality.

Table 3.1 – Performance of common image quality metrics on the IVC-Images database.

	PCC	**SCC**	**RMSE**
PSNR [43]	0.456	0.442	0.593
SSIM [44]	0.434	0.400	0.599
MS-SSIM [45]	**0.541**	**0.502**	**0.560**
IW-PSNR [46]	0.361	0.346	0.621
IW-SSIM [46]	0.533	0.479	0.563
Bliinds [48]	0.533	0.180	0.563
NIQE [47]	0.402	0.367	0.609

□ **Temporal Structure-Related Distortions within One viewpoint**

Temporal structure-related distortions in the form of temporal inconsistency within one viewpoints are challenging for commonly used metrics. IVC-Video dataset consists of sequences synthesized using seven different DIBR synthesized algorithms. Sequences in this database contain both structure-related spatial distortions, and temporal structure-related distortions observed at one viewpoint location. The detailed description of the dataset is given in section 4.2.1.2. The performances of commonly used image/video quality metrics on this dataset (reported in [49]) are concluded in Table 3.2. Tested metrics includes visual signal to noise (VSNR) [50], information fidelity criterion (IFC) [51], SSIM, visual information fidelity (VIF) [52], pixels version of VIF (VIFP) [52], noise quality measure (NQM) [53], PSNR, PSNR-human visual system masking model (PSNR-HVSM) [54], PSNR-human visual system (PSNR-HVS) [54], video quality metric (VQM) [55], video structural similarity measure (VSSIM) [56], weighted signal-to-noise ratio (WSNR) [57], MS-SSIM, and universal quality index (UQI) [58]. According to Table 3.2, it is obvious that those commonly used metrics are poorly correlated with the subjective scores. Among them, the PCC value of the best performing VSNR is less than 0.5.

Another example is shown in Table 3.3, this table reports the performances of commonly used image/video quality metrics on SIAT synthesized video quality database (as reported in [60]). The 140 videos in this database are synthesized from ten multi-view plus depth (MVD) based 3D sequences with different texture/depth quantization combinations. These sequences contain not only spatial structure distortions but also temporal structure-related artifacts (within one viewpoint). It

Table 3.2 – Performance of commonly used image/video quality metrics on the IRCCyN/IVC DIBR Videos database (IVC-Video) [59].

	VSNR [50]	VIFP [52]	IFC [51]	SSIM [44]	VIF [52]	NQM [53]	PSNR [43]
PCC	**0.46**	**0.46**	0.44	0.44	0.42	0.36	0.34
	PSNR-HVSM [54]	PSNR-HVS [54]	VQM [55]	VSSIM [56]	WSNR [57]	MS-SSIM [45]	UQI [58]
PCC	0.34	0.32	0.32	0.32	0.32	0.26	0.2

is claimed in [60] that this database is a supplement of IVC-Videos database. According to Table 3.3, it can be seen that both commonly used image quality metrics, including PSNR, WSNR [57], SSIM, MS-SSIM, and commonly used video quality metrics, including VQM, motion-based video integrity evaluation (MOVIE), fail to predict the perceived quality of synthesized views well. Among all these metrics, the PCC value of the best performing MS-SSIM is 0.703. On the one hand, there is still a room to improve the performance. On the other hand, metrics include MS-SSIM can obtain considerably good performance because this dataset contains mainly compression artifacts. Synthesized related artifacts are not obvious/dominant compared to the compression ones in this dataset. Therefore, in this study, this database is not used for performance evaluation.

Table 3.3 – Performance of commonly used image/video quality metrics on the SIAT synthesized video quality database [60].

	PCC	**SCC**	**RMSE**
PSNR [43]	0.648	0.627	0.097
SSIM [44]	0.608	0.598	0.101
MS-SSIM [45]	**0.703**	**0.731**	**0.091**
VQM [55]	0.669	0.655	0.095
WSNR [57]	0.605	0.589	0.102
MOVIE [61]	0.646	0.693	0.097

□ **Temporal Structure-Related Distortions among Viewpoints**

Temporal structure-related distortions in the form of temporal inconsistency among viewpoints (observed due to view switch) are challenging for commonly used metrics. Time freeze free-viewpoint-synthesized-video-dataset (FFV) consists of sequences synthesized using seven different DIBR synthesized algorithms. In this dataset, sequences were generated to mimic a smooth camera motion during a time freeze with synthesized sequences. As thus, sequences in this dataset contain mainly spatial structure-related distortions and temporal structural related distortions in the form of unsmooth transition among viewpoints. Detailed descriptions of the database are given in section 4.2.1.3. The performance of the commonly used image metrics on time freeze free-viewpoint-synthesized-video-dataset (FFV) are shown in Table 3.4 (reported in [62]). As it can be observed from Table 3.4, this new type of temporal structure-related distortion (caused by switches of viewpoints) is challenging for those commonly used video quality assessment metrics as they perform so poorly on this FFV database with PCC values less than 0.3.

■ **In VR Use Case**

Ghosting and structure inconsistency are challenging for commonly used metrics to quantify. In practical, references of stitched images are not available, and no reference metrics are thus needed. SIAQ dataset contains mainly structure distortions introduced by stitching algorithms. Detailed descriptions of the database are given in section 4.2.2. The performances of commonly used no reference image quality

Table 3.4 – Performance of commonly used image quality metrics on the time-freeze free-viewpoint-synthesized-video-database (FFV) [62].

	PCC	**SCC**	**RMSE**
PSNR [43]	**0.267**	**0.294**	**0.907**
SSIM [44]	0.000	0.000	0.941
MS-SSIM [45]	0.011	0.061	0.941
IFC [51]	0.128	0.065	0.934
VIF [52]	0.058	0.094	0.939
VIFP [52]	0.079	0.122	0.938
UQI [58]	0.000	0.000	0.941

assessment metrics on this database are summarized in Table 3.5 (reported in [63]). It is obvious that both Bliinds and distortion identification-based image verity and integrity evaluation (DIIVINE) index fail to evaluate the quality of stitched panoramic images (with PCC values lower than 0.3). One of the reason is that they are not able to localize and quantify dominant structure-related artifacts, e.g., ghosting artifact.

Table 3.5 – Performance of common image quality metrics on SIAQ stitched images database [36].

	PCC	**SCC**
Bliinds [48]	0.118	0.066
DIIVINE [64]	**0.258**	**0.145**

■ **In Texture Synthesis Use Case**

Texture synthesis-related distortions are challenging for commonly used metrics to quantify since synthesized texture images may contain structure as well as texture related distortions. SynTex dataset contains images generated using different texture synthesis algorithms. Details of this dataset are given in section 4.2.4. The performance of commonly used image quality assessment metrics on the SynTex dataset (reported in [7, 65]) are summarized in Table 3.6. According to Table 3.6, PSNR, SSIM, structural texture SSIM (ST-SSIM) [66], DIIVINE [64], and NIQE, fail to predict the perceived quality of synthesized texture images well. The PCC values of those metrics are all below 0.4.

Table 3.6 – Performance of commonly used image quality metrics on the SynTex database [7, 65].

	PCC	**SCC**	**RMSE**
PSNR [43]	0.237	0.345	1.210
ST-SSIM [66]	0.215	0.135	1.213
MS-SSIM [45]	0.293	0.122	1.105
DIIVINE [64]	**0.357**	**0.408**	**1.094**
NIQE [47]	0.253	0.218	1.154

■ **In Utility Assessment Use Case** Different levels of structure/texture-related distortions are challenging for commonly used metrics to handle regarding to different utility tasks. Cu-Nantes dataset is released for the task of utility assessment. Images in this database contain different level of structure, texture disruptions. Detailed introductions of this dataset are given in section 4.2.3. The performance of using commonly used image quality metrics for predicting utility scores on the Cu-Nantes dataset (reported in [67, 68]) is summarized in Table 3.7. Tested metrics include PSNR, SSIM, MS-SSIM, VSNR, WSNR, and NQM. As none of these metrics focus on quantifying the amount of structure disruptions according to the goals of the tasks, they fail to provide acceptable performance for utility assessment.

Table 3.7 – Performance of using commonly used image quality metrics as utility estimator on the Cu-Nantes database.

	PCC	**SCC**	**RMSE**
PSNR [43]	0.191	0.471	43.5
WSNR [69]	0.187	0.425	32.8
SSIM [44]	**0.749**	**0.871**	**18.2**
MS-SSIM [45]	0.566	0.726	23.9
NQM [53]	0.318	0.467	35.4
VSNR [50]	0.371	0.473	31.8

3.3 Limitations of Existing No Reference Metrics designed for Natural Images

On the one hand, NSS based no reference metrics including DIIVINE [64], Bliinds [48], and NIQE [47] are designed for capturing natural uniform distortions that distributed throughout the entire image/video (like the traditional gaussian blur, pepper noise or compression related distortions). Dealing with local non-uniform distortions that appear in images/videos are challenging for them. In most cases, those NSS based metrics underestimate the importance of these structure-related distortions.

On the other hand, deep learning based quality assessment models, like the one proposed in [70], assign image-level subjective scores for patches during training. This is questionable when it comes to applications where local non-uniform distortions are the dominant distortion. Furthermore, in the model proposed in [71], the final predicted quality score is obtained by calculating the statistics (e.g., mean and standard deviation) of local patches' features. By doing so, the impact of severe local distortions on perceived quality is 'averaged' and 'weaken'.

Taking the two points mentioned above into account, no reference metrics that are trained to learn the local non-natural structure (NNS) is needed.

3.4 Limitations of Existing Image Quality Assessment Metrics for FTV System

The very first full reference (FR) approach that designed for evaluating the quality of synthesized images is proposed by Bosc *et al.* [49] by applying some prior knowledge acquired through subjective tests (e.g., the common localization of view-synthesis artifacts along contours) to SSIM. Following this idea, Conze *et al.* [72], propose the view synthesis quality assessment (VSQA) metric, which improves SSIM with three visibility maps that characterizes the complexity of the images. Later, the '3D synthesized view image quality metric' (3DswIM) is proposed by Battisti *et. al.* [33]. This metric is based on statistical features of wavelet sub-bands. In addition, Tsai and Hang [73] propose a metric based on compensating the shifts of the objects that appear in synthesized views by calculating the noise around them. Considering the fact that using multi-resolution approaches could increase the performance of image quality metrics, Sandić-Stanković *et al.* develop the 'Morphological Wavelet PSNR' (MW-PSNR) using a morphological wavelet decomposition [74]. Later they extend the work by using a multi-scale decomposition based on morphological pyramids, which is called 'Morphological Pyramid PSNR' (MP-PSNR) [75]. Recently, Stanković *et.al.* [76] point out that PSNR is more consistent with human judgment

when it is calculated at higher morphological decomposition scales. They thus propose reduced versions of the morphological multi-scale measures called reduced MP-PSNR and reduced MW-PSNR correspondingly (denoted as MP-PSNR$_r$ and MW-PSNR$_r$). According to their experimental results, the reduced versions (i.e., MP-PSNR$_r$ and MW-PSNR$_r$) outperform the full versions (i.e., MP-PSNR$_f$ and MW-PSNR$_f$).

In real application, reference synthesized views are generally not available. Thus, No Reference (NR) metrics are more desirable. Nevertheless, compared to FR metrics mentioned above, only few NR metrics are designed for synthesized views in FTV scenario. In [77], NIQSV is proposed by hypothesizing that high quality images are consist of flat areas separated by edges. Later on, NIQSV+ is proposed in [42] to improve NIQSV by taking 'black holes' into account. Recently, a novel NR metric APT is proposed in [4] using the auto-regression (AR) based local image description.

All the FR/NR images metrics mentioned above suffers from at least one of the drawbacks mentioned below:

(1) Geometric distortions in a certain extent are acceptable for human observers and should be treated differently from those are unacceptable. For example, in Figure 3.1, the slightly expanded nose on the right in the third row is more acceptable for human observers than the twisted nose in the middle. Some of existing metrics fail to well predict good quality synthesized images as they over-penalized geometric distortions.

(2) The human visual system is sensitive to severe local artifacts [34, 78] local geometric distortions. The most upsetting artifacts in synthesized images are the inconsistent local geometric distortions instead of the consistent global shifting artifacts. These specific artifacts appear mainly around the disoccluded regions and thus are sparse. However, most of the existing metrics process the entire image equally. Thus, they are not able to locate and quantify local geometric distortions properly. Sensitive region selection should be considered as a pre-process module to select regions with structure-related distortions.

(3) Global shifting within certain limits is acceptable for human observers but is punished severely by point-to-point based metrics like PSNR. Due to equal-weighted pooling and point-wise comparison, some image quality assessment metrics mistakenly emphasize the consistent global shifting artifacts. For example, in Figure 3.1, it is obvious that the 'twisted nose' in the middle is more annoying than the 'slightly shifted nose' on the right. However, the PSNR score for the patch in the middle with its referred patch on the left is 20.2854 db while the one of the patch on the right is 18.6616 db, which incorrectly indicating that the quality of the 'twisted nose' is better.

(4) With the rapid development of machine learning technologies, many quality assessment models have been proposed recently. Many of these machine learning based models are trained based on the assumption that the perceived quality of any local regions in the image is the same as the one for the entire image. This assumption may work for images that contain uniform distortions but may not stand for those contain non-uniform distortions.

3.5 Limitations of Existing Video Quality Assessment Metrics for FTV system

The 'Peak Signal to Perceptible Temporal Noise Ratio' (PSPTNR) metric, introduced by Zhao and Yu [81], quantifies temporal artifacts that can be perceived by observers in the background regions of the synthesized videos. Similarly, Ekmekcioglu *et al.* [82] propose a video quality metric by using depth and motion information

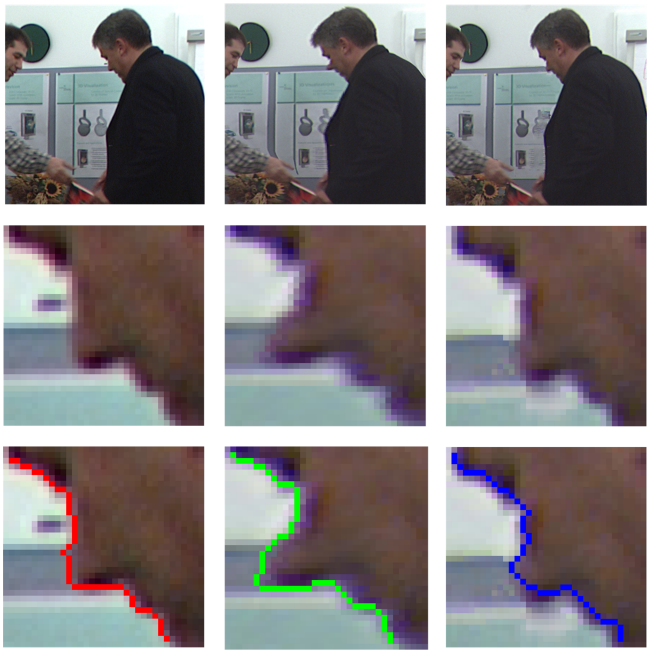


Figure 3.1 – Failure examples of using point-to-point metrics for synthesized views. Rows:(from up to down) : Part of the images for better observation; Patches from images; Extracted contours of patches. Columns: (from left to right) reference image, synthesized image obtained with algorithm proposed in [79], synthesized image obtained with algorithm proposed in [80]. PSNR(L, M)=20.2854 db, PSNR(L, R)=18.6616 db

to locate the degradations. The state-of-the-art video metric designed for free viewpoint videos is recently introduced by Liu *et al.* [60]. Their proposed metric considers the spatio-temporal activity and the temporal flickering that appears in synthesized video sequences.

However, none of the aforementioned video quality metrics is designed to quantify the unsmooth transition among views (temporal structure inconsistency observed during view switch).

3.6 Limitations of Existing Image Quality Assessment Metric for Stitched Panoramic Images

The model presented in [83] is an early work combining mainly low-level and high-level features. However, structure-related artifacts are not considered in this model. In [84] and [85], the authors focus mainly on color and intensity consistency. Nevertheless, color correction is no longer a problem for advanced stitching algorithms. MIQM [86] is proposed to quantify luminance, contrast, spatial motion and structure-related abrupt local changes in stitched images. Later, Qureshi *et al.* [87] proposed a SSIM-based metric to compute geometric distortions between the left and right reference images and the stitched ones. Recently, Yang *et al.* [36] proposed to compute the local difference of optical flow field energy between the stitched and reference images with the guidance of saliency map in order to quantify content-related structural loss. However, this metric is a full-reference metric. It is not practical since ground-truth reference images are rarely available in most practical cases.

In summary, none of the aforementioned metrics is capable of quantifying ghosting artifacts and structural inconsistencies without the need of any reference images. None of them is proposed based on learning the non-naturalness of these structure-related distortions. None of them can quantify and locate the local non-

uniform structure distortions. More importantly, none of them considers the impact of 'masking effect' (overlap of several different types of distortions) on the perceived quality.

3.7 Limitations of Existing Image Utility Assessment Metrics

In recent years, several metrics have been introduced for the purpose of predicting utility. The natural image contour evaluation (NICE) [67] metric is first proposed to predict utility by comparing contours of a test image and its reference. Later, the multi-scale version of NICE (MS-NICE) is proposed to improve the performance of NICE by comparing contours in each sub-band of a wavelet decomposition, with thresholding and normalization at each subband level. One of the most recent utility metric, the multi-scale difference of gaussian utility (MS-DGU), is proposed in [68] by comparing the number of extrema of a multi-scale difference of gaussian (DoG) decomposition between distorted and reference images.

However, all of these metrics focus only on using structure information and ignore the importance of texture. Texture also plays an important role in certain utility assessment tasks, e.g. material recognition. Moreover, none of these utility estimators are designed according to the goals of the task.

3.8 Limitations of Existing Image Quality Assessment Metrics for Synthesized Texture Images

In the past decade, many metrics have been proposed for the quality evaluation of synthesized texture images. The complex wavelet SSIM (CWSSIM) [88] is proposed to take advantage of the fact that image distortions could lead to changes of magnitude and/or phase of local wavelet coefficients. In [89], a multi-scale weighted variant of the complex wavelet SSIM (WCWSSIM) is proposed to improve CWSSIM with weights based on the human contrast sensitivity function. In [38], Swamy *et al.* propose to use parameters from a texture-synthesis algorithm. Synthesized texture quality assessment (STQA) index is proposed in [65] based on multi-scale spatial and statistical texture attributes. According to [65], CWSSIM, WCWSSIM, parametric metric proposed in [38] and STQA are the four most promising metrics. As discussed in the previous chapter, the acceptance of texture change is task dependent. For example, inpainting in high-frequency texture regions is acceptable in some cases but could be unacceptable in material recognition with the same amount of changes. None of these existing studies design a metric by exploring the role of structure and texture information with respect to the tasks.

3.9 Conclusion

In this chapter, failure cases of using commonly used metrics for evaluating the quality/utility of images/videos in different applications are given. Furthermore, limitations of existing metrics that designed for different applications are also introduced. It is shown that there are glaring needs to develop better image/video quality metrics by considering the characteristics of those local non-uniform structure-related distortions.

4

Relevant Datasets and Evaluation of Performance

4.1 Introduction

This chapter contains two main sub-parts. In the first part, the datasets used in this study are introduced. They are released for different use cases and contain images/videos with structure related-distortions. These datasets are used for 1) experimental comparison; 2) training the mid/higher-level representation based models; and 3) performance evaluation of the proposed models. In the second part of this chapter, the methodologies used for evaluating the performance of the proposed metrics are introduced.

4.2 Datasets

4.2.1 Datasets for Free-viewpoint TV Applications

In recent years, most of the research efforts in FTV scenario are spent on developing encoding approaches and view-synthesis algorithms [1], while the subjective evaluation of the QoE of such system is still limited [90]. For example, no subjective study considers the impact of content related navigation trajectories on perceived quality. This fact may be caused mainly by the technological aspects and novelties related to the visualization of FTV content, in the sense that it provides users with the possibility to freely watch different viewpoints of the scene. For this, displays offering motion parallax (at least horizontal, ideally also vertical) should be employed, such as SMV displays, light-field displays, etc. However, since these types of displays are still under development, and only some prototypes are available, other visualization techniques should be used. For example, conventional screens can be used together with some interactive interface allowing the user to select the desired viewpoint (e.g., browsers, head-tracking systems). This alternative also introduces difficulties in the subjective evaluation

of FTV content, since the provided interactivity makes it difficult to have reliable and reproducible results with traditional evaluation methodologies.

Apart from the preliminary subjective study carried out by Dricot *et al.* [91] that considers coding and view-synthesis artifacts using a light-field display, the majority of the experiments for FTV scenario are conducted using conventional screens and limiting the interactivity of the users. For example, some representative content or predefined trajectories simulating the movement of the observers are shown to the observers. In this way, it is possible to obtain more reliable results, as shown by the fact that MPEG has adopted this type of alternative for their recent standardization activities regarding the evaluation of compression techniques for FTV [92]. In particular, the adoption of this evaluation approach is based on previous subjective studies with SMV [93] through view sweep (i.e., generating videos in which a sweep across the different viewpoints is shown, as if the observer was moving his head horizontally from one viewpoint to another). These studies were carried out to study different aspects of this technology, such as smoothness in view transitions, comfortable view-sweep speed [94], and the impact of coding artifacts [95].

Although these studies provide some insights related to the effects of coding artifacts on quality, the evaluation of view-synthesis algorithms is still an open issue [96]. Therefore, some works that were carried out with previous technologies, like multi-view video, should be taken into account in the study of the effects of view-synthesis on the perceived quality in current FTV applications, such as FN and SMV. To this end, Bosc *et al.* conducted subjective studies to evaluate the visual quality of synthesized views using DIBR-based algorithms. In these studies, the quality performance of view synthesis was evaluated through different ways, such as by showing the observers: 1) synthesized still images [49], 2) synthesized views of Multi-View plus Depth (MVD) video sequences [97], and 3) a smooth sweep across the different viewpoints of a static scene [62]. These different approaches are illustrated in Figure 4.1 (a)-(c) correspondingly. In the following sections, details of the three datasets released along with these three studies are given.

4.2.1.1 IRCCyN/IVC DIBR Images (IVC-DIBR-I/IVC-Image)

The IVC-DIBR image dataset contains 96 images with a resolution of 1024×768. It contains both mean opinion score (MOS) obtained using ACR protocol and pair comparison results. The dataset was designed for benchmarking view synthesis algorithms. Images from this dataset [49, 98] were obtained from 3 multi-view video plus depth sequences. 7 DIBR algorithms (labeled as A1-A7) [3,32,79,80,99,100] were used to process the three sequences to generate four new virtual views for each of them. The dataset is composed of 84 synthesized views and 12 original frames extracted from the corresponding synthesized sequences. Images in this dataset contain only spatial synthesized artifacts caused by view synthesis.

4.2.1.2 IRCCyN/IVC DIBR Videos dataset (IVC-DIBR-V/IVC-Video)

The IVC-DIBR video dataset [59] consists of 102 videos with a resolution of 1024×768 generated with three multi-view plus depth contents. It contains MOS obtained using ACR-HR protocol. The dataset was designed for the evaluation of the reliability of DIBR algorithms by assessing the quality of the synthesized virtual views. Totally 7 DIBR related algorithms, which denoted as A1-A7 [3,32,79,80,99,100], are used to obtain 4 new virtual viewpoints for each content. Apart from the 9 original sequences and the 84 synthesized virtual viewpoints, there are also 9 sequences that contain only traditional compression artifacts obtained by encoding the texture

of the reference sequences. The sequences in this dataset contain only synthesized related spatial artifacts and temporal artifacts within one viewpoint, as there is no navigation among different viewpoints (switch of viewpoints) to mimic free navigation.

4.2.1.3 Time Freeze Free-Viewpoint Synthesized Video dataset (FFV)

The time freeze free-viewpoint synthesized video dataset [62] is composed of 264 videos sequences in resolution of 1024×768 / 1920×1080 generated with six multi-view plus depth original sequences. It contains MOS obtained using the ACR protocol. The dataset was released for the purpose of evaluating the impact of depth coding artifacts on the perceived quality. Since depth maps are important during the DIBR based rendering process, 7 codecs and 3 bitrates were adopted to encode the depth maps for later synthesis process. These 7 algorithms include 3D-HEVC [101], MVC [102], HM 6.1 [103], JPEG2000 [104], lossless-edge based codec [105], proposed in [106] using color frames' correlations, and Z-LAR-RP [107] using local information (they are labeled as C1 to C7 respectively). After generating the synthesized viewpoints between the reference views with a certain configuration, a sequence that navigates from one viewpoint to another (from left to right and vice versa) is generated with 100 key frames extracted from the synthesized viewpoints. Thus, the sequences in this dataset contain synthesized related spatial artifacts and temporal artifacts caused by views switch.

4.2.1.4 Limitations of the Three FTV Datasets

The three FTV datasets mentioned above are released to evaluate FTV contents that represents different degrees of navigation with respect to different type of distortions. Figure 4.1 (a)-(c) show the 'degree of navigation' in these three subjective studies. It is shown from Figure 4.1 (a)-(c) that the first subjective studies only considers spatial DIBR-related artifacts, the second study also considers temporal distortions within the synthesized view, and the third study considers spatial DIBR-related artifacts of all the views, but no temporal distortions. Therefore, a complete evaluation of spatial and temporal degradations caused by view synthesis, which takes content-related navigation trajectories into account, is still missing. It requires the use of content-related view sweep over the views in video sequences (similar to the trajectories designed in [93] but considering view-synthesis artifacts), as depicted in Figure 4.1 (d). To fill this need, a subjective study that considers all the factors as mentioned above is conducted in this thesis to further confirm the impact of content-related navigation trajectory on perceived quality. This novel subjective study is described in section 11.3.

4.2.2 Dataset for Stitched Panoramic Images (SIAQ)

The SIAQ dataset [36] consists of 1224 stitched images, in resolution of 2k $\times$ 3k. This dataset contains only subjective comparison scores. It was released for benchmarking the performance of different stitching algorithms. In total, 34 different contents (varying from scenery landscapes to indoor scenes) are included in the dataset. These stitched images are obtained by stitching adjacent left and right virtual views (totally 12 views for each content covering 360-degree surrounding views) with an off-the-shelf stitching software. The entire dataset can be divided into three groups: reference and two sets of stitched images generated using 2 sets of different parameters. The images in this dataset contain only spatial stitching artifacts.

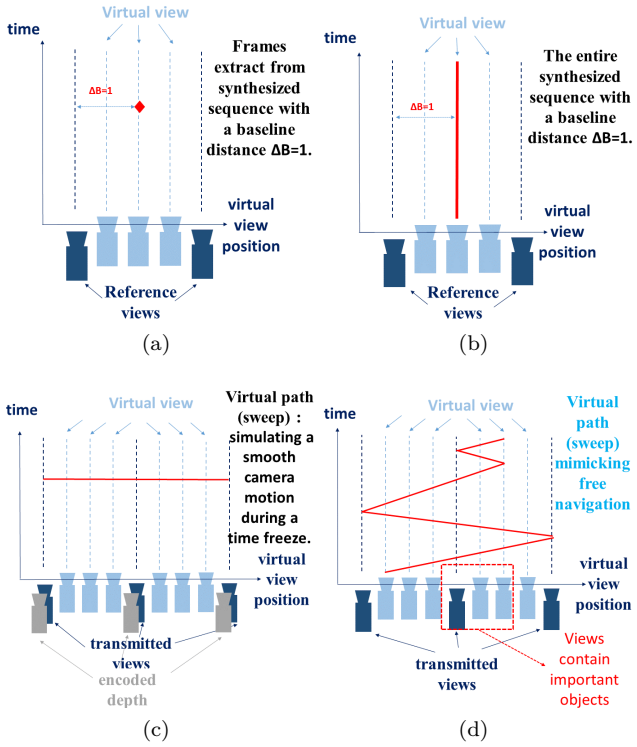


Figure 4.1 – Different possibilities to evaluate FTV content representing different degrees of navigation. (a) Synthesized image. (b) Video from a synthesized view (exploration along time). (c) Video containing a view sweep (exploration along views). (d) Video containing a view sweep from videos of various synthesized views (exploration along time and views)

4.2.3 Dataset for Utility Assessment (Cu-Nantes)

The CU-Nantes dataset [37] consists of 9 reference gray-scale images and 235 distorted images. This dataset contains both perceived quality and perceived utility scores for the distorted images. It was released for evaluating the performance of utility metrics. The quality scores were obtained using the SAMVIQ protocol while the utility scores were obtained using pair comparison. Each image in CU-Nantes is degraded by one of the five processes. These five processes include JPEG compression, blocking, JPEG2000 with dynamic contrast-based quantization, texture smoothing (TS) and texture smoothing with high pass filtering. The images in this dataset contain different level of structure loss due to degradation processes.

4.2.4 Dataset for Synthesis Texture Images (SynTex)

The SynTEX Granularity dataset [39–41] contains 105 synthesized texture images. This dataset is equipped with MOS obtained using ACR protocol. It was designed for evaluating the effect of granularity on synthesized textures, and benchmarking metrics designed for quality assessment of texture synthesized images. To prepare the dataset, 21 reference texture images that contain textures with low, medium and high granularity levels were selected. In this dataset, the synthesis texture images were generated utilizing five different texture synthesis algorithms, including parametric, non-parametric, statistical, and non-statistical approaches. The images in this dataset contain mainly texture synthesized related distortions. Both structure and texture of the synthesized texture images were modified to different extent compared to the reference images.

4.3 Performance Evaluation Methodology

In this section, the measures used for evaluating the performance of proposed objective metrics are described. To be compliant with the standard procedure [108, 109] for assessing the performance of objective image/video quality/utility assessment metrics, the accuracy, monotonicity, and consistency properties of the objective estimation of subjective score are considered. Measurements including pearson correlation coefficient (PCC), spearman rank order correlation coefficient (SCC), root mean square error (RMSE) and the 'Krasula' model [110, 111] are used for performance evaluation in the thesis.

Before calculating PCC, SCC, and RMSE, as recommended in [108, 109], a logistic regression is employed to map the predicted objective quality/utility scores to the subjective ones, with the constraint that the function is monotonic on the interval of perceived quality values:

$$obj_{fit}(obj) = a + \frac{b}{1 + exp\ [-c \cdot (obj - d)]}, \tag{4.1}$$

where a, b, c, d are the parameters of the fitting functions and obj, obj_{fit} are the predicted objective scores and the objective scores after fitting, respectively.

4.3.1 Pearsons Correlation Coefficient

The pearsons linear correlation coefficient (PCC) is computed between predicted objective quality scores and subjective scores to estimate the accuracy of the predicted scores, and is defined as

$$PCC = \frac{\sum_{i=1}^{M_s}(sub(i) - \bar{sub}(i))(obj(i) - \bar{obj}(i))}{\sqrt{\sum_{i=1}^{M_s}(sub(i) - \bar{sub}(i))^2}\sqrt{\sum_{i=1}^{M_s}(obj(i) - \bar{obj}(i))}}, \tag{4.2}$$

where $sub(i)$ and $obj(i)$ are the MOS for a sample i obtained from the observers and the predicted score predicted by the objective models correspondingly. $\bar{obj}(i)$ and $\bar{sub}(i)$ indicate the means of the objective and subjective scores respectively. M_s is the total number of samples.

4.3.2 Spearman Rank Order Correlation Coefficient

The spearman rank order correlation coefficient (SCC) is computed between predicted score and ground truth subjective score to estimate the monotonicity of the objective score:

$$SROCC = \frac{\sum_{i=1}^{M_s}(sub^r(i) - \bar{sub^r}(i))(obj^r(i) - \bar{obj^r}(i))}{\sqrt{\sum_{i=1}^{M_s}(sub^r(i) - \bar{sub^r}(i))^2}\sqrt{\sum_{i=1}^{M_s}(obj^r(i) - \bar{obj^r}(i))}}, \tag{4.3}$$

where $sub^r(i)$ and $obj^r(i)$ represent the rank of a sample i indicated by the MOS and the predicted subjective score respectively. $\bar{sub^r}(i)$, $\bar{obj^r}(i)$ represent the respective midranks and M_s is the total number of test samples.

4.3.3 Root Mean Square Error

For evaluating the performance of objective metrics, the root-mean-square error (RMSE) is also computed between quality scores predicted by the objective models and subjective scores. It is defined as

$$RMSE = \sqrt{\frac{1}{M_s - 1}\sum_{i=1}^{M_s}(sub(i) - obj_f(i))^2}, \tag{4.4}$$

where $sub(i)$ is the individual subjective score of the sample i and $obj_{fit}(i)$ is the predicted objective score of the sample after fitting. M_s is the number of total samples.

4.3.4 Krasula Model

The previous three methods are commonly used for evaluating the performance of metrics for objective quality assessment with respect to subjective ground truth. Nevertheless, those conventional methodologies suffer from at least one of the following disadvantages:

(1) These methodologies do not take into account the uncertainties in the subjective scores. Thus, certain decisions have to be made by the models without knowing the correct behaviors.

(2) These methodologies are vulnerable to the quality range of the stimulus in the experiments. So, mapping functions are required to pre-process the objective scores to compare the performance of the metrics, which are not the same as how the metrics are used in practice.

(3) These methodologies are not capable of dealing with pair comparison results directly, and other models should be used beforehand to pre-process the subjective comparison results.

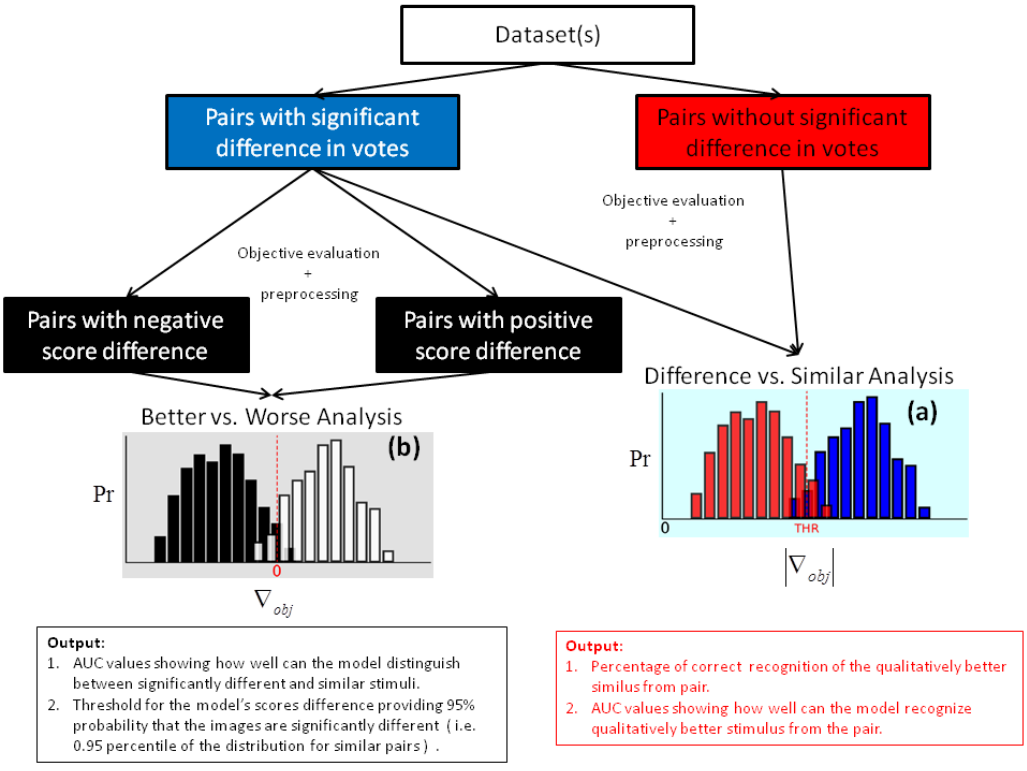


Figure 4.2 – Framework of the Krasula methodology for performance evaluation of objective metrics [110, 111].

In order to better evaluate the performance of different metrics, as well as to better assess the metrics

on datasets that contains only pair comparison results, in this thesis, the methodology proposed by Krasula *et al.* [110, 111] is used. In their model, it is assumed that the capability of an objective metric depends its capabilities of making reliable decisions about 1) when comparing two stimulus, whether they are qualitatively different and 2) if they are, which of them is of higher quality. In brief, the 'Krasula' model is based on determining the classification capabilities of the objective models considering 'Better or Worse' and 'Different or Similar' scenarios.

The entire framework can be divided into five steps and is depicted in Figure 4.2. To use the 'Krasula' model for performance evaluation, the first step is to preprocess the subjective results of the dataset to select pairs that are statistically significantly different from each other with respect to their subjective quality scores. Approaches like ANOVA test can be used for significance difference calculation as mentioned in [110]. After significance difference confirmation, the significantly different pairs are further divided into groups with positive and negative MOS difference for latter different/similar and better/worse analyses. For example, let $sub(i)$ and $obj(i)$ be the subjective score and objective predicted score of a sample i respectively, stimuli pairs that are significantly different can be defined as

$$Pr[sub(i) \neq sub(i)] > 1 - \alpha_{sig}, \tag{4.5}$$

where $Pr[sub(i) \neq sub(i)]$ denotes the probability that the stimuli i and j are qualitatively different and α_{sig} denotes the level of significance. Normally, the value of α_{sig} is set as 0.05 to ensure 95% probability. For subjective scores expressed in form of MOS, the probability of difference can be computed using the formula (4.6) defined in [112]

$$Pr[sub(i) \neq sub(i)] = \Psi\left(\frac{|sub(i) - sub(j)|}{\sqrt{v(i)/N_{obs} + v(j)/N_{obs}}}\right), \tag{4.6}$$

where $v(\cdot)$ is the variance of the stimulus and N_{obs} is the number of observers. For subjective scores expressed in the form of pair comparison scores, a statistical test like Barnard's test [113] can be used to determine whether the preference for one stimulus over another is statistically significant. In other words, the Krasula model is capable of dealing with subjective data containing only pair comparison results.

After categorizing pairs within one dataset into one group that contain pairs with significantly different quality scores and another one that contain pairs that are not significantly, the second step is to pre-process the predicted scores by calculating the difference between the predicted scores of each pair of stimuli i and j:

$$\nabla_{obj}(i, j) = obj(i) - obj(j). \tag{4.7}$$

With the pre-processed objective predicted and subjective scores, the **'Difference vs. Similar' Analysis** can be then conducted to check how well can the objective metric distinguish between significantly different and similar pairs. In this analysis, it is assumed that the difference of the objective scores predicted by a well-performing model, should be larger for significant pairs than for the non-significant ones. With this assumption, the objective metric can be indirectly considered as a binary classifier with categories 'Difference' versus 'Similar'. In detail, one dataset (in the case of analyzing several datasets, those datasets could be merged into one) can

be separated into two groups S and D as

$$\begin{aligned} |\nabla_{obj}(i,j)| \in S &\Longleftrightarrow Pr[sub(i) \neq sub(i)] \leq 1-\alpha_{sig}, \forall i < j \leq k, \\ |\nabla_{obj}(i,j)| \in D &\Longleftrightarrow Pr[sub(i) \neq sub(i)] > 1-\alpha_{sig}, \forall i < j \leq k \end{aligned} \tag{4.8}$$

The capability of the objective metric of categorizing similar and significantly different pairs can be determined by employing the receiver operating characteristic (ROC) analysis on these two sets (ROC quantifies how well are the two sets are separated). Then, the performance of the metric can be verified with the area under the ROC curve (AUC). In the following part of this manuscript, it is denoted as AUC_{DS}. Ideally, the different/similar ROC curve of a well-performing model will look approximately like the example shown in Figure 4.2 (a). As it can be observed, the two distributions are well separated from each other.

Another important analysis, which can be done with the pre-computed subjective and predicted scores, is the **Better Vs. Worse Analysis**. The goal of this analysis is to see whether the objective metric is capable of picking out stimuli that are of higher/lower quality. Similar to the previous analysis, one dataset can be divided into two sets B and W as

$$\begin{aligned} \nabla_{obj}(i,j) \in B &\Longleftrightarrow Pr[sub(i) > sub(i)] > 1-\alpha_{sig}, \forall i,j \leq k \wedge i \neq j, \\ \nabla_{obj}(i,j) \in W &\Longleftrightarrow Pr[sub(i) < sub(i)] > 1-\alpha_{sig}, \forall i,j \leq k \wedge i \neq j, \end{aligned} \tag{4.9}$$

Similarly, ROC is employed on the two sets. The performance of the under-test model can be then evaluated by checking the AUC value of the 'Better vs. Worse' ROC (denoted as AUC_{BW} in the following part of the thesis). Figure 4.2 (b) shows an example of the Better/Worse ROC curve of one well-performing objective model. Apart from AUC_{BW}, correct classification in 0 (CC) defined in [110, 111] is also used as another quantifier to evaluate the performance with respect to whether the stimuli of better quality are assigned with higher objective scores by the objective model. More specifically, CC is defined as

$$CC = \frac{\Omega_{B>0} + \Omega_{W>0}}{\Omega_B + \Omega_W}, \tag{4.10}$$

where Ω_B and Ω_W are cardinalities of sets B and W correspondingly, $\Omega_{B>0}$ is the number of positive members in the set B, and $\Omega_{W<0}$ is the number of negative members in the set W. The numerator of the equation is the total number of the members where the order is correct.

In summary, the advantages of using the Krasula model for performance evaluation can be summarized as below:

(1) The method does not require any mapping to enable numerical comparisons.

(2) The model takes into account the statistical significance of subjective scores and depends less on the quality range of the dataset.

(3) The model enables an easy combination of data from different subjective experiments and provides means to determine the statistical significance of the performance differences.

(4) The model makes it possible to evaluate the performance of one objective metric by considering certain factors.

(5) With this model, it is possible to evaluate the objective metric on databases that contain only pair

comparison results.

4.3.5 Maximum Likelihood Estimation (MLE) based Quality Recovery Model

Uncertainties and noise in raw subjective scores: in most of the studies, where objective quality assessment metrics are proposed, subjective scores in terms of raw mean opinion scores are used without any preprocessing procedure to remove the noise from the observers. However, it is described in [114] that observers' personal characteristics, including viewing experience, ages or their careers, may lead to uncertain ground truth driven by the number of panelists. As a result, the obtained subjective score may be noisy. If we can get rid of the uncertainties and noise before developing the objective quality metrics, the developed metrics could be more robust. To get rid of these uncertainties, a recently introduced recovery model based on maximum Likelihood estimation (MLE) [115] could be used to improve the discriminability of standard subjective quality assessment.

In the MLE model proposed in [115], the individual score of each subjective is considered as a combination of true subjective score and noise caused by subjects' biases and inconsistency. Based on this, individual subjective score of one processed video sequence (PVS) from a subject s_{obs} is defined as

$$X_{e_{obs},s_{obs}} = x_{e_{obs}} + B_{e_{obs},s_{obs}}, \tag{4.11}$$

where $x_{e_{obs}}$ is the true score of a sample e_{obs} and $B_{e_{obs},s_{obs}}$ is the noise from s_{obs}. The noise follows a Gaussian distribution $B_{e_{obs},s_{obs}} \backsim N(b_{s_{obs}}, v^2_{s_{obs}})$ with a mean of $b_{s_{obs}}$ (subject's bias) and variance of $v^2_{s_{obs}}$ (subject's inconsistency). The main goal of the model proposed in [115] is to use MLE model to jointly recover these three unknown parameters $\Theta = x_{e_{obs}}, b_{s_{obs}}, v_{s_{obs}}$. To this end, the log likelihood function is defined as

$$L = log\ P(X_{e_{obs},s_{obs}} | x_{e_{obs}}, b_{s_{obs}}, v_{s_{obs}}), \tag{4.12}$$

and the three parameters can be obtained by solving $\hat{\Theta} = argmax_{\Theta} L$. Each parameter's estimation is associated with a 95% confidence interval and is computed as

$$\hat{\Theta} \pm 1.96 \frac{1}{\sqrt{-\frac{\partial^2 L(\hat{\Theta})}{\partial \Theta^2}}}. \tag{4.13}$$

According to the experimental result reported in [114] on a usual ACR-dataset:

1. The uncertainties from subjects can be removed by using the quality recovery MLE model. Since most of the proposed models in this thesis are learning-based models, this MLE model is a useful tool to clean out the noise. After removing the uncertainty from the subjective scores, learning-based model can thus avoid being affected by the noise from the subjective scores and provide more robust performance.
2. Since the variance (in terms of confidence interval) of MOS could be reduced (proven in [114]), this model could be used as a pre-process procedure before using the 'Krasula' model. More specially, since traditional significant test like 't-test' and 'ANOVA' test may fail to select all the significant pairs from the data due to the large variance among observers, this model is of potential to be employed to select more significant pairs for the latter analysis by reducing uncertainties from observers.

As this is a recent study released not long ago, experimental results using this new model could not be finished by the time this thesis is finished. Therefore, relative analyses are considered as future work to improve

the proposed models in this thesis.

4.4 Execution Time

To better compare the complexity of metrics executed on different datasets with different machine, an execution time normalized based on PSNR is computed as done in [42]. For a given image I from a database, the normalized execution time is defined as

$$T_{nor}^{exe} = \frac{T_{obj}^{exe}}{T_{PSNR}^{exe}}, \tag{4.14}$$

where T_{obj}^{exe} is the execution time of using one test objective model to predict a quality score for image I, and T_{PSNR}^{exe} is the one for PSNR.

4.5 Conclusion

In respect of datasets that used for benchmarking/training different models in different applications, a total of 6 datasets are described in this chapter. It must be pointed out that although there are already some datasets designed for FTV scenarios, there is still a lack of dataset including sequences that contains both temporal synthesized artifacts within one viewpoint and temporal structure inconsistencies due to views switch, as described in section 4.2.1. With this objective, a new subjective experiment that takes both types of temporal artifacts into account is presented in section 11.3. The purpose of this experiment is to check the impact of navigation scan-paths on perceived quality.

In respect of performance evaluation measures, several methodologies are introduced. In this manuscript, PCC, SCC, and RMSE are used throughout all the experimental sections for objective metrics' performance evaluation. In certain cases, the 'Krasula' model is used for deeper and more reliable analysis. For example, for datasets like SIAQ (introduced in section 4.2.2), where only pair comparison results are provided, the 'Krasula' model provides advantages for better evaluating the performance of the metrics under-test.

5

From Existing Problems to Main Research Questions

In chapter 2, we highlighted specific structure-related distortions that are appearing in modern visual media technologies. In chapter 3, it has been illustrated that common used visual quality predictors fail to quantify the perceptual impact of those artifacts. As characteristics of those structure-related distortions are different in different applications and scenarios, it appears meaningful to consider application use case when targeting effective visual quality prediction. Such investigation requires relevant datasets (visual content with corresponding observers opinion on visual quality) aligned with the application use case. In chapter 4, we have seen that many of these datasets are available, nevertheless for some use case such navigation trajectories between views along time in video, dataset is missing. To solves these problems, this dissertation focuses on two research questions summarized in the following section.

5.1 Main Research Questions

- How to quantify **spatial** structure-related distortions based on the representative mechanism in HVS (low, mid, higher-level representations) ?

  As introduced in 2, spatial structure-related distortions are challenging to be captured and quantified due to their characteristics. Although these characteristics varies according to the applications under different scenarios, there are common properties among them (e.g., non-uniform, locally distributed, disrupt the semantics of images/videos to different levels). Models developed based on different level may have different representation capability and be suitable for different applications. Thus, how to develop a proper model for a target application in a certain scenario by using the right concept of perceptual representation (i.e., using low, middle, or higher-level representation) is one the main focuses of the thesis.

- How to quantify **temporal** structure-related distortions based on the representative mechanism in Human Visual System (low, mid, higher-level representations)?

  Temporal structure-related distortions in immersive multimedia applications are more difficult to quantify since 1) new factor 'navigation trajectory' is involved; 2) spatial structure-related distortions introduce temporal structure inconsistencies.

To study these research questions, effective image/video low, mid, and higher-level representations, especially structure-related representations, are investigated according to the characteristics of the distortions for a given use case. The rest of this dissertation is divided into three parts according to the three perceptual representation levels. The two main research questions are further decomposed into more specific questions in each part. As illustrated in Figure 5.1, in each part, quality assessment models are proposed and tested on different datasets considering different applications and scenarios:

Part II (low-level representation): models exploiting low-level representations are explored for image/video utility/quality assessment in different applications. Firstly, in order to check the roles of structure and texture information in different tasks, a bilateral-filter based model (BF-M) is proposed by utilizing bilateral filters to separate structure information from the texture. Secondly, in order to quantify the structural deformation of an image, an elastic metric based image quality assessment metric (EM-IQM) is proposed. It is then extended for video quality assessment as an elastic metric based video quality assessment metric (EM-VQM) by quantifying 1) the deformation between multi-scale motion trajectories in synthesized and original sequences; 2) the structure dissimilarities along the trajectories.

Part III (mid-level representation): models exploiting mid-level representations are explored for developing image/video quality assessment metrics. Mid-level contours 'encoding' methods are adopted to mimic the 'encoding' process of low-level structure information to mid-level structure representations. First of all, a sketch token based image quality assessment metric (ST-IQM) is proposed to quantify the geometric distortions by checking how the categories of contours change from a mid-level point of view (i.e., using a bag of word model to 'encode' contours into contour categories). Furthermore, since no existing database takes the impact of content related navigation scan-paths on perceived quality into account, a subjective study is conducted and introduced. It includes generated sequences with different content related viewing trajectories for FTV scenario. A novel **free viewpoint videos database (FVV)** is also released and presented. Sketch token based video quality assessment metric (ST-VQM), the extension of ST-IQM for video quality assessment, is introduced based on quantification of contours' classes temporal evolution.

Part IV (higher-level representation): higher-level representations are investigated for developing image/video quality assessment models. Considering that human visual system uses 'sparse mode' to compress low, mid-level information from V1, a no reference convolutional sparse coding based image quality metric (CSC-IQM) is proposed to quantify local non-uniform geometric distortions by learning the non-natural structures. Moreover, considering that 1) 'high-level' features extracted from deep neural network intermediate layers are well connected to semantics; 2) discriminator in generative adversarial network (GANs) model is usually trained to distinguish artificial images from the real ones with respect to the statistic of the training set, a no reference GAN based image quality assessment metric (GAN-IQM) is proposed to quantify local non-uniform inpainting related distortions.

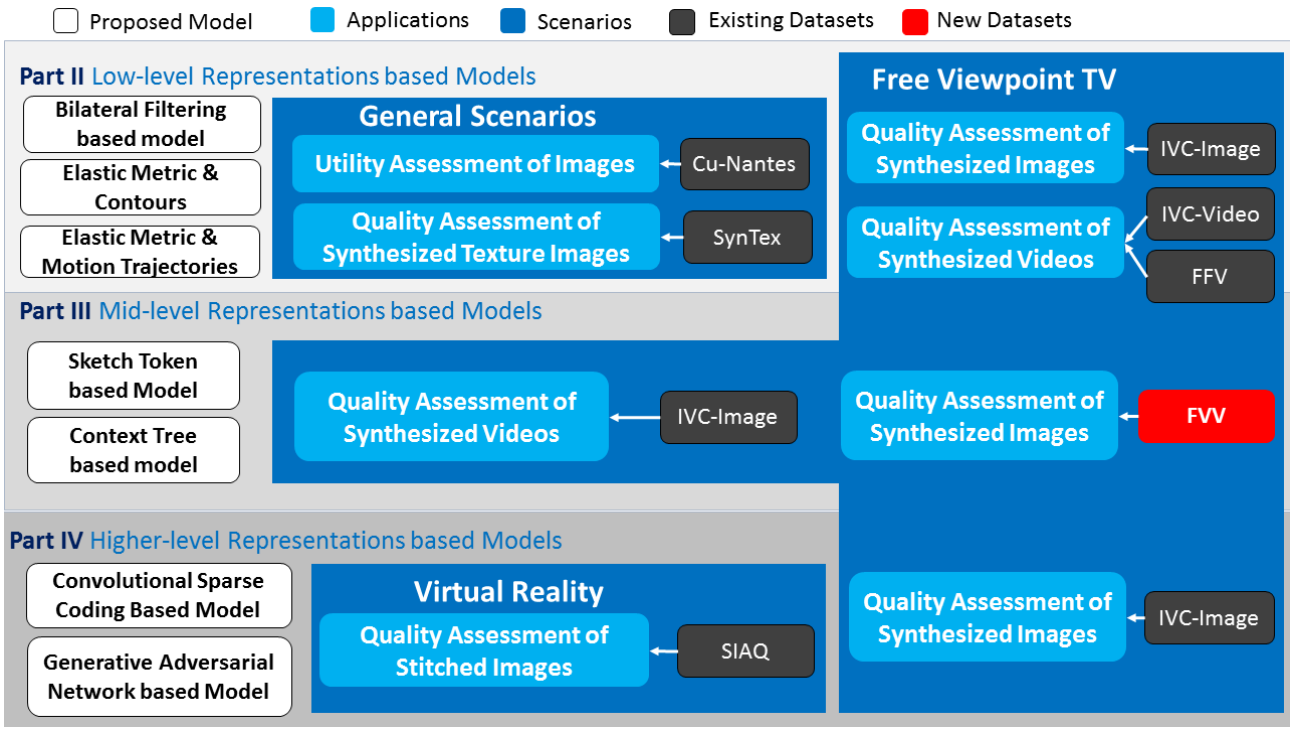
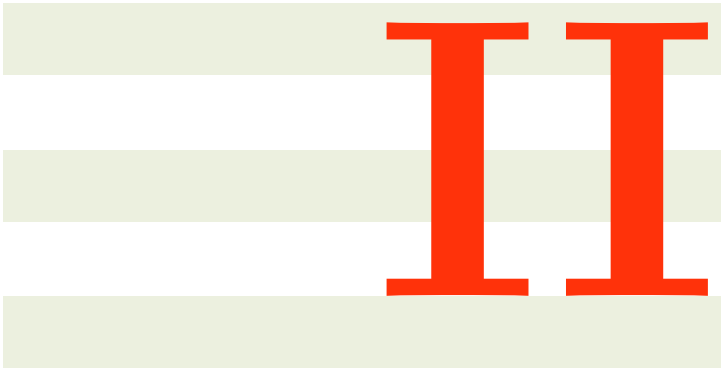


Figure 5.1 – Overview of the following parts of the dissertation: low, mid, and higher-level representation based models are proposed and tested in different applications under different scenarios on different relative datasets.



Exploring Low-Level Representation for Image/Video Quality Assessment

6

Introduction of Part 2

Low-level representations of images/videos are defined as local features and descriptors that represent local image/video basic information. In this part, based on the importance of structure and texture information in Human Visual System, two 'low-level representation' models are proposed using structure/texture features.

6.1 Low-Level Information in Human Visual System

6.1.1 Low-Level Structure Information

It is confirmed in [116] that low-level structure-related properties of the images are the foundation of categorical patterns of brain activity within scene-selective regions. The ability to perceive and recognize different visual scenes is essential for spatial navigation in the world. Although real-world scenes can be incredibly complex and heterogeneous, human observers are able to reliably recognize and categorize images of objects/scenes even when the images are shown briefly [117]. Human is able to extract important fundamental structure of the images/videos in form of low-level representations for latter higher semantic understanding. These studies have been taken to suggest that the initial perception of natural images is based on the global, visual properties 'the gist' of the scene [117]. Moreover, the human visual system tends to perceive global structure first and then fine-grained details of an image at the first glance. The procedure of processing a scene proceeds from the top of the hierarchy to the bottom (global-to-local) [118]. In other words, the global structure of a visual object within an observer's effective global span is comprehended before its local features. It has been pointed out in [118] that the global precedence accelerates several possible advantages including utilization of low-resolution information, the economy of processing resources, and disambiguation of indistinct details. Therefore, it is intuitively appealing to assume that structure information (i.e., edges, contours etc.) plays a greater role in tasks like utility assessment, where the objectiveness is to evaluate the usefulness instead of the perceived quality of a distorted natural image. As low-level structure representations (e.g., representations in form of edges/contours)

are important for human to extract basic information of an image/video on the first hand, especially in certain tasks, severe deformations and geometric transformations of structures may affect the process of information extraction and hence affect the perceived quality/utility later on.

6.1.2 Low-Level Texture Information

On the other hand, "visual texture" is usually defined as the portion of an image that is filled with repeated elements and often subject to some randomization in their location, size, orientation and so on [119]. The importance of texture related representations in early scene identification has also been proven in [120]. As mentioned in [120], a holistic cue is defined as a cue processed over the entire visual field and without requiring attention to analyze local features. Texture can be processed quickly and in parallel over the visual field [121], making it a candidate (of holistic cues) as well. Subjects can rapidly identify scenes without color. An image region with one texture seems to 'pop-out' or segregate easily from a background region with a perceptually different texture. Julesz *et al.* [122] claim that the first order statistics of 'textons' determine the strength of texture discrimination and make rapid discrimination possible. Textons are the elements that govern human's perception of texture. They are further described to be locally conspicuous features such as blobs, terminators and line crossings. Firstly, natural texture (texton) provides an important source of information of visible surfaces and details [123]. It is thus important for many tasks like material recognition and image/video quality assessment (where texture descriptors were usually utilized as a proxy to quantify blurriness). Secondly, texture cues in images provide human observers with a potentially rich source of surface and shapes of objects [124]. In summary, the texture is also important for human observers when viewing a scene as it also related to structure. Distortions like blurriness (e.g., blurriness introduced by inpainting in DIBR process) interrupt the characteristic of 'textons' and thus interrupt the 'pop-out' process.

6.1.3 Structure and Texture Information in Quality Assessment:

In the field of quality assessment, distortions on both structure and texture regions affect how human observers judge the quality of an image. For instance, a three-component weighted SSIM (3-SSIM) has been proposed in [125] by assigning different weights to the SSIM scores according to the type of local regions: edge, texture, or smooth areas. Another example is the quality assessment of synthesized views in FTV scenario. Different from common images/videos, synthesized views generated based on DIBR algorithms contain artifacts mainly around dis-occluded regions, including objects shifting, twisted shape of objects, and blurriness along edges. It can be visually observed that structure related distortions (e.g., geometric distortions) and texture related distortions (e.g., blurriness) affect unequally the process of evaluating the quality of the synthesized images/videos. Metrics that taking both information into account are needed for such tasks.

6.2 Research Questions Associated with Low-Level Representation Models Development

According to the discussion above, in this part, we explore low-level descriptors that can represent low-level information, as perceptual models. This investigation can be decomposed into more specific questions:

- Low-level representations of images/videos for quality/utility assessment in different tasks (Part II)

- ☐ How to confirm the roles of low-level structural and textural information in different tasks?

  As it has been discussed in chapter 2 that both texture and structure information is important for certain tasks, e.g., textures are important in the task of quality assessment of texture synthesized texture images while structure information is important in the task of image utility assessment. If one can tell which information plays a greater role than another in those tasks, the task can be then accomplished easier towards a correct direction by assigning higher weights/penalties to more critical information's degradation.

- ☐ How to separate low-level structural and textural information?

  As mentioned in chapter 2 and also in the previous research question, the roles of structural and textural information differ according to the tasks. One should first explore their roles in those tasks before designing a model to handle the task. However, before indicating which information is more important than another, one should be able to separate this two information properly.

- ☐ If structural and textural information play different roles in a task, do the roles change with the quality of the images?

  As mentioned in chapter 2, the relationship between perceived utility and quality is linear when the quality of an image is lower than a certain threshold [37], while the one between them is non-linear when the quality of the image is higher than that threshold. One inherent assumption may be that, for specific tasks, different information play different roles in different quality range.

- ☐ How to quantify distortions according to a specific task with low-level representations?

  As introduced in chapter 2, even though the structure related distortions appear in images/videos in different applications are similar, they are still not the same. For example, the geometric distortions in synthesized texture images and the ones in stitched images are different. Furthermore, as posted in previous questions, different information may play a different role in different tasks.

- ☐ How to quantify structural degradations regarding non-uniform contour deformations for image quality assessment without over-penalizing uniform global endurable distortions?

  As it has been presented in the examples shown in chapter 2, deformations of object shapes are one of the common structural distortions that exist in applications like FTV. This type of severe distortions is less acceptable for observers since they interrupt the structure of images/videos. Moreover, as pointed out in chapter 3 that, most of the point-to-point metrics over-penalize continuous global geometric distortions. Therefore, metrics designed for these cases should also be robust to global uniform distortions.

- ☐ How to quantify the structure related temporal distortions (introduced by spatial structure distortions) observed at one viewpoint location in FTV application?

  As discussed in chapter 2, in the case of FTV, there are spatial-temporal distortions in free viewpoint sequences as frames evolve due to the spatial geometric distortions. This kind of structure inconsistent within one viewpoint is different from the traditional temporal coding artifacts and are challenging. In addition, as concluded in chapter 3, there is still no efficient video metrics that are capable of quantifying this type of temporal artifacts and meet the need of the system.

7

The Roles of Structure and Texture information in Different Tasks

7.1 Introduction

This chapter introduces the first low-level representation based utility/quality assessment model. It is developed based on leveraging low-level edge/contour and texture based estimators with bilateral filtering. This proposed model is tested in task of utility assessment, quality assessment of synthesized texture images, and quality assessment of synthesized views.

The HVS of image perception is hierarchical. Humans tend to first perceive global structural information such as shapes and later focus on local details such as texture. Furthermore, it is widely believed that structure information plays the most important role in task of utility assessment and quality assessment [37], especially in new scenarios like free-viewpoint television, where the synthesized views contain geometric distortion around objects. We hypothesize that the degradation of structural information in an image is more annoying for human observers than the one of texture in certain application scenarios. To verify our hypothesis, a perceptually inspired bilateral filtering based model (BF-M) is proposed. In this scheme, a bilateral filter is first adopted to extract the structure and texture information separately based on a subjective study of Human Material Perception in [126], i.e., structural features are extracted from the filter response while the textural features are extracted from the residual after bilateral filtering. Then, 1) a'NICE' based edge estimator named 'bilateral natural image contour evaluation' (BI-NICE), 2) a shape related estimator named 'bilateral histogram of oriented gradients estimator' (BI-HOG), and 3) a texture estimator named 'bilateral local radius index estimator' (BI-LRI) are introduced by calculating the dissimilarity between the original and distorted images based on low-level features. Lastly, the model is designed by leveraging the weights of the three proposed basic estimators to yield the best performance in different tasks. By doing so, one can determine to what extent the disruption of different information in an image affects the perceived quality/utility of images in different tasks.

Figure 7.1 is an example explaining the fundamental idea of the proposed bilateral filtering based model: (1) By only observing the edge map of the response of bilateral filtering (fourth column in Figure 7.1), it is apparent that one can recognize the shape of the 'teddy bear' easily from the first image (first row,fourth column), while it is difficult to tell the second one (second row, fourth column) is an image of 'wood floor'. (2) For the third image from the IVC-Image dataset, one can observe not only the geometric distortions around objects but also the blurred regions. Apparently, the structure-related distortion is more annoying considering the fake edges and changes of shape around the women. (3) For the fourth image from the SynTEX dataset, one can see that the structure of the stones has been emphasized by comparing the edge maps of the original image (the fourth row, second column) and the one of the response of bilateral filtering (the fourth row, fourth column). The unrelated texture of the stones has been removed after bilateral filtering. It is thus more reasonable to extract structure-related features from the response instead of the original image. (4) The last two images in Figure 7.1 are from the CU-Nantes dataset with different levels of quality. The previous one (the fifth row) is the reference of the other one (the sixth row). By checking the last column of these two images (i.e., the residual obtained by subtracting the response of bilateral filter from the original image), one can see that there are more details/texture information maintained in the residual of the reference image. An intuitive assumption could be made based on this observation that texture plays a more important role for higher quality images in specific tasks.

As discussed above, it is evident that the effect of degradations on structural and texture regions differs with tasks. To the best of our knowledge, no related work explores the roles of structure and texture information in different tasks. In this study, our hypothesis and the model performance are verified on CU-Nantes database as a utility estimator and on SynTEX and IVC-Image database as a quality estimator. In the following sections, more details about the proposed model are given.

7.2 Hypothesis and Theoretical Foundation

As discussed in [127], on the one hand, structure information in a visual scene provide the human visual system (HVS) with more semantic information. Continuous edges/contours of an image could reveal the visual objects inside one image (e.g., people). Those structural edges are essential for the HVS and should be maintained as much as possible in digital image processing for tasks like object detection. On the other hand, textures are usually the surface of objects which can also be the material of the targets, such as texture patterns of clothes, grass, sea and buildings' surface. Texture contains details of objects, could thus further augment the objects with more appealing properties (as fine texture, smooth gray-scale transition and vibrant color) and make them vivid to human observers. In summary, structure and texture jointly construct and enrich the visual scenes. Structural contours provide the human visual system with most of the semantic information while textures provide the details [126]. Therefore, we hypothesize that features that focus on structural properties and features that measure details play different roles in different applications. To verify this, in the following sub-sections, we first explain the reason why local edges/contours can represent the structure and then further discuss how we separate texture from structure and extract different features separately afterward.

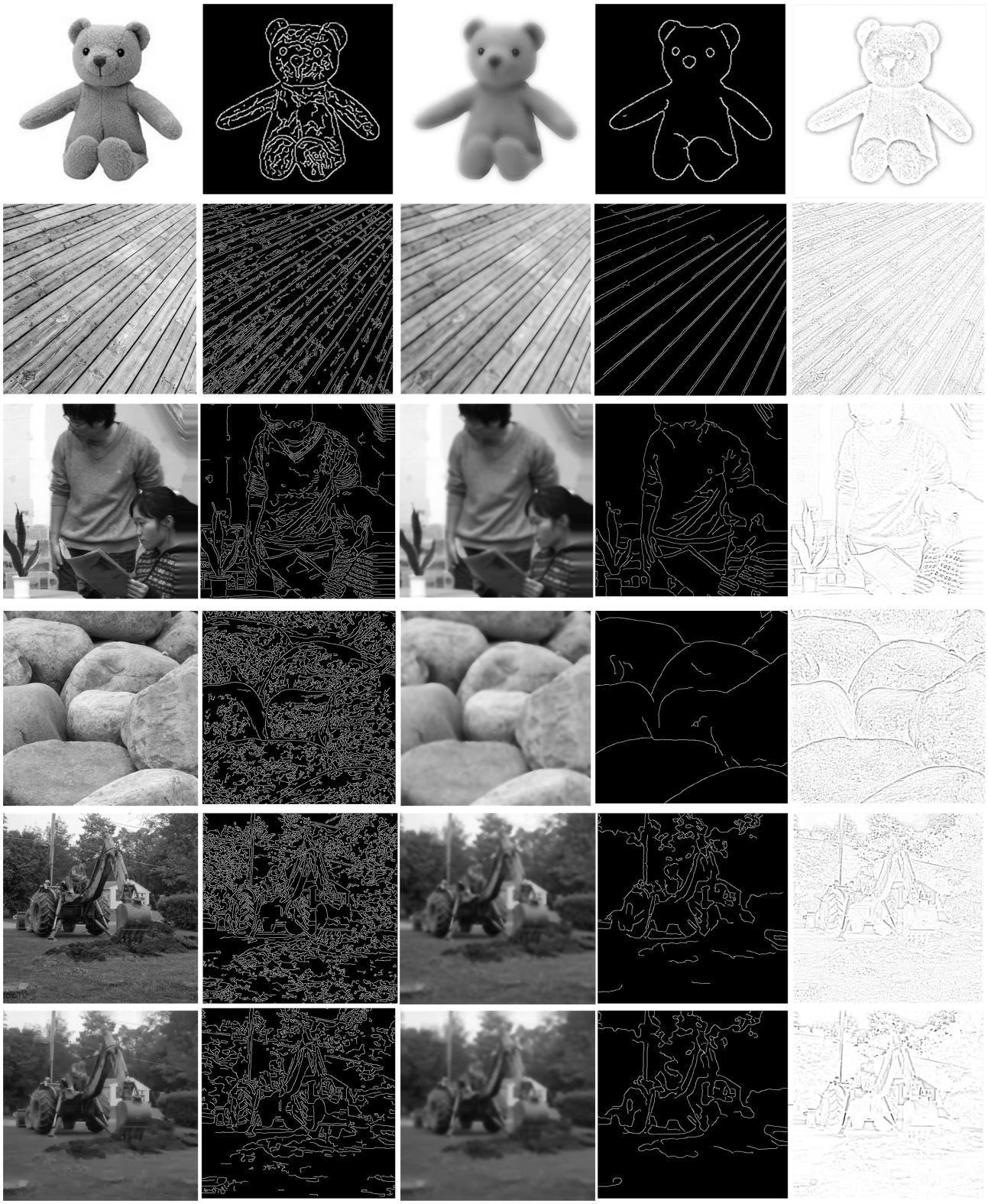


Figure 7.1 – Example of separating structure information from texture information. First column: original image; Second column: edge map of the original image; Third column: response of the of bilateral filter on the image; Forth column: edge map of the response of the bilateral filter; Fifth column: residual of bilateral filtering obtained by subtracting the original image with the response.

Local Edges/Contours Reveal Structure

According to [119], the perception of complex visual patterns and objects appears from neural activity as it is transformed through a cascade of areas in the cerebral cortex. Neurons in the primary visual cortex (V1) are selective for local orientation and spatial scale of visual input [128–130]. Downstream regions (V2-V4) contain neurons selective for more complex attributes, which is approximately achieved by assembling particular combinations of their upstream afferents. Considering the ubiquity of orientation selectivity in primary visual cortex [131], it is intuitive to assume that its computational purpose is to represent the orientation of edges.

Furthermore, over the past decades, the mainstream view in both biological and computational vision communities is that later stages of scene processing should somehow combine these local edge elements to construct more extensive contours, eventually leading to shapes, forms, and objects [132]. Until recently, most researches on object recognition were built around this paradigm, as well as much of the study of mid-level pattern perception, and physiological measurements in areas V2 and V4.

Local edges and contours, which are local structural information, are the vital foundation for the following processes of higher level semantic structural understanding of images. Edges and contours features are important elements that reveal the structure information of an image. Therefore, in the proposed model, a contour based estimator and a histogram of oriented gradient based estimator are designed to quantify the amount of structural changes due to disruptions. Detail of these estimators will be given in the following sections.

Separating Structure from Texture

Subjective test done in [126] about Human Material Perception provides us with important clues about how to extract structural and textural features separately. In [126], a subjective experiment is conducted in order to study which features are useful for the recognition of material categories. In this experiment, images emphasizing local surface information and global structure information were generated separately with bilateral filtering, which is usually used as a non-linear, edge-preserving and noise-reducing smoothing filter for images.

More specifically, Sharan *et al.* followed the idea of Bae *et al.* [133] to extract the micro-structure of the surface by smoothing an image with bilateral filtering. Afterward, they utilized the residual image for further texture analysis. The residual image was obtained by subtracting the bilateral filtering results from the gray-scale versions of the original images to emphasize details of surface structure, which is an operation similar to high-pass filtering. In their subjective test, observers were asked to categorize those distorted images into ten material categories. Based on their results, they concluded that texture is an important attribute of material appearance, while information about surface micro-structure are often related to certain categories (higher level semantics). These analyses are conducted based on using bilateral filter to leverage the two information.

Based on their study, bilateral filtering is used as a proxy to separate structure and texture information. In the proposed model, structure features related to shape are extracted from the response of bilateral filtering while texture features are extracted from the residual.

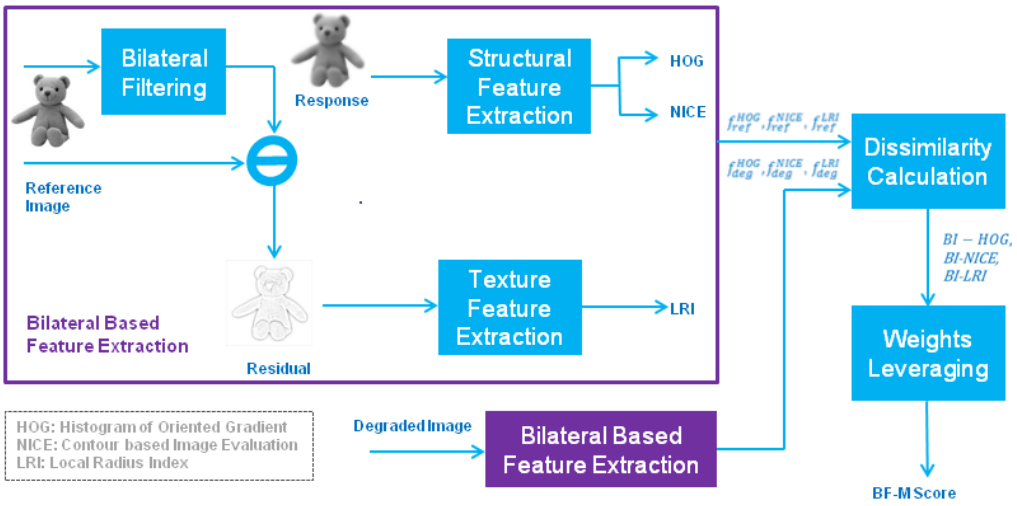


Figure 7.2 – Overall framework of the proposed model based on separating structure and texture information using bilateral filtering

7.3 The Proposed BF-M Model for Validating the Proposed Hypothesis

In order to verify the roles that structure and texture information are playing in different perceptual tasks, a model based on separating this two information is proposed in this section. Figure 7.2 is the overall framework of our proposed model. First and foremost, structure and details related features are extracted separately with bilateral filtering from both of the original and the degraded images. More specifically, images are first separated into a base image (bilateral responses) and the residuals after bilateral filtering [126]. In order to generate a response more efficiently, a faster approximated bilateral filter is used [134] . The scale σ_s of the spatial kernel and the range value σ_r are set according to the tasks [135]. Structure-related features, including the histogram of oriented gradients estimator (HOG) and the natural image contour evaluation estimator (NICE), are calculated with the base image,whereas texture related feature local radius index (LRI) are extracted from the residual image. With the extracted feature sets f^{HOG}, f^{NICE} and f^{LRI} from both the reference and degraded images, dissimilarity scores are then calculated. After normalization, the three estimators, $BI\text{-}NICE$, $BI\text{-}HOG$, and $BI\text{-}LRI$ are combined with different assigned weights according to different applications. Finally, the roles of different information can be investigated by checking the optimized weights

A. Bilateral Filtering based Contours based Image Evaluation Estimator (BI-NICE)

As discussed in section 7.2 that local edge/contours provide important structural information to observers, in this section, a contour based estimator is thus introduced. It has been confirmed that fragments of contours can be used to successfully identify semantics in images [136–138]. This further showcases the importance of structure information in semantics related tasks. Since contours are important for global structure understanding, the NICE estimator is improved by using a bilateral filter to emphasize important structural local elements. First of all, the edge maps are generated only on the responses of bilateral filters using the Canny edge detector. For reference and degraded images, the obtained contour maps are then denoted as C^{BI} and $C^{\hat{BI}}$ respectively.

To probe and expand the shapes contained in the images, contour maps are subjected to morphological dilation with a 3×3 'plus-sign' shaped structuring element E_{se}. In line with the one-scale NICE estimator, the

object score was computed by comparing the binary contours maps of the reference and the test images. Then, the final contour error map is obtained by exerting point-wise exclusive-or (XOR) operation of the dilated binary images, since XOR is the commonly used operation for contours maps comparison. The overall $BI\text{-}NICE$ score for a test image is defined as

$$BI\text{-}NICE = \frac{D_H(C^{BI} \otimes E_{se}, \hat{C^{BI}} \otimes E_{se})}{N_C^{BI}}, \tag{7.1}$$

where N_C^{BI} is the number of contours elements, $D_H(X,Y)$ denotes the Hamming distance between the X and Y, and $C^{BI} \otimes E$ denotes the dilation operation of the contour map C^{BI} with the morphological structuring element E.

B. Bilateral Filtering based Histogram of Oriented Gradients Estimator (BI-HOG)

Considering the fact that histogram of oriented gradients (HOG) [139] is a powerful shape related descriptor used in computer vision and image processing for object detection, action recognition etc., we extract HOG features from each response of bilateral filters as a higher level structure feature. First, each image is divided into 8×8 cells/blocks. After calculating the histogram of each cell, spatial pooling strategy based on visual importance [137] is utilized to pool the dissimilarity values. This pooling strategy is presented based on the perception study that humans tend to perceive 'poor' regions (i.e., regions with visible/severe distortions) in an image more severely than the 'good' ones (i.e., regions without visible distortions).

Finally, the shape related estimator named as bilateral HOG estimator ($BI\text{-}HOG$) is then defined as

$$BI\text{-}HOG = \frac{1}{|b_{ij} \in B_p|} \sum_{b_{ij} \in B_p} D_e(H\text{-}HOG_{ij}^R, H\text{-}HOG_{ij}^D), \tag{7.2}$$

where $H\text{-}HOG_{ij}^R$ and $H\text{-}HOG_{ij}^D$ denote the histograms corresponding to the cell at the i_{th} row and j_{th} column of the bilateral response of both the reference and distorted images. B_p is the lowest 60% cells ranked according to the dissimilarity values (lowest quality/ highest dissimilarity values). $D_e(X,Y)$ denotes the euclidean distance between the two vectors X and Y.

C. Bilateral Filtering based local radius index Estimator (BI-LRI)

To represent detailed information in images, texture related features are considered in this section. Different from [126], instead of extracting micro-jet and micro-SIFT features, the local radius index (LRI) [140] texture descriptors are extracted in this chapter with a size limit of $K = 4$ and a threshold T_{LRI} equals to the standard deviation of the image divided by 2. Similar to $BI\text{-}HOG$, LRI texture descriptor are extracted based on 8×8 cells/blocks. After extracting the LRI descriptors from the residual of bilateral filtering from both of the reference and degraded images, the texture based estimator named as bilateral LRI estimator ($BI\text{-}LRI$) is then defined as

$$BI\text{-}LRI = \frac{1}{|b_{ij} \in B_p|} \sum_{b_{ij} \in B_p} D_e(H\text{-}LRI_{ij}^R, H\text{-}LRI_{ij}^D), \tag{7.3}$$

where $H\text{-}LRI_{ij}^R$ and $H\text{-}LRI_{ij}^D$ denote the LRI feature histograms corresponding to the cell at the i_{th} row and j_{th} column of the bilateral residual of both reference and distorted images.

D. The Final Bilateral based Model

As discussed at the beginning of this chapter, structural and textural information plays different roles in different applications scenarios. Therefore, we combine the three proposed estimators, and weights of them are tuned as parameters according to applications. The output of each estimator, which is the dissimilarity value calculated based on each feature, is normalized to a range of $[0, 1]$. Finally, the proposed $BF\text{-}M$ model, which can also be utilized as a task-based parametric image metric, is designed as

$$\begin{aligned} &BF\text{-}M = 1 - (\alpha_{BI} \cdot BI\text{-}NICE + \beta_{BI} \cdot BI\text{-}HOG + \gamma_{BI} \cdot BI\text{-}LRI) \\ &s.t. \;\; \alpha_{BI} + \beta_{BI} + \gamma_{BI} = 1, \end{aligned} \tag{7.4}$$

where α_{BI}, β_{BI} , γ_{BI} are the aforementioned weights used for fine-tuning the roles of the contour, shape, and texture based estimators respectively, $\alpha_{BI} + \beta_{BI} + \gamma_{BI} = 1$. The configurations of these three estimators are set differently according to the specific task in our experiments and are further discussed to investigate the functionality of different information in images in the following section.

7.4 Results and Analysis

To verify the assumption that structure information like edges/contours do not play the same roles as detail information like texture in different tasks, the proposed $BF\text{-}M$ model described in the previous section serves as an utility estimator on the CU-Nantes dataset [37] and as a quality estimator on both the SynTex dataset [39–41] and the IVC-Image dataset [49, 98]. Details of these three datasets are given in section 4.2. With the best-fit weights assigned to $BI\text{-}NICE$, $BI\text{-}HOG$ and $BI\text{-}LRI$, the roles of both structure and texture information in the correspondence tasks can be uncovered.

The proposed model is applied to different tasks by tuning the weights, performances of the metrics and roles of different information in different tasks are analyzed in each of the following sub-section for each application. Model performances are evaluated according to the PCC, SCC, and the RMSE as introduced in section 4.3.

A. Results: Objective Estimates of Perceived Utility

In utility assessment task, observers estimate the usefulness of a natural image as a substitute for a reference. In such a task, structure information is important since the primary purpose is to quantify the amount of useful information from an image instead of evaluating its quality. According to what has been analyzed in [37] based on the results obtained on the CU-Nantes database, there is a linear relationship between perceived quality score and perceived utility score for images with quality scores under 30. On the contrary, the relationship between quality and utility score is non-linear for those whose quality scores are higher. It is concluded that observers evaluate very low quality images by checking whether the content is interpretable. About why the relationship between perceived quality and utility of higher quality images is non-linear, one possible explanation could be that texture information play different roles in the task of utility and quality assessment for higher quality images. Because the higher the quality, the more details will be maintained. Disruption of texture (blurriness) is annoying for human observers while judging the quality of the image. For example, in Figure 7.1, the image in the last row is one degraded image while the one in the fifth row is its reference image. It can

be observed from the last columns of the two rows (the residual of the correspondence images) that there is more texture information in the residual of the reference image than the one of the degraded image. In the case of evaluating the quality of higher quality images, details are important to distinguish one quality level form another. On the contrary, in some utility task, the perceived utility remain the same after reaching a certain quality threshold. Therefore, we also hypothesize that the roles of structure and texture information in the task of utility assessment vary with the quality of the image.

To confirm the assumption that (1) structure information play a main role in utility assessment, (2) the role of texture and structure differs in different quality ranges, the proposed *BF-M* model is utilized as utility estimator and is tested on the CU-Nantes database [37] introduced in section 4.2.3. To further check how the weights of different information vary with quality, one best configuration will be selected for each sub-interval divided according to the perceived quality scores. To confirm the feasibility of using the proposed model as utility estimator as well as to check the roles of different informations, ReDLOG [141], most apparent distortion (MAD) [142] metric, multi-scale SSIM (MS-SSIM) [45], the visual information fidelity criterion (VIF) [52], the contours based image evaluation (NICE) [67] metric, the multi-scale version of NICE (MS-NICE) and the Multi-Scale Difference of Gaussian Utility (MS-DGU) [68] metrics are chosen for utility prediction performance evaluation.

Table 7.1 – Performance of the proposed parametric metric with different parameters in different quality ranges.

SCC	**Quality Range**			
$\alpha_{BI}, \beta_{BI}, \gamma_{BI}$	**[1 , 2]**	**[2 , 3]**	**[3 , 4]**	**[4 , 5]**
1 ,0 ,0	0.687	0.752	0.659	0.661
0.9,0.1,0	**0.696**	**0.756**	0.719	0.854
0.8, 0.1,0.1	0.681	0.743	**0.737**	0.831
0.7, 0.1,0.2	0.694	0.755	0.728	**0.888**

In the experiment, since each sample in the database is labeled not only with a utility score but also with a quality score ranging from 1 to 5, we divide the whole range into quarters and optimize one configuration for each subrange. Table 7.1 illustrates the correlation between objective and subjective scores in different quality ranges along with the relative weights configuration. As it can be observed from Table 7.1, for images locates in quality range of [1,3], the proposed model performs the best with a configuration of $\alpha_{BI} = 0.9$, $\beta_{BI} = 0.1, \gamma_{BI} = 0$, while for higher quality images in range of [3,5], the model performs better with a higher weight for the texture estimator. Overall speaking, it can be concluded that structure plays a vital role in utility assessment, especially for lower quality images. Furthermore, it is apparent that textures play certain roles in evaluating the utility of higher quality images. It has been verified that the roles of structure and texture information are different among different quality ranges in utility evaluation task.

For performance evaluation, best weights are selected for the three basic estimators for images with different quality according to Table 7.1. The overall performances of the metrics are concluded in Table 7.2. Among the compared metrics, the proposed *BF-M* performs the best. We demonstrate that the proposed model is qualified for the task of utility assessment.

To better understand why structure-related information is important for utility assessment, the edge maps and the extracted HOG descriptors are visualized in the second and third row of Figure 7.3. The first column shows the reference image while the other two display the degraded images. The degraded image in the second column has a higher utility score than the third one ($10.528 > -47.638$). By only observing the edges and HOG

Table 7.2 – Performances of various estimators as utility estimator.

	PCC	**SCC**	**RMSE**
ReDLOG [141]	0.7575	0.7757	39.89
MAD [142]	0.7241	0.7303	42.1
MS-SSIM [45]	0.833	0.8510	33.8
VIF [52]	0.943	0.959	12.4
$NICE_{canny}$ [67]	0.935	0.937	13.3
MS-NICE [67]	0.911	0.959	15.4
MS-DGU [68]	**0.960**	**0.961**	**10.3**
BF-M	**0.961**	**0.961**	**10.2**

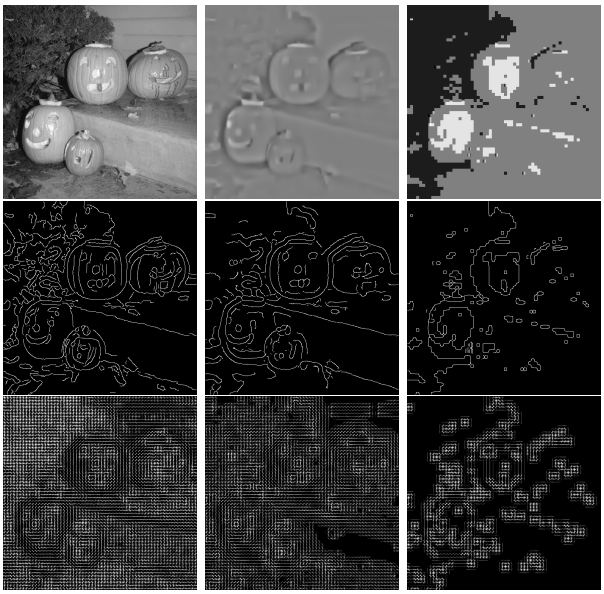


Figure 7.3 – Examples explaining why structure-related information play greater tole in task of utility assessment

maps, one can notice that the shapes of the 'pumpkin lanterns' on the floor in the first and second column are recognizable while the ones in the third column are not. It can be concluded that, for low quality distorted images, where most of the texture information is lost, structure is the most important information for judging its utility.

B. Results: Objective Estimates of Perceived Quality for Synthesized Texture Images

Texture-synthesis is a broadly and commonly used technique for bit-rate saving in images, videos compression, in-painting (e.g., used for error concealment or dis-occluded regions filling for view synthesis in FTV system), etc. The purpose of quality assessment for texture synthesized images is to estimate the perceived quality of the synthesized textures referring to the original textures in images. The role of texture information is more important than the one of structure information in such case.

In verifying what has been discussed above, we test the proposed *BF-M* model on the SynTEX Granularity dataset [39–41] introduced in section 4.2.4. According to [65], CWSSIM [88], WCWSSIM [89], parametric metric proposed in [38] and STQA [65] are the 4 most promising metrics on the SynTEX Granularity dataset for evaluating the quality of synthesized texture images. Therefore, the performance of the proposed model used as an estimator of perceived quality for texture synthesized images is tested on the same database and

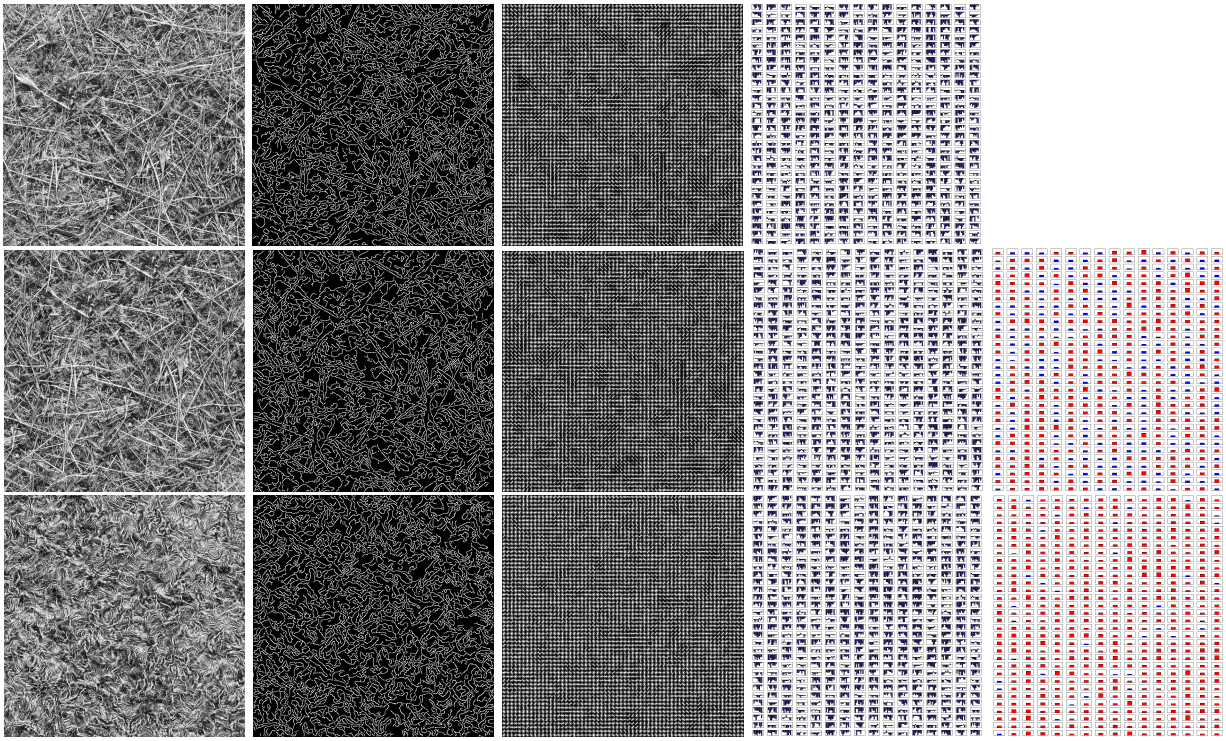


Figure 7.4 – Examples explaining why texture related information play a more significant role in the task of quality assessment for synthesis texture.

compared to these four methods.

Table 7.3 – Performances of various metrics as quality estimator for synthesized texture.

	PCC	**SCC**	**RMSE**
WCWSSIM [89]	0.546	0.497	0.170
CWSSIM [88]	0.663	0.644	0.198
Parametric [38]	0.412	0.481	0.253
STQA [65]	**0.766**	**0.755**	**0.799**
BF-M	0.708	0.719	0.162

During the experiment, by setting $\alpha_{BI} = 0.2, \beta_{BI} = 0.2, \gamma_{BI} = 0.6$, the proposed model gets the best performance. Since the weight for the texture estimator (i.e., BI-LRI) accounts for the greatest proportion, we can conclude that texture is more important than structure in the task of quality assessment for synthesized texture images. In addition, the overall performance of the model applied as a quality estimator for synthesized texture images is reported in Table 7.3. Although $BF\text{-}M$ does not outperform STQA, the performance is still comparable to the others. This result demonstrates the feasibility of using the proposed model as a quality estimator for synthesized texture images.

To further interpret why texture information is most important for the quality assessment of synthesized texture images, we visualize the edge, HOG, LRI maps and the error map between LRI maps of the reference and the synthesized texture images in the second to fourth columns in Figure 7.4. In this figure, the first row corresponds to the reference image while the second and third row correspond to synthesized texture images. The image in the second one has a higher perceived quality score ($4.647 > 1.235$). For better observation, when generating the visualized LRI maps, we select a slightly larger block size 16×16 and crop only the top left part of the image. In the visualized LRI maps, each sub-figure is a LRI histogram representing the texture

information of the local block. LRI is a statistical texture feature that considers inter-edge distance distribution along different angles, i.e., 8 directions by comparing the current pixel value to the closest edge pixel value along each direction. The magnitude of each bin in the histogram is decided by the pixel number between the current pixel and the closest edge pixel along the direction. The sign of the bin is decided by comparing the two pixels' value. Therefore, the more saturated the histogram, the smoother the block. By comparing the edges, HOG maps of the two synthesized texture images with the ones of the original image, it is almost impossible for human observers to tell the difference between them. On the contrary, the LRI map can provide more clues about the statistical difference between the texture in images. By comparing the error maps calculated using euclidean distance between LRI histograms of the original and the synthesized image, it can be observed that the overall error of the synthesized texture image in the third row is larger than the one in the second row, which is consistent with perceived quality score (bins in error map that are larger than a value of 1.2 are labeled with red color). It can be concluded from this sub-section that texture information is more important when the task involves mainly fine-granular texture in images. Since there is no clear main structure (e.g., boundaries of objects) in these texture images, detail of these images (i.e., the texture) is the dominant factor in the task.

C. Results: Objective Estimates of Perceived Quality for DIBR based Synthesized Views

Views that are synthesized with DIBR based techniques contain specific distortions (e.g., object shifting, incorrect rendering, flickering, blurriness, and geometry distortion around disoccluded regions). Since the human visual system is more sensitive to severe local disruptions than the consistent global ones [78], we hypothesize that structure information play a greater role than texture information during the process of assessing quality of synthesized views.

To verify our hypothesis, the proposed model is applied as a quality estimator and tested on the IVC-Image dataset [49,98]. In [143,144], images synthesized with A1 is excluded from the experiment due to the significant shifting artifacts compared to the others. However, according to the MOS, images synthesized with A1 [79] have better quality compared to others. It is more similar to advanced synthesized algorithms. Since the main purpose of developing a full reference image/video quality metric in an FTV system is to evaluate the performance of synthesis algorithms, the tested dataset should be in line with the images/videos synthesized with the more advanced synthesis algorithms to follow the trend. In our experiments, we include the image set generated by A1 and check the performance on the full IVC-Image dataset. As claimed in [76, 143, 144], MP-PSNR and MW-PSNR performed the best among the existing metrics designed for synthesized views. According to Dragana *et al.* [76], PSNR is more consistent with human judgment when calculated at higher morphological decomposition scales. They thus proposed reduced versions of the morphological multi-scale measures: reduced MP-PSNR, and reduced MW-PSNR. The reduced versions outperform the full ones. In this section, we compare our proposed model with MW-PSNR$_f$, MP-PSNR$_f$ MW-PSNR$_r$ and MP-PSNR$_r$. To obtain the best performance from them, a 5×5 size of SE is used for MP-PSNR, and a min Haar wavelet decomposition is used for MW-PSNR as reported in [76].

The overall performances of the metrics are reported in Table 7.4. In our experiment, by setting $\alpha_{BI}, \beta_{BI}, \gamma_{BI}$ to 0.5, 0.2 and 0.3, the performance of the proposed model peaks. This configuration indicates that both structure and texture information play a role in evaluating the quality of synthesized views. However, structural

Table 7.4 – Performance of the proposed metric compared with existing metrics for synthesized views.

	PCC	**SCC**	**RMSE**
MP-PSNR$_f$ [75]	0.6553	0.6239	0.5029
MP-PSNR$_r$ [76]	0.6733	0.66	0.4923
MW-PSNR$_f$ [74]	0.6089	0.5738	0.4948
MW-PSNR$_r$ [76]	0.6444	0.6218	0.5091
BF-M	**0.6980**	0.5885	**0.4768**

information are more important than textural information. In other words, artifacts that degrade the structure of the view are more annoying for the human visual system, which verifies our previous assumption. Moreover, according to Table 7.4, the proposed BF-M achieves 0.6980 value of PCC, which outperforms all of the compared metrics designed for synthesis images. Compared to the second best performing MP-PSNR$_r$ our proposed model obtains a gain of 0.0247 in PCC, which verifies its capability of assessing the perceived quality of synthesized views.

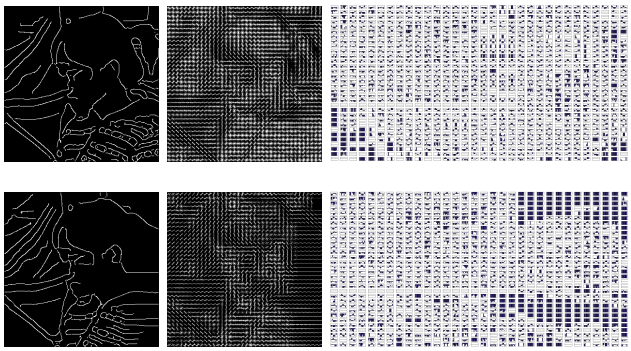


Figure 7.5 – Examples explaining why both structure and texture information play a considerable role in the task of quality assessment synthesized views.

In order to understand how different information loss affects the perceived quality of synthesized views, the edges, HOG and LRI maps of the synthesized image in the third row of Figure 7.1 and the ones of its reference are shown in Figure 7.5. By only comparing the edges and HOG maps, one can easily notice the geometric distortions around the face of the women, especially in the right part where the entire regions are blurred. Therefore, it is evident that structure-related information is more important in this task since the deformation of objects' shapes caused by synthesized algorithms are more annoying for human observers and can be well captured by structure-related descriptors. More interestingly, by comparing the right part of the two LRI maps, one can easily notice the differences between the histograms in blurred regions within the synthesized image compared to the ones in the reference. Due to the blurriness introduced by the DIBR algorithms, texture information has been modified, and start to become annoying. That is why BI-LRI accounts for 20% weights in this task.

Execution Time

In verification of the efficiency of the proposed BF-M, execution time of the metrics normalized by PSNR as introduced in section 4.4 are listed in Table 7.5. According to the table, even though our proposed metric is a bit slower than MW-PSNR and MW-PSNR$_r$.

Table 7.5 – Normalized execution time of proposed metric compare to the state-of-the-art metrics

Metric	MW-PSNR	MW-PSNR$_r$	MP-PSNR	MP-PSNR$_r$	BF-M
Normalized time	12.4	**9.6**	100	35	17

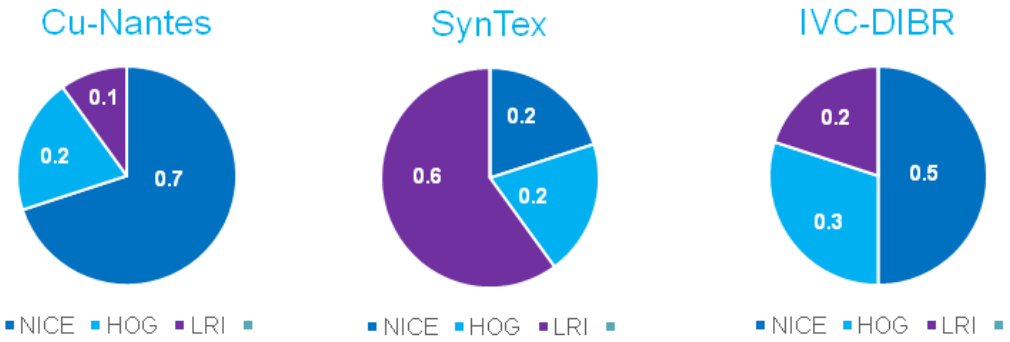


Figure 7.6 – The configurations of α_{BI}, β_{BI} and γ_{BI} which yield the best performances in the corresponding tasks on relative datasets.

7.5 Conclusion

In summary, the optimized configurations of *BI-NICE*, *BI-HOG*, and *BI-LRI* in tasks of utility assessment, quality assessment of synthesized texture images and synthesized views are concluded in Figure 7.6. Weights are selected according to the best performance of the proposed model tested on the Cu-Nantes, SynTEX, and IVC-Image datasets. According to the optimized settings, two main conclusions can be made:

(1) Our hypothesis has been verified : It is obvious that structure information does play greater role than texture information in tasks like utility assessment and quality assessment for synthesized views. Nevertheless, in the context of synthesized texture images quality assessment, texture information is more important.

(2) In utility assessment tasks, interesting result can be found: the role of structure and texture informations varies with quality of images. It can be concluded from the experiments that textures are also useful for images with higher quality.

Human observers tend to perceive global structure first then finer details like texture. We hypothesize that structure and texture information play different roles in visual tasks according to the characteristics of the tasks. To validate this hypothesis, a contours, a coarse-grained structure-related, and a texture estimator are introduced using bilateral filtering. A bilateral filtering based model (BF-M) is then designed to combine the three estimators according to different applications. Experiments are conducted on three different datasets for different tasks. The optimized configuration of the model serves as a proxy for checking the roles of structure and texture information in those tasks. According to the experimental results, our hypothesis has been verified and the performance of the proposed model applied as a tasks-based parametric metric is proven to be comparable to the state of the art utility/quality metrics.

8

Quantifying Structure Deformation with Elastic Metric

8.1 Introduction

In this chapter, the second low-level representation based metric is presented. In order to quantify the deformation of 'curves', an elastic metric based model is proposed for image/video quality assessment. This proposed model is tested FTV scenario for image/video quality assessment.

Images/Videos in applications like FTV contain mainly local non-uniform geometric distortions. As described in section 2.2, observers are more sensitive to such local severe deformations than consistent shifting artifacts which are penalized by most of the point-to-point metrics. Elastic metric is capable of measuring the difference in stretching or bending between two curves and thus is suitable for evaluating such geometric distortions. Examples shown in Figure 8.1. 1) The elastic metric can be used as a distance measure for curves (i.e., measure for quantifying the amount of deformations of contours), whose output is the dissimilarity value between curves ranging from 0 to 1. The larger the value, the more severe the curve is deformed compared to the reference one. Let $D_{EM}(c_1, c_2)$ be the function to calculate the distance between two curves based on the elastic metric. In Figure 8.1, the difference between the extracted curves of the reference patch (red) and the 'twisted nose' (green), i.e., $D_{EM}(L, M)$, equals to 0.1926 while the one $D_{EM}(L, R)$ for the 'slightly shifted nose' equals to 0.1781. This is in line with the fact that the local crooked contours are more annoying. 2) Consistent global shifting artifacts are not over-penalized by elastic metric with curves matching. For example, in Figure 8.1, the second patch in the second row contains annoying geometric artifacts where the shape of the nose is changed significantly while the one on the right is expanded and shifted slightly. It is obvious that the 'twisted nose' in the middle is more annoying than the 'slightly shifted nose' on the right. However, the PSNR value for the patch in the middle with its reference patch on the left $PSNR(L, M)$ is 20.2854 db while the one $PSNR(L, R)$ of the patch on the right is 18.6616 db, incorrectly indicating that the quality of 'twisted nose' is

better. Details of the proposed model are given in the following sections.

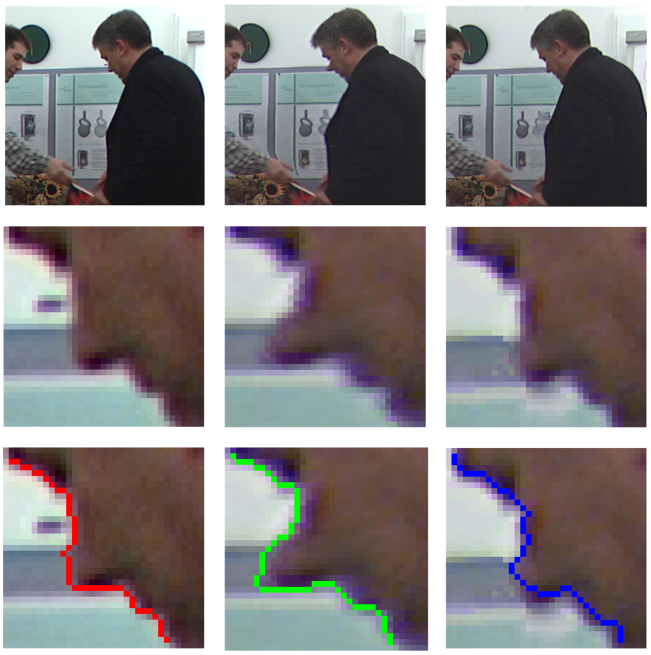


Figure 8.1 – Examples of advantages of elastic metric and disadvantages of commonly used metric PSNR Rows:(from up to down) : Part of the images for better observation; Patches from images; Extracted contours of patches. Columns: (from left to right) reference image, a synthesized image obtained with A2, a synthesized image obtained with A5. PSNR(L, M)=20.2854 db, PSNR(L, R)=18.6616 db, D_{EM}(L,M)=0.1926 , D_{EM}(L,R)= 0.1781.

8.2 Elastic Metric based Image Quality Assessment Metric (EM-IQM)

In this section, the elastic metric based image quality assessment is described by taking the quality assessment of synthesized views in FTV use case as an example. The overall framework of the proposed scheme is shown in Figure 8.2. Firstly, to select local regions where human observers are sensitive to, and to ensure shifting resilience, speeded-up robust features (SURF) [145] descriptors are first extracted and matched from the reference to the synthesized images. After matching the detected interest points from the reference images to the synthesized ones, simple linear iterative clustering (SLIC) [146, 147] is used for contour extraction on the matched patches centering at the matched features points. Before calculating the dissimilarity D_{EM} on the matched contours set (C_{ori}, C_{syn}) with elastic metric, contours inside the matched patches are matched based on the features of the superpixels. Finally, the predicted objective score S_{EM} for one synthesized image is obtained by spatially pooling the elastic scores of all pairs of matched contours.

8.2.1 Local Sensitive Regions Selection based on Interest Points Matching

In FTV scenarios, it is observed that :

1. DIBR based synthesized images contain mainly local non-uniform distortions around disoccluded regions instead of uniform artifacts throughout the entire image due to the DIBR process. These geometric distortions are around important regions (interest of regions), e.g., regions at the center of the image and

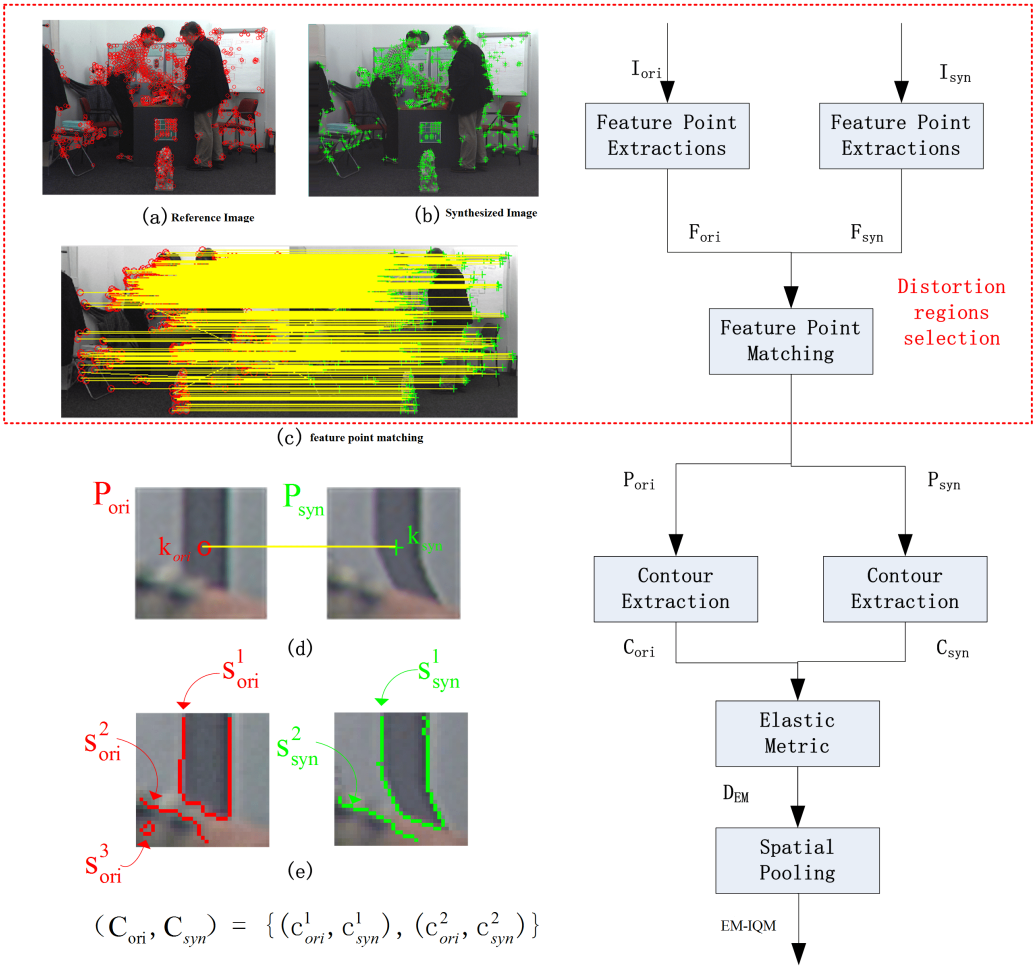


Figure 8.2 – Framework of the proposed elastic metric based image quality assessment model.

regions around the boundaries of objects.

2. While observing an image, artifacts that are located around/within regions of interest are much more annoying than those are located around inconspicuous area [34]. Meanwhile, 'poor' regions are more likely to be perceived by humans in an image with more severity than the 'good' ones. Thus, an image with even a small number of 'poor' regions is penalized more gravely.

3. The typical DIBR artifacts, such as 'Shift of objects', is a big challenge for point to point metrics like PSNR due to the mismatched correspondences.

Based on this observation, it is important to select local regions to locate geometric distortions and avoid over-penalize acceptable global shifting.

Speeded up robust features (SURF) [145], which is a local feature detector and descriptor, is frequently used in tasks like object recognition, image registration, as well as 3D reconstruction. It uses an integer approximation of the determinant of Hessian blob detector to detect interest/feature points. Those detected interest/feature points are normally key points of objects that reveal images' local properties and local shape information of objects. With these key points, same/similar interest points in two images can be then matched by calculating structure-related similarity between them.

Based on the characteristic of SURF, it is a good candidate for selecting the local sensitive regions, where geometric artifacts for observers. After interest points matching, key points of objects are better aligned. Therefore, interest point matching can also compensate the consistent 'Shift of Objects' artifacts, which is to

some extent acceptable for the human visual system.

To confirm the feasibility of using SURF for sensitive regions selection, we check the overlap areas between the matched interest points regions and the local severely distorted regions indicated by error maps. For better understanding, an example is visualized in Figure 8.3. The error maps are generated with the synthesized and the reference images as introduced in [144]. The darker the regions the more distortions appear in the regions as shown 8.3 (c). Interest-point regions are covered by matched patches centering at the matched interest points as the green bounding boxes shown in 8.3 (c). In this study, the size of the patch is empirically set as 35×35. It can be observed that the majority of regions that with severe local distortions are covered by the matched interest-point regions. It demonstrates the feasibility of utilizing SURF for sensitive regions selection.

Figure 8.3 – Example of sensitive regions selection based on interest point detection. Left: Reference image; Middle: Synthesized image with A2; Right: Matched SURF points regions on the error map.

According to the analysis above, SURF descriptors (interest/feature points) are first extracted from both the reference and synthesized images. Interest points in reference image are then matched to the ones in synthesized images to get corresponding patches (i.e., patches that are center at the matched feature points). This regions selection process can be summarized by the red dash bounding box in Figure 8.2. In Figure 8.2, (a) is an original image I_{ori} and (b) is a synthesized image I_{syn}. The extracted surf descriptors (key/interest point) k_{ori} of I_{ori} are labeled with red circle and the ones k_{syn} in I_{syn} are labeled with green cross. Then, SURF interest points are matched k_{ori} to k_{syn}, as shown in Figure 8.2 (c) where matched feature points are connected with yellow lines. Pairs of interest points that have significantly different coordinate (x or y) values are discarded (difference between one of the coordinate values is larger than a threshold). In the following process, closed curves/contours are extracted from patches centering at the matched features points in both the synthesized and reference images. For example, in Figure 8.2 (d), k_{ori}, k_{syn} are one pair of matched SURF points and P_{ori}, P_{syn} are the corresponding patches centering at k_{ori}, k_{syn}. In the following process, only the selected regions are considered.

8.2.2 Curve Extraction based on Patch Segmentation

With the matched selected regions P_{ori} and P_{syn}, image segmentation approach SLIC proposed in [147] is then utilized to further segment the patches into superpixels, whose boundaries are considered as closed curves for later matched curves comparison. As shown in [147], SLIC, which clusters pixels in the combined five-dimensional color and image plane space, outperforms the mainstream superpixel methods in boundary adherence, segmentation speed, and performance. Therefore, it is selected for its simplicity and efficiency.

The proposed algorithm for superpixels matching is shown in Algorithm 1. For each corresponding P_{ori}

Algorithm 1 Superpixels matching algorithm

procedure S MATCHING(S_{ori}, S_{syn})
  for $i \in [1, ..., n]$ **do**
    while $\frac{|s^i_{ori}|}{|P_{ori}|}, \frac{|s^j_{syn}|}{|P^j_{syn}|} > \epsilon_{EM}$ **do**
      $c^i_{ori} = f_{curve}(s^i_{ori})$
      for $j \in [1, ..., m_s]$ **do**
        $f^i_{ori} = \left[\frac{x^i_{ori}}{P_{width}}, \frac{y^i_{ori}}{P_{heigth}}, \frac{|s^i_{ori}|}{|P_{ori}|}\right]$
        $f^j_{syn} = \left[\frac{x^j_{syn}}{P_{width}}, \frac{y^j_{syn}}{P_{heigth}}, \frac{|s^j_{syn}|}{|P_{syn}|}\right]$
        $M_{distance}(i, j) = D_e(f^i_{ori}, f^j_{syn})$
      end for

$$c^j_{syn} = f_{curve}\left(\underset{s^j_{syn} \in P_{syn}}{\arg\min}\ M_{distance}(i, j)\right)$$

    end while
  end for
  return(C_{ori}, C_{syn})
end procedure

and P_{syn}, two sets of superpixels S_{ori} and S_{syn} are obtained with SLIC. Here, the goal is to extract fragments of objects' boundaries. Superpixels that are too small may not be useful (may not be a part of and object boundary). For instance, in Figure 8.2 (e), s^3_{ori} is discarded as it is too small, and it does not represent meaningful shape. To do so, before curves matching, superpixels that are too small are discarded basing on a ratio. This ratio is defined as the number of pixels in the superpixels to the one in the patch $\frac{|s|}{|P|}$, where $|R|$ represents the number of pixels in a region R. If the ratio of a superpixes is smaller than a threshold ϵ_{EM}, it would be discarded. Let n_s and m_s be the number of superpixels in P_{ori} and P_{syn}. The boundary of each superpixel $s^i_{ori}, i \in 1, ..., n_s$ is considered as the closed curve c^i_{ori}, which needs to be matched from S_{syn}. In Algorithm 1, $f_{curve}(s)$ is the function of getting the boundary of the superpixels s. For each candidate $s^j_{syn}, j \in 1, ..., m_s$, features that reflect the location and the size of superpixels are concatenated as a feature vector for comparing the similarity among superpixels. Here, $\overline{x}$, $\overline{y}$ are the mean of the row and column values of the pixels inside the superpixels. They are normalized by the width P_{width} and the height of the patch P_{height} separately. The last dimension of the feature vector is the number of pixels inside the superpixels normalized by the size of the patch and denoted as $\frac{|s|}{|P|}$. For each possible pair s^i_{ori} and s^j_{syn}, they are represented as feature vectors $f^i_{ori} = [\frac{\overline{x^i_{ori}}}{P_{width}}, \frac{\overline{y^i_{ori}}}{P_{heigth}}, \frac{|s^i_{ori}|}{|P_{ori}|}]$ and $f^j_{syn} = [\frac{\overline{x^j_{syn}}}{P_{width}}, \frac{\overline{y^j_{syn}}}{P_{heigth}}, \frac{|s^j_{syn}|}{|P_{syn}|}]$ correspondingly. Dissimilarity between superpixels is computed using euclidean distances $D_e(f^i_{ori}, f^j_{syn})$, and stored in the matrix $M_{distance}$. The matched superpixels of s^i_{ori}, i.e., $s^j_{ori} = \arg\min_{s^j_{syn} \in P_{syn}} M_{distance}(i, j)$ is the one that minimize $M_{distance}$. Then, the matched closed curved c^j_{syn} is the boundary of the matched superpixels s^j_{ori}. For example, in Figure 8.2 (e), s^1_{ori} is matched with s^1_{syn} when s^2_{ori} is matched withs^2_{syn}. Finally, the set of matched closed curves (C_{ori}, C_{syn}) is obtained and is used for curves comparison with elastic metric.

8.2.3 Curve Comparison based on Elastic Metric in Euclidean Spaces

With the matched closed curves generated with the aforementioned sensitive regions selection and curves extraction stages, the framework proposed in [148, 149] is then used for measuring the amount of 'stretching' and 'bending' between two matched curves from the original and synthesized images.

Given a parametrized curve c along with its curve parameter $t \in D$, it is first defined as

$$c : D \to (x, y) \in \mathbb{R}^n, \tag{8.1}$$

where (x,y) represents the coordinates of each point in the curve. For general case, $D = [0, 1]$, but for closed curves $D = \mathbb{S}^1$. Then, the parameterized curve can be further represented with square-root velocity (SRV) function defined by $q : D \to (x, y) \in \mathbb{R}^n$, where

$$q(t) \equiv F(\dot{c}(t)) = \dot{c}(t)/\sqrt{\|\dot{c}(t)\|} \tag{8.2}$$

In (8.2), $\|\cdot\|$ represent the euclidean 2-norm in $\mathbb{R}^n$ and $\dot{c(t)} = \frac{dc}{dt}$. It is reversible that one can obtain the curve with the equation: $c(t) = \int_0^t q(s)\|q(s)\|ds$.

To completely specify curve c as well as to quantify deformations of the curves, Srivastava then defined $\phi : D \to \mathbb{R}$ by $\phi(t) = ln(\|\dot{c}(t)\|)$ and $\theta : D \to \mathbb{S}^{n-1}$ by $\theta(t) = \dot{c}(t)/\|\dot{c}(t)\|$ in [149]. Therefore, a riemannian metric named as 'Elastic Metric' on a tangent space of $\Phi \times \Theta$ is then defined based on calculating an inner product:

$$\begin{aligned} D_{EM} =& \langle (u_1, v_1), (u_2, v_2) \rangle_{(\phi,\theta)} \\ &= a^2 \int_D u_1(t)u_2(t)e^{\phi(t)}dt + b^2 \int_D v_1(t)v_2(t)e^{\phi(t)}dt, \end{aligned} \tag{8.3}$$

where $\langle\cdot\rangle$ denotes the standard dot product in $\mathbb{R}^n$ and (u_1, v_1), $(u_2, v_2) \in T_{\phi,\theta}(\Phi \times \Theta)$. As explained in [148,149], u_1 and u_2 in the first integral are variations of the log speed ϕ of the curves while v_1 and v_2 in the second integral are the variations of the direction θ of the curves. The first and second integrals can be interpreted to measure the amount of 'stretching' and 'bending' correspondingly and a^2,b^2 are the weights chosen to penalize these two types of deformations. In order to compute geodesics with equation (8.3) in the pre-shape and shape spaces more efficiently, the SRV formulation (8.2) was used and adjusted in terms of (ϕ, θ) by $q(t) = e^{\frac{1}{2}\phi(t)}\theta(t)$. Afterwards, the tangent vectors to $\mathbb{L}^2(D, \mathbb{R}^n)$ at q is obtained with $f = \frac{1}{2}e^{\frac{1}{2}\phi}u\theta + e^{\frac{1}{2}\phi}v$. For two elements f_1 and f_2 of $T_{\phi,\theta}(\Phi \times \Theta)$, computing the $\mathbb{L}^2$ -metric (elastic metric) of them yields

$$\begin{aligned} D_{EM} =& \langle f_1, f_2 \rangle = \int_D \langle \frac{1}{2}e^{\frac{1}{2}\phi}u_1\theta + e^{\frac{1}{2}\phi}v_1, \frac{1}{2}e^{\frac{1}{2}\phi}u_1\theta + e^{\frac{1}{2}\phi}v_2 \rangle dt \\ &= \int_D (\frac{1}{4}e^{\theta}u_1u_2 + e^{\theta}\langle v_1, v_2 \rangle)dt \end{aligned} \tag{8.4}$$

8.2.4 Pooling Stage

As discussed in previous sections, human observers tend to perceive 'poor' regions than the 'good' ones within an image. For DIBR based synthesized images, the disoccluded regions are the 'poor' regions and should be penalized during the quality assessment.

After sensitive regions selection, the curves are only extracted from the selected regions where the annoying local distortions could be unacceptable. Moreover, due to local sensitive regions selection, artifacts in local important disoccluded regions are penalized sufficiently, and at the same time, the consistent global artifacts

are not over penalized. Hence, the final object score is calculated by simply summing out all the elastic dissimilarities values without applying any specific pooling strategies as defined below

$$EM - IQM = \sum D_{EM}(c^i_{ori}, c^j_{syn}), \tag{8.5}$$

where $(c^i_{ori}, c^j_{syn}) \subset (C_{ori}, C_{syn})$.

8.2.5 Experimental Results

The performance of elastic curve based synthesized image quality assessment metric (EM-IQM) is evaluated on the entire IVC-Image dataset [49, 98] as described in section 4.2.1.1. To compare the performances between existing metrics designed for synthesized images, the widely employed criteria PCC, SCC, and RMSE as described in section 4.3 are considered with non-linear mapping between the subjective scores and objective measures.

8.2.5.1 Performance Comparison

As claimed in [76, 143, 144], MP-PSNR MW-PSNR and their reduced versions perform the best among existing metrics designed for synthesized images. Therefore, in this section we mainly compare our proposed metric with MW-PSNR$_f$, MP-PSNR$_f$ MW-PSNR$_r$ and MP-PSNR$_r$.

The overall result is concluded in Table 8.1. According to Table 8.1, the proposed EM-IQM achieves 0.7430, 0.6626 and 0.4455 value of PCC, SCC and RMSE correspondingly, which outperforms all of the compared metrics designed for synthesis images. Compared to the second best performing MP-PSNR$_r$, our proposed metric achieves a gain of 25% in PCC. To further check whether the proposed EM-IQM significantly outperforms the second best performing MP-PSNR$_r$, a t-test is conducted taking the difference between the subjective scores and the objective scores predicted using MP-PSNR$_r$ and the ones using EM-IQM as input. According to the t-test result, EM-IQM significantly outperforms MP-PSNR$_r$ (P-value=10_{-5}).

Table 8.1 – Performance comparison of the proposed metric with state-of-the-art metrics

	PCC	**SCC**	**RMSE**
MP-PSNR$_f$ [75]	0.6553	0.6239	0.5029
MP-PSNR$_r$ [76]	0.6733	0.6600	0.4923
MW-PSNR$_f$ [74]	0.6089	0.5738	0.4948
MW-PSNR$_r$ [76]	0.6444	0.6218	0.5091
EM-IQM	**0.7430**	**0.6626**	**0.4455**

To demonstrates the advantage of the proposed metric with PSNR based metric like MP-PSNR, the scatter plots of subjective DMOS values versus MP-PSNR and the proposed EM-IQM (for better observation, the figures are zoomed) are illustrated in Figure 8.4 and Figure 8.5 respectively. In the figures, each stimulus that generated with different algorithms is labeled with different colors and shapes. It is obvious that the performance of MP-PSNR on the set of images obtained with A1 is poor since most of the blue cross points (A1) are outliers. It can be concluded that PSNR based metrics like MP-PSNR are not robust to global acceptable 'objects shifting' artifacts. On the contrary, according to Figure 8.5, the performance of EM-IQM on the same subset is much better as most of the blue cross points are gathered along the diagonal line. The reason is that the proposed metric penalizes only the local annoying artifacts and compensates the shifting artifact with the sensitive regions selection process.

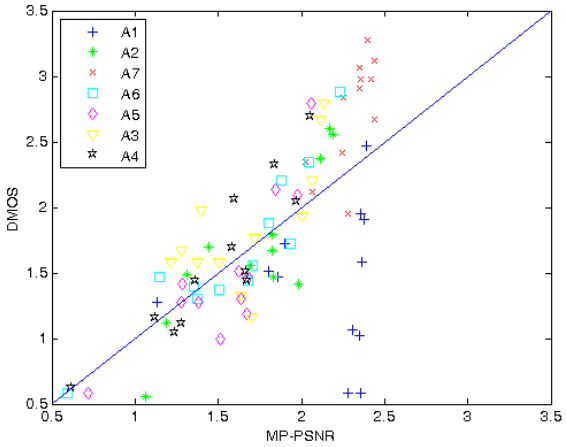


Figure 8.4 – Scatter plots of MOS versus MP-PSNR, the blue diagonal line represents the perfect prediction

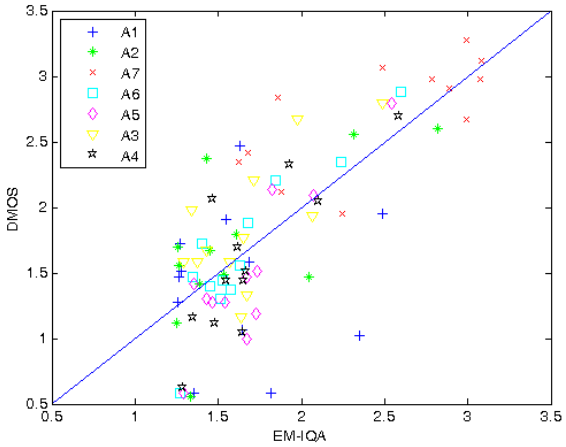


Figure 8.5 – Scatter plots of MOS versus EM-IQM, the blue diagonal line represents the perfect prediction.

To check the efficiency of the proposed EM-IQM, execution time of the metrics normalized by PSNR as introduced in section 4.4 are listed in Table 8.2. According to the table, the proposed EM-IQM is slower than MW-PSNR and MW-PSNR$_r$. However, since the gain compared to the second best performing metric is significant, it is acceptable if the running time is a bit longer.

Table 8.2 – Normalized execution time of proposed metric compare to the state-of-the-art metrics

Metric	MW-PSNR	MW-PSNR$_r$	MP-PSNR	MP-PSNR$_r$	EM-IQM
Normalized time	12.4	**9.6**	100	35	127

8.3 Elastic Metric based Video Quality Assessment Metric (EM-VQM)

As described in the section 2.2, temporal structure-related distortions within one viewpoint and among different viewpoints (observed due to views switch) are difficult for conventional video quality assessment metrics to capture. There is an obvious lack of such video quality metrics in predicting the perceived quality of sequences in FTV system. Targeting at solving this problem, here, the EM-IQM is extended to VQM to

quantify the structure-related temporal distortions in FTV system, including spatial geometric distortions, temporal structure inconsistency, and unsmooth viewpoints changing.

The framework of EM-VQM is summarized in Figure 8.6. As motion trajectory reveals important structural-motion information, local structure disruptions that affect the quality of the synthesized sequences could be quantified based on the multi-scale trajectory representations. In the proposed scheme, synthesized sequences Seq_{syn} and their reference sequences Seq_{ref} are firstly represented as a set of multi-scale trajectories $Tra_{syn}^{s_{tra}}$ and $Tra_{ref}^{s_{tra}}$, where s_{tra} indicates different scales. Considering the special characteristics of DIBR based synthesized techniques, neighborhoods around the trajectories could be considered as the sensitive regions(candidates of possible disturbing distorted regions), where local non-uniform distortions are less endurable for observers. With the trajectory representations, spatial-temporal related features, i.e., $F_{syn}^{s_{tra}}$ and $F_{ref}^{s_{tra}}$, along the trajectories are extracted. Afterwards, 1) the structure deformations of the moving objects, which can be represented by the deformation of motion trajectory D_{def}^{tem}, could be quantified using elastic metric with $Tra_{syn}^{s_{tra}}$ and $Tra_{ref}^{s_{tra}}$; 2) the temporal structural losses along trajectories D_{SL}^{tem} could be quantified with the temporal structural features in the form of histograms $H_{syn}^{s_{tra}}$ and $H_{ref}^{s_{tra}}$. Details of the computation of temporal structural dissimilarity between a synthesized sequence and its reference is given in the following sections. Finally, SRV is used to obtain one final quality score for the synthesized sequence by combining these two type of temporal dissimilarity values between the synthesized and reference sequences at all the scales.

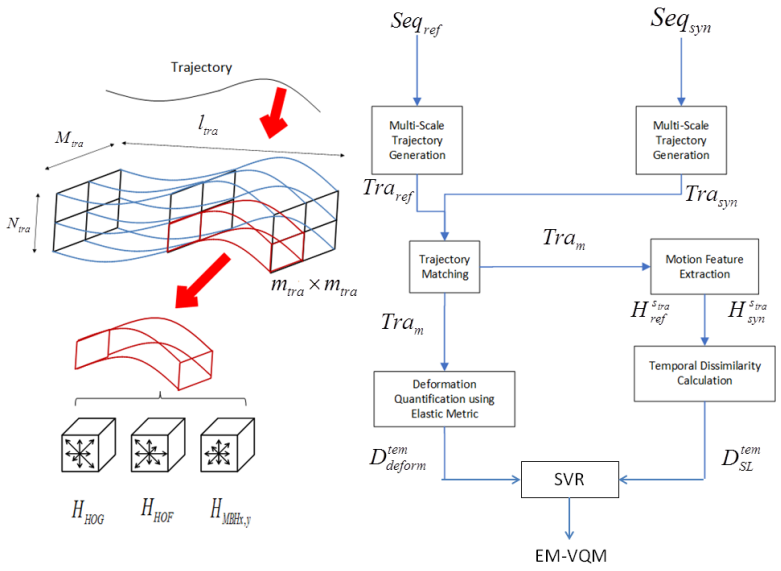


Figure 8.6 – Framework of Temporal Structural loss computation

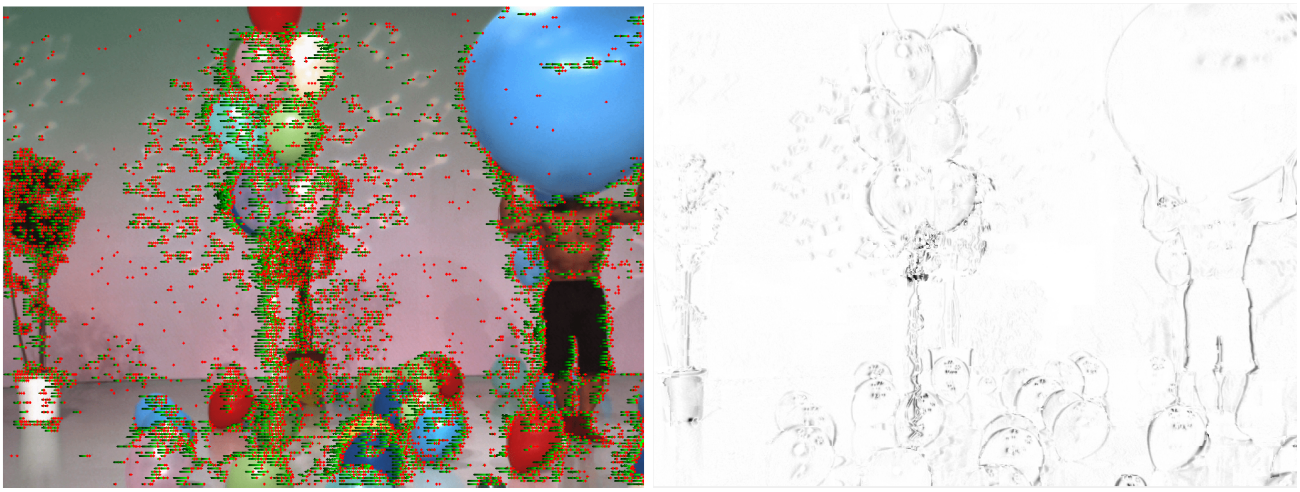
8.3.0.1 Multi-scale Motion Trajectory Representation as Spatial-Temporal Sensitive Regions Selection

Dense motions trajectory, which is first proposed in [150] by Wang *et al.*, is utilized to represent a free viewpoint sequence. It is a spatial-temporal representation for video in the form of multi-scale dense trajectories and motion boundary descriptors along the trajectories.

After generating the multi-scale representation of a sequence Seq with several spatial scales s_{tra}, feature points are sampled on each spatial scale with a sampling step of W (in this study, totally 7 scales are considered). Considering that most of the local annoying geometric distortions are located around the boundaries of the

objects instead of homogeneous texture regions, points within regions that do not contain any structure are thus removed. Sampled points on each spatial scale are then tracked by using large displacement optical flow algorithm (LDOF) proposed in [151]. Each trajectory $tra_{s_{tra}}$ obtained at a certain scale s_{tra} can be represented as a sequence of points $(p_1, ..., p_f, ..., p_{l_{seq}})$ with a length of l_{seq} (equals to the frame number of the sequence). In $tra_{s_{tra}}$, p_f is a feature point at frame f, which is spatially-temporally related to feature points in previous and later frames, i.e., p_{f-1} and p_{f+1}. As human observers are more sensitive to structure-related distortions in moving structural regions, e.g., moving objects, static trajectories that do not contain any motion are thus pruned.

It is worth mentioning that the process of generating trajectories could be served as a proxy to select sensitive regions where local synthesized distortions are less endurable. An example is shown in Figure 8.7. In Figure 8.7 (a), the points of optical flow on current frame are marked with red color. The green lines connect the corresponding points between the previous frame and the ones in the current frame. Figure 8.7 (b) is the error map generated with the frames extracted from a synthesized sequence (i.e. sequence 'C1-balloons-R47-view-3') and its reference sequence (i.e. sequence 'Original-balloons-view-3') from FFV dataset (introduced in section 4.2.1.3). In the error map, the darker the color, the more errors are in the regions. It can be observed that most of the error regions have been covered by the detected motion trajectories (most error regions are covered by the neighborhoods of detected motion trajectories). It is thus feasible to employ multi-scale motion trajectory generation as a proxy for spatial-temporal sensitive regions selection.



(a) (b)

Figure 8.7 – (a) Example of dense motion trajectory. (b) Error map between frames extracted from reference and the synthesized views.

8.3.0.2 Motion Structure-Related Trajectory Descriptor

As mentioned in the previous sections, the dominant non-uniform distortions are mainly located around the boundaries of objects. Boundaries of objects provide shape information of the objects and thus are important structure information. In order to better quantify the changes of structure as well as motion information along trajectories due to synthesizing process, three motion-structure-related descriptors are extracted from each trajectory at each scale [151]. They are the histogram of oriented gradient (HOG) [152], the histogram of optical flow (HOF) [153] and the motion boundary histogram (MBH) [154], which are all extracted within a spatial-

temporal volume that is aligned with a trajectory $Tra_{s_{tra}}$. The size of the temporal volume is $M_{tra} \times N_{tra} \times l_{tra}$ as shown in the left part of Figure 8.6, where l_{tra} is the length of the trajectory and $M_{tra} \times N_{tra}$ is the spatial block size. During feature extractions, each spatial block is further divided into $m_{tra} \times m_{tra}$ sub-blocks for histogram based feature extraction. Among those structure-related features, MBH is computed with the derivatives of both the horizontal and vertical element of optical flow, which further ends up into two histograms as MBH_h and MBH_v respectively. Both MBH_h and MBH_v are normalized with L_2 norm. In conclusion, for each trajectory at scale s_{tra}, four structural histograms $H^{s_{tra}}_{HOG}$, $H^{s_{tra}}_{HOF}$, $H^{s_{tra}}_{MBHx}$ and $H^{s_{tra}}_{MBHy}$ are obtained after feature extraction.

8.3.0.3 Temporal Structure Dissimilarity

After getting the trajectory representations along with the extracted structural features, trajectories at each scale in synthesized and reference sequences are matched according to their averaged horizontal and vertical coordinates. Only the matched trajectory pairs $(tra^{s_{tra}}_{ori}, tra^{s_{tra}}_{syn})$ in the matched trajectory set Tra_m are maintained for later deformation quantification and structure loss computation. To quantify the two typical temporal distortions (i.e., 1) deformation of motion trajectories, and 2) unsmooth transition of structures) mentioned in section 2.2, two main aspects are taken into consideration:

1. First, the temporal evolution of spatial local structure-related distortions may result in deformation of motion trajectories within the sequences, e.g., the motion trajectories distributed along boundaries of foreground objects may fluctuate and result in changes of the shapes of the trajectories. This changes of trajectories in term of global motion trajectory deformation could be quantified by using elastic metric D_{EM} described in section 8.2.3. More specifically, the entire deformable change of trajectories between the synthesized and the reference sequences at a scale s_{tra} is calculated by accumulating all the elastic errors calculated with the matched trajectories at this scale using elastic metric:

$$D^{tem,s_{tra}}_{def}(Tra_m) = \sum_{n_{tra}} D_{EM}(tra^{s_{tra}}_{ori}, tra^{s_{tra}}_{syn}), \tag{8.6}$$

   where n_{tra} is the number of matched trajectory pairs $(tra^{s_{tra}}_{ori}, tra^{s_{tra}}_{syn}) \in Tra_m$. Since $D^{tem,s_{tra}}_{def}(\cdot)$ computes the amount of deformations between trajectories, ideally, it is able to capture not only the temporal flickering within one viewpoint but also the one among viewpoints due to views switch (reflecting the smoothness of the transition among frames at one viewpoint position as well as the smoothness of the transition among viewpoints). By doing so, local structure-related severe temporal distortions are well captured, while the global uniform distortions are not over-penalized.

2. To further quantify the unsmooth transition of structures from one frame to another, structural statistical dissimilarities along trajectories are computed with the four extracted motion structure-related descriptors described in section 8.3.0.2. More specifically, each type of temporal structural statistical loss $D^{ij,s_{tra}}_{SL}$ at a scale of s_{tra} is defined as the distance between one type of extracted features vectors H^i of the synthesized sequence and the one of its reference using a certain distance measure D^j:

$$D^{ij,s_{tra}}_{SL} = D_j(H^{i,s_{tra}}_{ref}, H^{i,s_{tra}}_{syn}), \tag{8.7}$$

   where $H^{1,s_{tra}} = H^{s_{tra}}_{HOG}$, $H^{2,s_{tra}} = H^{s_{tra}}_{HOF}$, $H^{3,s_{tra}} = H^{s_{tra}}_{MBHx}$ and $H^{4,s_{tra}} = H^{s_{tra}}_{MBHy}$ indicates the four

motion descriptors and D_j denotes one type of distance measures. In this study, four distance measures are considered including D_1 = jensen-shannon divergence (JSD), D_2 = euclidean distance, D_3 = cosine distance and D_4 = minkowski summation.

8.3.1 Spatial-Temporal Scores Aggregation

Finally, in order to predict the final quality score, SVR is employed to aggregate the calculated temporal $D_{def}^{tem,s_{tra}}$ and the 16 temporal structural error $D_{SL}^{ij,s_{tra}}, i,j = 1,...,4$ at all scales with a linear kernel. Intuitively, SVR serves as a distance measure and trajectory scale selector for predicting perceived quality. In this study, 7 scales are considered, the dimension of the final vector of each sequence is 119 (16 dimensions for $D_{SL}^{tem,s_{tra}}$ and 1 dimension for $D_{def}^{tem,s_{tra}}$ at each scale). The SVR model training process is conducted according to [155–157] by employing a 1000-fold cross-validation. Each dataset is randomly divided into 80% for training and 20% for testing, without overlap between them. To evaluate the median value of the performance estimation benchmark (e.g., PCC) is reported across 1000 runs for performance evaluation.

8.3.2 Experimental Results

The performance of EM-VQM is evaluated on the IVC-Video and FFV dataset described in section 4.2.1.2 and 4.2.1.3. Apart from synthesis related spatial distortions, IVC-Video database contains temporal distortions within one viewpoint while FFV dataset contains temporal distortions among different viewpoints (artifact observed during the switch of viewpoints). To evaluate the performances, PCC, SCC, RMSE and AOC_{DS}, AOC_{BW}, CC introduced in section 4.3, are utilized. In section 3.2, it has been pointed out that commonly used metrics designed for capturing compression artifacts fail to correctly predict the perceived quality of synthesized sequences in FTV scenario. Therefore, in this section, only image/video metrics (described in section 3.4 and 3.5) designed for DIBR based synthesized views quality evaluation are considered. For those image metrics, predicted quality score for each frame is averaged to obtain one score for the entire sequence.

Table 8.3 – Performance Comparison of the Proposed EM-VQM with Existing Metrics Designed for FTV Scenario

Database	IVC-Video						FFV					
Metric	**PCC**	**SCC**	**RMSE**	**AOC_{DS}**	**AOC_{BW}**	**CC**	**PCC**	**SCC**	**RMSE**	**AUC_{DS}**	**AUC_{BW}**	**CC**
Image Quality Metrics Designed for Synthesized Views												
MW-PSNR$_f$	0.448	0.425	0.612	0.523	0.692	0.665	0.429	0.291	0.662	0.497	0.647	0.615
MW-PSNR$_r$	0.450	0.439	0.590	**0.537**	0.704	0.671	0.430	0.296	0.610	0.508	0.653	0.621
MP-PSNR$_f$	0.523	0.542	0.564	0.531	0.754	0.723	0.440	0.318	0.609	0.508	0.660	0.623
MP-PSNR$_r$	0.461	0.496	0.587	0.521	0.739	0.704	0.410	0.287	0.617	0.5102	0.659	0.625
EM-IQM	**0.666**	**0.647**	**0.493**	0.493	**0.830**	**0.739**	**0.522**	**0.556**	**0.575**	**0.551**	**0.781**	**0.722**
Video Quality Metrics Designed for Synthesized Views												
Liu-VQA	0.617	0.609	0.521	0.507	0.704	0.692	0.574	0.629	0.552	0.559	0.799	0.760
EM-VQM	**0.848**	**0.806**	**0.248**	**0.796**	**0.883**	**0.815**	**0.802**	**0.782**	**0.289**	**0.745**	**0.799**	**0.763**

The overall performance of the metrics is summarized in Table 8.3. According to the table, among the image quality metrics designed for synthesized images, EM-IQM performs the best. Among the video quality metrics designed for synthesized videos, the proposed EM-VQM outperforms the other. Compared to the second best performing video quality metric Liu-VQM, EM-VQM obtains a gain of 37% and 39% in terms of PCC values on IVC-Video and FFV database, respectively. It is proven that the proposed EM-VQM can capture not only temporal artifacts within one viewpoint but also the ones among viewpoints.

Moreover, to check the time complexity of the proposed EM-VQM compared to the second best performing video quality metric Liu-VQM, execution time of the metrics normalized by PSNR as introduced in section 4.4 are listed in Table 8.4. It can be observed from the table that the proposed EM-VQM is slower than Liu-VQM. However, since the gain of the EM-VQM compared to the second best performing metric is significant, it is acceptable if the running time is a bit longer.

Table 8.4 – Normalized execution time of proposed metric compare to the state-of-the-art metric

Metric	Liu-VQM	EM-VQM
Normalized time	**20K+**	48K+

8.4 Conclusion

EM-IQM: Basing on the fact that DIBR based synthesis algorithms mainly introduce local geometric distortions and humans are more sensitive to severe local artifacts, an elastic metric based image quality assessment metric is first proposed in this chapter. In the proposed scheme, a SURF based sensitive regions selection process is incorporated to penalize only annoying local artifacts but to compensate shifting artifacts. The core concept of the proposed metric is to use the elastic metric to quantify the deformation dissimilarities between curves from reference and synthesized images. Among the compared metrics, the proposed EM-IQM metric provides the best performance.

EM-VQM: Targeting at quantifying both the structure-related temporal artifacts within one viewpoint and among viewpoints, the EM-IQM metric is extended for video quality assessment by first representing videos with multi-scale dense trajectories and then quantifying spatial-temporal artifacts based on 1) the deformation dissimilarity between trajectories in reference and synthesized sequences calculated using elastic metric; 2) spatial-temporal structure dissimilarity calculated based on motion descriptors extracted along the trajectories. Experimental results have proven its capability of quantifying not only the unique temporal artifacts within one viewpoint but also the ones among viewpoints.

9

Conclusion of Part 2

In this part, low-level representations have been explored for images utility/quality assessment. Two low-level based models have been proposed.

9.1 Answers to Research Questions

- Low-level representations of images/videos for quality/utility assessment in different tasks:
  - Verification of the roles of low-level structural and textural information in different tasks.

    In order to verify what are the roles of structure and texture information in different tasks, a bilateral filtering based model (BF-M) has been proposed in chapter 7 to first separate structure and texture information. Afterwards, the roles of these information according to given visual content usage have been illustrated.
  - A bilateral filtering based metric is proposed to leverage structure and texture related distortions by using low-level features.

    A bilateral filtering based model has been proposed in chapter 7. Since the respective importance of structure and texture estimators can be easily leveraged, a parametric metric has been defined to balance the roles of the two information according to the visual content usage.
  - An elastic metric based image quality metric (EM-IQM) has been proposed to quantify the structural degradation in terms of curves/contours deformations.

    The elastic metric based image quality metric has been proposed in section 8.2 based on low-level curves representation. In this model, the elastic metric is used to quantify the amount of stretched/bent between curves in distorted images to the ones in reference. To avoid over-penalizing acceptable continuous distortion, a sensitive region selection approach has been proposed along with a fast curves matching algorithm. According to the experimental result, EM-IQM is robust to global shifting and is able to quantify local structure deformation.
  - An elastic metric based video quality assessment metric (EM-VQM) has been proposed in based of

elastic metric and multi-scale motion trajectory to quantify the temporal structure related artifacts in FTV scenario.

As structure related distortions introduce not only new types of structure inconsistencies in terms of non smooth transition of structures along time intra view but also non smooth transition among view, the EM-IQM has been extended to quantify both types of temporal distortions in section 8.3. Temporal artifacts are quantified with the amount of deformation between matched multi-scale trajectories using elastic metric and the amount of structure disruption using structure related descriptors along trajectories. Experimental results have highlighted the efficiency of EM-VQM compared to existing video quality models designed for synthesized views.

9.2 Overall Performance on Tested Datasets

The performance and executing time of the proposed models on all the tested datasets are respectively summarized in Table 9.1 and 9.2.

Table 9.1 – Summary of performance and discussion

PCC		Low-level		
Related Task	Related Database	BF-M	EM-IQM	EM-VQM
Utility Assessment	CU-Nantes	0.961		
IQM of synthesis texture image	SynTex	0.708		
IQA & VQA in FTV	IVC-Image	0.698	0.743	
	IVC-Video			0.847
	FFV			0.801

BF-M has been tested on CU-Nantes for utility assessment, on SynTex for quality assessment of synthesized texture images and on IVC-Image for quality assessment of synthesized views. The performance of this metric on the three datasets are comparable to other metrics designed for the corresponding task. However, the gains are not significant. Since the main purpose of BF-M is to provide users with a parametric tool that could be used to decide which information is more important than another, it is not further extended for video quality assessment. In addition, the performance of BF-M on CU-Nantes has already reached a PCC value of 0.961. Even though there is still a small room to improve the performance, the database is too limited to be used for further exploration (it is designed for general utility task and thus not practical enough to be used as a benchmark for new applications; for example, to select useful training samples for machine learning models). Hence, other metrics developed in this study are not tested on this database.

Table 9.2 – Summarization of executing time of the low-level representation based models on different datasets

Normalized time		Low-level		
Related Task	Related Database	BF-M	EM-IQM	EM-VQM
Utility Assessment	CU-Nantes	17		
IQA of synthesis texture image	SynTex	17		
IQA & VQA in FTV	IVC-Image	17	127	
	IVC-Video			48k+
	FFV			48k+

Moreover, it has been verified in chapter 7 that textures play the dominant role in the task of quality assessment of synthesized texture images. As the purpose of this thesis is to deal with images/videos where

structure related distortions are the dominant distortions in immersive multimedia use cases, other metrics other than BF-M have not been tested on SynTex database.

Both BF-M, EM-IQM have been tested on IVC-Image database. According to Table 9.1 and t-test analysis, EM-IQM outperforms BF-M significantly. There are mainly three reasons : 1) EM-IQM incorporates a sensitive regions selection procedure to ensure that endurable 'global artifacts' are not over penalized; 2) even though both of them deal with low-level features directly, EM-IQM makes use of elastic metric to compute the amount of deformation between curves, which is of better capacity to quantify the amount of geometric distortions in terms of contours' deformation. It is straightforward designed according to the characteristics of the task and thus may not be used for a task like the quality assessment of synthesized texture images. In another word, EM-IQM is more 'problem-focus'; 3) EM-IQM is locally-focused, which means that it penalizes more severe local distortions with the sensitive regions selections.

Considering its performance, EM-IQM has been further extended as EM-VQM. EM-VQM has been tested on IVC-Video and FFV databases for quality assessment of synthesized videos by making use of low-level temporal structure related representations. EM-VQM is able to well predict perceived quality of free viewpoint videos mainly because 1) dense motion trajectory is used as a proxy to select temporal sensitive regions to avoid over-penalize 'global shifting'; 2) unsmooth transitions among frames at one viewpoint position and unsmooth transitions among viewpoints could be quantified by the changes of multi-scale trajectories; 3) structure disruptions along trajectories can be quantified with the structure dissimilarity values calculated with the structure and motion related descriptors along the trajectories. It is confirmed that multi-scale trajectory and low-level structure-motion descriptors along trajectories are suitable low-level representations for video, being able to capture temporal structure related distortions.

Last but not least, according to Table 9.2, even though there is still a big room for the two mid-level based models to improve (in term of performance), their complexities are considerably low since no learning or optimized processes are involved.

9.3 Summary

One of the big advantage of low-level based models is their simplicity in terms of executing time. Low-level models that are more 'problem-focus', e.g., using problem-oriented distance measures and representations, obtain better performance. The representative ability of low-level representations is weak (are not linked to quality directly), as performance still depends on certain 'distance measure' to quantify perceived quality. Last but not least, these low-level representations do not represent enough information of the quality of the images/videos and thus are difficult to be used to develop powerful no reference metrics.

Exploring Mid-Level Representation based Models for Image/Video Quality Assessment

10

Introduction of Part 3

Mid-level representations of images/videos are defined as intermediate 'pattern-based encoded feature', where the patterns are learned by summarizing regularity/characteristics/properties of local low-level information. These representations are between low-level and high-level representations, obtained by simplifying/encoding low-level information. Inspired by the encoding strategy in HVS, two mid-level representation based models are proposed by defining the 'category' and 'entropy' as patterns.

10.1 Mid-level Encoding Strategy in HVS

As mentioned in [158], human visual system is very efficient in encoding the properties of stimulus by utilizing available regularities, e.g., shape of objects, from the inputs. Here, inputs are mainly low-level representations of the perceived contents, e.g., contours. Human brain is subject to processing great amounts of information, and efficiency in information encoding is hence often postulated as one of the major organizing principles in the brain [159]. According to the previous study, efficiency has been observed at many levels, including highly optimized information transmission and redundancy in the retinal ganglion cells, sparse encoding strategy of natural images in V1 and utilization of higher-order stimulus regularities in mid-level and high-level vision [160]. It is claimed in [158] that efficient representations would be maximally informative with respect to the actual inputs in the world. In particular, stimuli that are more likely to occur should be encoded more compactly. The primate visual system has long been known to utilize such perceptual regularities [161]. Another mid-level strategy, known as norm-based encoding [162], utilizes one particular regularity of the distribution of encountered exemplars from a given category, namely the center of this distribution. Leopold and colleagues argue in [163] that such strategy minimizes resources the system needs to learn and store stimulus.

10.2 Research Questions Associated with Mid-Level Representation Models Development

According to the discussion above, in this part, we explore mid-level representations that learn patterns to encode structure, as perceptual models. This investigation can be decomposed into more specific questions:

- ■ Mid-level representations of images/videos for quality assessment in different tasks (Part III)
  - □ What type of mid-level representations of images could be used for capturing structure related distortions for image/video quality assessment to mimic the 'encoding strategy' in the human visual system?

    As introduced at the beginning in chapter 1 that human visual system is able to encode low-level features extracted from images/videos efficiently for later interpretation of the scene, one may thus be curious about whether those structural distortions shown in chapter 2 can be 'encoded' too with some mid-level representations.
  - □ How to quantify the change of contours/curves from a higher semantic level for image/video quality assessment?

    As presented in chapter 2, some of the structure related distortions disrupt the 'categories' of the contours and thus are annoying for observers. For instance, a 'L' shape contour may be changed into an 'I' shape contour. This type of structural change can be considered as mid-level semantic change and are less acceptable compared to common compression artifacts. Therefore, a metric that is capable of quantifying the changes contours' categories from a higher level can be a solution to the problem.
  - □ Would how observers navigate among different viewpoints (when viewing a free-viewpoint video) affects the perceived quality?

    As emphasized in section 4.2, there is no existing subjective studies considering how observers navigate among viewpoints (especially in the form of content related navigation trajectories) effects perceived quality. In practice, it is common for human observers to stop navigating and stay in one viewpoint that contains important objects when viewing a free-viewpoints content. When an observer stops at one viewpoint and observes the moving objects, geometric distortions around those objects are easier to be noticed. Furthermore, as introduced in section 1.1, with the rapid development of immersive multimedia, more applications allow users to navigate in the virtual world. It is thus of great research value to explore whether content related navigation trajectory is one important factor that affects user visual experience.
  - □ If content related navigation scan-path matters, how to quantify the temporal artifacts appear due to views switch in applications where multi-views are available?

    As it has been pointed out in section 2.2, transitions of spatial structure related distortions among different viewpoints introduce a new type of structure related temporal distortions in terms of unsmooth structural transition among viewpoints. Also, according to what has been discussed in chapter 3, there is no video quality metric is designed for sequences where both temporal distortions observed within viewpoints and among viewpoints exist.
  - □ Is it possible to mimic the concept of 'encoding of configural regularity' in the human visual system for image quality assessment?

As mentioned in section 1.2.1, the visual system is very efficient in encoding stimuli properties by utilizing available regularities. Therefore, it is worth trying to find a way to model/mimic how structure related low-level configural regularity (e.g., contour configural regularity) is encoded in the visual system.

11

Encoding Contours with Sketch-Token Categories

11.1 Introduction

In this chapter, the first mid-level representation based model is presented and tested in the scenario of FTV. More specifically, a sketch-token based synthesized image quality assessment metric (ST-IQM) is proposed. In this model, contours are first 'encoded' as a vector of contour categories likelihood values. Then the perceived quality is predicted by quantifying to what extent the classes of contours change due to structure related distortions by comparing the 'encoded' contour category vectors.

According to the experimental results, the performance of ST-IQM is desirable and is of potential to be extended for video quality assessment on a more practical dataset that consider most of the important factors that affect perceived quality. However, as discussed in section 4.2.1, there is no subjective studies considers the impact of content-related navigations trajectories on perceived quality. To verify this impact, a subjective study is conducted and presented in this chapter. A free viewpoint video dataset (FVV) is released along with the subjective study. Afterwards, ST-IQM is extended for video quality assessment ST-VQM and tested on the new FVV dataset.

It is discussed in section 3 that structure related distortions contained in nowadays immersive multimedia are difficult for commonly used metrics to quantify. Distinguishable contour descriptors which capture edge structures have great potential in evaluating geometric transformations around disoccluded regions after synthesis. Bag of words based contours descriptor, sketch-token (ST) [164], trains a codebook for representing the categories of contours. This model could be used to mimic the 'encoding' process as the human visual system tends to encode low-level information for following higher-level process. With a ST-codebook, for each pixel in a test image, the probability that a patch centered at this pixel contain a certain category of contours could be predicted. In other words, contour within a patch could be represented as a vector of contour category.

For example, Figure 11.1 (c) shows a pair of patches with a part of human face and their extracted ST descriptors V_{ori}, V_{syn} from the reference and synthesized images. By observing the synthesized patch in Figure 11.1 (b), it is found that the boundary of the face is twisted and the shape has changed due to synthesis. Each dimension of ST vector is a probability value indicating how likely the current patch belongs to one certain category of contour from the codebook. In Figure 11.1 (c), each contour class from the codebook is visualized as edge patch, where edge pixels are labeled with white color. The positions of the elements (ST classes) in one ST vector are sorted according to the probability values for better observation. It can be observed that the most likely contour class that the original patch belongs to is a vertical straight line (the 64_{th} class in the codebook) while the one of the synthesized patch is a crooked line (the 105_{th} class in the codebook). The probability p_{64} for the straight token to exist in the original and synthesized patch are 10% and 1% respectively. Both of the two respective 'tokens' manage to reveal the basic shape of the patches, and the geometric transformation can be assessed based on the comparison between the extracted vectors.

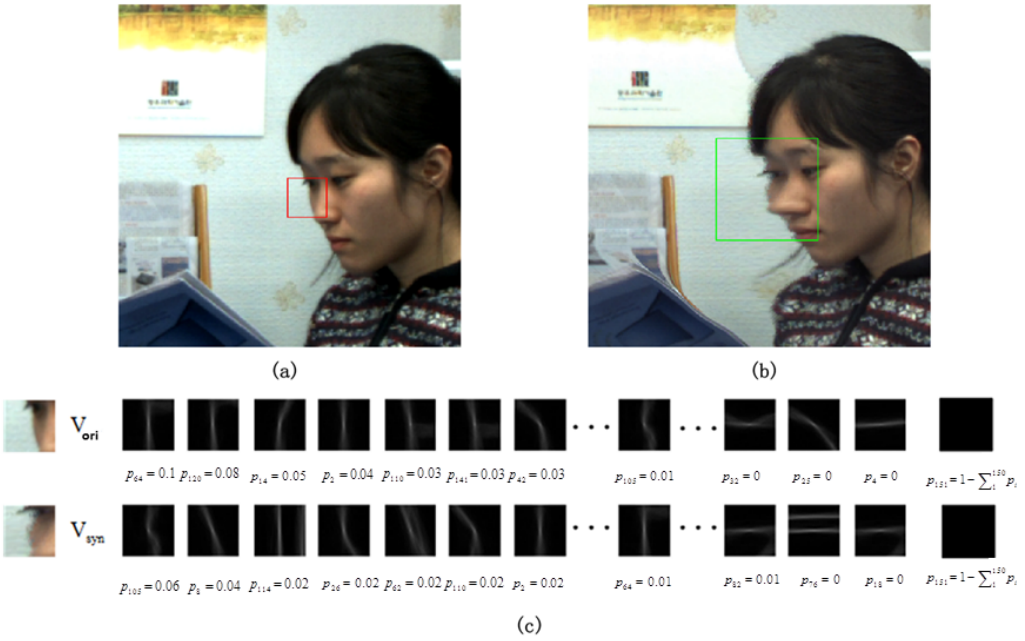


Figure 11.1 – Example explaining the principle of the proposed metric. (a) A patch in the reference image labeled with a red bounding box. (b) A searching window in synthesized image labeled with a green bounding box. (c) A pair of patches and their corresponding ST descriptors from the reference and synthesized images.

11.2 Sketch-Token based Image Quality Assessment Metric (ST-IQM)

In this section, the proposed ST-IQM is described. First and foremost, a registration step based on normalized cross-correlation [165] is incorporated to ensure shifting resilience and return a set of match patches. Then, mid-level contours feature called 'sketch-token' are extracted from both of the original and synthesized images in parallel with the registration step. An 'sketch-token' descriptor represents each pixel centering at a patch in the image with a vector showing the likeliness of the existence of each contour class in the patch. Dissimilarity among each matched patches is then calculated based on the contour feature vector, and a dissimilarity map

between the original and synthesized image is obtained. Finally, the objective scores of the synthesized images are estimated by pooling the dissimilarity map using Minkowski summation.

11.2.1 Registration Stage

Aiming at tackling the shifting artifacts described in section 2.2, the image registration approach proposed in [165] is utilized here for matching template centers at each pixel in the reference image from the searching window which centers at the same coordinate in the synthesized image. For example, the red bounding box (e.i. template) in Figure 11.1 (a) is matched to another patch from the green bounding box (e.i., searching window) in Figure 11.1 (b) during the registration stage. In order to match patches along borders of the images, both reference and synthesized images are padded with extra regions along the boundaries (e.g., dotted bordered rectangle in Figure 11.2) according to the size of the searching windows. The matching process is illustrated in Figure 11.2. For each pixel in the reference, a pixel (x_r, y_r) is considered as the central point of the template (blue square) which is needed to be matched in the corresponding searching windows (green rectangle) centralizing at the same position (x_s, y_s) in the synthesized image. This process involves the calculation of each position of the searching windows under examination a distortion function that measures the degree of dissimilarity between the template and alternative patches in the searching window. Then, the minimum distortion/maximum correlation position (x_m, y_m) is taken as the matched path (red square) from corresponding searching window in the synthesized image and is stored in the mapping matrix M_{match}, where $M_{match}(x_r, y_r) = (x_m, y_m)$.

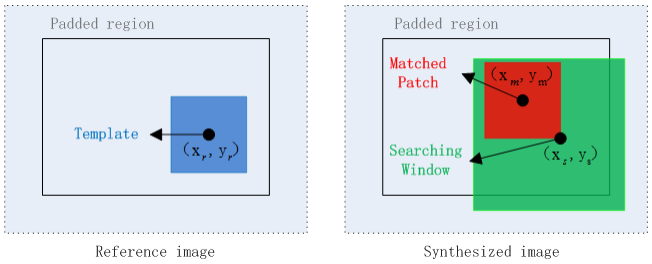


Figure 11.2 – Registration between reference and synthesized images

Normalized cross correlation(NCC) is employed for similarity measure [165, 166] in order to achieve better robustness. For a given template tem_{ref} that is located at a certain pixel in the reference image I_{ref} and its corresponding searching windows win_{syn} at the same position in synthesized image I_{syn}, the normalized cross covariance is defined as

$$NCC = \frac{C_{IM}(tem_{ref}, win_{syn})}{[\sum\sum win_{syn}^2(x + u_{shift}, y + u_{shift})]^{\frac{1}{2}}}, \tag{11.1}$$

where u_{shift} and v_{shift} are variables representing shift components along x-direction and y-direction respectively. $C_{IM}(tem_{ref}, win_{syn})$ is defined as (11.2) with n_{pixel} equaling to the number of pixels in the template:

$$C_{IM}(tem_{ref}, win_{syn}) = \frac{n_{pixel}\sum tem_{ref} win_{syn} - \sum tem_{ref}\sum win_{syn}}{\sqrt{(n_{pixel}\sum tem_{ref}^2 - (\sum tem_{ref})^2)(n_{pixel}\sum win_{syn}^2 - (\sum win_{syn})^2)}}. \tag{11.2}$$

11.2.2 Sketch-Token Descriptors Extraction

The local mid-level features called 'sketch-token' is adapted here to better capture how contour boundaries change for the sake of predicting the quality of synthesized images. To obtain the sketch-token classes, Joseph J. Lim *et al.* [164] first asked human subjects to generate sketches for each training image as the structural contours. Then, the sketch-token categories set were defined by clustering patches with a fixed size of 35×35 pixels. After the clustering process, only 151 of tokens which capture the most commonly occurred edges were maintained. Random decision forests model was then used to trained classifiers for each image patch with a set of low-level features including oriented gradient channels [167], color channels, and self-similarity channels [168]. Each output of these 151 classifiers corresponds the possibility p_i of the existence of corresponding token i in that patch and $\sum_i p_i = 1$. The dimension of a ST contour descriptor for one pixel in an image is 151 including an extra dimension indicating how likely this patch does not contain any tokens (no contour class). For instance, in Figure 11.1 (c), the ST descriptor of the patch that is located at (x, y) can be represented as $V(x, y) = (p_1, p_2, ..., p_{151})$. At the end of feature extraction stage, the contour feature maps M_{ref} and M_{syn} are obtained for both reference and synthesized images.

11.2.3 Distortion and Pooling Stage

After obtaining the contour features map M_{ref}, M_{syn} and the mapping matrix M_{match}, the distance between each matched contour vectors is then calculated with certain distance measure. Considering the fact that the sum over the 151 dimensions of each contour vector $\sum_i p_i$ equals to 1, Jensen – Shannon divergence is used here for calculating the distance between two contour vectors. Similar to Kullback – Leibler divergence, Jensen – Shannon divergence is an approach which can be used as similarity measurement between two probability distributions. Other distance measures are also tested and further described in the next section. For each pixel (x_r, y_r) in the reference image, the corresponding center coordinate of its matched patch in the synthesized image is given by $M_{match}(x_r, y_r) = (x_m, y_m)$. The contour descriptor of each pixel is stored in the aforementioned contour feature maps, where $M_{ref}(x_r, y_r) = V_{ori}(x_r, y_r)$ and $M_{syn}(x_m, y_m) = V_{syn}(x_m, y_m)$. Then, the dissimilarity between the matched patches centering at (x_r, y_r) and (x_m, y_m) respectively is calculated as

$$\begin{aligned} D_{JSD}(s) = &\frac{1}{2} D_{KLD}(V_{ori}(x_r, y_r), A) \\ &+ \frac{1}{2} D_{KLD}(V_{syn}(x_m, y_m), A), \end{aligned} \tag{11.3}$$

where $A = \frac{1}{2}(V_{ori}(x_r, y_r) + V_{syn}(x_m, y_m))$, and D_{KLD} is the Kullback– Leibler divergence defined as

$$D_{KLD}(V_{ori}, V_{syn}) = \sum_i V_{ref}(i) log \frac{V_{ori}(i)}{V_{syn}(i)}. \tag{11.4}$$

As mentioned before, since $\sum_i p_i = 1$ and p_{151} corresponds to the category of non-contour.The majority pixels belong to non-contour pixel are with high p_{151} values while all the other elements in the feature vector are around zero. The dissimilarity values for non-contour regions in M_{match} are also around zeros since $D(V_{ori}, V_{syn}) \approx D(p_{151}^{ref}, p_{151}^{syn})$. The dissimilarity matrix is commonly a sparse matrix because there is few differences in non-contour regions. Most of the non-zero elements for most of the dissimilarity maps in the dataset is lower than 0.5 (D_{JSD} range from 0 to 1). In order to amplify error regions along the contours, the minkowski distance

measure is used as pooling strategy to pool dissimilarity values to get the final objective score. The proposed metric ST-IQM is then defined as

$$ST - IQM = \frac{[\sum_{N_{pixel}} D_{JSD}(V_{ori}(x_r, y_r), V_{syn}(x_m, y_m))^{\beta_{ST}}]^{\frac{1}{\beta_{ST}}}}{N_{pixel}}, \tag{11.5}$$

where N_{pixel} is the number of pixels in the image and β_{ST} is a parameter corresponds to the $\beta - norm$ defining the L^{β}_{ST} vector space. The selections of β_{ST} and distance metric for calculating distance between contour vectors in (11.6) are further discussed in the section 11.2.4.

11.2.4 Experimental Results

The performance of sketch-token based synthesized image quality assessment is evaluated using the IVC-Image dataset [49, 98] described in section 4.2.1.1. As mentioned in [144], images synthesized with algorithms A1 is excluded from the experiment due to the significant shifting artifacts compared to other algorithms. In order to be consistent with their experiment, the proposed metric is evaluated not only on the entire dataset but also on the subset where images generated with A1 are excluded.

To compare the performances between existing metrics designed for synthesized images summarized in section 3.4 and the proposed metric, the following widely employed criteria PCC, SCC, and RMSE introduced in section 4.3 are utilized with non-linear mapping between the subjective scores and objective measures.

The performance dependency of the proposed algorithm on the exponent variable β_{ST} in equation (11.6) is examined on the subset where images generated by A1 are excluded. The result is provided in Figure 11.3. According to Figure 11.3, with increasing value of β_{ST}, the performance of the proposed metric increases significantly and peaks with $\beta_{ST} = 4$. Afterwards, the performance drops steadily.

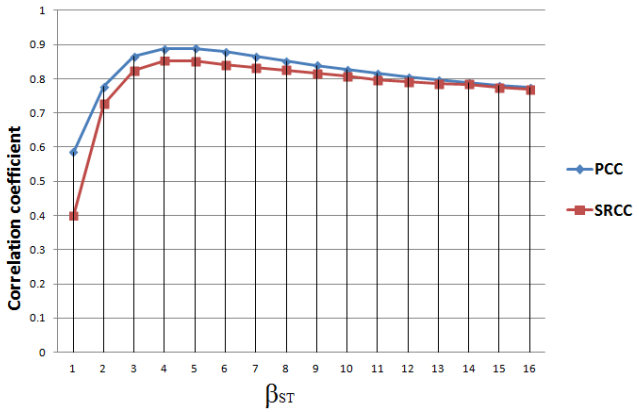


Figure 11.3 – Performance dependency of the proposed ST-IQM metric with changing β_{ST}.

Distance metrics are also explored for calculating the dissimilarity between each ST contour vector utilized in (11.6), and to check the dependency between the performance and the distance metric. Different distance metrics including Jensen – shannon divergence, cosine distance, euclidean distance, and chi-squared distance are tested. This test is also conducted on the subset where images generated by A1 are excluded. The results are illustrated in Table 11.1. Table 11.1 shows that the performance of the metric does not vary significantly with the change of using different distance metrics. The proposed ST-IQM metric acquires the best performance

with Jensen – shannon divergence distance measure by achieving 0.8877, 0.8525 and 0.3070 values of PCC, SCC, and RMSE respectively. This outcome proves the feasibility of choosing Jensen – shannon divergence metric for dissimilarity evaluation.

Table 11.1 – Performance of ST-IQM with different distance approaches

	PCC	**SCC**	**RMSE**
Jensen– Shannon divergence	**0.8877**	0.8525	**0.3070**
Cosine Distance	0.8680	0.8419	0.3312
Euclidean Distance	0.8584	0.8024	0.3422
Chi-Squared Distance	0.8829	**0.8531**	0.3132

Based on the experimental results and analyses described above, we have fixed β_{ST} to be 4 and selected Jensen – Shannon divergence measure for calculating the distance between ST contour vectors of each matched pair of matched patches from reference and synthesized images. During the registration stage, the size of the template in original images and the one of the searching windows in the synthesized images are set empirically as 35×35 and 90×90 respectively.

The overall result of performance comparison among the image metrics is concluded in Table 11.2. According to Table 11.2, the proposed ST-IQM achieves 0.8877, 0.8525 and 0.3070 value of PCC, SCC and RMSE correspondingly on the subset of the dataset (where images generated with A1 are excluded) and 0.8217, 0.7710 and 0.3929 on the entire dataset. According to t-test results, it outperforms the compared image metrics significantly (with P-value smaller than 0.05).

Table 11.2 – Performance comparison of the proposed metric with the state-of-the-art metrics.

	PCC	**SCC**	**RMSE**
Subset without images generated with A1			
MP-PSNR$_f$ [75]	0.8874	0.8175	0.3165
MW-PSNR$_f$ [74]	0.8855	0.8298	0.3188
ST-IQM	**0.8877**	**0.8525**	**0.3070**
Entire dataset			
MP-PSNR$_f$ [75]	0.6553	0.6239	0.5029
MW-PSNR$_f$ [74]	0.6089	0.5738	0.4948
ST-IQM	**0.8217**	**0.7710**	**0.3929**

For the purpose of checking the capacity of ST-IQM to detect specific artifacts generated with DIBR algorithms, the dissimilarity maps M_{dis} calculated with Jensen– shannon divergence metric are visualized according to the dissimilarity values. Figure 11.4 is an example showing some regions of the obtained dissimilarity maps along with their corresponding regions from original and synthesized images. The regions are generated by enlarging the aforementioned template for better observation. By observing these error maps, it is found that:

1. Incorrect rendering/texture stretching: For the first row, by checking the respective visualized dissimilarity map, it could be found that both the 'fake edge' generated by blurred region and the disappearance of the missing 'hair boundary' are well captured.
2. Blurry regions: The synthesized region in the second row is from 'Book Arrival' sequence synthesized with A5. Blurred regions along the objects' boundaries as well as the missing hands of the clock are emphasized with darker colors indicating higher dissimilarity values.
3. Dark holes: For the third row, in the respective dissimilarity map region, the dark hole regions are assigned with larger dissimilarity values (darker color) verifying the capacity of our metric to detect these kinds of artifacts.

4. Geometry distortion (twisted shape): The main problems in the synthesized regions in the last row are twisted shape of the face, especially the right part of it, and the missing left ear of the girl. According to the respective dissimilarity map, these main distorted regions are also well detected, especially the left ear region which is depicted with the darkest color.

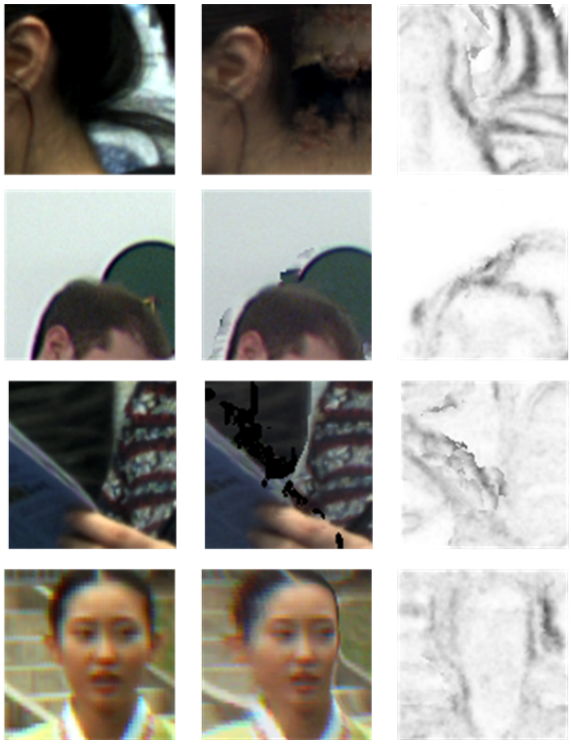


Figure 11.4 – First column: regions from original images; second column: matched regions in synthesized images; third column: corresponding regions in dissimilarity maps obtained from ST-IQM (the darker the color the higher the dissimilarity value)

Furthermore, to check the complexity of the metric compared to other metrics designed for synthesized views, the execution time of the metrics normalized by PSNR as introduced in section 4.4. It is shown that the proposed ST-IQM is much slower than the second best performing MP-PSNR. However, since the gain of performance on the entire dataset is significant compared to the second best performing metric, i.e., 18 % of gain in PCC values, it is acceptable if the running time is longer.

Table 11.3 – Normalized execution time of proposed metric compare to the state-of-the-art metrics.

Metric	MW-PSNR	MP-PSNR	ST-IQM
Normalized time	**12.4**	100	1324

11.3 Impact of Navigations Scan-Path on Perceived Quality: Free Navigation vs. Predefined Trajectories

Immersive media technologies provide the users with more freedom to explore the content allowing more interactive experiences than with traditional media. These new possibilities introduce the observers' behavior as an important factor for the perceived quality [96]. Given the fact that each observer can explore the content differently, there are two approaches that can be adapted to practically study this factor: 1) let the observers

navigate the content freely; 2) let the observer watch the sequences in the form of certain pre-defined navigation trajectories. By employing the first approach, a common trajectory could be obtained according to all the observers' data. However, this common trajectory does not necessarily represent the critical one that will stress the system to the worse case. Moreover, if observers are allowed to navigate freely during a test, it may become a new factor that increases the variability of the MOS (despite observer's variability in forming a quality judgment). In order to obtain MOS that can distinguish one system from another statistically significantly. more observers are required. The second approach (predefined trajectories) is not affected by this trajectory-source of variability but it comes with the challenge of selecting the 'right' trajectory. In case of system benchmark, the 'right' trajectory could be defined as the most critical one or weakest link (e.g., the one that leads to the lowest perceived quality). Nevertheless, there is a high possibility that this trajectory-effect is highly dependent on content, some being more sensitive than some others to the choice of trajectory. Identifying the impact of navigation trajectory among different viewpoints on perceived quality for a given content is then of particular interest. For quality evaluation, it may be useful to know how navigation affects the visual experience and which are the 'worst' trajectories for the system, to carry out performance evaluations of the system under study in the most stressful cases. Consequently, the availability of computational tools to select the critical trajectories would be extremely useful.

As discussed at the beginning of the thesis in chapter 4, no existing subjective study is conducted to check how observers navigate the free-viewing content affect the quality of it. To meet this need, a subjective study is conducted by designing content related trajectories to mimic the worse cases. A video quality dataset for FVV scenarios named as 'image, perception and interaction group free-viewpoint video dataset' (FVV) is built. This dataset consists of sequences that contain both compression, view-synthesis artifacts, and temporal structure consistencies. More specifically, the videos of this dataset are generated by simulating exploring trajectories that the observers may use in real scenarios, which are set by the hypothetical rendering trajectory (HRT), defined in the section 11.3.1.

11.3.1 Hypothetical Rendering Trajectory

A commonly used naming convention for subjective quality assessment studies was provided by the video quality experts group [169], including: SRC (i.e., source or original sequences), HRC (i.e., hypothetical reference circuit or processing applied to the SRC to obtain the test sequences, such as compression techniques), and PVS (i.e., processed video sequence or the resulting test sequence from applying an HRC to a SRC). In the context of free navigation, another dimension of the system under test related to the interactivity part should be reflected (e.g. the use of exploration trajectories in the quality evaluation of immersive media). Towards this goal, the term hypothetical rendering trajectories (HRT) is introduced, to reference the simulated exploration trajectory that is applied to a PVS (as the result of an HRC on a give SRC) for rendering. It is worth mentioning that the generality of this term is applicable to all immersive media from multi-view video, VR, light fields, AR to point clouds.

11.3.2 Test Material

Three different super multi-views sequences are utilized in this study. These three sequences are 'champagne tower' (CT), 'pantomime' (P) and 'big buck bunny flowers' (BBBF). The description of the three SMV sequences

is summarized in Table 11.4. They were also selected as test materials in [170]. For each of the 3 SRC sequences, 20 HRCs, are selected, covering 5 baselines and 4 rate-points (RP). In addition, 2 HRTs are also included to generate 120 PVSs. Details on these parameters (selected after a pretest with expert viewers) are described in the following subsections.

Table 11.4 – Information of the sequences, including properties and selected configuration (rate-point and baseline distance).

Name	Views	Resolution	Fps	Seconds	Frames	QP values				Baseline Distance
						RP1	RP2	RP3	RP4	
BBBF	91	1280 x 768	24	5	121	35	-	45	50	B0, B2, B5, B9, B13
CT	80	1280 x 960	29.4	10	300	37	43	-	50	B0, B4, B8, B12, B16
P	80	1280 x 960	29.4	10	300	37	43	-	50	B0, B2, B6, B12, B16

11.3.2.1 Camera Configuration

For each SRC, 5 stereo baseline values, as summarized in Table 11.4, are selected in the test including the setting Set_{b_0} without using synthesized views. The baseline is measured based on the camera distance/gap between the left and right real views. Here, B_i or b_i represents the stereo baseline distances that were settled to generate the synthesized virtual views, where i is the number of synthesized views between two reference views. Figure 11.5 illustrates the baseline setting for synthesized views generation in the subjective study. For instance, for camera setting Set_{b4} in the upper part of Figure 11.5, between each pair of views that captured by original cameras (indicated by two closest black cameras in the figure) there are four virtual views that are synthesized using them as left and right reference. In this case, the baseline distance is 4, denoted as b_4. For example, in the lower part of Figure 11.5, for Set_{b4}^{R1}, between each two transmitted encoded views, there are totally 4 virtual synthesized views are generated.

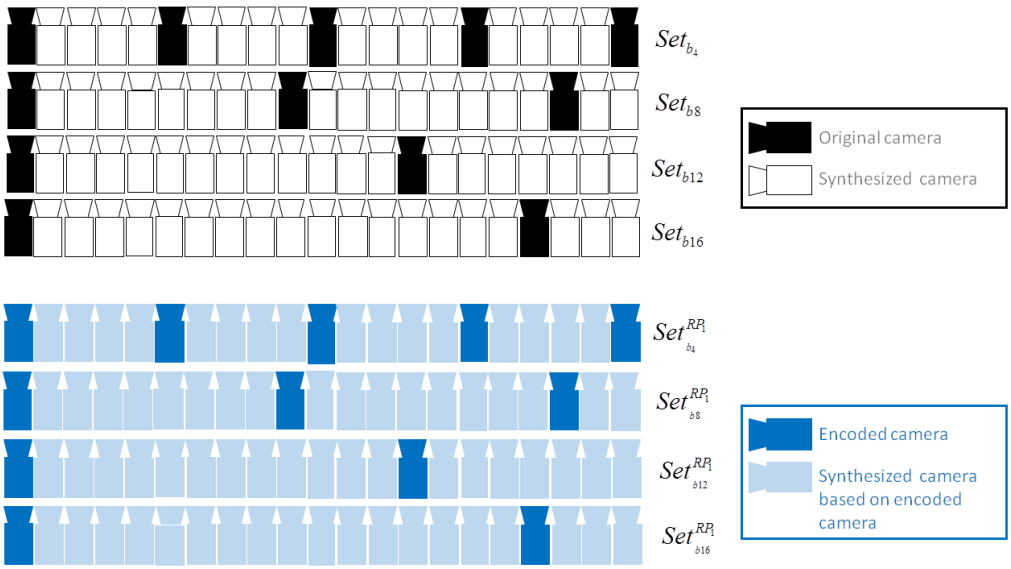


Figure 11.5 – Camera arrangements (1) The upper part of the figure is the configuration designed in [93, 94] where the black cameras represent the sequences taken with real original cameras while the white ones indicate the synthesized view using the original ones as references. (2) The lower part of the figure is the camera configuration in our experiment, where the deep blue camera represents the encoded/transmitted sequences taken from the corresponding original camera and the lighter blue ones indicate the synthesized ones using the encoded ones as references.

11.3.2.2 3D-HEVC Configuration

In this experiment, HTM 13.0 in 3D high-efficiency video coding (3D-HEVC) mode is used to encode all the views of the three selected SMV sequences. These encoded views along with the selected original ones are used as the reference views in the following synthesis process, which are also named as 'anchors'. The configuration of the 3D-HEVC encoder recommended in [170] is adopted in this experiment. Specifically, taking into account the contents and the limitations of the duration of subjective experiment tests, 3 rate-points, as summarized in Table 11.4, are selected for each SRC according to the results of the pretest. For each content, the original sequences without compression are included in the experiment and are denoted as RP_0.

11.3.2.3 Depth Maps and Virtual Views Generation

In this study, reference software tools are used for the preparation of the synthesized views, including depth estimation reference software (DERS) and view synthesis reference software (VSRS). Both of them have been developed throughout the MPEG-FTV video coding exploration and standardization activities. To generate virtual views with reference sequences taken by real cameras, depth maps, and related camera parameters are required. For sequences 'CT' and 'P' [171], since original depth maps were not provided, DERS v6.1 is used to generate depth maps for each corresponding view. Relative parameters are set as recommended in [172, 173]. For synthesized views-generation, the version 4.1 of VSRS is applied. For each corresponding content, the configuration of the relative parameters is set according to [173].

11.3.2.4 Navigation Trajectory Generation

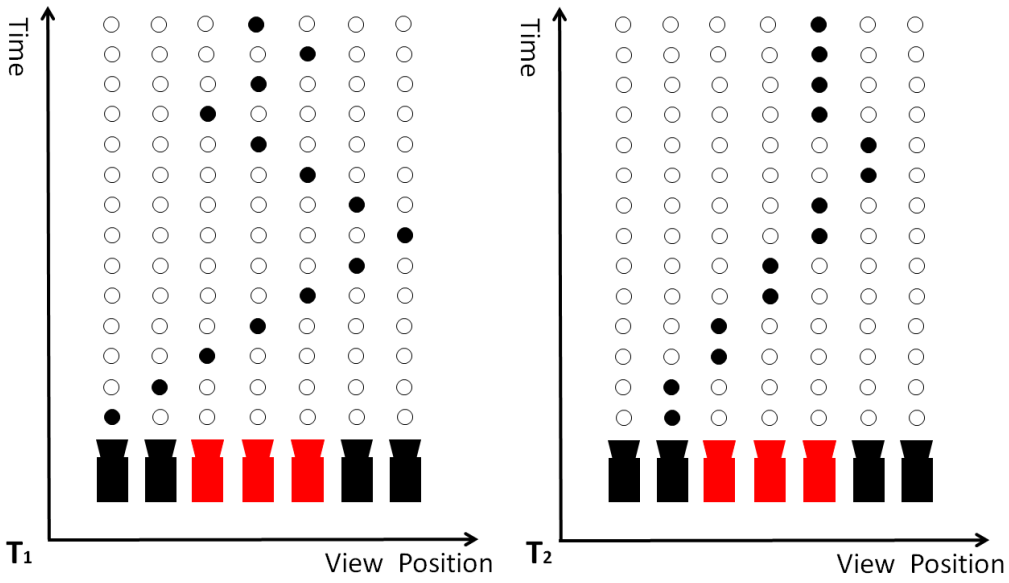


Figure 11.6 – Description of generated trajectories. In the figure, red cameras indicate views contain important objects while the black ones represent the one mainly contain background (1) **Left** T_1 : Sweeps (navigation path) are constructed at a speed of one frame per view (as what is done in MPEG) (2) **Right** T_2 : Sweeps (navigation path) are constructed at a speed of two frames per view.

One of the purposes of this study is to check whether semantic contents (e.g. moving objects) of the videos and how the navigation trajectories among views will affect the perceived quality. Therefore, different HRTs are considered in this study, generating sweeps that focus more on important objects since human visual system

tends to attach greater interest on 'regions of interest' (ROI) [174] that contain important objects. Specifically, the following two HRTs are chosen from the pretest session (because human observers may pay more attention and even stop navigating to observe targeted objects in the video). These two HRTs are denoted with T_1 and T_2 as depicted in Figure 11.6: (T_1) An 'important-objects HRT' that first scans from the left-most to the right-most views to observe the overall contents in the video, then scans back to the views that contain the main objects and looking left and right around the central view that contain the objects several times at a velocity of one frame per view (1fpv); (T_2) An 'important-objects-stay HRT' that first scans from the left-most to the right-most views to observe the overall content in the video, then scans back to the views that contain main objects at a velocity of 2fpv and finally stays in the central view that contains the main object. Due to the limitation of resources, only two trajectories are considered in this study as initial exploration.

11.3.3 Test Methodology

Absolute category rating with hidden reference (ACR-HR) [175] is adopted for this subjective experiment. The observers watch the test videos sequentially, and after each one, they provide a score using the five-level quality scale. For this purpose, an interface with adjectives representing the whole scale is shown until the score is provided before next text video is displayed. Additional, the test videos are shown to each observer in different random orders, and each of them is shown only once. At the beginning of the test session, an initial explanation is given to each participant indicating the purpose and how to accomplish the test. Then, a set of training videos are presented to the observers to familiarize them with the quality range of the content. The entire session for each observer lasts 30 minutes.

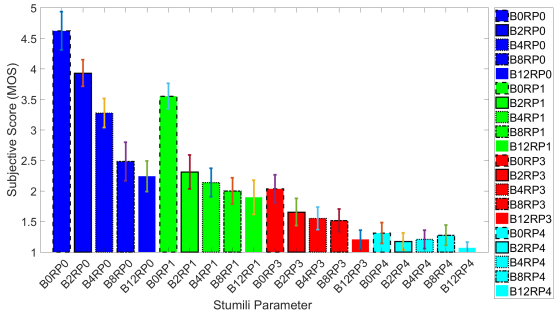
11.3.4 Environment and Observers

The test sequences are displayed on a professional screen TVLogic LVM401W, using a high-performance computer. Observers are provided with a tablet connected to the displayed computer for voting. The test room is set up according to the ITU recommendation BT.500 [176]. The walls are covered by gray-color curtains, and the lighting conditions are regulated accordingly to avoid annoying reflections. Moreover, a viewing distance of 3H (H being the height of the screen) is chosen.

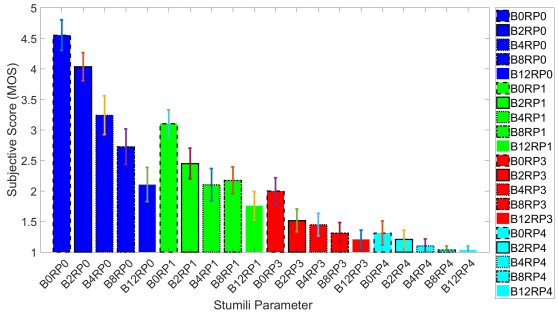
There are 33 participants in the subjective test, including 21 females and 12 males, with ages varying from 19 to 42 (average age of 24). Before each experimental trial, observers are screened for correct visual acuity and color vision using the Snellen chart and Ishihara test, respectively. All of them report normal or corrected-to-normal vision. After the subjective test, the obtained scores are screened according to the procedure recommended by the ITU-R BT.500 [176] and the VQEG [169]. As a result of this screening, data of four observers were removed.

11.3.5 Subjective Experiment Results and Analysis

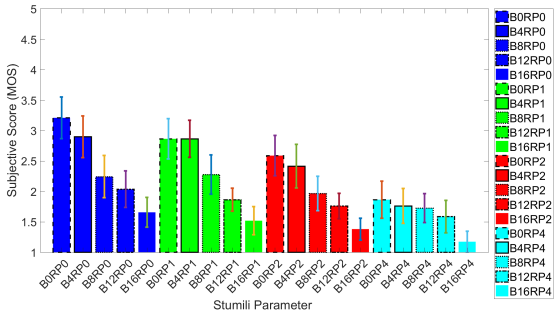
The result of the subjective test is depicted in Figure 11.7, where each sub-graph summarizes the mean opinion score (MOS) (with confidence intervals [176]) for each content in each virtual sweep. Apart from MOS, the differential mean opinion score (DMOS) is also provided along with the dataset, computed from the hidden references according to [175]. As required for a quality dataset, the MOS values are well distributed covering almost the whole rating scale. In addition, in order to verify whether different baselines (B), rate-points (RP) and, especially, virtual trajectories (T) have significant impacts on perceived quality, a three-way analysis of



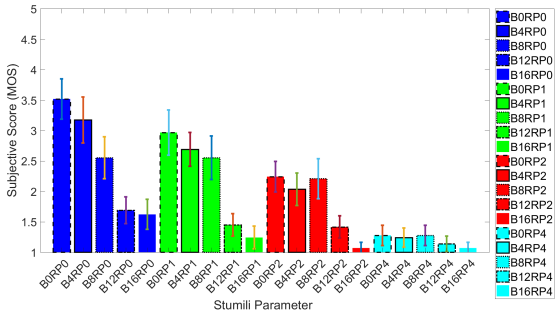
(a) Sequence BBB Flowers with T_1



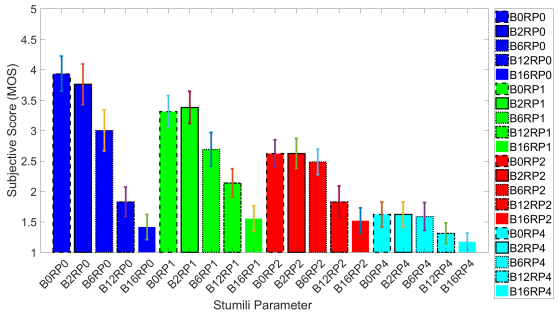
(b) Sequence BBB Flowers with T_2



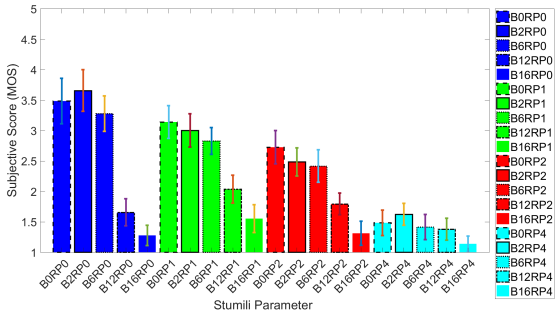
(c) Sequence Champagne with T_1



(d) Sequence Champagne with T_2



(e) Sequence Pantomime with T_1



(f) Sequence Pantomime with T_2

Figure 11.7 – MOS of the sweeping sequences with different rate-points (RP), different baselines (B) and different sweeping trajectories (T) in the FVV dataset.

variance (ANOVA) is performed. Considering the results of this test and the results illustrated in Figure 11.7, the following main conclusions could be made:

- With a same configuration (i.e., baseline, rate-point, and trajectory), the quality obtained with different contents are significantly different.
- The effects of view-synthesis and compression artifacts are obvious, as shown when considering how the perceived quality changes with only baseline (for a given RP), or with only bitrate (fixing the baseline). The accumulation of the effects can be also observed in the scores for the tests sequences with combined degradations.
- The three considered factors, specially trajectory T, have significant impact on the perceived quality ($p = 0$ for B and RP, and $p = 0.038$ for T).
- Regarding interaction among the considered factors, the interaction between baseline distance and coding quality has a significant effect on the MOS scores ($p = 0$).

Following are more detailed analysis of the impact of trajectory on perceived quality:

1. The averaged MOS values (averaged contents 'CT', 'P', 'BBBF' and conditions) of sequences in form of T_2 are smaller than the ones of T_1. Apart from ANOVA test, to further confirm the impact of the navigation trajectory on perceived quality, the dataset is divided into two sets based on which trajectory the sequences are generated with. A t-test is conducted by taking the pairs of sequences in form of T_1 and T_2 with same baseline, rate-point configuration as input. According to the result, there is a significant difference between the quality of these two sets (i.e., T_1 and T_2).

2. Certain contents are more sensitive to certain trajectories. To further check whether the impact of certain trajectories depend on the content of the sequences, another t-test is conducted. More specifically, for each content, pairs of sequences that generated with the same baseline and ratepoint but different trajectory are first formed. Then, a t-test is conducted by taking the individual subjective scores (opinion scores from all the observers) of each pair of these sequences as input. According to the t-test result, for content 'C', 50 % of the pairs are of significantly different perceived quality. However, for content 'CT' and 'BBBF', only around 10% of pairs are of significantly different quality. It is proven that the impact of the trajectory on quality is content dependent. In other words, 'extreme trajectory' of videos with different contents are different.

3. Whether the quality of sequence in the form of one trajectory is higher than another depends also on the quality range (regarding baseline and rate-point setting). The results of t-test taking individual subjective score of each trajectory pair as input also shows that, for content 'C', videos that in the form of T_2 are of better quality than the ones in T_1 when quality is higher than a certain threshold (smaller baseline or smaller rate-point) and vise versa. For example, for content 'C' with rate-point larger than RP_2, sequence in form of T_1 is better than the one in form of T_2.

In conclusion, it is confirmed by the subjective study that there is an impact on perceived quality from navigation trajectory. It is found that content related trajectory can stress the system one step further for a more extreme situation. Therefore, image/video objective metrics that are able to indicate sequences in the form of one trajectory is of better quality than another is required to push the system to its limit according to the contents. To meet this need, a video quality metric is introduced in the next section by extending ST-IQM.

11.4 Sketch-Token based Video Quality Assessment Metric (ST-VQM)

Considering the facts that 1) content related trajectory is able to stress the system; 2) content is related to structure; 3) geometric distortions are the most disturbing degradations that interrupt structure introduced by view synthesis, in this section, ST-IQM metric is extended for video quality assessment as ST-VQM. Since temporal structure inconsistencies are the most upsetting degradations in synthesized sequences, the main idea of the proposed method is to assess the quality of the sweeping videos by quantifying to the changes of contour category change due to synthesis/compression distortions or transition among views.

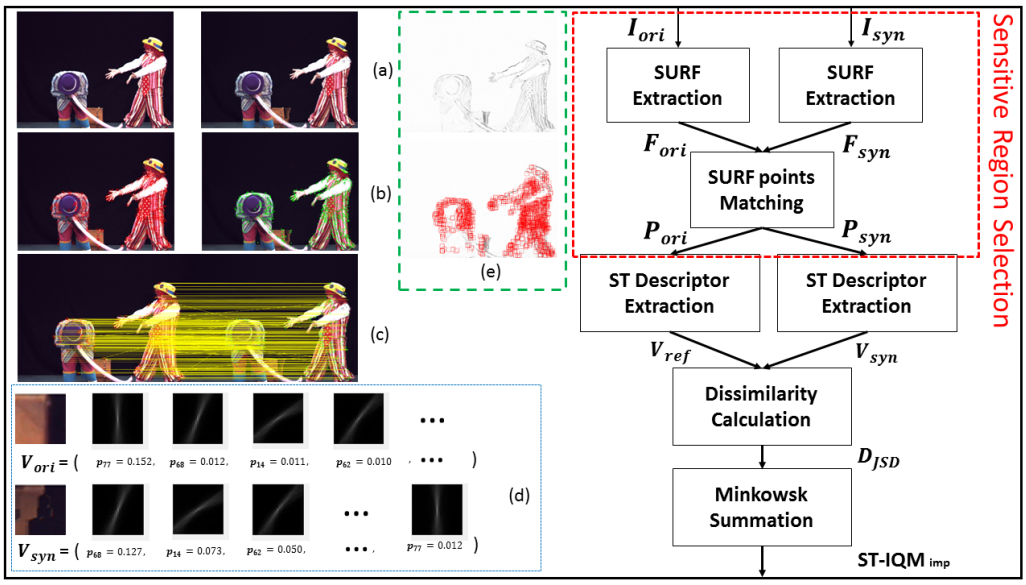


Figure 11.8 – Overall framework of the proposed objective metric: (a) Reference image (on the left) and synthesized image (on the right); (b) Extracted SURF key-points of the reference and synthesized images; (c) Matched key-points from the reference to the synthesized image; (d) Extracted ST feature vector of the corresponding patches and its visualization of each contour category.

As mentioned in section 11.2, one of the disadvantages of ST-IQM is related to its complexity. Image registration stage is the bottleneck of the metric. In order to improve the efficiency of ST-IQM for efficient video quality assessment, ST-IQM$_{imp}$, an improved version of ST-IQM, is proposed in this section by replacing image registration stage with the sensitive/distortion regions selection methodology proposed in section 8.2.1. Then, ST-VQM is proposed by combining the improved spatial estimator ST-IQM$_{imp}$ and one sketch-token based temporal estimator (ST-T). The framework of ST-IQM$_{imp}$ and ST-T are represented in Figure 11.8 and Figure 11.9 respectively. For ST-IQM$_{imp}$, after incorporating sensitive regions selection procedure proposed in 8.2.1 to ensure shifting resilience. The sensitive regions selection process is based on SURF points detection and matching as depicted in section 8.2.1. Similarly, the obtained selected sensitive regions are actually matched patches whose centers are the matched SURF points in synthesized and reference frames. With the selected sensitive regions, ST descriptors are then extracted from these matched patches. In particular, ST descriptor represents each matched patch with a vector showing the likeliness of the existence of each contour class in the dictionary. Once the contour feature vectors from the reference and the distorted frames are obtained, a dissimilarity value is computed for each matched patches. Then, by pooling all the dissimilarity values

(calculated based on the matched patches) using a minkowski summation, a global value representing the spatial dissimilarity of the whole frame can be then obtained as ST-IQM$_{imp}$. For ST-T, a temporal vector is computed by concatenating the spatial difference between each frame pair in the synthesized and reference sequence. Afterwards, the temporal inconsistency could be quantified by calculating the distance between the obtained temporal vectors of the synthesized and reference sequences. In the end, the spatial (ST-IQM$_{imp}$) and temporal (ST-T) structure dissimilarities are combined to get the final objective score for the sequence. Details of each part of the proposed model are given in the following subsections.

11.4.1 Local Sensitive Regions Selection

Local regions selection is essential for the later evaluation of the quality of DIRB-based synthesized views as already described in section 8.2.1 (mainly three reasons). Therefore, the same local distortion regions selection method proposed in section 8.2.1 is incorporated in this model to improve its efficiency.

The process of sensitive regions selection is summarized by the red dash bounding box in Figure 11.8. First SURF feature points (i.e. F_{ori} and F_{syn}) are extracted from both original I_{ori} and synthesized frames I_{syn}. Then SURF points matching between the two frames is achieved following the reference method in [145] (the original frame is considered as the reference for this matching process). Pairs of interest points that have significantly different x and y values are discarded. They are considered as not plausible matched regions from the synthesis process. The patches P_{ori}, P_{syn} centered at the corresponding matched SURF points in synthesized and original images are considered. The size of these patches is set as 35×35 to match ST formalism as introduced by [164]. The matching relation for all patches is encoded in a matching matrix $M^{RS}_{match}(x_r, y_r) = (x_m, y_m)$, where (x_r, y_r) corresponds to the coordinate of one SURF point of the patch in the reference frame and (x_m, y_m) is the coordinate of its matched SURF point of the patch in the synthesized frame. Different from the matrix M_{match} in section 11.2, M^{RS}_{match} is obtained by sensitive regions selection.

To illustrate the capability of SURF for selecting sensitive regions, an example is presented in Figure 11.8 (e). The error maps are generated with the synthesized and the reference images as introduced in [144]. The darker the region, the more distortions it contains, as depicted in the top part of the dashed bounding green box in Figure 11.8 (e). The red bounding box in the lower part of Figure 11.8 (e) represents the sensitive regions as extracted by the proposed process. It can be observed that, the majority of regions containing severe local distortions are well identified by this process.

11.4.2 Improved Sketch-Token based Spatial Dissimilarity

Similar to ST-IQM, the dissimilarity between each matched contour vectors is computed with M_{ref}, M_{syn} and M^{RS}_{match} using Jensen–Shannon divergence. For each matched interesting key point pixel (x_r, y_r) in the reference image, the corresponding coordinate of its matched key point in the synthesized image is given by $M^{RS}_{match}(x_r, y_r) = (x_m, y_m)$. The contour descriptor of each pixel after region selection is stored in the aforementioned contour feature maps, where $M^{RS}_{ref}(x_r, y_r) = V^{RS}_{ref}(x_r, y_r)$ and $M^{RS}_{syn}(x_m, y_m) = V^{RS}_{syn}(x_m, y_m)$. Then, the dissimilarity between the matched patches centering at (x_r, y_r) and (x_m, y_m) respectively can be calculated using equation (11.3) and the final improved ST-IQM$_{imp}$ can be calculated using

$$ST\text{-}IQM_{imp}(I_{ref}, I_{syn}) = \frac{[\sum_{N_{pixel}} D_{JSD}(V_{ref}^{RS}(x_r, y_r), V_{syn}^{RS}(x_m, y_m))^{\beta_{ST}}]^{\frac{1}{\beta_{ST}}}}{N_{pixel}}, \tag{11.6}$$

where N_{pixel} is the total number of pixels in the frame and β_{ST} is a parameter corresponds to the $\beta - norm$ defining the L^{β} vector space.

11.4.3 Sketch-Token based Temporal Dissimilarity

Sweeping between views introduces and amplifies specific structure related temporal artifacts, including flickering, temporal structure inconsistency, etc. Among them, temporal structure inconsistency is usually the most sensitive artifacts for human observers since they are usually located around important moving objects and are easier to be noticed compared to other temporal artifacts. To quantify temporal structure inconsistency,

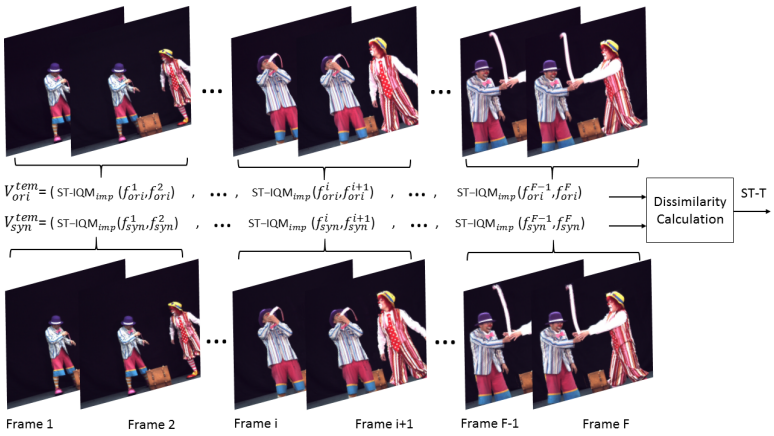


Figure 11.9 – Diagram of sketch-token based temporal distortion computation, where F is the total frame number of the sequence.

the dissimilarity score between each pair of continuous frames is computed using the proposed sketch-token model introduced in section 11.4.2. In section 11.4.2, ST-IQM$_{imp}$ is used to quantify the difference of structure between synthesized and its reference frames (the original purpose of this framework). It can also be used to encode and describe how structures are evolving from one frame to another along a given sequence. Temporal structure changes as observed in FVV should affect this description. This approach is exploited to refine the quality estimation in case of FVV in order to capture temporal structure inconsistency.

How the sketch-token based temporal distortions are quantified is explained in Figure 11.9. More specifically, for each pair of continuous frames of a sequence S, f^i and f^{i+1}, $ST - IQM_{imp}(f^i, f^{i+1})$ is computed using equation (11.6). A temporal vector V^{tem} can be formed considering all frames of the sequence (each component of the vector corresponding to $ST - IQM_{imp}(f^i, f^{i+1})$). The sketch-token based temporal dissimilarity (ST-T) between the original and the synthesized sequences is defined as the euclidean distance between the two temporal vectors of the original and the synthesized sequence:

$$ST\text{-}T(S_{ori}, S_{syn}) = D_e(V_{ori}^{tem}, V_{syn}^{tem}) \tag{11.7}$$

where $D_e(\cdot)$ is the euclidean distance function.

11.4.4 Pooling

With the improved spatial sketch-token based score ($ST - IQM_{imp}$) and the temporal sketch-token based score (ST-T), they are then combined to produce an overall quality score defined as:

$$ST\text{-}VQM = w_S \cdot ST\text{-}IQM_{imp} + w_T \cdot ST\text{-}T + \gamma_{ST} \tag{11.8}$$

where w_S, w_T are two parameters used to balance the relative contributions of the spatial and temporal scores with a bias term γ_{ST}. The selection and the influence of the related parameters will be given in section 11.4.5.

11.4.5 Experiment Results of the Proposed ST-VQM

The FVV dataset described in section 11.3 is adopted for the evaluation of ST-VQM's performance. For comparison, only image/video measures designed for quality evaluation of view-synthesis artifacts are tested since commonly used metrics fail to quantify geometric distortions as already reported in section 3.2. To compare the performance of the proposed metric with other metrics, the standard criteria including PCC, SCC, and RMSE between the subjective scores and the objective ones are used after applying a non-linear mapping over the measures [169]. For image quality metrics, their corresponding spatial objective scores are first calculated frame-wise, and the final object score is computed by averaging the spatial scores.

Table 11.5 – Performance comparison of the proposed metric with metrics designed for synthesized views in FTV scenario

	PCC	**SCC**	**RMSE**
Image Quality Metrics			
3DSwIM [33]	0.5230	0.5649	0.8640
MW-PSNR [144]	0.5705	0.8192	0.8304
MW-PSNRr [76]	0.5779	0.8295	0.8252
MP-PSNR [143]	0.5706	0.8299	0.8304
MP-PSNRr [76]	0.5603	0.8319	0.8377
ST-IQM$_{imp}$	0.8805	0.8511	0.4793
Video Quality Metrics			
Liu-VQM [60]	0.9286	0.9288	0.3753
ST-T	0.8336	0.8926	0.4837
ST-VQM	**0.9509**	**0.9420**	**0.3131**

The overall results are summarized in Table 11.5 and the best performance values are marked in bold. As it can be observed from Table 11.5, ST-VQM, Liu-VQM are the two best-performing metrics, with PCC equals to 0.9509, 0.9286 correspondingly. To check whether the differences between those values are significant, a t-test is carried out taking the difference of the predicted score between DMOS and Liu-VQM, and the one between DMOS and ST-VQM as inputs. The results show that our proposed metric significantly outperforms the second best performing Liu-VQM. As it can be observed from Table 11.5, the performances of the image metrics, including MW-PSNR and MP-PSNR, are very limited. These results could be explained by: (1) they over-penalize the consistent shifting artifacts, and (2) these measures do not take into account the temporal distortions.

Table 11.6 – Performance comparison of metrics for distinguishing sequence in different trajectories

	AUC$_{DS}$	**AUC$_{BW}$**	**CC**
Image Quality Metrics			
3DSwIM [33]	0.4603	0.8311	**0.8667**
MW-PSNR [74]	0.5571	0.6889	0.6000
MW-PSNRr [76]	0.5317	0.6933	0.6667
MP-PSNR [75]	0.5079	0.7022	0.6667
MP-PSNRr [76]	0.5238	0.6933	0.6667
ST-IQM$_{imp}$	0.5016	0.7244	0.6000
Video Quality Metrics			
Liu-VQM [60]	0.6270	0.8311	0.7333
ST-T	0.5857	0.8800	0.8000
ST-VQM	**0.6762**	**0.8933**	**0.8667**

As it has been verified in the subjective experimental results in the previous section, navigation scan-paths affect the perceived quality. Therefore, it is important for an objective metric to point out whether the perceived quality using a given trajectory is better than using other trajectories. As thus, the metric can be used to evaluate the limit of the system in worse navigation situations. To this end, the Krasula performance criteria [110,111] is used to assess the ability of objective quality metrics to estimate whether one trajectory is better than another with the same rate-point and baseline configurations regarding perceived quality. Pairs of sequences generated with the same configurations but in form of T_1 and T_2 in the dataset are selected to calculate the area under the ROC curve of the 'Better vs. Worse' categories (AUC$_{BW}$), area under the ROC curve of the 'Different vs. Similar' category (AUC$_{DS}$), and percentage of correct classification (CC) (see [110,111] for more details). More specifically, since pairs are collected in the form of (T_1, T_2) with other parameters fixed, if one metric obtain higher AUC$_{BW}$, it shows more capability to indicate that sequences with certain trajectory are better/worse than sequences with another trajectory. Similarly, if the metric obtain higher AUC$_{DS}$, then it can better tell whether the quality of one sequence in the form of one trajectory is different/similar to the one in the form of another trajectory. Results are reported in Table 11.6. As it can be observed, the proposed metric obtains the best performance regarding the three evaluation measures. It is proven that the proposed ST-VQM is capable of quantifying temporal artifacts introduced by views switch. More importantly, ST-VQM is the most promising metric in decidng sequence generated in the form of which trajectory is of better quality than another.

11.4.6 Selection of Parameters

The performance of a reliable VQM should not vary significantly with a slight change of the parameters. In this section, an analysis of the selection of the parameters of the proposed metric is presented. In order to properly select w_S, w_T and γ_{ST} in equation (11.8), as well as to check the performance dependency of the parameters, a 1000 times cross-validation is conducted. More specifically, the entire dataset is separated into a training set (80%) and testing set (20%) 1000 times, and the most frequently occurred value will be selected for the corresponding parameter. Before the validation test, we first multiply *ST-IQM* by 10^{10} and *ST-T* by 10^5 so that the difference between the corresponding parameter w_S, w_T will be smaller making it easier for later visualization (it has to be pointed out that this operation does not change the performance). The values of the three parameters with the corresponding PCC values across of 1000 times cross-validation are shown in Figure 11.10 (d). It can be observed that both the values of the three parameters and the performance do not change significantly throughout 1000 times, which verifies the fact that the performance of the metric does not

change dramatically along with the modification of the parameters. Figure 11.10 (a)-(b) depicts the histograms of frequencies of the three parameters' values respectively. As it can be observed that $w_S = 0.28, w_T = -0.43$ and $\gamma_{ST} = 3.26$ are the three most frequent value among 1000 times. They are thus selected and fixed for reporting the final performance in Table 11.5 and 11.6. The mean value of PCC, SROCC, and RMSE of the proposed metric across the 1000 times is 0.9513, 0.9264 and 0.2895 correspondingly, which are close to the performance values reported in Table 11.5 with the selected configuration.

Subsequently, the performance dependency of the proposed algorithm on the exponent variable β_{ST} in equation (11.6) and the distance approaches has been reported and examined in [177]. Therefore, in this section, the same $\beta_{ST} = 4$ and the Jensen Shannon divergence are selected.

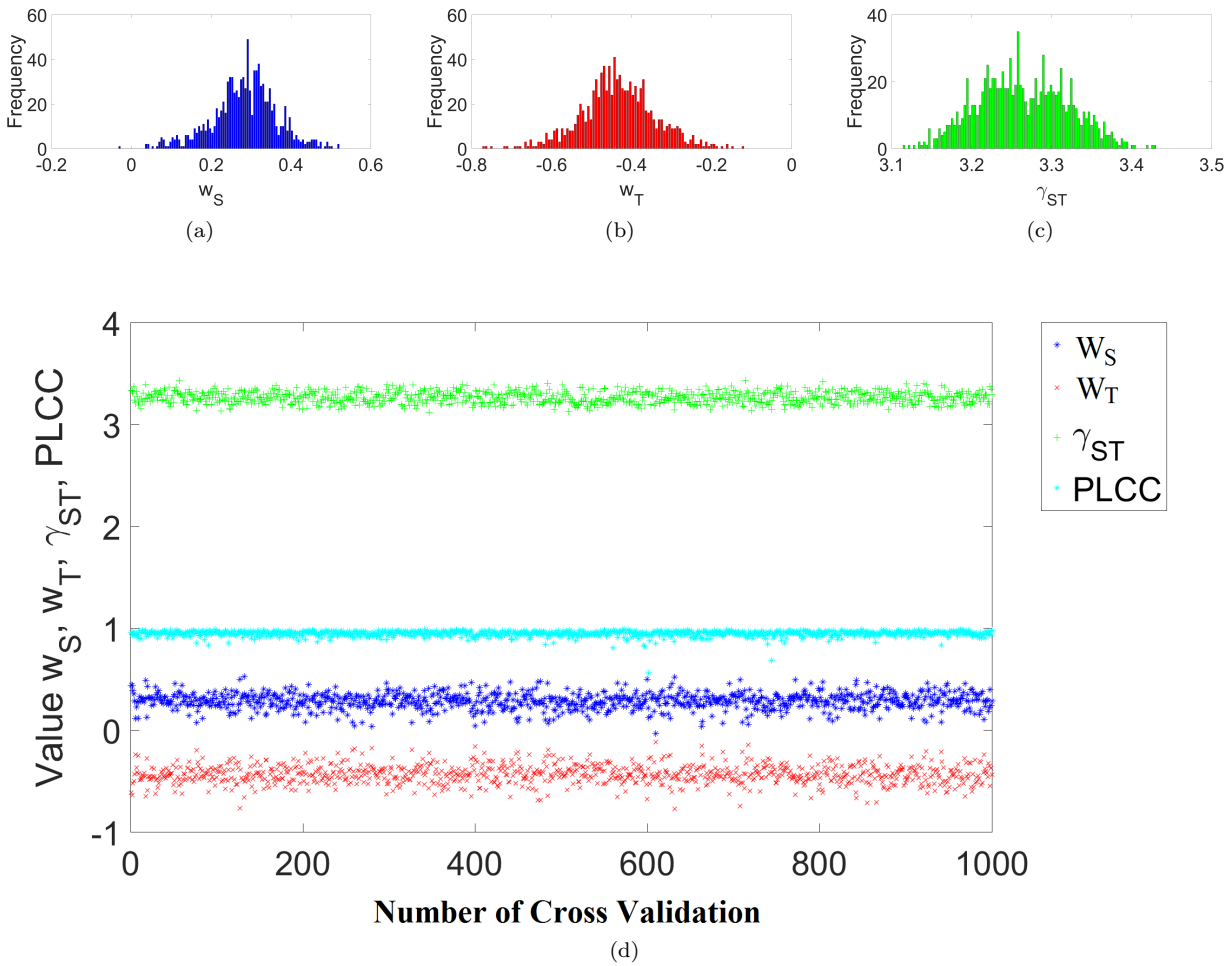


Figure 11.10 – (d) Values of w_S, w_T, γ_{ST} and their corresponding PCC values across 1000 times cross validation.

11.4.7 Execution Time

Moreover, to check the time complexity of the proposed ST-VQM compared to the second best performing video quality metric Liu-VQM, execution time of the metrics normalized by PSNR as introduced in section 4.4 are listed in Table 11.7. It can be observed from the table that the proposed ST-VQM is slower than Liu-VQM. However, since the gain of the ST-VQM compared to the second best performing metric is significant (according to t-test result: p-value=10^{-4}), it is acceptable if the running time is longer.

Table 11.7 – Execution time of proposed metric compared to state-of-the-art metric

Metric	Liu-VQM	ST-VQM
Normalized time	**20K+**	62K+

11.5 Conclusion

11.5.1 Conclusion of Subjective Study on Navigation Trajectories' Impact on Perceived Quality

Compression and views synthesis are the two main sources of degradations in the FVV scenario. Therefore, in the subjective study of this chapter, different configurations of the state-of-the-art software for views compression and view-synthesis have been considered. In addition, following the approach of using simulating navigation trajectories of immersive media that the users may employ to explore the content, two different trajectories (referred as hypothetical rendering trajectories) have been used to study their impacts on the perceived quality. Knowing these possible effects may help on the identification of critical trajectories that may be more suitable to carry out quality evaluation studies related to the benchmark of systems in the worst cases. Also, it must be pointed out that the sweeps that generated in this test focus more on views that contain regions of interest (e.g., moving objects) in videos since human observers are more interested in them and even stop navigating after these regions show up. By analyzing the subjective results, we find that the way of how the trajectories are generated affects the perceived quality. In addition, the dataset generated for the subjective tests (FVV), along with the obtained subjective scores, is made available for the research community of the field.

11.5.2 Conclusion of ST-IQM and ST-VQM

ST-IQM At the beginning of this chapter, a sketch-token based image quality assessment metric has been proposed by quantifying the change of contours' classes. Among the compared metrics including MW-PSNR and MP-PSNR, the proposed ST-IQM metric performs the best and shows great improvement. Visualized results have also showcased the capacity of the proposed metric to detect as well as to quantify the amount of structure related artifacts generated during DIBR process around disoccluded regions.

ST-VQM Aiming at better quantifying the structure related distortions in sequences generated in FVV systems, a sketch-token based video quality metric is proposed by checking how the classes of contours change between the reference and the degraded sequences spatially and temporally. The results of the experiments conducted on the FVV dataset has shown that the performance of proposed ST-VQM is promising. More importantly, ST-VQM is the best performing metric in predicting if sequences that are in the form of a given trajectory are of higher/lower quality than sequences that are in the form of other trajectories, with respect to subjective scores.

12

Encoding Structure Information with Context Tree Encoder

12.1 Introduction

In this chapter, the second mid-level representation based model is presented. As mentioned in section 10, there is an 'entropy encoding' like mechanism in human brain to proceed low-level features into mid-level representations. Hence, entropy-based contour encoder is explored in this chapter for quality assessment of synthesized views in FTV scenario in this chapter.

Recently, Zheng *et al.* [178] proposed to use geometric prior with context tree (CT) for contour coding. In their proposed scheme, object contours composed of contiguous between-pixel edges were first converted into sequences of directional symbols by using differential chain code (DCC) [179]. For example, in Figure 12.1, a contour that has N_{ep} contiguous edge pixels can be represented with a symbol string as $c = [x_0, ..., x_{N_{ep}}]$. For the starting point of the contours x_0, the possibility of four directions, including north, east, south, and west were assumed to be equal. These four directions are shown in Figure 12.1(f). For any other subsequent DCC symbol $x_i, i > 1$, only three relative directions are possible with respect to the previous symbol x_{i-1}: left (l), straight (s) and right (r), as shown in Figure 12.1(e). With one generated training DCC set (e.g., all the contours represented with DCC), an optimal variable-length context tree (CT) [180, 181] T_{CT} is computed and all the symbols of contours could be then encoded with arithmetic coding [182] using the trained context tree. To obtain the context tree, a maximum a posterior (MAP) formulation for estimating symbol conditional probability was designed in [178]. Since the encoding scheme is designed based on contours, it can be utilized and served as a tool for quantifying perceptual annoyance of structure-related distortions reflecting in quality score. For example, suppose there are a great amount of geometric distortions within one synthesized view. When using this model to encode the contours, the bits assignment are supposed to be different from their reference view since the sequences of symbols (contours) are different. Ideally, the more structure-related

distortions/loss (i.e., contour changes) there are, the larger the gap there is between the encoding cost of the synthesized and the reference view. Based on the discussion above, in this part of the thesis, we investigate the relationship between the dissimilarity in encoding cost and the perceptual annoyance. By hypothesizing that the dissimilarity in encoding cost computed using the context tree based contour coding model is relevant to perceptual quality score, we propose a context tree based image quality assessment metric (CT-IQM) in the context of free-viewpoint TV.

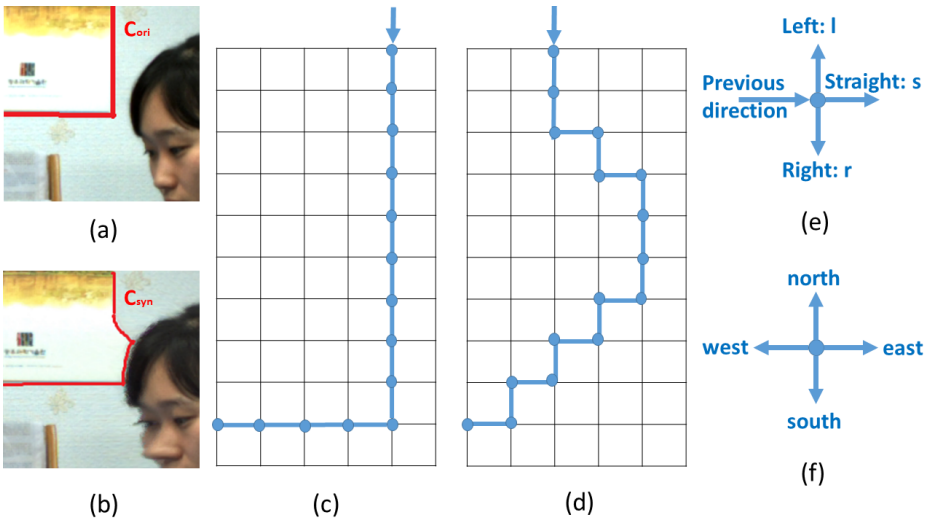


Figure 12.1 – (a) A patch from one reference image. (b) A patch from one synthesized image. (c) Contour in reference patch represented by four-connected chain code. (d) Contour in synthesized patch represented by four-connected chain code. (e) Direction code for the non-starting point (f) Direction code for the starting point.

In general, the strength of the proposed metric is three folds:

(1) It is capable of quantifying overall structure/contour loss (due to geometric distortions) by firstly representing contours with differential chain code (DCC) and secondly calculating the difference between encoding cost of contours in synthesized and reference images with the learned CT: For example, in Figure 12.1, (c) is the rough DCC representation (actual chain is much longer since each node in the subfigure corresponds to one pixel in the image) of the contour C_{ori} in an original image patch as shown in (a), while (d) is the one of the corresponding contour C_{syn} in a synthesized image patch as shown in (b). It can be observed that the chain code of C_{ori} : $south - s - s - l - r - l - r - s - s - r - l - r - l - r - l - r$ is totally different from the one of C_{syn} : $south - s - s - s - s - s - s - s - s - s - r - s - s - s$. By using CT tree, the change of contours, as well as structure, could be quantified indirectly. Further introduction of DCC and CT is given in section 12.2.

(2) It is robust to consistent global shifting: For example, suppose C_{ori} in Figure 12.1 (a) has shifted slightly to one direction, its corresponding chain code remain unchanged as $'south - s - s - l - r - l - r - s - s - r - l - r - l - r - l - r'$. In our proposed scheme, shifting artifacts are captured but not be over-penalized by dissimilarity in contour characteristics (i.e., dissimilarity in the coordinates of contours' starting points).

(3) Deformations of straight contours are penalized sufficiently. In [178], to avoid overfitting due to the lack of enough training data, a geometric prior was proposed and formulated as straightness of all the contexts in one learned context tree. For instance, for one straight contour in one reference view, the curvier the corresponding contours in the synthesized view, the more different the encoding cost will be. Since deformations of straight contours are more severe for human observers, the metric can be benefited by the design of the context tree.

12.2 Context Tree based Image Quality Assessment Metric (CT-IQM)

The overall framwork of the proposed scheme is illustrated in Figure 12.2. First of all, edges in reference and synthesized views E_{ref}, E_{syn} are detected with gradient based approach proposed in [183]. Afterwards, as showed in Figure 12.2, the following process is composed of two parts (bounded by blue and red dashed box correspondingly):

(1) Overall structure/contour dissimilarity is computed based on variable-length context tree: The detected edge/contour sets E_{ref}, E_{syn} are first converted into differential chain code (DCC) and represented by sets of symbol strings X_{ref}, X_{syn}. For each set of synthesized images whose reference is X_{ref}, one optimal context tree T^*_{CT} is learned with X_{ref}. After getting T^*_{CT}, the encoding cost of the target synthesized view EC_{syn} as well as the one of its reference EC_{ref} are calculted using the scheme proposed in [178]. Then the structure dissimilarity D_{sl} between the original and the synthesized view is obtained by subtracting the normalized EC_{ref} from the normalized EC_{syn}.

(2) Overall dissimilarity in contour characteristics is computed by considering contour statistics: After edge detection and DCC conversion, contour characteristics including coordinates of contours' starting points, the total number of contours and the total number of symbols from both original and synthesized views are calculated. The dissimilarities between these statistical information D_{cs} are then computed.

Finally, the proposed context tree based image quality assessment metric (CT-IQM) is designed by combining D_{sl} and D_{cs}. More details are given in the following subsections.

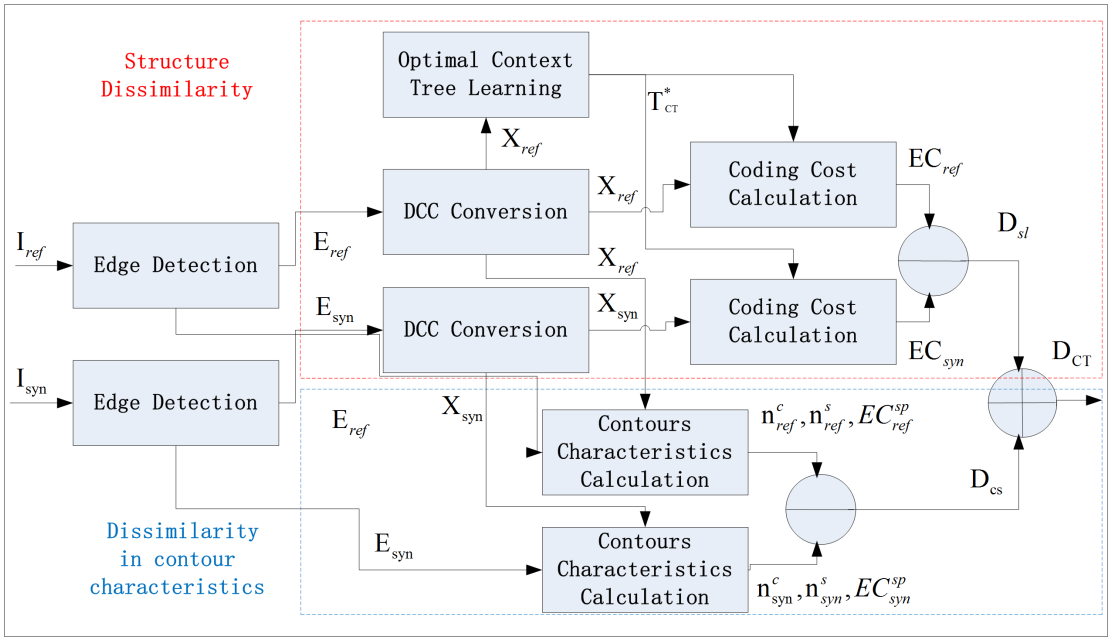


Figure 12.2 – Overview framework of the proposed metric

12.2.1 Context Tree based Overall Structure Dissimilarity

The overall scheme of the context tree based structure dissimilarity calculation can be concluded as below. In order to quantify the overall difference in contour structure, the dissimilarity values of encoding cost between the reference and synthesized views are calculated referring to the context tree based encoding scheme proposed

in [178]. Different from [178], for each set of synthesized images created with the same source, one independent optimal context tree is trained using only the reference view (e.g., for synthesized images $I_{A_1}, ..., I_{A_7}$, whose correspondence image is I_{ref}, the optimal context tree is trained using all the contours detected from I_{ref}) since the purpose is to check how the structure of the image changes after synthesis. With the trained tree obtained using contours in I_{ref}, one can then calculate the encoding cost of the synthesized images I_{syn} as well as the one of I_{ref}. In the beginning, each detected contour in the reference image and the target synthesized image is converted into a differential chain code (DCC) [179](each differential chain code is a string of symbols). Here, each reference image is considered as the training image for its corresponding synthesized images. Each symbol in the string is chosen from a size-three alphabet (left, straight, right) as described in [178]. For each reference image, an optimal context tree T^*_{CT} is then constructed with the DCC strings in this original image via solving a maximum a posterior (MAP) problem referring to [178]. With the context tree, the conditional probability distribution of each symbol in an input DCC string (corresponding to a contour in the target image) can be identified. Afterward, the coding costs of an image could be obtained by computing the sum of all the encoding cost of all the symbols within it. The encoding cost of symbols are computed using the arithmetic coding with the context trees T^*_{CT} by taking each DCC as input. Then, the global structure difference between a synthesized image and its reference is calculated by subtracting the encoding cost of the synthesized image from the one of the reference. Details of related definitions and the process of training a context tree are given below.

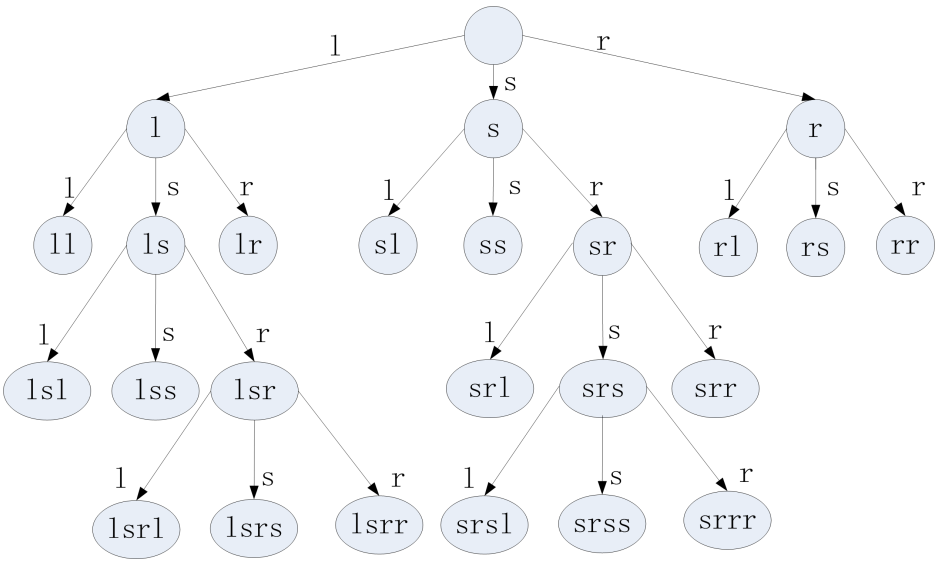


Figure 12.3 – An example of a context tree where each node is a sub-string and the root node is an empty substring. The contexts are all the end nodes of the tree: $T_{CT} = \{ll, lsl, lss, lsrl, lsrs, lsrr, lr, sl, ss, srl, srsl, srsl, srss, srrr\ , srr, rl, rs, rr\}$

Suppose there are M DCC strings (contours) in a given reference image I_{ref}, then all these strings constitute a training set $X = \{x(1), .., x(m), .., x(M)\}$. Each string is composed of a series of symbols and the length of $x(m)$ is denoted as l_m. The total number of symbols in each I_{ref} is equal to $L = \sum_{m=1}^{M} l_m$. Given a sub-string $x_a^b = [x_b, x_{b-1}, ..., x_a]$, where $a < b$ and $a, b \in \mathbb{Z}^+$, the number of occurrences of sub-string in the training set is defined in [178] as:

$$N(x_a^b) = \sum_{m=1}^{M} \sum_{i=1}^{l_m - |x_a^b| + 1} \mathbf{1}\left(x(m)_i^{i+|x_a^b|-1} = x_a^b\right), \tag{12.1}$$

where $|x_a^b| = b - a + 1$ is the length of the sub-string, and $\mathbf{1}(e)$ is an indicator function that equals to 1 if the expression e returns 'true' and 0 otherwise. Given I_{ref} and its corresponding strings set X_{ref}, $P(x|x_a^b)$ can be estimated by,

$$P(x|x_a^b) = \frac{N(< x, x_a^b >)}{N(x_a^b)}, \tag{12.2}$$

where $N(x_a^b)$ denotes the occurrence of substring x_a^b in X_{ref} and $< x, x_a^b >$ denotes the concatenation of sub-string x and x_a^b. In order to calculate the conditional probability $P(x_i \mid x_1^{i-1})$ of any symbol x_i given with all its previous symbols x_1^{i-1}, a model needed to be trained to determine a context w. In [178] , $w = x_{i-l}^{i-1}$ was defined as an l long prefix of the sub-string of x_1^{i-1} so that

$$P(x_i \mid x_1^{i-1}) = P(x_i \mid w), \tag{12.3}$$

where $P(x_i \mid w)$ is calculated with (12.2). All the possible mappings from x_1^{i-1} to w can be then represented as a context tree T_{CT}. This context tree T_{CT} is a full ternary tree where each node of the tree has either zero children or all three children. The sub-strings of the end nodes are the contexts of the tree. For each x_1^{i-1}, a context w can be obtained by traversing T_{CT} from the root node until an end node to get a matching path with a series of symbols $x_{i-1}, x_{i-2}...x_2$. Figure 12.3 is an example explaining a context tree T_{CT}. For example, context 'lsrl' can be obtained by traversing from the root node to the leaf node 'lsrl' by passing symbols 'l','s','r' and 'l' one by one.

With a set of all the DCC strings X_{ref} in the reference image I_{ref}, the purpose is to select the optimal tree T from all the possible context trees for later evaluation of structure/contour loss between synthesized images and their corresponding source. First of all, the posterior probability of T_{CT} given X_{ref} is given via Bayes' rule:

$$P(T_{CT}|X_{ref}) = \frac{P(X_{ref}|T_{CT})P(T_{CT})}{P(X_{ref})}, \tag{12.4}$$

where $P(T_{CT})$ is the prior that describes a priori knowledge about the context tree. The likelihood term is defined as the joint conditional probability of all the symbols in X_{ref} with their past and T. It is reformulated as (12.5) since the prefix w of the past symbols $x(m)_1^{i-1}$ of each symbol $x(m)_i$ can be settled to calculate the conditional probability:

$$P(X \mid T_{CT}) = \prod_{w \in T_{CT}} \prod_{x \in D} P(x \mid w)^{N(<x,w>)} \tag{12.5}$$

In [178], to avoid over-fitting, a geometric prior is proposed and defined as the sum of straightness $s(w)$ of all the contexts w in one tree T_{CT} based on the assumption that contours in natural image are more likely to be straight than curvy.

The context w to a shape segment on a 2D grid with $|w| + 2$ points is mapped from the most recent symbol $w_{|w|}$ to the symbol w_1 furthest in the past plus an initial edge. Afterwards, the straightness of w is computed as the maximum distance from any point $p_k, 1 \leqq k \leqq |w| + 2$ to a straight line $f(p_1, p_{|w|+2})$ that connect point p_1 and $p_{|w|+2}$:

$$s(w) = maxdist(p_k, f(p_1, p_{|w|+2})). \tag{12.6}$$

Then, the prior term $P(T_{CT})$ in (12.3) is defined based on the sum of the straightness $s(w)$ of all contexts w in

the context tree as below:

$$P(T) = exp\left(-\theta_{CT} \sum_{w \in T_{CT}} s(w)\right), \tag{12.7}$$

where θ_{CT} is a constant defined in [178]. Finally, the optimal context tree problem can be solved by MAP estimation formulated as:

$$T^*_{CT} = \underset{T_{CT} \in F_{CT}}{\arg\min}\ P(X_{ref}|T_{CT}), \tag{12.8}$$

where F_{CT} is a context forest constituted of all possible context trees.

During the optimization procedure, an initial tree T^0_{CT} is first constructed with the reference image. Then T^0_{CT} is pruned to obtain the optimal context tree T^*_{CT} as described in [178].

After obtaining the optimal context tree T^*_{CT} for each reference image, one can calculate the structure/contour loss between one synthesized image and its reference image by calculating the difference in coding cost using adaptive arithmetic coding with T^*_{CT}. For each symbol x_i in the contours represented by DCC strings of the target synthesized image, the matched context $w = x^{i-1}_{i-|w|}$ of x_i is found first. With w, the corresponding conditional probability distribution $P(x_i|w)$ is then computed with T^*_{CT}. Then, the distribution is inputted into an adaptive arithmetic coder to encode x_i. The encoding cost of all the symbols in the synthesized image EC^{ct}_{syn} can be then acquired and used to compare to the one of the reference. To get the final structure loss, the context tree based encoding cost of both the synthesized and reference images are normalized by the corresponding number of symbols. The context tree based overall structure/contour dissimilarity is defined as

$$D_{sl} = |\frac{EC_{ref}}{|X_{ref}|} - \frac{EC_{syn}}{|X_{syn}|}|, \tag{12.9}$$

where $|X_{ref}|$ and $|X_{syn}|$ is the total number of symbols in the reference and synthesized views correspondingly. It has to be pointed out that the value of D_{sl} between a synthesized view with a considerable amount of disturbing geometric distortions and its reference could be large. Due to geometric distortion, the matched context w_{syn} of one symbol x_{syn} is different from the one of the correspondence symbol x_{ori} in the reference image. Therefore, the encoding cost is different. From an overall point of view, the more geometric distortions there are, the larger the difference between the encoding cost of the reference and the synthesized image.

12.2.2 Overall Dissimilarity in Contour Characteristics

As discussed in the previous section, the starting point of a contour is represented by one of the four symbols including 'north', 'east', 'south' and 'west' along with its 2D coordinate. This information reveals the spatial distribution of contours in an image and thus is important structure information of the image. By checking the dissimilarity in contours characteristics between the reference and synthesized images, one can get the structural gap between them. Therefore, for both of the reference and synthesized views, we accumulate the total contour number n^c_{ref}, n^c_{syn} , the total amount of contours' starting point information and the total number of symbols n^s_{ref}, n^s_{syn} correspondingly.

For the measurement of difference in contours' starting point information, one cannot compare starting points in one synthesized image to the ones in its reference directly, due to synthesized artifacts like object shifting. Coordinate comparison without contour matching is impracticable in this case. In order to quantify the amount of dissimilarity considering starting points of contours in synthesized and reference views, we refer

to [178] and calculate the encoding cost of all the starting points' coordinates. Considering the fact that the difference among neighboring points in coordinates is small, the mixed-Golomb (M-Golomb) [184] algorithm is used. In [178], all the starting points in one image are first sorted in ascending order according to one chosen coordinate component. The difference of one coordinate of the neighboring points is coded with Golomb coding while the other components are coded with fixed length binary coding. To distinguish the encoding cost of starting points of the contours and the encoding cost of the DCC strings in images, here the encoding cost of contour starting points in reference and synthesized images are denoted as EC^{sp}_{ref} and EC^{sp}_{syn}, respectively.

we denote the encoding cost of contour starting points in reference and synthesized images with EC^{sp}_{ref} and EC^{sp}_{syn}. Then, the overall dissimilarity in contour characteristics between one synthesized image and its reference D_{cs} can be written as :

$$D_{cs} = |sum(n^c_{ref}, n^s_{ref}, EC^{sp}_{ref}) - sum(n^c_{syn}, n^s_{syn}, EC^{sp}_{syn})|. \tag{12.10}$$

12.2.3 The Final Proposed Metric

Finally, the context tree based structure loss metric for one synthesized image with its corresponding reference image is formulated by leveraging the aforementioned context tree based overall structure dissimilarity value and the overall dissimilarity value in contour characteristics as defined below:

$$D_{CT} = \alpha_{CT} \cdot D_{ls} + \beta_{CT} \cdot D_{cs}, \tag{12.11}$$

where α_{CT}, β_{CT} are two weights for the two corresponding dissimilarity values and $\alpha_{CT} + \beta_{CT} = 1$. The setting of the two parameters is further discussed in the next section.

12.3 Experimental Results

The performance of context tree based image quality assessment metric is evaluated on the IVC-Image dataset described in section 4.2.1.1 and compared to the best performing full reference image quality metrics developed for synthesized images summarized in section 3.4.

To compare the performances existing metrics designed for synthesized images, the widely employed criteria PCC, SCC, and RMSE described in section 4.3 are utilized with a non-linear mapping between the subjective scores and objective.

Table 12.1 – Performance comparison of the proposed metric with existing metrics designed for synthesized views

	PCC	**SCC**	**RMSE**
MP-PSNR$_f$ [75]	0.6553	0.6239	0.5029
MP-PSNR$_r$ [76]	0.6733	0.66	0.4923
MW-PSNR$_f$ [74]	0.6089	0.5738	0.4948
MW-PSNR$_r$ [76]	0.6444	0.6218	0.5091
CT-IQM	**0.6809**	**0.6312**	**0.4877**

The result is concluded in Table 12.1. As it can be observed from Table 12.1, the proposed CT-IQM achieves 0.6809 and 0.4877 value of PCC and RMSE correspondingly, which slightly outperforms all of the compared metrics designed for synthesis images. In this experiment, for the proposed CT-IQM, a configuration

of $\alpha_{CT} = 0.9, \beta_{CT} = 0.1$ yields the highest correlation with subjective utility scores. The selected configuration shows that the first part (encoding cost based structure/contour dissimilarity) of the metric plays a more significant role in predicting the perceived quality. This configuration also proves the hypothesis that there is a relationship between the encoding cost computed using the context tree based contour coding model and the perceived quality.

Finally, to check the efficiency of the proposed CT-IQM, the execution time of the metrics normalized by PSNR as introduced in section 4.4 are listed in Table 12.2. According to the table, CT-IQM is the most complex metric.

Table 12.2 – Normalized execution time of proposed metric compare to the state-of-the-art metrics.

Metric	MW-PSNR	MW-PSNR$_r$	MP-PSNR	MP-PSNR$_r$	CT-IQM
Normalized time	12.4	**9.6**	100	35	458

12.4 Conclusion

In this chapter, the relation between encoding cost (calculated by a context tree based contour coding scheme) and perceived quality is investigated. Based on an assumption of such relation, a variable-length context tree based image quality assessment metric (CT-IQM) is proposed to measure how the structure change due to synthesized artifacts. CT-IQM is consist of two main parts. The first part is the dissimilarity in encoding cost between the original and synthesized views using the context tree based model. The second part is the dissimilarities in various contour characteristics. The experimental results have confirmed our hypothesis. However, compared to the performance of ST-IQM, its performance is less desirable, and it is more time-consuming. Therefore, CT-IQM is neither further tested in other applications nor extended for video quality assessment.

13

Conclusions of Part 3

In this part, mid-level representations have been explored for images/videos quality assessment. Two mid-level based models have been proposed, and certain research questions that posed in section 10.2 are answered.

13.1 Answers to Research Questions

- Mid-level representations of image/video for quality assessment in different tasks (Part III)
  - A sketch token based image quality assessment metric (ST-IQM) is proposed for quantifying the structure related distortions.

    Borrowing the concept of 'encoding' low-level feature into the mid-level representation to carry more semantic information in the human visual system, ST-IQM is proposed in section 11.2 to quantify geometric distortions by firstly encoding contours into 'patterns' of contours' categories and secondly checking how the categories of the contours change after degradations. It outperforms all the compared full reference image quality metrics that designed for synthesized views.
  - The impact of navigation scan-paths on perceived quality is verified.

    As there is no existing study that has a deep analysis on how human observer navigates among viewpoints affect on perceived quality in the context of FTV, especially in the case when observers focus on 'regions of interest'. To this end, a subjective experiment is conducted and introduced in section 11.3. Two trajectories are considered in the experiment, including one tracing the main moving object (content related) and the other just repeatably navigating from left to right then back forward. Subjective results indicate that there are trajectory related impacts on perceived quality. Moreover, a video database named as free viewpoint videos database (FVV) is released along with subjective scores. Different from most of the existing datasets, this dataset contains compression, synthesis and views transition artifacts.
  - A sketch token based video quality assessment metric (ST-VQM) is proposed for quality evaluation of videos where spatial-temporal structure distortions exist.

To quantify the structure related temporal artifacts within one viewpoint and among viewpoints based on mid-level contours representation, ST-IQM is extended and improved by incorporating another ST based temporal estimator to quantify temporal structure inconsistency. In the framework, sensitive regions selection process introduced in section 8.2 is utilized to improve the metric in terms of complexity. The extended ST based video quality assessment metric achieves the best performance on FVV database compared to other existing metrics designed for synthesized views as reported in section 11.4. The feasibility of extending a mid-level representation based image metric into a video one is verified.

- A context tree based image quality metric (CT-IQM) is introduced by encoding contours in images using pre-trained context tree.

  Borrowing the concept of encoding more frequently appear 'item' with more compact 'code' strategy in the human visual system (tested in psychological experiments), the context tree based contour encoder is used in chapter 12 to quantify structure loss regarding encoded contours dissimilarity. Although the gain of performance is not significant, the proposed CT-IQA still outperform most of the existing metrics designed for synthesized images. The feasibility of using contour encoder for quantifying of structure disruption is confirmed.

13.2 Summary of performance and discussion

The performance and executing time of the proposed mid-level representation based models on all the tested datasets are summarized in Table 13.1 and 13.2.

Both ST-IQM and CT-IQM have been examined on the IVC-Image dataset. According to Table 13.1 and the t-test results, ST-IQM significantly outperform CT-IQM. There are mainly three reasons for ST-IQM to obtain better performance: (1) Even though they are both mid-level representation based models using the concept of encoding low-level structure information, the representative (semantic) level of ST-IQM (i.e., contour categories) is higher than the one of CT-IQM (i.e., frequencies of particular type of contour appear in distorted image with respect to the ones in the reference). (2) Registration stage incorporated in ST-IQM helps to better avoid over penalizing global acceptable continuous distortions. (3) ST-IQM first obtains local dissimilarity value and then employs pooling strategy to yield the final quality score, while CT-IQM calculates one final score in one time. In such a way, ST-IQM is better in capturing local structure dissimilarity since it does not only deal with detected edges like CT-IQM.

The feasibility of predicting perceived quality by quantifying the change of contours' categories from a mid-level point of view is verified in the experiment conducted on the IVC-Image database. Therefore, ST-IQM is further extended as video quality metric ST-VQM and tested on the new presented FVV database by incorporating a sketch token based temporal estimator. Considering the fact that FVV database contains sequences with compression distortions, spatial synthesized related distortions, temporal synthesized related distortions, and temporal inconsistency of structure due to view switch, ST-VQM is only tested on this new database (most of the conditions considered in IVC-Video database and FVV database are considered in this one too).

By comparing the performance of mid-level representation based models with the ones of low-level representation based models on IVC-Image database (based on Table 13.1, 9.1, 9.2 and 13.2), it is found that ST-IQM

outperforms EM-IQM. Nevertheless, mid-level based model CT-IQM does not outperform EM-IQM. On one hand, it is because that models employ higher level representation does not necessarily ensure better performance. Only when the mid-level representation reaches a certain semantic level, e.g., categories of contours in ST-IQM, models proposed based on this representation can be benefited from it. On the other hand, it is also because that choosing a proper distance measure for quantifying the changes of structures, e.g., elastic metric in EM-IQM, is also important. By comparing the executing time of low-level and mid-level based representation models, it is found that mid-level representation based models are more complex. For CT-IQM, a context tree has to be learned first with the reference, which increases its complexity. For ST-IQM, the registration stage is the bottleneck of the model.

Table 13.1 – Summarization of performance of mid-level representation based models tested on different databases

PCC		Mid-level		
Related Task	Related Database	ST-IQM	ST-VQM	CT-IQM
Image quality assessment in FTV	IVC-Image	0.821		0.680
Video quality assessment in FTV	FNV		0.950	

Table 13.2 – Summarization of execution time of mid-level representation based models tested on different databases

Normalized time		Mid-level		
Related Task	Related Database	ST-IQM	ST-VQM	CT-IQM
Image quality assessment in FTV	IVC-Image	1324		458
Video quality assessment in FTV	FNV		62k+	

13.3 Summary

Generally speaking, the mid-level representation based models are of higher representative capability compared to the low-level ones. However, if the mid-level representation is not rich enough regarding representing more meaningful information, it can not reach better performance. Structure dissimilarity measure like elastic metric is useful for image/video quality assessment models in applications where geometric distortion is the dominant distortion. Last but not least, mid-level representation based models proposed in this part are still not linked to the quality directly. In another word, such learning process seems still not be sufficiently directly related to perceptual quality.

Exploring Higher-Level Representation based Models for Image/Video Quality Assessment

14

Introduction of Part 4

Higher-level representations of images/videos are defined as 'task-related abstraction', which learns a set of meaningful abstract patterns reflecting the characteristics of the task. Those representations are always related to the semantics of tasks. In this part, 'abstraction' is defined by the abstract 'distortions' learned according to the tasks. Based on this concept, two higher-level representation based models are proposed.

14.1 Higher-Level Sparse Representation in Human Visual system:

On one hand, with respect to high-level sparse coding in the human visual system as mentioned in section 1.2.1, human observers are capable of seeking semantic information from the scene with sparse representation, where each item in the representation represents certain meaningful 'characteristics' captured from the world. In inferotemporal cortex (IT), a wide range of studies support the notion that neurons are selective to high-level object dimensions, and to features such as faces and hands. It is expected for these neurons to show a high degree of sparseness for efficiency. Based on these perceptual study, it is reasonable to assume that how the observer detects structure related distortion follows such similar sparse strategy. Instead of getting semantic related sparse elements for sparse coding (e.g. 'face' or 'hand'), the human visual system may try to get quality related elements (e.g. 'ghosting edge' or 'shading edge') for judging the quality of one image. For example, it has been discussed in [185] that functions of different intermediate feature layers in deep convolutional network models capture different levels of information from the input. The latter the layer, the higher-level of semantic information can be extracted depending on the target task. Therefore, it is reasonable to use deep learning models for representing images/videos from a higher level.

Higher-Level Representations Related to Semantic Tasks On another hand, as introduced in section 1.2.1, high-level vision transfer input signals into categorical or semantic representations that enable later classification or identification. Nowadays, deep learning is commonly used for semantic tasks. More importantly, features extracted from intermediate layers of a deep neural network are commonly used as high-level features for

different tasks including sequence classification [186], speech emotion recognition [187], object detection [188], image quality assessment [71] and so on. Semantic information like the category of objects are the high-level representations. Therefore, models that designed to learn the semantic information are of potential to be used to obtain high-level representations of the input signals.

14.2 Research Questions Associated with Higher-Level Representation Models Development

According to the discussion above, in this part, we explore higher-level representations that learns a set of meaningful abstract patterns reflecting the characteristics of the task, as perceptual models. This investigation can be decomposed into more specific questions:

- ■ Higher-level representations of images/videos for quality assessment in different tasks (Part IV)
  - □ How to quantify local non-uniform distortion related to structure degradation using the concept of sparse coding in the human visual system?

    As discussed in section 1.2.1, a considerable amount of neurophysiological data from high-level visual cortex support Barlow's hypothesis that the neural codes are sparse and also those sparse elements of the codes stand for meaningful features of the world [28,29] (e.g., complex shapes, object-components). Therefore, it is worth trying to find a learning-based model to learn the effect of non-uniform structure related distortions on perceived quality with a meaningful sparse codebook containing 'understandable distortion' elements.
  - □ In the case where several different structure related distortions appear at the same location, how to learn this type of 'masking effect' of different distortions?

    As shown in section 2.3, images/videos in some applications, e.g., stitched panoramic images, may contain more than one type of structure related distortions. In another word, different types of structural distortions will overlap at the same locations and give rise to the 'masking effect' phenomenon.
  - □ In some cases, incorrect inpainting may cause annoying structure related distortions too. How to quantify local non-uniform distortion incurred by inpainting using advanced machine learning techniques?

    As shown in section 2.2, in case of FTV, inpainting related artifacts are another disturbing distortions that degrade structures of images/videos. As those local structure degradation are usually accompanied with blurriness, they may become the dominant artifacts that affect the perceived quality directly. Therefore, if there are machine learning based models that could be used to capture and quantify this type of artifacts, the quality of services where inpainting is required can be improved significantly.

15

From Natural Scenes Statistics to Non Natural Structure: Learning Structure-Related Distortions with Convolutional Sparse Coding

15.1 Introduction

In this chapter, the first higher-level representation based model is presented. As described in 3.3, the existing natural scene statistics (NSS) based models fail to predict the quality of image/video that contains non-natural structure (NNS). In this chapter, an advanced machine learning model, convolutional sparse coding, is adapted to learn the NNS within images and predict their perceived quality. More specifically, this model tries to train a sparse codebook where each item corresponds to one type of 'meaningful' geometric distorted element (e.g., twisted vertical curve) and quantify the local non-uniform distortions by checking the amount of activated non-natural elements within the test image with the codebook.

In the cognitive psychology domain, sparse coding is suggested to be an underlying strategy of the brain's neural system, for instance, examples of sparse coding in brain regions is given by Olshausen and Field [189]. Sparse coding is also widely used in the computer vision domain for semantically related tasks [190]. Further, in [191], the process of image quality assessment is also assumed to adhere to such a strategy. As mentioned in section 1.2.1, if each item within the sparse codebook is relevant to higher 'semantic' meaning, a higher-level representation could be obtained for a given image/video using the sparse codebook. In this thesis, higher 'semantic' meaning could be considered as 'distortion' category (e.g., a type of ghosting artifact), and sparse coding could be used to learn the non-natural structure.

Unlike conventional sparse coding, convolutional sparse coding (CSC), first introduced in [192], computes a sparse representation for an entire image with the sum of a set of convolutions with dictionary filters, instead of the linear combination of a set of dictionary atoms. Briefly, instead of independently computing sparse representations for a set of overlapping patches, CSC only computes one for the entire image. Thus, by using it for a no-reference model, it is possible to: 1) get one score from the model for one image instead of pooling local score calculated patch-wise; 2) better locate certain types of artifacts locally by checking the pixels that are activated by the corresponding learned convolutional filters; and 3) better determine whether the co-occurrences of certain types of artifacts will amplify or weaken the visual distortion effect. Therefore, CSC model is used for no reference image quality assessment of synthesized images in FTV and stitched images in VR scenarios in the following sections.

15.2 CSC based No Reference Metric for Synthesized Views

In this section, the details of the proposed convolutional sparse coding based image quality metric (CSC-IQM) is given and tested for quality assessment of synthesized views. The overall framework of the proposed scheme is illustrated in Fig. 15.1. First, a set of convolutional kernels D_{CSC} are learned using the improved fast CSC algorithms [193] with a training set I_{train} composed of a set manually selected patches, which contains obvious synthesized local distortion. Afterwards, CSC based feature vectors v_{csc} of images I in IVC-Image dataset [49] (i.e., testing database) are extracted. Finally, the objective score S_{CSC} is obtained with linear support vector Regression.

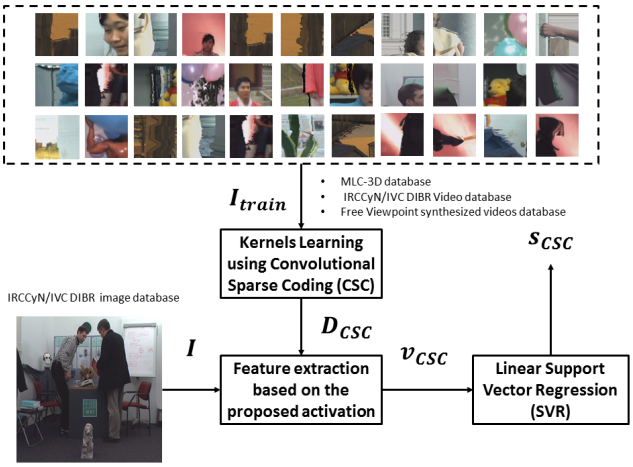


Figure 15.1 – Diagram of the proposed model

15.2.1 Mid-Level Features Extraction with CSC using the Proposed Activated Function

Convolutional sparse coding explicitly models local interactions through the convolution operator. In general, as introduced in [193], the CSC is commonly defined as in equation (15.1), since it has been proven that sparse

representation is recoverable using l_1 norm:

$$\begin{aligned} & min_{D,Z}\frac{1}{2}\|y-\sum_{k=1}^{K}D_k \circledast Z_k\|^2+\beta_{CSC}\sum_{k=1}^{K}\|Z_k\|_1 \\ & s.t.\|D_k\| \leqslant 1 \\ & , \end{aligned} \tag{15.1}$$

where $\circledast$ denotes the convolution operation, y denotes observed samples, Z_k represents sparse feature maps and D_k are the convolution kernels. K is the number of convolution kernels and β_{CSC} is a positive scalar used for balancing model accuracy and sparsity of feature maps, which can be tuned accordingly. By using CSC, there are two modes commonly involved:

1. The convolution kernels learning step by solving the optimization problem (15.1) with training data that can be written as:
$$min_{\{D_k\}}\frac{1}{2}\|y-\sum_{k=1}^{K}D_k z_k\|^2, s.t.\|D_k\|_2^2 \leqslant 1, \tag{15.2}$$
where z_k denotes the operators of convolution with feature maps Z_k;

2. The feature extraction step where the kernels are settled and features are extracted with minimization over feature maps:
$$min_z\frac{1}{2}\|y-d\cdot z\|^2+\beta_{CSC}\|z\|_1, \tag{15.3}$$
where $d=[d_1,...,d_k]$ represents the operator consists of convolutions with K kernels D_k, and $z=[Z_1^T,..,Z_K^T]$ is the vectorized feature maps vector.

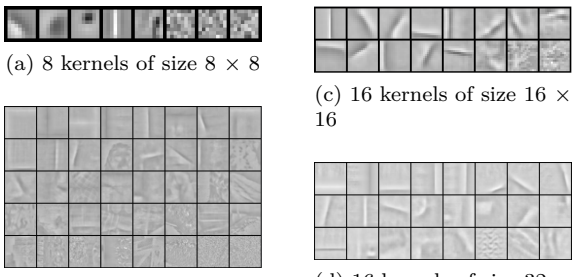
Considering the efficiency of the training phase, the state-of-the-art model proposed in [193] is adapted to learn the convolutional kernels for obtaining the mid-level structure descriptors for synthesized views. In order to speed up the inversion step in algorithms proposed in [192, 194], Sorel *et al.* propose to compute this most time-consuming step in [194] non-iteratively in the Fourier domain with the matrix inversion lemma. With a training set I_{train}, their improved approach is employed for convolution kernels learning to obtain a dictionary of convolutional kernels $D_{CSC}=D_k, k\in[1,...,K]$. In our experiment, three scales of kernels are chosen, which are filters with sizes of $8\times 8, 16\times 16$, and 32×32. The relative numbers of kernels for each scale are 8, 16, and 64, correspondingly. Therefore, the total number of kernels is $K=88$. Figure 15.2 shows the kernels learned in our experiment. For better comparison of more meaningful kernels to the others, we separate the 64 kernels of size 32×32 into sub-figures (d) and (d). The kernels in Figure 15.2 are of higher energy. One can easily notice that kernels in (d) are more structure-related while the ones in (b) are with more noise. Details of the images set I_{train} that we utilize are given in section 15.2.2. With the trained kernels D_{CSC}, for a $M\times N$ test image I, a set of feature maps Z can be obtained with the feature extraction process proposed in [193]. So that the image can be sparsely represented as $S=\sum_{k=1}D_k\circledast Z_k$. $Z=[Z_1;...;Z_k;...;Z_K]$ is a $M\times N\times K$ matrix of K feature maps, where each Z^k corresponds to one kernel. Afterward, a feature vector v_{csc} extracted based on CSC for an image I can be generated with

$$v_{csc}=(f_{act}(Z_1),...,f_{act}(Z_K)), \tag{15.4}$$

where activated function f_{act} is defined as

$$f_{act}(Z_k) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\mathbf{1}(Z_k(i,j) > \varepsilon_{CSC})}{M \cdot N}. \tag{15.5}$$

In equation (15.5), $\mathbf{1}(c)$ is an indicator function that equals to 1 if the specified binary clause c is true and 0 otherwise, and ε_{CSC} is a threshold for selecting activated pixels. Function $f_{act}(\cdot)$ accumulates the number of pixels which are above the threshold ε_{CSC} in each sparse feature map Z_k that corresponds to each kernel D_k. Intuitively, this function checks the number of the pixels that are activated by the corresponding kernel as an activated function. For example, let kernel D_k represent a certain type of synthesized artifact, if certain regions of one image contain the same artifact, then these regions in the corresponding feature map Z_k of the image will be activated. This activated function can be interpreted as the ratio of the area of non-uniform distorted regions versus the entire image.



(a) 8 kernels of size 8 × 8

(b) 48 kernels of size 32 × 32

(c) 16 kernels of size 16 × 16

(d) 16 kernels of size 32 × 32

Figure 15.2 – Kernels learned by the convolutional sparse coding [193] on three different scales. Kernels are sorted by energy of the corresponding feature map in a descending order from top-left to bottom-right.

15.2.2 Convolution Kernels Learning

During the training process, two major factors are considered here. First, in order to learn a set of convolution kernels that is capable of capturing local non-uniform distortions introduced by DIBR based algorithms, the training data set must contain typical local synthesized artifacts including geometric distortions. Second, to learn a general codebook that can represent general DIBR based relative artifacts, the training set should not be limited to one dataset that consists of images synthesized with limited DIBR algorithms. To this end, we manually select a total of 366 images that contain significant DIBR related synthesis artifacts from three datasets including the FFV [62] dataset described in 4.2.1.3, IVC-Videos [59] dataset described in section 4.2.1.2, and the MCL-3D [195] image database. For the previous two videos datasets, we extract one frame from each video. Especially, we extract different frames from the ones that were selected in the IVC-Image dataset, which is used for performance testing, to avoid overlap of training and testing sets. After labeling the locations of synthesis artifacts in the images, two professional observers were asked to pick up the most annoying patches centering at pre-labeled locations; only patches agreed by both of the observers are maintained for training. Examples of selected patches of size 128 × 128 are shown in the dashed bounding box in Figure 15.1. It can be observed that all the remaining patches contain obvious synthesized-related distortions.

Table 15.1 – Performance comparison of the proposed metric with existing metrics designed for synthesized views.

	PCC	**SCC**	**RMSE**
Full Reference Metric (FR)			
3DSwIM [33]	0.6864	0.4842	0.6125
MP-PSNR$_r$ [75]	0.6954	0.4784	0.6606
MW-PSNR$_r$ [74]	0.6637	0.4921	0.6293
CT-IQM (section 12.2)	0.6809	0.6626	0.4877
EM-IQM (section 8.2)	0.7430	0.6726	0.4455
ST-IQM (section 11.2)	**0.8217**	**0.7710**	**0.3929**
NO Reference Metric (NR)			
NIQSV [77]	0.6346	0.5146	0.6167
APT [4]	0.7307	0.7140	0.4622
CSC-IQM	**0.8302**	**0.7827**	**0.3233**

15.2.3 Prediction Module

After extracting the feature vector v_{csc} based on the proposed model, referring to [157], support vector machine regression is then used on v_{csc} with a linear kernel. During the training procedure, a 1000-fold cross-validation is applied. For each fold, the dataset is randomly split into 80% of the images for training and 20% for testing, with no overlap between them. After doing so, the image contents on which the model is tested are different from the ones on which the model is trained, to ensure the robustness of the trained models [155]. The median PCC , median SCC, and median RMSE between subjective and objective scores are reported across the 1000 runs for performance evaluation.

15.2.4 Experimental Results

The performance of the proposed CSC-IQM is evaluated on the IVC-Image dataset [49] as described in section 4.2.1.1. It is only compared to the full reference and no reference metrics designed for quality assessment of synthesized images summarized in section 3.4 and other image models proposed in this study.

The results of performance are concluded in Table 15.1. According to Table 15.1, ST-IQM performs the best among the full reference DIBR quality metrics, while the proposed CSC-IQM performs the best among the no reference ones in terms of PCC, SCC, and RMSE. Overall, CSC-NRM is the most consistent objective metric with the subjective scores and it even slightly outperforms the best performing full reference metric ST-IQM. In our experiment, due to the limitation of RAM, testing images are scaled with ratios for efficiency and the largest ratio that has been tested in our experiment is 0.8. We have also found that the performance of the proposed metric decreases with a reducing scaling ratio, which is reasonable since downscaling an image will inevitably degrade its quality. Therefore, CBC-IQM has the potential to achieve better performance with larger scaling ratio.

To better understand why and how the proposed model is capable of learning the effect of non-uniform artifacts on perceived quality in the case of synthesized views, the linear coefficients (the weights of dimensions in the feature vector that corresponds to the kernels) of the support vector regression model that yields the median PCC during cross-validation are visualized in Figure 15.3 and the largest 8 coefficients are represented as red bars. The larger the absolute value of the coefficient, the more important the role of the correspondence feature is in predicting the objective quality score, and thus the corresponding kernels are more important in capturing related local distortions. In the figure, the 8 largest weights are labeled with the corresponding visualized kernels bounded with red boxes while the smallest one is bounded with cyan color. By comparing

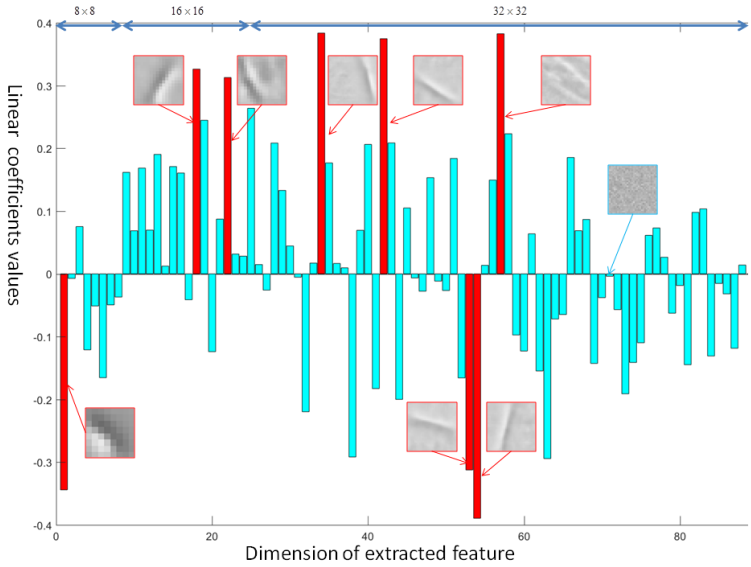


Figure 15.3 – Linear coefficients of learned SVR model.

them, one can find that most of the useful kernels contain structure information, specially double-edges-like structure, while the useless ones contain mainly noise. Furthermore, the color points labeled on the error map in Figure 15.4 are examples of the pixels activated by the top 3 filters selected according to the weights. Different colors corresponds to different kernels. It can be observed that those points are well distributed around the non-uniform error regions, meaning that the corresponding filters are capable of capturing these local geometric distortions. According to the analysis above, the feasibility of using CSC for revealing the impact of non-uniform structure-related distortions on the perceived quality, in the context of view synthesis, has been verified.

Figure 15.4 – Example of non-uniform distortion, Left: Reference, Middle: Synthesized view, Right: Error map

Last but not least, to meet the need of real-time computation for applications like multi-view live match broadcasting, the efficiency of the no reference quality assessment metric should be high enough. In order to check the efficiency of the proposed metric, the execution time of the metrics normalized by PSNR as introduced in section 4.4 are listed in Table 15.2. Here, only the no reference metrics are reported. According to the table, our proposed metric is slower than NIQSV. However, the gain of the performance of CSC-IQM is significant compared to NIQSV in PCC value (i.e., 30 % gains in PCC value, t-test result: P-value=10^{-4}), this advantage outweighs the disadvantage of its complexity.

Table 15.2 – Normalized execution time of proposed metric compare to the state-of-the-art metrics.

Metric	NIQSV	APT	CSC-IQM
Normalized time	**18**	13k+	985

15.3 CSC based No Reference Metric for Stitched Panoramic Image

In this section, the proposed CSC-based no-reference metric CSC-IQM is utilized to quantify ghosting and structure inconsistency artifacts that are specific to stitched panoramic images in the context of VR applications. Considering the masking effect, i.e., overlap of two types of artifacts that may amplify the annoyance level, a compound feature selection algorithm is further proposed and incorporated into the model. The contributions of this section compared to the previous one are twofold: 1) produce and release a training database labeled with location information of the stitching artifacts; and 2) propose an efficient compound feature selection algorithm by only considering the useful combinations of feature maps obtained in the previous iteration.

In general, Figure 15.5 shows the overall framework for quality assessment of stitching images. First, a set of training images I_{train} that contain obvious stitching-related artifacts are collected; more information is given in section 15.3.4. With the training database, a CSC dictionary D_{CSC} is learned using an available fast CSC implementation [193] as done in the previous section. Then, the learned dictionary is used to generate representations of the input testing images I_{test}, from which we form initial feature vectors F_{init} by aggregating activated pixels in each feature map. To make better use of location information and investigate compound feature effects, we design a sequential feature selection algorithm to select meaningful filters and discover impactful combinations of simultaneously activated filters. Finally, with the selected feature set FS_{final}, support vector regression (SVR) is adapted to learn the final quality score.

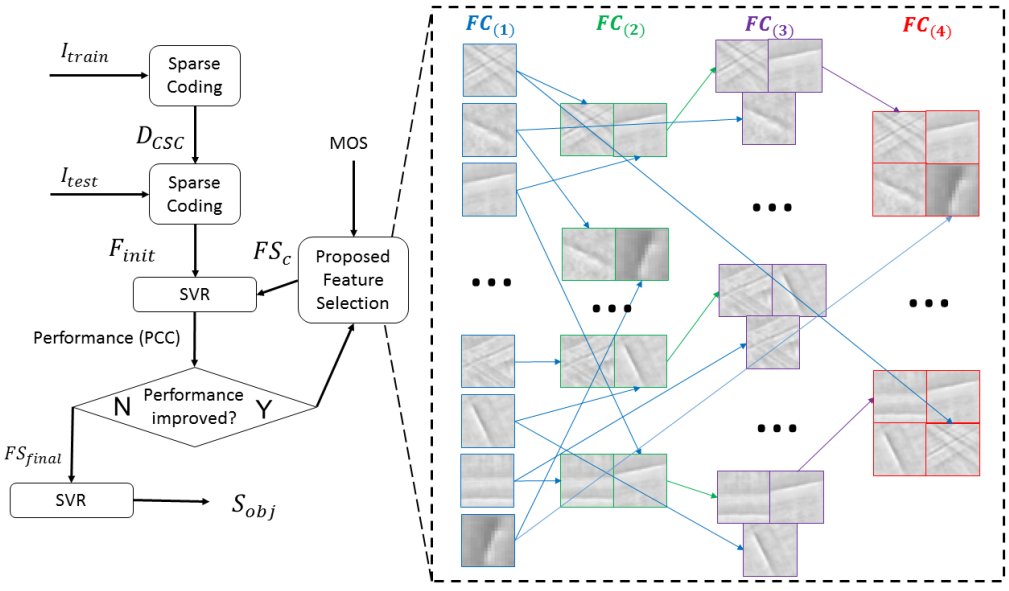


Figure 15.5 – Diagram of the proposed scheme

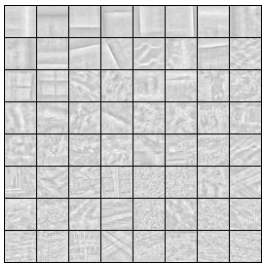
15.3.1 Kernels Training and Feature Extraction with CSC

A Detailed definition of CSC has been given in section 15.2. Hence, a simpler definition is given in this section as shown in equation (15.6), where sparse representation of an image is recovered using l_1-norm:

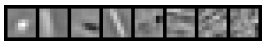
$$\begin{aligned} &\min_{\{D_k, Z_k\}} \frac{1}{2}\|y - \sum_{k=1}^{K} D_k \circledast Z_k\|^2 + \beta_{CSC} \sum_{k=1}^{K} \|Z_k\|_1 \\ &\text{s.t.} \quad \|D_k\| \leqslant 1, \end{aligned} \tag{15.6}$$

where $\circledast$ denotes the convolution operation, y denotes observed samples, Z_k represents sparse feature maps and D_k are the convolution kernels as introduced in the previous section. Similarly, K is the number of convolution kernels, and β_{CSC} is a positive scalar variable used for balancing the model accuracy and the sparsity of feature maps. They are parameters that can be tuned.

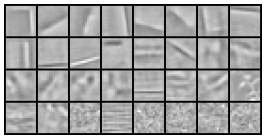
Similarly, considering the efficiency of the training phase, the state-of-the-art implementation proposed in [193] is adapted in order to learn the convolutional kernels to obtain the mid-level structure descriptors for stitched images.



(a) 64 kernels of size 32 × 32



(b) 8 kernels of size 8 × 8



(c) 32 kernels of size 16 × 16

Figure 15.6 – Kernels learned by the convolutional sparse coding [193] on three different scales. Kernels are sorted by the energy of the corresponding feature map in descending order from top-left to bottom-right.

With a training set I_{train}, their improved approach is employed to train a set of convolution kernels $D_{CSC} = \{D_k\}, k \in \{1, ..., K\}$. In our experiment, three scales of kernels, which are filters with sizes of $8\times8, 16\times16, 32\times32$, are chosen, and the relative numbers of kernels for each scale are 8, 32, and 64 respectively. Hence, the total number of kernels is $K = 104$ in this case. Figure 15.6 shows the kernels learned in our experiment. One can observe that characteristics of ghosting artifacts like 'double edges' are well captured by the dictionary. With the trained kernels D_{CSC}, for a $M \times N$ test image I, a sparse representation Z_I for each filter given an input image can be obtained. $Z_I = [Z_1; \ldots; Z_k; \ldots; Z_K]$ is a $M \times N \times K$ tensor of feature maps for I, where each map Z_k is the response of using kernel D_k. Then, a mid-level feature vector v_{csc} for image I can be generated using CSC as:

$$v_{csc} = (f_{act}(Z_1), ..., f_{act}(Z_K)), \tag{15.7}$$

where f_{act} is defined as

$$f_{act}(Z_k) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} \mathbf{1}(Z_k(i,j) > \varepsilon_{CSC})}{M \times N}. \tag{15.8}$$

As introduced in section 15.2, $\mathbf{1}(c)$ is an indicator function that equals to 1 if the specified binary clause c is true and 0 otherwise, and ε_{CSC} is a threshold for selecting activated pixels. Function $f_{act}(\cdot)$ aggregates the number of pixels which are above the threshold ε_{csc} in each sparse feature map Z_k corresponding to each kernel

Algorithm 2 Feature Selection for Evaluation of Simultaneously Activated Feature Maps

1: **Initialization:** $i = 1$; $c = 0$; $opt_{overall} = 0$;
2: $FS = \emptyset$; $FC_{(0)} = \{f_1, ..., f_k\}$;
3: **while** $c <= C$ **do**
4: $opt^1 = opt_{overall}$
5: $d = size(FC_{(c)})$;
6: **while** $i < d$ and $opt^i > opt^{i-1}$ **do**
7: $max = 0$;
8: **for** j from 1 to d **do**
9: **if** $f_j \notin FS$ **then**
10: $score = eval(f_{tmp} \longleftarrow FS + f_j)$
11: **if** $score > max$ **then**
12: $max = score; f_{max} = f_j$;
13: **end if**
14: **end if**
15: **end for**
16: **if** $max > opt^i$ **then**
17: $opt^i = max$;
18: **end if**
19: $FS \longleftarrow FS + f_{max}$
20: $i = i + 1$;
21: **end while**
22: **if** $opt^i > opt_{overall}$ **then**
23: $opt_{overall} = opt^i$;
24: **end if**
25: $c = c + 1$;
26: **if** c $= 1$ **then**
27: $FC_{(c)} = FS$;
28: **else**
29: $FC_{(c)} = F_{ext}(Cmb(FC_{(1)}) \cup Cmb(FC_{(c-1)}))$
30: **end if**
31: **end while**

D_k. Intuitively, this function counts the number of pixels that are activated by the corresponding kernel. In other words, since the kernels are trained to capture stitching-related artifacts, this process can be interpreted as the computation of certain types of artifacts in the entire image and thus can be used to indicate perceived quality.

15.3.2 Adjusted Forward Feature Selection for Evaluation of the Interplay among Feature Maps

In the trained dictionary, ideally, each filter in the dictionary reveals a certain type of distortion introduced by stitching algorithms. However, in practice, only a subset play active roles in reflecting distortions. Further, observing that several convolutional filters may be activated at the same local neighborhood, simultaneous activation of different distortion types (e.g., ghosting artifacts and structure inconsistencies) may amplify the extent of visual annoyance. It is thus reasonable to quantify the amplification effect of simultaneously activated feature maps.

To this end, we propose a new sequential forward feature selection algorithm. This algorithm improves the traditional sequential forward selection (SFS) by accounting for simultaneously activated filters. It is summarized in Algorithm 2. In detail, the inside loop (i.e., line 5-17) selects one best feature at a time from the current candidate set $FC_{(c)}$, which is updated by the outside loop, until the performance starts to decrease.

Specifically, the algorithm starts with an empty selected feature set FS and adds one feature from the candidate feature set $FC(c)$, which contains d candidates, for the first step which gives the highest value for the target evaluation function $eval(\cdot)$. In the experiment, PCC is used for feature selection. From the second step onwards, the remaining features are added individually to the current subset, and the new subset is evaluated. The individual feature is permanently included in the subset $FS \leftarrow FS + f_{max}$ if it gives the maximum performance score ($opt^i = max$). The process is repeated until the current maximum performance opt^i does not improve, i.e. $opt^i < opt^{i-1}$.

The outside loop (i.e. line 3-25) generates a new candidate feature set $FC_{(c)}$. Each iteration considers c number of feature maps, which are selected in the previous iteration $c-1$ until the maximum number C is reached or the overall performance $opt_{overall}$ does not improve anymore. In the experiment, C is set as 5, since it is observed that no pixel is activated by more than 5 filters at the same time. The loop starts with $c = 0$, where $FC_{(0)}$ is the set of the original 104 features from v_{CSC} without feature selection. The dashed bounding box in Figure 15.5 shows an example of how the combinations are selected. For better visualization, we use the corresponding convolutional filters instead of the activated feature map to represent one feature in the selected feature vector. The example starts after one inner loop, and the first candidate set $FC_{(1)}$ is shown as the first column. Then all the elements (green color) in $FC_{(2)}$ are obtained by selecting two filters from $FC_{(1)}$ (blue color). Similarly, to set up $FC_{(3)}$, all the candidates are obtained by selecting 1 element from $FC_{(1)}$ and other combined elements (2 filters) from $FC_{(2)}$. This operation can be represented as $Cmb(FC_{(1)}) \cup Cmb(FC_{(2)})$, where $Cmb(F)$ is the function of selecting one possible compound item with $c-1$ kernels selected from the set of F in the previous iteration without repetition. More specifically, when $c > 1$, one of the kernels is always selected from $FC_{(1)}$, when the other $c-1$ kernels are selected by $Cmb(FC_{(c-1)})$ as one compound element, with the restriction that there are no repetitions among the currently selected kernels. For instance, the first candidate (purple color) in the third column in Figure 15.5 is obtained with the second candidate (blue) in the first column and another two bounded filters (green) in the first row of the second column in the figure. The reason why the other $c-1$ filters are selected from $FC_{(c-1)}$ is that, for one image, if one location is not activated by $c-1$ filters selected in the previous iteration, it is unlikely to be activated by another $c-1$ filters that have not been selected before.

By doing so, the obtained set of candidates can be represented as $Cmb(FC_{(1)}) \cup Cmb(FC_{(c-1)})$, which is a set of n possible combinations $l = \{(l_1^1, .., l_j^1, .., l_c^1), .., (l_1^i, .., l_j^i, .., l_c^i), .., (l_1^n, .., l_j^n, .., l_c^n)\}$, where $(l_1^i, .., l_j^i, .., l_c^i)$ is the i_{th} combination with c elements. Finally, F_{ext} defined below is used to extract the feature,

$$\begin{aligned} F_{ext}(l) = \{&f_{act}^{mlt}((l_1^1, .., l_j^1, .., l_c^1), ..., \\ &f_{act}^{mlt}(l_1^i, .., l_j^i, .., l_c^i), ..., f_{act}^{mlt}(l_1^n, .., l_j^n, .., l_c^n)\}. \end{aligned} \tag{15.9}$$

Similar to the activation function (15.8), here $f_{act}^{mlt}(\cdot)$ is a multi-element activation function accumulating the number of pixels that are activated simultaneously by the corresponding filters of the feature map $Z_{l_1^i}, ..., Z_{l_c^i}$:

$$\begin{aligned} &f_{act}^{mlt}(l_1^i, .., l_j^i, .., l_c^i) = \\ &\frac{\sum_{x=1}^{M}\sum_{y=1}^{N} \mathbf{1}(Z_{l_1^i}(x,y) > \varepsilon_{CSC} \quad \& \quad ... \quad \& \quad Z_{l_c^i}(x,y) > \varepsilon_{CSC})}{M \times N} \end{aligned} \tag{15.10}$$

For example, if $c = 4$, and the i-th combination is $(l_1^i = 5, l_2^i = 45, l_3^i = 66, l_4^i = 92)$, then the 5-th, 45-th, 66-th

and 92-th feature maps will be used to check the simultaneously activated pixels to compute feature f_i. Finally, the entire feature candidate set is obtained by $F_{ext}(l) = \{f_1, .., f_i, f_n\}$.

15.3.3 Experimental Result

To validate the performance of the improved CSC-IQM with compound feature selection CSC-IQM$_{FS}$, we use the publicly available SIQA database released in [36] as introduced in section 4.2.2. Since the SIAQ database is only equipped with pairwise comparison results [196], a pre-processing step is needed to acquire scalar scores for later training. The least squares complete matrix solution with Bradley Terry's logistic model [196] is used in this step to scale the pairwise comparison scores from 28 observers by assuming that the reference image is always better than the stitched images. After linear mapping, the scaled mean opinion scores (MOS) lie in the range of $[0, 100]$. To better evaluate the performance of our proposed metric from pair comparison subjective scores, besides the commonly used PCC, SCC, and RMSE, we also utilize the 'Krasula' evaluation model proposed in [111]. Specifically, as introduced in section 4.3.4, the area under the curve for the Better/Worse (AUC$_{BW}$) receiver operating characteristic (ROC) analysis and correct classification rate (CC) are calculated. Further, in order to compare with the existing objective quality models designed for stitched images summarized in [36], the precision is also calculated by comparing the subjective and objective predicted scores.

15.3.4 Training Set Collection

Training process is vital for CSC. The use of the trained kernels to correctly capture the specific distortions introduced by stitching algorithms depends on whether the training dataset contains enough well represented distortions. Towards this goal, we collected more than two hundred images (from APAP [197] and PTIS [198] database) and used advanced stitching algorithms including APAP [197], PTIS [198] and simple stitching with global homography [199] to get more than 100 stitched images. Afterwards, stitched images that contain significant stitching-related artifacts were selected, and coordinates of the center of the distorted regions were manually labeled. Finally, two another professional observers were asked to pick up the most annoying patches centering at the pre-labeled locations; only patches agreed by both of the observers were maintained for training. Examples of selected patches of size 256×256 are shown in Figure 15.7. It can be observed that all the remained patches contain obvious ghosting or structure inconsistency artifacts. The training database can be downloaded from: ftp:\\ftp.ivc.polytech.univ-nantes.fr\LS2N_IPI_Stitched_Patches_Database\.

Figure 15.7 – Examples of patches selected manually by observers contain annoying stitching related artifact

15.3.5 Result and Analysis

First, to confirm our hypothesis that the adverse visual effects of different filters will be amplified or weakened by others (and to validate our proposed algorithm), the performance of our proposed metric with different numbers of combinations of feature maps are examined and reported in Table 15.3. In this table, the first column contains the numbers of feature maps combinations. For example, FS-4 means four feature maps are selected to generate the new features; FS-0 means no feature selection is adopted. The second column shows the number of the feature dimensions after feature selection. The third column is the change of the dimension after considering another number of the combination. Further, the dimension is accumulated here as the number of combination c increases. For example, by considering the co-activation of three feature maps FS-3, the dimension increases from 49 to 51 compared to FS-2 where two more dimensions are added. Overall, with an augmented number of considered combination numbers among kernels, the performance increases steadily, which confirms our hypothesis and validates our proposed feature selection scheme.

Table 15.3 – Performance of the proposed metric with different combination numbers of Features Map and theirs corresponding optimized dimension numbers.

Z.num	Dim.num	△ Dim	Precision	RMSE
FS-0	104	0	0,8871	0,3234
FS-1	40	-64	0,8947	0,3164
FS-2	49	+9	0,8947	0,3168
FS-3	51	+2	0,8980	0,3165
FS-4	54	+3	0,8992	0,3163
FS-5	55	+1	**0,9000**	**0,3161**

Table 15.4 – Results summarizing performance of the proposed metric and the compared full reference metrics

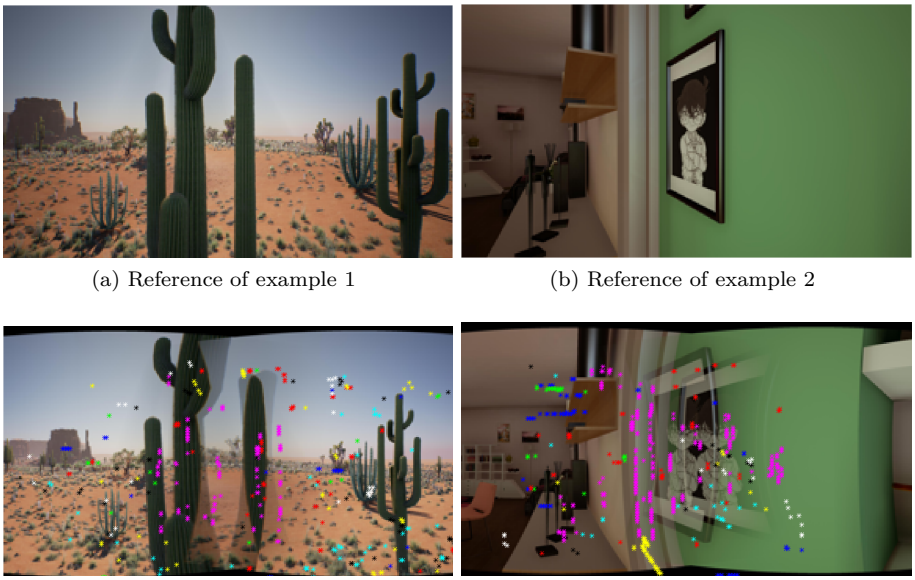
Metric	Precision with subjective PC score	RMSE
Conventional IQA		
FSIM [200]	0.8162	0.4287
SR-SIM [201]	0.8333	0.4082
SIQA		
Quereshi [87]	0.5343	0.6824
Solh [86]	0.8947	0.3803
Yang [36]	**0.9436**	**0.2374**
Ours	0.9000	0.3161

The overall performances of the proposed metric and the compared metrics are summarized in Table 15.4 and 15.5. In the previous table, performance is evaluated with the precision score and RMSE obtained by comparing the predicted score and the subjective pair comparison (PC) scores as reported in [36]. The compared metrics can be categorized into two groups: the conventional image quality assessment metric group and the stitched image quality assessment metric (SIQA) group. Comparing to the first group, we observe that our proposed metric outperforms all of the conventional IQA metrics. More importantly, in the SIQA group, the CSC based metric is the second best performing metric and is comparable to the best performing full reference metric proposed by Yang [36]. In the latter table, our proposed metric has significant gains compared with another three commonly used no-reference metrics. Since the distortions within the stitched images are non-uniform and these local severe distortions are more eye-catching than minor distortions spread globally, the assumption that all the regions share the same subjective score labeled at the image level is no longer sufficient. This makes models trained based on this assumption questionable in this case. In the contrary, the proposed model

is designed to quantify local non-uniform distortions and is thus more promising.

Table 15.5 – Results summarizing performance of the proposed metric and the compared no reference metrics

	PCC	SCC	AUC_{BW}	CC
Full reference metrics				
Solh [86]	**0.9533**	0.7161	**0.9515**	**0.9076**
Yang [36]	0.9102	**0.8437**	0.9082	0.8725
No reference metrics				
Bliinds [48]	0,1184	0,066	0,5167	0,4461
DIIVINE [64]	0,2582	0,1451	0,5405	0,1577
CNN [202]	0,2261	0,2428	0,5512	0,4798
Ours	**0,8574**	**0,7295**	**0,9427**	**0,8643**



(a) Reference of example 1 (b) Reference of example 2

(c) Stitched image of example 1 (d) Stitched image of example 2

Figure 15.8 – Visualization of activated points for stitched related artifacts' regions detection

In order to demonstrate the superiority of our proposed metric, the obtained activated maps of two examples are shown in Figure 15.8. The first row are the original images and the second row is the corresponding stitched images. The points labeled on the stitched images are the activated pixels by specific filters (the best-selected filters after feature selection), where different colors correspond to different filters. First, we observe that the pink points are well distributed around regions where ghosting artifacts are obvious, meaning that the corresponding filters are capable of capturing these vertical ghosting artifacts (double vertical curves). Second, we observe that different colors of points overlap around the distorted regions, which also proves the fact that different filters can amplify the visual distortion in local areas, adversely affecting perceived image quality.

15.4 Conclusion

CSC-IQM in FTV scenario: To learn the impact of non-uniform artifacts on visual quality, a convolutional sparse coding model is first studied in the case of 3-D synthesized views quality assessment in this chapter. Taking advantage of the characteristics of CSC, which can be used to learn from local regions and generate one sparse representation for the entire image, a referenceless metric is proposed by designing an activated function. According to the experimental results, the proposed CSC-IQM performs the best among FR/NR metrics

dedicated to evaluating the quality of synthesized views and its capability of revealing the effect of non-uniform artifacts on perceived quality has been verified.

CSC-IQM in VR scenario: Latter in this chapter, CSC model is further tested on quantifying structure-related distortions (e.g. ghosting and structure inconsistency) in stitched images. Different from the CSC based metric used in FTV application, a layered feature selection algorithm is proposed to quantify the amplification effects of simultaneously activated distortion filters by exploiting the local characteristics of CSC. Extensive experiments have validated that our proposed no-reference metric has competitive performance compared to the state-of-the-art full reference metric designed for stitched images.

16

Learning Synthesized Structure-Related Distortion with Generative Adversarial Network

16.1 Introduction

In this chapter, the second higher-level representation based model is presented. It aims to extract higher-level features from advanced new deep neural network to learn meaningful distortion representation.

As summarized in chapter 14, it is common to use deep learning models for representing images/videos from a higher semantic level in many domains. However, among the existing deep learning models, which model is suitable for learning the effects of non-uniform local structure related distortions on perceived quality? Generative Adversarial Network (GANs), first proposed by Goodfellow *et al.* [203], might be an answer to this question. It has been widely and successfully used in solving problems in computer vision domain, such as super-resolution [204], semantic inpainting (i.e. context encoder) [205–207], and scene images generation [208]. The main idea of the adversarial nets framework is to train a generator (G) and a discriminator (D) simultaneously: a generative model that capture the data distribution and a discriminative model [203] that can tell 'real' image from the generated one. They are trained together so that the discriminator can keep pitting against the generator until the discriminator cannot distinguish the counterfeits generated by the generator. By doing so, both of them are driven to improve their performance until the probability of D making a mistake is maximized.

Considering the case of synthesizing views obtained with DIBR methods, the most annoying non-uniform distortion could be the non-continuous inpainted regions that introduced by the hole filling stage. Examples are shown in the second row of Figure 16.1. In Figure 16.1, sub-figure (d) is a patch containing dark holes (dis-occluded regions) that needed to be filled, while sub-figures (e)-(g) are results after filling up the holes with

different inpainting algorithms. If the generator in GANs is trained to inpaint mimicked dis-occluded regions (i.e., similar darks holes as shown in (d) of Figure. 16.1), and the discriminator is trained along with the generator to tell whether the input is an inpainted image or not. An intuitive assumption is that if the GANs is trained to inpaint similar dis-occluded holes that may be generated during the DIBR process, the discriminator is then trained to identify whether the DIBR synthesized views are of good quality or not. More specifically, on one hand, the intermediate architecture of discriminator is to some extent trained to relate the input to the quality concerning the ground truth. On the other hand, the output of the discriminator can be used as an indicator for selecting local poorly inpainted regions from the entire image.

(a) Original image (b) Synthesized views with A_7

(c) Green box (d) Red box (e) A6 (f) A4 (g) A3

(h) Exp. 1 (i) Exp. 2 (j) Exp. 3 (k) Exp. 4 (l) Exp. 5

Figure 16.1 – Examples of non-uniform distortions in DIBR based synthesized views and examples of results of using different inpainting algorithms for dis-occluded regions filling.

Based on this assumption, in this chapter, a NR quality metric is proposed for evaluating synthesized images. There are mainly three contributions in this work: 1) A GANs based context inpainter/encoder is retrained by designing special masks that are similar to 'dis-occluded regions' induced by DIBR algorithms. This inpainter can thus be used in DIBR framework for 'dis-occluded regions' inpainting; 2) A local non-uniform distortion regions detection strategy is proposed based on the pre-trained discriminator. 3) A quality-aware 'bag of distortion words' is proposed to obtain new quality related representation for each synthesis image by extracting higher-level quality relevant features from the retrained discriminator.

16.1.1 Generative Adversarial Networks based Semantic Inpainting

In the field of computer vision, semantic inpainting is a brand new application, where the goal is to infer missing regions within images according to the semantics of the image. Unlike traditional inpainting or texture synthesis methodologies, semantic inpainting [206, 207, 209–212] aims at filling the missing parts by using

statistical information from external dataset instead of only making use of the internal properties of the image needed.

Among the existing semantic context inpainters, the ones proposed in [206,207], which are based on GANs, provide the best performance. As introduced before, during the training procedure of GANs, the goal is to train two networks at the same time, including a Generator (G) and a Discriminator (D). The role of G is to produce artificial images that look real by mapping a noise sample z from Z_n distribution to an image training data distribution I_{train}, while the role of D is to distinguish the generated images from the real ones. More specifically, (1) after taking both the generated and the real images as input, the duty of the adversarial discriminator D is to be able to tell the difference between them; (2) the duty of Generator G is to be able to 'fool' D by offering as real as possible images. They are commonly trained with the following adversarial loss function:

$$\min_{G} \max_{D} \mathbb{E}_{i \in I_{train}}[log(D(i)] + \mathbb{E}_{z \in Z_n}[log(1 - D(G(x))]. \tag{16.1}$$

In [209], the proposed context encoder is designed as an autoencoder with the unfilled images as the condition. In detail, to ensure continuity within the context, L_2 norm reconstruction is defined in (16.2) to regress the missing parts to the ground truth content.

$$\mathcal{L}_{rec}(i) = \| M \odot (i - G((1 - M) \odot I_{train})) \|_2^2, \tag{16.2}$$

where M denote the binary mask indicating the missing regions needed to be inpainted and $G(\cdot)$ is the (Generator) function representing the autoencoder, which generates the inpainted image with an input i. To overcome the blurry preference problem aroused by L_2 loss (i.e, it tends to predict the mean of the distribution resulting in an averaged blurry image), the adversarial loss is also introduced to jointly optimize both G and D as formalized in equation (16.3):

$$\mathcal{L}_{joint} = \lambda \mathcal{L}_{rec} + (1 - \lambda)\mathcal{L}_{adv}, \tag{16.3}$$

where λ is a hyper-parameter to balance the weights between the two losses. $\mathcal{L}_{adv}$ is further defined in equation (16.4), which is derived from equation (16.1) by customizing GANs for the context encoder task with the mask M:

$$\mathcal{L}_{adv} = \max_{D} \mathbb{E}_{i \in I_{train}}[log(D(i) + log(1 - D(G((1 - M) \odot I_{train})))] \tag{16.4}$$

In this chapter, based on the common GANs formulation (16.4) used for semantic inpainting, specific masks M are designed, which mimics the 'dis-occluded' regions appear during DIBR process, to retrain a new 'context inpainter'. Then, we explore using the pre-trained discriminator to evaluate the quality of synthesized images. Details of the proposed model are given in the following sections.

16.2 The Proposed GAN-IQM Model

In this section, the proposed GANs based No-reference Quality Metric for synthesized views (GANs-NQM) is described in detail.

The overall framework of the proposed model is illustrated in Figure 16.2. The entire scheme can be divided into three sub-procedures, which are 1) Pretraining of the GANs based context encoder; 2) 'bags of distortions

words' code-book training; and 3) quality prediction. They are bounded by the blue, green and red dashed boxes in Figure 16.2 correspondingly.

First and foremost, during the context encoder (inpianter) pre-training procedure, special masks that locate at possible dis-occluded regions or possible distorted regions introduced by DIBR based algorithms are designed. As thus, the inpainter is trained to generate inpainted images that are similar to the synthesized views generated by DIBR based methodologies. Then, during the code-book training process, the validation image set is divided into a set of overlapping patches. After feeding those patches into the pre-trained discriminator, a 'bag of distortion codebook' is obtained by clustering all the patches into K clusters. Ideally, each cluster represents a category of patches that with a similar type of distortion. To predict the perceived quality, images are first represented as a set of overlapped patches. Afterward, the discriminator is used as a distortion regions selector, only patches with values smaller than a certain threshold (i.e., indicating the regions are not well inpainted) are remained for the next process. After distortion regions selection, the image under test is then represented as a histogram of distortion categories with the distortions related code-book. Finally, with each image represented as h_{adv}, quality scores are predicted with SVR. Details of each sub-procedure are given in the following sub-sections.

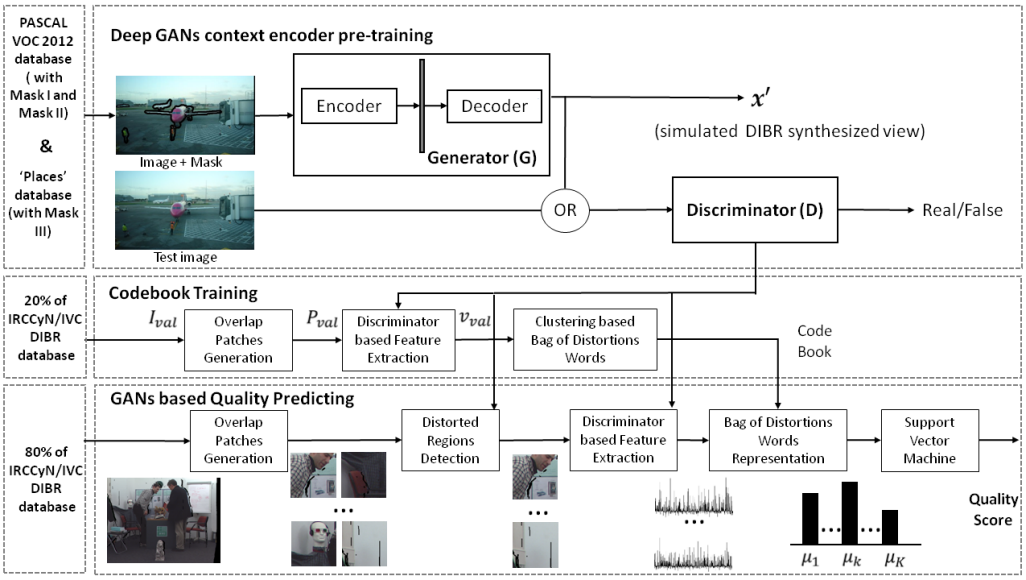


Figure 16.2 – Diagram of the proposed model: (1) Deep GANs context encoder pre-training; (2) Distortion codebook training; (3) Quality predicting.

16.2.1 Pre-training of GANs for inpainting of RGB-D synthesis view

16.2.1.1 Design of masks

As introduced before, dis-occluded regions are introduced during the DIBR synthesis process. There are mainly two types of dis-occluded regions: 1) edge-like holes that are located along the boundaries of the foreground objects as shown in Figure 16.3a, and 2) small or medium size of holes that are distributed throughout the entire images as shown in Figure 16.3b. The shapes of these regions are normally related to the shapes of objects. These regions can be filled with certain inpainting algorithms. However, inpainting-related artifacts may also be introduced.

Generally, dis-occluded regions that are located along the border between the foreground and the background

are challenging for existing inpainting algorithms. It is often to see that foreground regions are inpainted with background pixels or vice versa. As a result, the structures of objects are disrupted. Structure related degradation around foreground objects, accompanying with inter-view inconsistency on depth, might then cause binocular rivalry, binocular suppression, or binocular superposition [213, 214] which eventually lead to visual discomfort. Concerning the issues above, and to train a new context encoder that is capable of inpainting the distortions mentioned above, two types of masks M are designed:

(a) (b)

Figure 16.3 – Examples of typical dis-occluded regions introduced during the process of DIBR based views synthesis. (a) Examples of dis-occluded regions that are around foreground objects' boundaries (bounded by green boxes); (b) Examples of small and median size of dis-occluded regions (bounded by red and blue bounding boxes correspondingly) that distributed throughout the image;

- **Mask I**: to mimic the holes in dis-occluded regions, which is generally around foreground objects' boundaries. The mask is designed as the dilated object boundaries. An example is shown in Figure 16.4c.
- **Mask II**: to mimic the shifted objects' boundaries in the synthesized views induced by compression on depth map [49]. We generate the second type of masks by simply shifting the first type of mask with certain pixels as shown in Figure 16.4d.

Generally, it is easier to inpaint smooth regions with homogeneous textures than complicated regions with non-homogeneous textures, as the context around a smoother region is more 'copyable' and less structure are involved. Hence, the quality of smooth inpainted regions within homogeneous texture is generally better than the non-homogeneous ones. If one wants to train a more powerful context encoder, the selected masks should contain contents/structures that can not be replicated from the surroundings. In addition, unlike the big connected region masks with arbitrary location introduced in [207], dis-occluded regions or missing parts in a virtual view are generally disconnected, and the shapes of these regions are always related to the foreground objects (i.e., related to the depth map). With these two concerns, the third mask is proposed:

- **Mask III**: The SLIC super-pixels algorithm [147] is used to select regions where masks should be located for later training. More specifically, an image is first segmented into a set of super-pixels as shown in Fig. 16.5a and 16.5d. Then, **two mask sizes are considered**. Super-pixels that contains less than 100 pixels are considered as small size mask, while super-pixel contains 200 to 1000 pixels are considered as medium size mask. Examples are presented in Fig. 16.5. The black holes in Fig. 16.5b and 16.5e are small size masks which are similar to the small holes shown in Fig. 16.3a. The holes in Fig. 16.5c and 16.5f are with medium size. By doing so, 1) the selected masks are separately distributed in the entire image; 2) the shape of the masks are related to objects; 3) the content within each mask region is more

independent from its neighborhood.

Since the three masks designed in this work are to mimic the dis-occluded regions that appear during DIBR process, where they are generally black holes or boundaries, M is thus with value 0 while in [206,207] the mask is white with value 1. The corresponding calculation on loss shown in Equation (16.2) and Equation (16.4), used in [206,207], are therefore changed to:

$$\mathcal{L}_{rec}(x) = \| (1 - M) \odot (x - G(M \odot x)) \|_2^2 \tag{16.5}$$

$$\mathcal{L}_{adv} = \max_D \mathbb{E}_{x \sim p_{\mathbf{x}}}[log(D(x)) + log(1 - D(G(M \odot x)))] \tag{16.6}$$

16.2.1.2 Training data

To generate a new dataset with the three masks mentioned above, we collect images from the PASCAL VOC 2012 [215] and the Places database [216]. There are in total 10K training images in this study.

- **PASCAL VOC 2012 database**: The original objective of this database is for a challenge to recognize objects from a number of visual object classes in realistic scenes. It contains 3K images with twenty object classes, which diverse from people, animals to vehicles and indoor scenes. One of the merits of this database is that it provides us with pixel-wise segmentation labels, which gives the boundary of 'objects' against the 'background' label. An example is given in Fig. 16.4a and Fig. 16.4b. In our study, we utilize this segmentation label to generate Mask I and Mask II mentioned above, which leads to 6K training data.
- **Places database**: To have a balanced dataset with mask I and II, the validation set from the 'Places Challenge 2017', which contains around 2K images, are selected as a part of the training set in this study with mask III mentioned above. This dataset contains images with diverse contents, which varies from outdoor landscapes, cities views to indoor people portrait images. As there are two mask sizes in Mask III, this leads to 4K training images.

(a) Image (b) Labeled map (c) Mask I (d) Mask II

Figure 16.4 – Example of images in the training set and with mask I and II.

16.2.1.3 GANs training process

The framework of the 'context inpainter' is implemented based on the pipeline developed by Pathak *et al.* [207] with Caffe and Torch packages. The commonly used stochastic gradient descent method Adam [217] is used for optimization. We start with a learning rate of 0.0002, as set in DCGAN [218], but a different bottleneck of 4000 units. In our experiment, the impact of trade-off between G and D, *i.e.*, different λ in Equation (16.3), on the performance of the proposed metric has been tested.

(a) Supper-pixel map (b) Small size mask (c) Medium size mask

(d) Supper-pixel map (e) Small size mask (f) Medium size mask

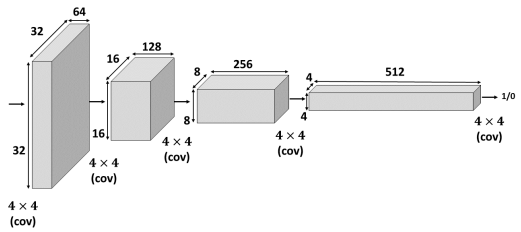
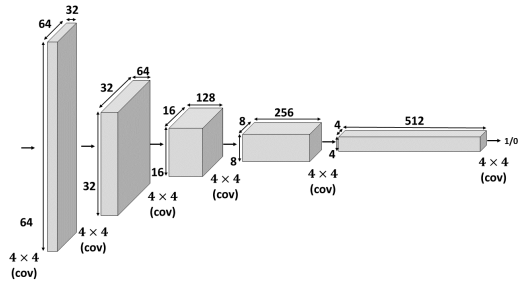
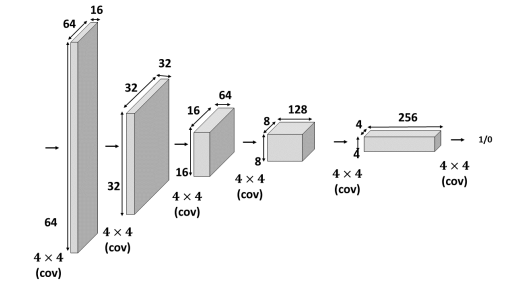
Figure 16.5 – Example of images in the training set and designed mask III. Two mask sizes are considered.

For the architecture of the GANs network, as it has been tested in [207] that finer inpainting results can be obtained by replacing pooling layers with the convolutional ones, in this study, the pool-free structure remains. Furthermore, since the main focus of this section is to explore the discriminator for quality assessment of synthesized views with local non-uniform distortions, we only change the architecture of the discriminator. Details of all the discriminator architectures that have been tested in this study are summarized in Table 16.1. The main difference between D_1 and the other two architectures is the size of images that can be fed into. D_2 is of less complex structures than D_3 and D_1, where the number of convolutional kernels is halved in each layer. With such design, we could check how the input size and complexity of the discriminator influence the performance of the proposed scheme.

16.2.2 Bag-of-Distortion-Words (BDW) codebook learning with pre-trained discriminator

As discussed before, the discriminator serves as an indicator telling whether a patch is well inpainted or not. Thus the output of the discriminator is related to the quality of the patch. Therefore, it is reasonable to hypothesize that the intermediate output of D is strongly related to inpainting related distortions which affect the perceived quality significantly. Based on this hypothesis, we propose to use the discriminator to get a latent codebook with 'codewords' that represent different types of distortions. This codebook is trained with a validation set I_{val} that contains real DIBR-based images (in this study, it is 20% of the IRCCyN/IVC DIBR database [49], details can be found in Section 16.2.4 and Section 16.3). With this codebook, a higher-level representation could be obtained for each image. Details are illustrated below.

Table 16.1 – Different discriminator architectures tested in this study, In is the input of each layer, $InSize$ is the input size of each layer, k is the kernel size, **s** is the stride, $OutL$ is the output channels for each layer and Act is the activation function of each layer.

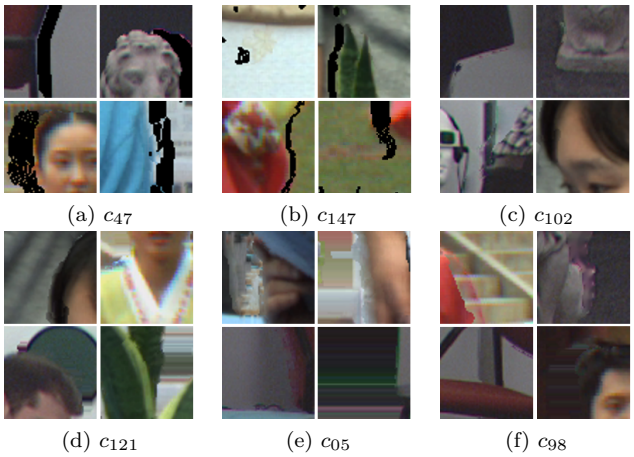
Layer	In	InSize	k	s	OutL	Act	Visualization
Discriminator architecture D_1							
conv_1	image	64×64	4	1	64	Leaky ReLU	
conv_2	conv_1	32×32	4	1	128	Leaky ReLU	
conv_3	conv_2	16×16	4	1	256	Leaky ReLU	
conv_4	conv_3	8×8	4	1	512	Leaky ReLU	
conv_5	conv_4	4×4	4	1	1	Sig moid	
Discriminator architecture D_2							
conv_1	image	128×128	4	1	32	Leaky ReLU	
conv_2	conv_1	64×64	4	1	64	Leaky ReLU	
conv_3	conv_2	32×32	4	1	128	Leaky ReLU	
conv_4	conv_3	16×16	4	1	256	Leaky ReLU	
conv_5	conv_4	8×8	4	1	512	Leaky ReLU	
conv_6	conv_5	4×4	4	1	1	Sig moid	
Discriminator architecture D_3							
conv_1	image	128×128	4	1	16	Leaky ReLU	
conv_2	conv_1	64×64	4	1	32	Leaky ReLU	
conv_3	conv_2	32×32	4	1	64	Leaky ReLU	
conv_4	conv_3	16×16	4	1	128	Leaky ReLU	
conv_5	conv_4	8×8	4	1	256	Leaky ReLU	
conv_6	conv_5	4×4	4	1	1	Sig moid	

To predict the image level quality by considering local distortion, the image needed to be processed locally. Therefore, a set of multiple overlapping patches $P_i = \{p_{ij} | j = 1, \cdots, n\}$, where n is the total number of patches, is used to represent the image x_i as done in [71]. In this study, the overlap size is selected as half of the patch size, and the patches are sampled over the whole image (along both the horizontal and vertical direction) to maintain as much structural information as possible. Afterwards, with the pre-trained GANs model, these patches are fed into the adversarial discriminator to extract higher-level features for later patches categorizations. For each patch p_{ij} in the entire validation dataset I_{val}, its corresponding feature vector v_{ij} is extracted from the l_{th} layer in the discriminator as:

$$v_{ij} = D(p_{ij}, l) \tag{16.7}$$

In this study, the feature vector is extracted from the last convolutional layer of the discriminator (Details are

shown in Section 16.2.3). Finally, $m \times n$ vectors can be obtained for the m images in the validation set I_{val}.



(a) c_{47} (b) c_{147} (c) c_{102}

(d) c_{121} (e) c_{05} (f) c_{98}

Figure 16.6 – Selected 'Words' in the learned BDW Codebook.

With the set of extracted features in correspondence to their patches, now we want to look for a new representation of the entire image by taking the intermediate output of the discriminator into account so that this new representation is able to link the local information with the entire image quality.

Intuitively, the idea is to categorize image patches into different clusters that can be representatives of perceived quality, so the quality of the tested image can be quantified by checking how many 'good' or 'poor' patches it has. Formally, the $m \times n$ patches v_{ij}, $i = \{1, \cdots, m\}$, $j = \{1, \cdots, n\}$ are reshaped to $v_o, o = \{1, \cdots, n \times m\}$. Then the v are clustered into K clusters $\{c_1, \cdots, c_K\}$ using an advanced clustering algorithm [**?**], which is a fast nearest neighbor algorithm robust in high dimensional vectors matching. Selected cluster results are shown in Fig. 16.6. It can be observed that patches with similar type of distortions are gathered in the same cluster as a 'distortion word'. For example, both of the cluster c_{47} and c_{147} are consist of patches with 'dark holes', and the ones in cluster c_{47} are obviously larger than that of c_{147}, which indicates worse quality. For other clusters shown in the figure, the distortions of c_{102} is imperceivable (guarantee good quality), while c_{121}, c_{05} and c_{98} are with more obvious inpainting related artifacts. Naturally, different 'codeword' in the clustered 'codebook' actually represents a certain level of quality with respect to the types and magnitudes of distortions, which is in consistent with our hypothesis. Based on this observation, in this study, the trained codebook is named after 'bag of distortion words' (BDW). With the BDW codebook, each image x_i can then be encoded as a histogram $h_{adv}(i) = \{\mu_{i1}, \cdots, \mu_{iK}\}$, where each μ_{ik} is defined as

$$\mu_{ik} = \frac{\sum_{j=1}^{n} \mathbf{1}(p_{ij} \subset c_k)}{n} \tag{16.8}$$

$\mathbf{1}(c)$ is an indicator function that equals to 1 if the specified binary clause c is true. An intuitive interpretation of this BDW based representation of the image is that the histogram statistically quantifies how many 'good quality' and 'poor quality' patches that a synthesized image has. As local significant synthesized distortion is more annoying than the global uniform one, this new representation is a higher-level quality descriptor which can indirectly predict the overall quality of one image. During clustering, K is an important parameter that will have an impact on the final performance. Therefore, further discussion about the selection of K is given in Section 16.3.

16.2.3 Local distortion regions selection

Generally, artifacts located at a region of interest is much more annoying than those located at an inconspicuous area when observing an image [34]. In our case, 'poor' quality regions (*i.e.*, holes and inpainting artifacts) are generally in the regions of interest (such as the foreground object), thus, they are more likely to be attracted by observers than the 'good' ones. Therefore, images with even a small number of 'poor' regions are penalized more gravely by the observers. Accordingly, it is reasonable to do the same penalization in the objective model as well.

Moreover, as discussed before, the discriminator is trained to distinguish artificial generated picture (inpainted images in this case) from the real one. A well-trained discriminator is supposed to be able to indicate poor inpainted regions. The output of the discriminator D is a boolean value indicating whether the input patch p_{ij} is an inpainted or not, where '1' for real patches and '0' for generated patches. It is intuitive to hypothesize that patches assigned with '0' by the discriminator are those with poor quality. Hence, the discriminator is further utilized as a 'poor' quality patches selector. As thus, Equation (16.8) could be modified to:

$$\mu_{ik} = \frac{\sum_{j=1}^{n} \mathbf{1}(p_{ij} \subset c_k) \cdot XOR(D(p_{ij}), 1))}{n} \tag{16.9}$$

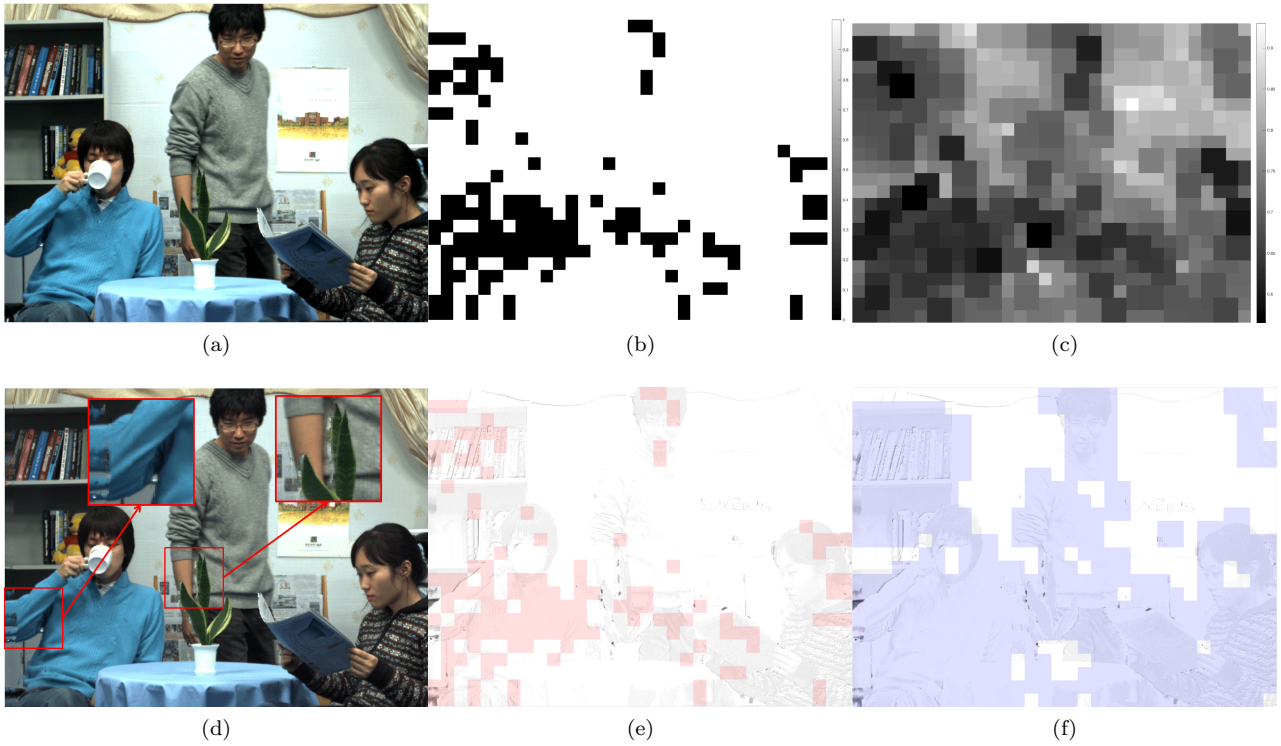
where $D(\cdot)$ is the direct boolean output of the pre-trained discriminator when taking a patch p_{ij} as the input. $XOR(\cdot)$ is the exclusive OR operation, $XOR(D(p_{ij}), 1))$ equals to 1 if $D(p_{ij}) = 0$.

Apart from using the final boolean output of the discriminator for selecting the possible inpainted regions, another possibility is to use the output just before the final sigmoid layer (*i.e.*, the last convolutional layer) with normalization. To do this, the output of the last convolutional layer of the discriminator for all the training patches $p_{ij}, i = \{1, \cdots, m\}, j = \{1, \cdots, n\}$ are collected and normalized into a range of $[0, 1]$. After the normalization, the output of the last convolutional layer serves as a probability value indicating that whether the test patch is natural (non-inpainted) or not. A smaller value represents a higher probability that this patch is inpainted and with a greater magnitude of distortions. Afterwards, patches that with a certain magnitude of in-painting distortions can be selected according to a certain threshold ε, meaning that only poorly inpainted regions with certain low-quality level are selected for the final quality decision. By doing so, Equation (16.9) could be further rewritten as:

$$\mu_{ik} = \frac{\sum_{j=1}^{n} \mathbf{1}(p_{ij} \subset c_k) \cdot \mathbf{1}(D_{BS}(p_{ij}) < \varepsilon)}{n} \tag{16.10}$$

where $D_{BS}(\cdot)$ means we only consider the output of the last convolutional layer in the discriminator with a patch p_{ij} as input. ε is a threshold for poor-quality patches selection. The setting of threshold ε is discussed in Section 16.3.

Examples of possible distortion regions predicted by the discriminator are shown in Fig. 16.7. Fig. 16.7 (b) is the distortion map obtained using the direct output of the discriminator, i.e., $D(I)$. Dark color patches are the regions predicted to be inpainted while the white ones are predicted to be the real. Fig. 16.7 (c) is the distortion map $D_{BS}(I)$. To check whether distortion regions within a synthesized image are well indicated by the output of the discriminator, we plot $D(I)$ and $D_{BS}(I < 0.7)$ (red and blue color respectively) on the ground truth error map in 16.7 (e) and (f) respectively. According to Figure 16.7 (e) and (f), the distortion regions are better covered by $D_{BS}(I < 0.7)$ than $D(I)$. For $D_{BS}(I)$, it is observed that a threshold of 0.7 could make most of the severely distorted regions detected.



(a) (b) (c)

(d) (e) (f)

Figure 16.7 – Example of possible distortion regions selected by the pre-trained discriminator (better see in color). (a) original image I; (b) distortion map $D(I)$ generated directly using the 0/1 output of the discriminator for all the patches within one image; (c) distortion map $D_{BS}(I)$ generated with the normalized output of the previous layer before the last sigmoid layer of the discriminator (the darker the color, the more likely the distortions exist); (d) synthesized image with zoomed-in regions that contain severe inpainting-related distortions ; (e) possible synthesized regions (labeled with red color) indicated by $D(I)$ with ground truth error map as reference (obtained with the reference (a) and synthesized image (d)). (f) possible synthesized regions (labeled with red color) indicated by $D_{BS}(I)$ (with a threshold $\varepsilon = 0.7$, meaning that patches with a normalized value smaller than 0.7 are plotted) with ground truth error map as reference.

16.2.4 Final Score Prediction

After extracting the histogram h_{gan} , SVR is then applied on h_{gan} with a linear kernel to predict the final quality score. In the experiment, the entire database is divided into 20% validation set for model parameters selection (e.g., codebook training) and 80% for performance evaluation. During the performance evaluation procedure, a 1000-fold cross-validation is applied. For each fold, the remaining 80% of the dataset is further randomly split into 80% of the images for SVR training and 20% for testing, with no overlap between them [155]. The median PCC, SCC, and RMSM between subjective and objective scores are reported across the 1000 runs for performance evaluation.

16.3 Experimental Result

The performance of the proposed GAN-IQM is evaluated on the IVC-Image database [49] as described in section 4.2.1.1. To the best of our knowledge, this is the only existing image database designed for comparing different DIBR synthesis algorithms, and released with subjective scores. To provide more robust performances evaluation data augmentation is conducted by rotating each image by 90°, 180° and 270° counterclockwise successively, which ends up into totally 384 images. Unlike other data augmentation methodology, such as scaling, rotation does not introduce distortions. We thus assume the quality of the augmented image remains unchanged.

16.3.1 Performance Dependency of Utilized Parameters

16.3.1.1 Number of 'Distortion Words' K in BDW

To check if the performance of the proposed GAN-IQM is sensitive to the cluster number K, different numbers of K for the quality-aware dictionary training are tested on the validation set. The results are shown in Figure 16.8. The corresponding PCC/SCC curves are obtained by fixing other related parameters. As it can be observed that, the performance of GAN-IQM in PCC/SCC raises gradually along with the increase of K at the beginning. After the performance peaks at a certain number of K (160), it starts to drop gradually. In this study, we set $K = 160$.

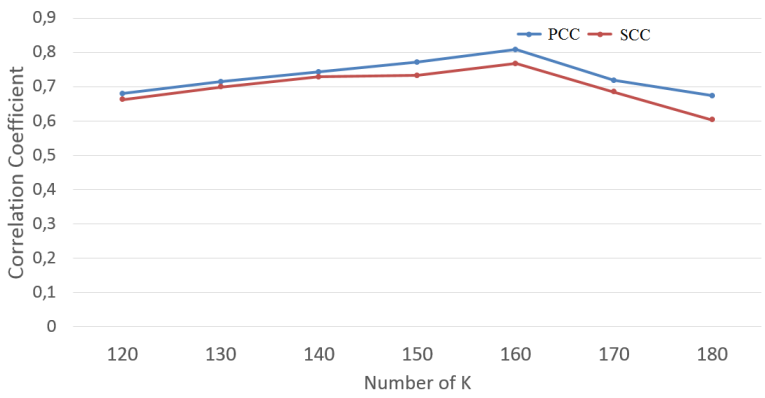


Figure 16.8 – Performance dependency of proposed metric with changing K number

Table 16.2 – Performance dependency of proposed metric with different solver hyper-parameters λ

PCC	$\lambda = 0.5$	$\lambda = 0.9$	$\lambda = 0.999$
D_1	0.7802	**0.8083**	0.7821
D_2	0.7377	0.7536	0.7339
D_3	0.7266	0.7280	0.7273

Table 16.3 – Performance dependency of proposed metric with different Threshold ε

	PCC	**SCC**	**RMSE**
no selection	0.7691	0.7268	0.4782
direct output	0.8083	0.7669	0.4214
$\varepsilon = 0.3$	0.7525	0.7259	0.4995
$\varepsilon = 0.4$	0.7631	0.7649	0.4593
$\varepsilon = 0.5$	0.7889	0.7600	0.4546
$\varepsilon = 0.6$	0.7963	0.7798	0.4176
$\varepsilon = 0.7$	**0.8195**	**0.7920**	**0.4016**
$\varepsilon = 0.8$	0.7704	0.7248	0.4723

16.3.1.2 Different Solver Hyper-Parameters λ

As introduced in [207,218], the solver hyper-parameter λ in equation (16.3), is suggested to be set as 0.999. It is a tunable parameter balancing the reconstruction loss and the adversarial loss during training. Since the discriminator is utilized for both distortion regions selection, and higher level feature extraction in this study, higher weight for the adversarial loss is tested, i.e. lower λ in equation (16.3). The performances of the proposed model with different λ are reported in Table 16.2. The performance reported in this table is obtained by fixing $K = 160$ and using the direct output of the discriminator for distortion region selection. By comparing the performances with λ equaling to 0.5 and 0.9, it is obvious that with larger $\lambda = 0.9$ the proposed metric achieve better PCC value despite what discriminator architectures are using. Interestingly, it is found that the PCC drops when $\lambda = 0.999$. In this study, we set $\lambda = 0.9$.

16.3.1.3 Different Discriminator Architecture

The performances of the proposed model equipped with different discriminator architectures, which are described in Table 16.1, are also reported in Table 16.2. It is found that, with any chosen λ, the proposed method always attains better PCC value with architecture D_1 than with D_2 or D_3. In the proposed model, we finally choose architecture D_1 for discriminator.

16.3.1.4 Threshold ε Setting

The influence of the threshold ε on the performance of GAN-IQM is illustrated in Table 16.3. The performance of using a strategy of selecting a proper threshold in equation (16.10) for distortion regions selections is better than using the direct output of the discriminator. The performance climbs with an increasing ε until it reaches to 0.7, then the performance declines.

Based on this observation, conclusions can be made that 1) regions that contain considerably serious local distortion do play a major role in predicting the perceived quality score for synthesized images; 2) the direct output of the discriminator may fail to detect the regions with better quality but still contain synthesized related artifacts, which also play certain role in deciding the quality. .

16.3.2 Overall Performance

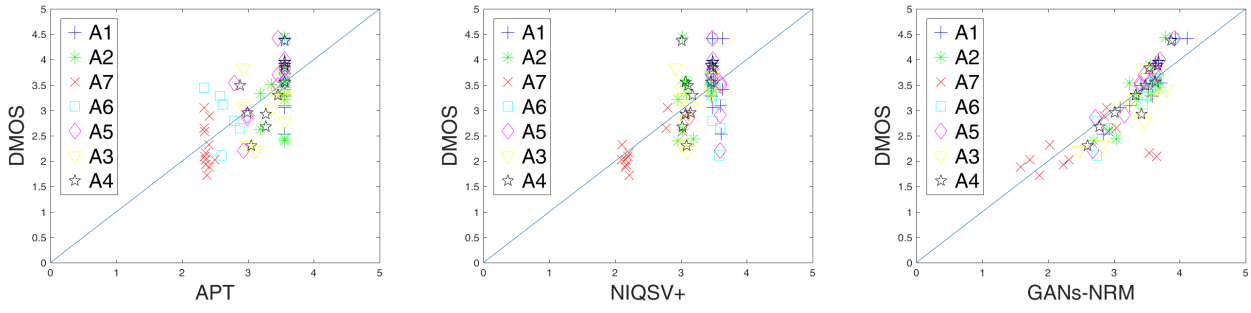
The performance of the proposed model is compared with all the 3D quality metrics that are developed for assessing synthesized views summarized in chapter 3. For fair comparison, the median performance of the compared metrics are also reported under a 1000-fold cross-validation.

Performance results are summarized in Table 16.4. These metrics can be divided into two groups, which are the full reference (FR) metrics and the no reference (NR) ones. Parameters of GAN-IQM that yield the best performance are selected according to the previous discussion. It can be seen from Table 16.4 that our proposed method attains the best performance within the group of NR metrics in terms of PCC, SCC and RMSE. The gain of GAN-IQM compared to the second best NR metric APT is 17% in PCC. Furthermore, even compared to FR metrics, its performance is comparable to the best performing ST-IQM.

Table 16.4 – Performance of the proposed metric an the state-of-the-art metrics designed for synthesized views

	PCC	**SCC**	**RMSE**
Full Reference Metric (FR)			
3DSwIM	0.7266	0.6421	0.4304
VSQA	0.5096	0.5064	0.5336
MP-PSNR$_r$	0.7489	0.7011	0.4148
MP-PSNR$_f$	0.7336	0.6634	0.4199
MW-PSNR$_r$	0.7400	0.6836	0.4240
MW-PSNR$_f$	0.7183	0.6419	0.4401
CT-IQM	0.7107	0.6151	0.4481
EM-IQM	0.7599	0.7012	0.4038
ST-IQM	0.8462	0.7681	0.3415
NO Reference Metric (NR)			
NIQSV+	0.7010	0.5158	0.4553
APT	0.7046	0.7198	0.4993
GAN-IQA(NR)	**0.8262**	**0.8072**	**0.3861**

The scatter plots of all the NR quality metrics versus DMOS are provided in Fig. 16.9. By comparing the scatter plots of GAN-IQM with others, we can notice that most of the objective scores that predicted by the proposed metric are well distributed along the diagonal of the plot. In the scatter plot of APT and NIQSV+, images that synthesized using same DIBR algorithm are predicted with similar objective scores, which leads to a 'vertical line' as shown in Figure 16.9 (a) and (b).



(a) (b) (c)

Figure 16.9 – Scatter plots of the three blind quality metrics versus DMOS on IVC-Image database. (a) APT. (b) NIQSV+. (c) GAN-IQM.

Last but not least, to meet the need of real-time computation for real applications such as multi-views live match broadcasting, the time cost of the quality assessment metric should be low enough. In verification of the efficiency of the proposed metric, the execution time of the metrics normalized by PSNR as introduced in

section 4.4 are listed in Table 16.5. Due to the lack of reference views in most of the practical situation, here only the no reference metrics are reported. According to the table, even though our proposed metric is a bit slower than NIQSV+, it is still much faster than second best performing APT.

Table 16.5 – Normalized execution time of proposed metric compare to the state-of-the-art metrics.

Metric	NIQSV+	APT	GANs-NRM
Normalized time	**21**	13k+	157

16.3.3 Inpainting results

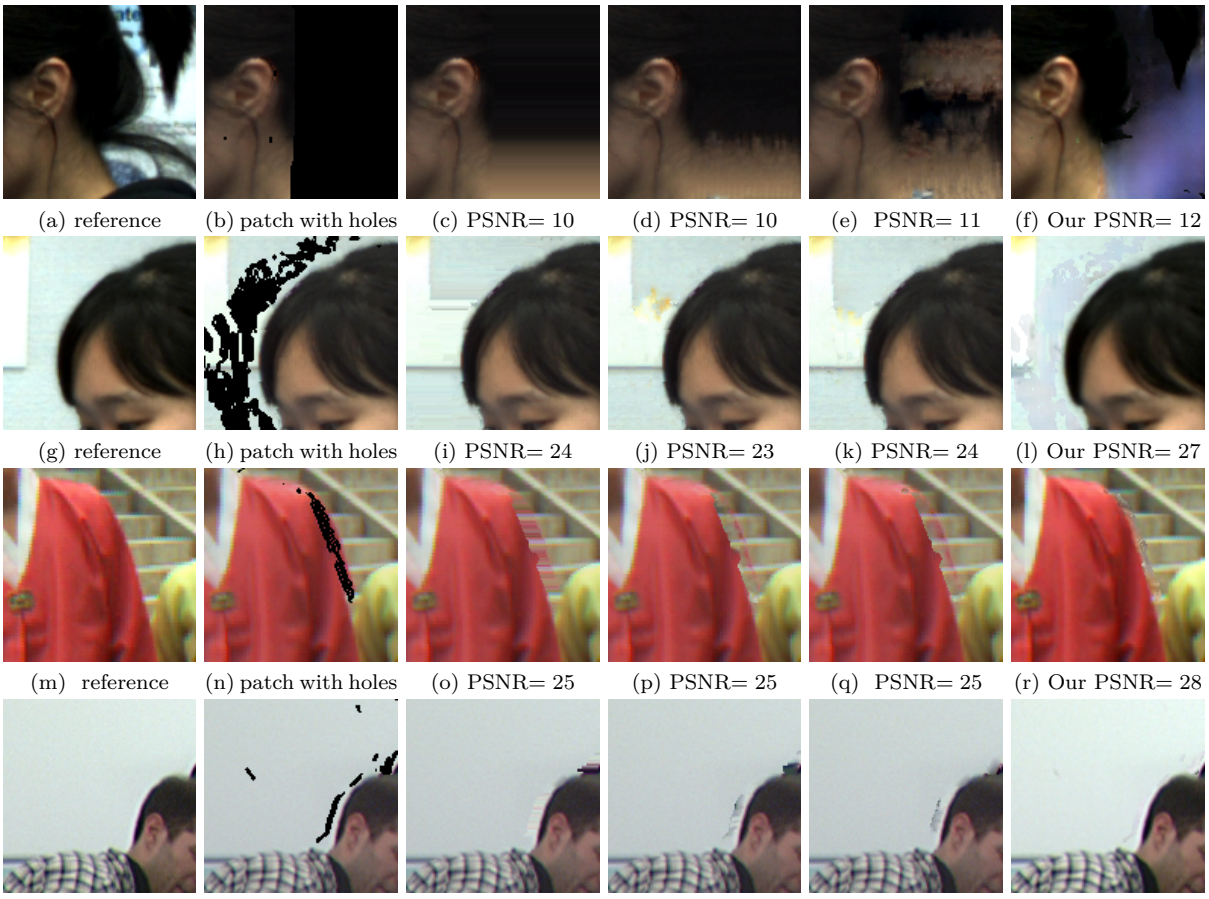
The theoretical assumption of this study is that the generator/discriminator are simultaneously trained to inpaint/evaluate the RGB-D dis-occluded regions/RGB-D synthesis views. The performance of utilizing discriminator to predict the quality of the RGB-D synthesis views has been demonstrated in the previous section. As a side outcome of this study, it would be interesting to evaluate the performance of the pre-trained context inpainter (generator) on the same database, *i.e.*, the synthesized views that contain dis-occluded regions in the IRCCyN/IVC DIBR images database.

PSNR between the reference and the inpainted image is calculated for evaluation. Three inpainting algorithms [80] [99] [100] are used for comparison. Due to the limitation of space, selected results are shown in Fig.16.10.

Based on the results, it is observed that 1) By comparing our inpainted result in Fig. 16.10f to the others with respect to the reference, the shape of the braid of the girl is better remained by our model. Similar results could also be observed in Fig. 16.10l where the corner of the poster is better preserved compared to the others; 2) The shape of the dis-occluded regions in Fig. 16.10n are better inpainted by the proposed models as shown in Fig. 16.10r. There are obvious 'double-edge like' shapes, *i.e.* ghosting artifacts, along the objects after being inpainted by other methods; 3) In the condition that holes appear in homogeneous texture regions which are also close to the borders of foreground objects, our inpainted result is with higher texture consistency than the others as shown in Fig. 16.10x.

In conclusion, our proposed context inpainter could maintain the structures of the dis-occluded regions, especially when the dis-occluded regions are large. For the challenging dis-occluded regions that lie on the border of foregrounds and backgrounds, as well as in the homogeneous texture regions close to the border of foreground objects, the proposed inpainter performs better than the others.

The appealing performance of our pre-trained context inpainter (generator) on RGB-D dis-occluded regions validates the effectiveness of the proposed training strategy, which uses specific designed masks to mimic the typical black-hole artifacts induced in DIBR process. The proposed strategy is more flexible in using the large-scale image databases in the computer vision domain rather than the RGB-D datasets where the depth information might be noisy. It should also be noted that the training data scale in our study is only 10K, which could be definitely further augmented by employing the existing datasets. Therefore, there is still improvement space for our current trained model, no matter for quality assessment or for hole filling of RGB-D synthesis view.



(a) reference (b) patch with holes (c) PSNR= 10 (d) PSNR= 10 (e) PSNR= 11 (f) Our PSNR= 12

(g) reference (h) patch with holes (i) PSNR= 24 (j) PSNR= 23 (k) PSNR= 24 (l) Our PSNR= 27

(m) reference (n) patch with holes (o) PSNR= 25 (p) PSNR= 25 (q) PSNR= 25 (r) Our PSNR= 28

(s) reference (t) patch with holes (u) PSNR= 23 (v) PSNR= 23 (w) PSNR= 23.9 (x) Our PSNR= 27

Figure 16.10 – Results of using our re-trained generator to inpaint the dis-occluded regions. First column: reference patches; Second column: patches with dis-occluded regions; Third column: inpainted results using algorithm proposed in [99]; Forth column: inpainted results using algorithm proposed in [80]; Fifth column:inpainted results using algorithm proposed in [100]; Sixth column: inpaintd results using our retrained generator;

16.4 Conclusion

In this section, we proposed a GANs-based NR quality metric, GANs-NQM, to evaluate the perceptual quality of RGB-D synthesis views. To resolve the challenges of the training data scales in DNNs, a novel strategy is proposed which exploits the current existing large-scale 2D computer vision datasets rather than RGB-D datasets, where depth data may be unreliable. The spirit of the strategy can be easily applied to other applications in RGB-D domain or even other community. Based on the assumption that if a generator of a GANs could be trained to inpaint the dis-occluded regions then the discriminator could be used to predict the quality, in this study, we learned a 'Bag of Distortions Word' (BDW) codebook, proposed a local distortion region selector from the discriminator, and eventually mapped the non-uniform inpainting related artifacts to perceptual quality through SVR. According to experimental results, the proposed GANs-NQM provides the best performance compared to the state-of-the-art FR/NR quality metrics for RGB-D synthesized views. As a side outcome, the pre-trained inpainter also shows an appealing performance in inpainting the challenging holes in RGB-D synthesis view.

17

Conclusion of Part 4

In this part, higher-level representations have been explored for image quality assessment. Based on this study, two higher-level based models have been proposed, and certain research questions that posed in section 14.2 are answered. Based on the performance and complexity of the proposed models, some conclusions are made.

17.1 Answers to Research Questions

- ■ Higher-level representations of image/video for quality assessment in different tasks (Part IV)
  - □ A convolutional sparse coding based no reference image quality assessment metric (CSC-IQM) is proposed for capturing local structural non-uniform distortions in both the use-cases of FTV and VR.

    Referring to the fact that there is a 'sparse mode' in the human visual system that transmits low, mid-level image/video representations to higher semantic representations, where each item within the mode is 'semantic' relative, in section 15.2 and 15.3, a convolutional sparse coding no reference image quality metric is proposed. A set of convolutional 'items' are trained using manually labeled patches that contain non-natural structure according to specific applications, where each 'item' are related to a certain type of non-natural structure, e.g., double edges kernel. According to experimental results and visualized analysis, it is proven that the proposed metric is able to localize and quantify local non-uniform structure distortions.
  - □ An adjusted forward feature selection algorithm is proposed for quantifying the interplay relation among different structure related artifacts in the case of quality assessment of stitched panoramic images.

    Considering the fact that there are certain interplay relationships between different structure related distortions, i.e., the 'masking effect' happens when more than one type of structure related distortions appear at the same location. To handle and quantify this effect, a compound feature selection

algorithm is proposed in section 15.3.2 by selecting useful feature maps corresponding to different convolutional kernels (different kernels correspond to different distorted structures/non-natural structure). According to experimental results, the performance of CSC-IQM is improved with the incorporation of the proposed feature selection algorithm. Visualized results have also proven its capability of localizing distortions' overlap regions and quantifying the overlapping effects.

- A generative adversarial network based no reference image quality assessment metric (GAN-IQM) is proposed to quantify local non-uniform inpainting related artifacts.

  Discriminator in a generative adversarial network is trained to distinguish artificial images from the real one according to the goal of the task. In order to learn the effect of inpainting related artifact (causing changes of structure) on perceived quality, a new GANs model is retrained in chapter 16 to inpaint regions that similar to dis-occluded regions that are introduced in DIBR based synthesis procedure. As the goal is to well inpaint those regions, the discriminator is thus trained to tell whether the input image is good enough to be real. It is then assumed that the discriminator judges the input based on its quality. Based on which, the GA-IQM is further proposed by using the discriminator as higher-level feature extractors. According to the experimental results, the proposed metric is able to quantify local inpainting structure related distortions.

17.2 Performance summary and discussion

The performance and executing time of the proposed higher-level representation based models on all the tested datasets are summarized in Table 17.1 and 17.2 respectively.

Both CSC-IQM and GAN-IQM have been tested on the IVC-Image dataset for evaluating the quality of synthesized frames in FTV scenario. The performance of CSC-IQM is slightly higher than the one of GAN-IQM (PCC value). However, GAN-IQM is much faster regarding complexity since it mainly needs to fit the input image (a set of patches) into the pre-trained discriminator while CSC-IQM needs to optimize the features maps for the input image with the pre-trained codebook. CSC-IQM is more visual friendly than GAN-IQM since the learned convolutional kernels can be easily visualized to show the learned un-natural structure. Both of these two higher-level representation based models maintain well local information. For CSC-IQM, both global and local information can be accessed easily with the feature maps. Since each kernel corresponds to one potential type of non-natural structure, this type of local non-uniform distortion could be localized with the pixels coordinates activated with the corresponding kernel. With this advantage, one can then further check the interplay relations among different types of non-uniform structure distortions with the feature maps as shown in section 15.3.2. For GAN-IQM, an image is first divided into patches, those patches are then fitted to the pre-trained discriminator and are further assigned to a particular 'word' (in the trained 'bag of distortions word' codebook) with the extracted high-level feature. As thus, whether local patches contain severe inpainting related distortion could be indicated by either the normalized output of the discriminator or the 'word' within the trained 'bag of distortions word' codebook that the patches belong to.

Compared to low, mid-level representation based models presented in the previous parts, the two higher-level representation based models proposed are of higher representation capability, which generate representations that quantify directly the amount of structure disruptions (i.e. (1) number of activated pixels where non-uniform structure are located at in CSC-IQM; (2) number of patches contain non-uniform distortions in GAN-IQM).

More importantly, with better representative power, distance measures are not required in these two models, and they are no reference metrics. The performances of these two metrics are comparable to the best performing mid-level model ST-IQM. In terms of complexity, GAN-IQA is faster than the mid-level based metrics proposed in this thesis and slightly slower than the low-level ones. Since they are newly developed metrics, they will be further extended for video quality assessment in the future.

Table 17.1 – Summarization of performance of higher-level representation based models

PCC		Higher-level	
Related Task	Related Database	CSC-IQM	GAN-IQM
IQM in FTV	IVC-Image	0.830	0.826
IQM of Stitched panoramic image	SIQA	0.857	

Table 17.2 – Summarization of executing time of higher-level representation based models

Normalized time		Higher-level	
Related Task	Related Database	CSC-IQM	GAN-IQM
IQM in FTV	IVC-Image	985	157
IQM of Stitched panoramic image	SIQA	672	

Higher-level representation based models are of better capacity in representing images/videos. Compared to low, mid-level representation based models, higher-level representation based models learn to represent and quantify structure related distortions by taking advantage of the characteristics of advanced machine learning based models according to the task. As thus, representations obtained from these models have a stronger link to perceived quality in terms of showing the portion of regions that contain non-natural structure. Unlike other 'black box' deep learning based models, the learning procedures of these two models are highly related to the task (task-oriented), e.g., in order to train a discriminator to reveal the quality of inpainted images, the GANs network is designed as a 'context inpainter' with designed dis-occluded regions to be inpainted. Therefore, the promising performance is guaranteed by how the learning process is designed according to the task. As mentioned in section 3.3, natural scene statistics (NSS) based models fail to handle local geometric distortions, these two models meet this need by learning the non-natural structure (NNS). Last but not least, higher-level representation based model GAN-IQM is efficient enough to be executed in real-time.

18

Conclusions and Perspectives

In this dissertation, a research effort has been conducted to explore different levels of image/video representations, referring to human visual representation mechanism, for image/video quality assessment in immersive multimedia applications. In those applications, local non-uniform structure-related distortions are one of the toughest distortions that commonly used metrics fail to deal with.

As all of the proposed metrics have been tested on the IVC-Image dataset, their performances (in terms of PCC value) and the execution time (in terms of normalized time with the executing time of PSNR) are summarized in Table 18.1. The following conclusions are drawn based on their performance and complexity (Table 18.1).

Table 18.1 – Summarization of performance of all the proposed models on IVC-Image database

	Low-level		Mid-level		Higher-level	
	BF-M	EM-IQM	ST-IQM	CT-IQM	CSC-IQM	GAN-IQM
PCC	0.6980	0.7430	0.8219	0.6809	**0.8302**	0.8262
Executing time	**17**	127	1324	458	985	157

The proposed low-level representation based models (Part II) evaluate the quality/utility of image based on the dissimilarities between low-level representations of the reference and the distorted images/videos. These low-level representations ('white box' approaches) show almost no intermediate visual pattern or any semantics related to the tasks. As a result, the performances of the two proposed low-level representation based models are less promising compared to other metrics (except for CT-IQM). However, simplicity is one of the advantages of these two models. As shown in Table 18.1, BF-M and EM-IQM are two of the most effective metrics among all the proposed models. Since BF-M is proposed to check the role of structure and texture information in different tasks, its advantages outweigh its disadvantages in such a case (e.g., used as a guide for improving another learning-based model). Last but not least, EM-IQM outperforms one of the mid-level representation based model CT-IQM. As discussed before, it is because 1) sensitive regions selection process is incorporated 2) elastic metric is suitable for quantifying geometric distortions in FTV use case. One of the

most important conclusion that could be made from this part is that: if the right distance measures are selected according to the problem that needs to be solved, low-level representation based models could be powerful too.

The proposed mid-level representation based models (Part III) evaluate the quality of images/videos by checking how intermediate patterns (e.g., contours' categories and entropies of contours) change. The proposed mid-level representation based models are of higher representative capability compared to the low-level ones. According to Table 18.1, ST-IQM obtains one of the best performance among all proposed full reference metrics. However, CT-IQM is the worse performing metric. It proves that if a mid-level representation is not rich enough regarding its capability to represent meaningful information according to the task, it could not guarantee better performances compared to low-level representation based models. In the case of quality assessment of synthesized views, observers are sensitive to degradations of structure, ST-IQM is able to quantify changes of contours from a higher semantic level than CT-IQM. One of the reasons why CT-IQM is not performing well could be that for certain types of contours, they may have the same entropy. For example, for 'T' and 'L' shape contours, their frequencies of occurrence in the an image could be the equal, which may result in same entropy for the two types of contours. Furthermore, the complexity of these two models are higher compared to others since registration stage (in ST-IQM) or learning stage (i.e., context tree learning in CT-IQM) is involved.

The proposed higher-level representation based models (Part IV) evaluate the quality of images/videos by learning to represent and quantify structure-related distortions according to the tasks. In this study, the two proposed higher-level representation based models are trained to learn the non-natural structure within images. Thus, they are of better ability in representing images/videos for the task. In other words, representations obtained from these models have a stronger connection with the perceived quality of images that contain structure-related distortions. Unlike other 'black box' deep learning based models, the learning procedure is highly task-oriented. The promising performances are guaranteed by how the learning process is designed according to the task. Unlike natural scenes statistics (NSS) based models, these two models are capable of predicting the perceived quality of images/videos in immersive multimedia use cases by learning the non-natural structures (NNS). According to Table 18.1, GAN-IQM is the second lest time-consuming model, which confirms that using more a complex network does not necessarily lead to higher complexity.

Comparisons of different level representation based models:

The proposed low-level models are the least representative models, while the higher-level representation based models are the most representative ones. Low, mid-level representation based models proposed in this study are not directly linked to quality. They rely on using certain distance measures to compute the final quality scores with the images/videos representations. Unlike higher-level models, even the learning process is involved in these models, they are not trying to learn the distortion/unnaturalness. Compared to low, mid-level based models, higher-level representation based model learn to represent and quantify distortions by designing the learning process according to the tasks.

18.1 Perspectives

Throughout the thesis, different level of visual representations have been explored for image/video quality/utility assessment in different use cases. The final goal of this dissertation is to highlight insights obtained during the exploration and the potential usages/improvements of the proposed models.

Better usage of distance measures: Considering the fact that low-level representation based model EM-

IQM outperforms mid-level representation based model CT-IQM, it is essential to use proper distance measures according to the characteristics of the task.

'White box' vs. 'Black box' approaches: On one hand, the 'black box' methods do not necessarily outperform 'white box' methods. For example, ST-IQM obtains performance close to GAN-IQM. The key point is to design and learn a model which is suitable for the task. On another hand, the 'black box' method does not necessarily mean less comprehensible. For example, the learned kernels of CSC-IQM model are visualized friendly, where most of the learned kernels correspond to certain types of non-natural structures, e.g., 'double line shape' curve. One should choose and design a model properly according to the characteristics of the task.

Take advantages of 'Black Box' method by designing the learning process carefully according to the task: GAN-IQM model is a successful example of using deep learning approach by mimicking 'similar problems' for the 'black box' to solve. Compared to those deep learning based quality models, which fine-tunes an existing deep net with MOS as label, our proposed model trains the network by forcing it to accomplish the same task by generating artificially generated training samples (e.g., it is forced to well inpaint the artificial dis-occluded regions). By doing so, the model is more task-oriented and can obtain more promising and understandable performance.

Using deep learning to speed up optimization process: As reported before, the cumbersome optimization process is one of the main reason that makes those learning based models slow. For instance, for the proposed CSC-IQM, the optimization procedure of getting the feature maps with a learned codebook could be improved significantly by employing deep learning based unrolling methods (e.g., algorithm proposed in [219]).

Combining different level representation based models: In this work, different level representation based models are proposed separately. As concluded above, they are of different superiorities depending on the use cases. Therefore, combining different level representation based models according to the characteristics of the task may yield a more robust hierarchical model.

Following is an example of combing BF-M with ST-IQM for the task of quality assessment of synthesized images in FTV scenario. As the primary goal of proposing BF-M is to explore the roles of structure and texture information in different tasks by first separating the structural and textural information. By using the bilateral filter, structures within an image can be emphasized, e.g., obtaining clearer contours at the same time better separating texture from the structure. One possible improvement is to equip BF-M with more powerful structure and texture features or measures. ST-IQM could be that structure measure.

In order to check how BF-M can be combined with ST-IQM and with another more advanced texture descriptors, we replace the simple descriptors in BF-M with ST-IQA and a perceptual inspired texture features that we proposed in [220].

More specifically, the ST-IQM is used to replace the structure-related estimator $BI\text{-}NICE$ and $BI\text{-}HOG$ as $BI\text{-}ST$, while the spatial contrast sensitivity (CSF) based texture descriptor [220] is used to replace the texture-related estimator $BI\text{-}LRI$ as $BI\text{-}CSF$. As thus, equation (7.4) can be modified as

$$\begin{aligned} &BF\text{-}M_{new} = 1 - [(\alpha_{BI} + \beta_{BI}) \cdot BI\text{-}ST + \gamma_{BI} \cdot BI\text{-}CSF] \\ &s.t. \;\; \alpha_{BI} + \beta_{BI} + \gamma_{BI} = 1. \end{aligned} \tag{18.1}$$

The performance of the $BF\text{-}M_{new}$ is summarized in Table 18.2. The optimized performance of $BF\text{-}M_{new}$ is obtained while $(\alpha_{BI} + \beta_{BI})$ is set as 0.8 and γ_{BI} as 0.2. This best-fit configuration is similar to the optimum

configuration of $BF\text{-}M$ as described in chapter 7, conclusion that structural information plays the major role in task of synthesized view quality evaluation is still solid in this case. It can be observed from the table that the performance of $BF\text{-}M_{new}$ with more advanced structural/texture estimators outperforms the one of $BF\text{-}M$. It can be concluded that the performances of the proposed models can be improved by combining different level representation based models proposed in this dissertation.

Table 18.2 – Performance of the improved BF-M.

	PCC	**SCC**	**RMSE**
ST-IQM	0.8217	0.7827	0.3233
BI-ST	0.8313	0.7956	0.3183
BI-CSF	0.7279	0.6409	0.4565
BF-M	0.6980	0.5885	0.4768
BF-M$_{new}$	0.8405	0.8072	0.3052

List of Publications

■ Conferences

- ☐ Yashas Rai, Ahmed Aldahdooh, Suiyi Ling, Marcus Barkowsky, Patrick Le Callet. Effect of content features on short-term video quality in the visual periphery. 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2016 [220].
- ☐ Suiyi Ling, Patrick Le Callet and Zitong Yu. The Role of Structure and Textural Information in Image Utility and Quality Assessment Tasks. Electronic Imaging 2018 [221].
- ☐ Suiyi Ling, and Patrick Le Callet. Image Quality Assessment for DIBR Synthesized Views using Elastic Metric. Proceedings of the 2017 ACM on Multimedia Conference (ACMMM). ACM, 2017 [222].
- ☐ Suiyi Ling, and Patrick Le Callet. Image quality assessment for free viewpoint video based on mid-level contours feature. 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017 [177].
- ☐ Suiyi Ling, Patrick Le Callet, and Gene Cheung. Quality assessment for synthesized view based on variable-length context tree. 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2017 [223].
- ☐ Suiyi Ling, Patrick Le Callet. How to learn the effect of non-uniform distortion on Perceived Visual Quality ? Case study using Convolutional Sparse Coding for quality assessment of synthesized views. 2018 IEEE International Conference on Image Processing (ICIP). IEEE, 2018 [224].
- ☐ Suiyi Ling, Gene Cheung, Patrick Le Callet. No Reference Quality Assessment for Stitched Panoramic Images Using Convolutional Sparse Coding and Compound Feature Selection. 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018 [63].
- ☐ Li J, Mantiuk R, Wang J, Ling S, Le Callet P. Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation. InAdvances in neural information processing systems 2018 [225].
- ☐ Ling, Suiyi, Jing Li, Patrick Le Callet, and Junle Wang. "Perceptual representations of structural information in images: application to quality assessment of synthesized view in ftv scenario." In 2019 IEEE International Conference on Image Processing (ICIP), pp. 1735-1739 [226].

■ Journals

- ☐ Suiyi Ling, Patrick Le Callet and Zitong Yu. The Role of Structure and Textural Information in Image Utility and Quality Assessment Tasks. Journal of Perceptual Imaging 2018.
- ☐ Ling, S., Gutiérrez, J., Gu, K. and Le Callet, P., 2019. Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 9(1), pp.204-216 [227].
- ☐ Zhou, Yu, Leida Li, Suiyi Ling, and Patrick Le Callet. "Quality assessment for view synthesis using low-level and mid-level structural representation." Signal Processing: Image Communication 74 (2019): 309-321 [228].
- ☐ Ling, Suiyi, Jing Li, Zhaohui Che, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. "Quality Assessment of Free-viewpoint Videos by Quantifying the Elastic Changes of Multi-Scale Motion Trajectories." arXiv preprint arXiv:1903.12107 (2019) [229].
- ☐ Ling, Suiyi, Jing Li, Junle Wang, and Patrick Le Callet. "GANs-NQM: A generative adversarial networks based no reference quality assessment metric for RGB-D synthesized views." arXiv preprint

arXiv:1903.12088 (2019) [230].

Bibliography

[1] M. Tanimoto, "Ftv: Free-viewpoint television," *Signal Processing: Image Communication*, vol. 27, no. 6, pp. 555–570, 2012. 15, 37

[2] C. Fehn, R. De La Barre, and S. Pastoor, "Interactive 3-dtv-concepts and key technologies," *Proceedings of the IEEE*, vol. 94, no. 3, pp. 524–538, 2006. 15

[3] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 93–104. 15, 38

[4] K. Gu, V. Jakhetiya, J.-F. Qiao, X. Li, W. Lin, and D. Thalmann, "Model-based referenceless quality metric of 3d synthesized images using local image description," *IEEE Transactions on Image Processing*, 2017. 15, 34, 135

[5] M. Xu, C. Li, Z. Wang, and Z. Chen, "Visual quality assessment of panoramic video," *arXiv preprint arXiv:1709.06342*, 2017. 16

[6] R. Konrad, E. A. Cooper, and G. Wetzstein, "Novel optical configurations for virtual reality: evaluating user preference and performance with focus-tunable and monovision near-eye displays," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 1211–1220. 16

[7] S. A. Golestaneh and L. J. Karam, "Synthesized texture quality assessment via multi-scale spatial and statistical texture attributes of image and gradient magnitude coefficients," *arXiv preprint arXiv:1804.08020*, 2018. 16, 32, 189

[8] M. Wang, B. Yan, and K. N. Ngan, "An efficient framework for image/video inpainting," *Signal Processing: Image Communication*, vol. 28, no. 7, pp. 753–762, 2013. 16

[9] T. Matsuyama and T. Takai, "Generation, visualization, and editing of 3d video," in *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*. IEEE, 2002, pp. 234–245. 16

[10] T. N. Pappas, J. Zujovic, and D. L. Neuhoff, "Image analysis and compression: Renewed focus on texture," in *Visual Information Processing and Communication*, vol. 7543. International Society for Optics and Photonics, 2010, p. 75430N. 16

[11] B. Furht, S. W. Smoliar, and H. Zhang, *Video and image processing in multimedia systems*. Springer Science & Business Media, 2012, vol. 326. 16

[12] A. R. Rao, *A taxonomy for texture description and identification*. Springer Science & Business Media, 2012. 16

[13] B. B. Chaudhuri and N. Sarkar, "Texture segmentation using fractal dimension," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 17, no. 1, pp. 72–77, 1995. 16

[14] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transactions on circuits and systems for video technology*, vol. 11, no. 6, pp. 703–715, 2001. 16

[15] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 341–346. 16

[16] I. I. Groen, E. H. Silson, and C. I. Baker, "Contributions of low-and high-level properties to neural processing of visual scenes in the human brain," *Phil. Trans. R. Soc. B*, vol. 372, no. 1714, p. 20160102, 2017. 17, 191

[17] T. J. Andrews, D. M. Watson, G. E. Rice, and T. Hartley, "Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway," *Journal of Vision*, vol. 15, no. 7, pp. 3–3, 2015. 17, 191

[18] J. W. Peirce, "Understanding mid-level representations in visual processing," *Journal of Vision*, vol. 15, no. 7, pp. 5–5, 2015. 17, 18

[19] R. Veale, Z. M. Hafed, and M. Yoshida, "How is visual salience computed in the brain? insights from behaviour, neurobiology and modelling," *Phil. Trans. R. Soc. B*, vol. 372, no. 1714, p. 20160113, 2017. 17

[20] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust, "Do we know what the early visual system does?" *Journal of Neuroscience*, vol. 25, no. 46, pp. 10 577–10 597, 2005. 18

[21] A. E. Welchman, A. Deubelius, V. Conrad, H. H. Bülthoff, and Z. Kourtzi, "3d shape perception from combined depth cues in human visual cortex," *Nature neuroscience*, vol. 8, no. 6, p. 820, 2005. 18

[22] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon, "A functional and perceptual signature of the second visual area in primates," *Nature neuroscience*, vol. 16, no. 7, p. 974, 2013. 18

[23] A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, J. L. Gallant, H. Lu, and W. Vanduffel, "Toward a unified theory of visual area v4," *Neuron*, vol. 74, no. 1, pp. 12–29, 2012. 18

[24] S.-M. Khaligh-Razavi and N. Kriegeskorte, "Deep supervised, but not unsupervised, models may explain it cortical representation," *PLoS computational biology*, vol. 10, no. 11, p. e1003915, 2014. 18

[25] U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 005–10 014, 2015. 18

[26] K. Ramakrishnan, H. S. Scholte, I. I. Groen, A. W. Smeulders, and S. Ghebreab, "Visual dictionaries as intermediate features in the human brain," *Frontiers in computational neuroscience*, vol. 8, p. 168, 2015. 18

[27] N. Kanwisher and D. D. Dilks, "The functional organization of the ventral visual pathway in humans," *The new visual neurosciences*, pp. 733–748, 2013. 18

[28] P. Foldiak, "Sparse coding in the primate cortex," *The handbook of brain theory and neural networks*, 2003. 18, 130

[29] C. G. Gross, C. d. Rocha-Miranda, and D. B. Bender, "Visual properties of neurons in inferotemporal cortex of the macaque." *Journal of neurophysiology*, vol. 35, no. 1, pp. 96–111, 1972. 18, 130

[30] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*. IEEE, 2017, pp. 1–7. 19

[31] M. Tanimoto, T. Fujii, and K. Suzuki, "View synthesis algorithm in view synthesis reference software 2.0 (vsrs2. 0)," *ISO/IEC JTC1/SC29/WG11 M*, vol. 16090, p. 2009, 2009. 21, 22, 191

[32] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3d warping using depth information for ftv," *Signal Processing: Image Communication*, vol. 24, no. 1, pp. 65–72, 2009. 22, 38

[33] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3d synthesized views," *Signal Processing: Image Communication*, vol. 30, pp. 78–88, 2015. 23, 33, 111, 112, 135

[34] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 2. IEEE, 2007, pp. II–169. 23, 25, 34, 73, 154

[35] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the multiview and 3d extensions of high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 35–49, 2016. 24

[36] L. Yang, Z. Tan, Z. Huang, and G. Cheung, "A content-aware metric for stitched panoramic image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2487–2494. 25, 32, 35, 39, 141, 142, 143, 189

[37] D. M. Rouse, R. Pepion, S. S. Hemami, and P. Le Callet, "Image utility assessment and a relationship with image quality assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009, pp. 724 010–724 010. 26, 40, 55, 57, 63, 64

[38] D. Siddalinga Swamy, "Quality assessment of synthesized textures," Ph.D. dissertation, Oklahoma State University, 2011. 26, 36, 65, 66

[39] S. Varadarajan and L. J. Karam, "A reduced-reference perceptual quality metric for texture synthesis," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 531–535. 27, 40, 63, 65

[40] S. A. Golestaneh, M. M. Subedar, and L. J. Karam, "The effect of texture granularity on texture synthesis quality," *SPIE Optical Engineering+ Applications*, pp. 959 912–959 912, 2015. 27, 40, 63, 65

[41] S. Varadarajan and L. J. Karam, "A no-reference perceptual texture regularity metric," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1894–1898. 27, 40, 63, 65

[42] S. Tian, L. Zhang, L. Morin, and O. Déforges, "Niqsv+: A no-reference synthesized view quality assessment metric," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1652–1664, 2018. 30, 34, 46

[43] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008. 30, 31, 32, 33

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 30, 31, 32, 33

[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. Ieee, 2003, pp. 1398–1402. 30, 31, 32, 33, 64, 65

[46] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011. 30

[47] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a" completely blind" image quality analyzer." *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013. 30, 32, 33

[48] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012. 30, 32, 33, 143

[49] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-d synthesized view assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332–1343, 2011. 30, 33, 38, 63, 67, 77, 99, 132, 135, 149, 151, 156

[50] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE transactions on image processing*, vol. 16, no. 9, pp. 2284–2298, 2007. 30, 31, 33

[51] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005. 30, 31, 32

[52] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006. 30, 31, 32, 64, 65

[53] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE transactions on image processing*, vol. 9, no. 4, pp. 636–650, 2000. 30, 31, 33

[54] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Proceedings of the third international workshop on video processing and quality metrics*, vol. 4, 2007. 30, 31

[55] "Video quality metric (vqm) software," http://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx. 30, 31

[56] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004. 30, 31

[57] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE transactions on image processing*, vol. 9, no. 4, pp. 636–650, 2000. 30, 31

[58] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002. 30, 31, 32

[59] E. Bosc, R. Pépion, P. Le Callet, M. Köppel, P. Ndjiki-Nya, L. Morin, and M. Pressigout, "Perceived quality of dibr-based synthesized views," in *Applications of Digital Image Processing XXXIV*, vol. 8135. International Society for Optics and Photonics, 2011, p. 81350I. 31, 38, 134, 189

[60] X. Liu, Y. Zhang, S. Hu, S. Kwong, C.-C. J. Kuo, and Q. Peng, "Subjective and objective video quality assessment of 3d synthesized views with texture/depth compression distortion," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4847–4861, 2015. 30, 31, 35, 111, 112, 189

[61] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2010. 31

[62] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed depth data," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt, Germany, Jul. 2013, pp. 100–105. 31, 32, 38, 39, 134, 189

[63] C. G. Ling, Suiyi and P. Le Callet, "No reference quality assessment for stitched panoramic images using convolutional sparse coding and compound feature selection," in *Multimedia and Expo (ICME), 2018 IEEE International Conference on*. IEEE, 2018. 32, 171

[64] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011. 32, 33, 143

[65] S. A. Golestaneh and L. J. Karam, "Reduced-reference synthesized-texture quality assessment based on multi-scale spatial and statistical texture attributes," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3783–3786. 32, 36, 65, 66, 189

[66] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2545–2558, 2013. 32

[67] D. M. Rouse and S. S. Hemami, "Natural image utility assessment using image contours," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 2217–2220. 32, 36, 64, 65

[68] E. T. Scott and S. S. Hemami, "Image utility estimation using difference-of-gaussian scale space," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 101–105. 32, 36, 64, 65

[69] T. D. Kite, B. L. Evans, and A. C. Bovik, "Modeling and quality assessment of halftoning by error diffusion," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 909–922, 2000. 33

[70] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740. 33

[71] D. Li, T. Jiang, and M. Jiang, "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 378–386. 33, 130, 152

[72] P. H. Conze, "Objective view synthesis quality assessment," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8288, p. 53, 2012. 33

[73] C.-T. Tsai and H.-M. Hang, "Quality assessment of 3d synthesized views with depth map distortion," in *Visual Communications and Image Processing (VCIP), 2013*. IEEE, 2013, pp. 1–6. 33

[74] D. Sandić-Stanković, D. Kukolj, and P. L. Callet, "Dibr synthesized image quality assessment based on morphological wavelets," in *International Workshop on Quality of Multimedia Experience*, 2015, pp. 1–4. 33, 68, 77, 100, 112, 121, 135

[75] ——, "Dibr-synthesized image quality assessment based on morphological multi-scale approach," *Eurasip Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–23, 2016. 33, 68, 77, 100, 112, 121, 135

[76] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Dibr-synthesized image quality assessment based on morphological multi-scale approach," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 4, 2016. 33, 67, 68, 77, 111, 112, 121

[77] S. Tian, L. Zhang, L. Morin, and O. Deforges, "Niqsv: A no reference image quality assessment metric for 3d synthesized views," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1248–1252. 34, 135

[78] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE journal of selected topics in signal processing*, vol. 3, no. 2, pp. 193–201, 2009. 34, 67

[79] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004. 35, 38, 67, 191

[80] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand, "Depth image-based rendering with advanced texture synthesis for 3-d video," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 453–465, 2011. 35, 38, 159, 160, 191, 194

[81] Y. Zhao and L. Yu, "A perceptual metric for evaluating quality of synthesized sequences in 3dv system," in *Proc. SPIE*, vol. 7744, 2010, p. 77440X. 34

[82] E. Ekmekcioglu, S. Worrall, D. De Silva, A. Fernando, and A. M. Kondoz, "Depth based perceptual quality assessment for synthesised camera viewpoints," in *International Conference on User Centric Media*. Springer, 2010, pp. 76–83. 34

[83] S. Leorin, L. Lucchese, and R. G. Cutler, "Quality assessment of panorama video for videoconferencing applications," in *IEEE 7th Workshop on Multimedia Signal Processing, 2005*. IEEE, 2005, pp. 1–4. 35

[84] P. Paalanen, J.-K. Kämäräinen, and H. Kälviäinen, "Image based quantitative mosaic evaluation with artificial video," *Image Analysis*, pp. 470–479, 2009. 35

[85] W. Xu and J. Mulligan, "Performance evaluation of color correction approaches for automatic multi-view image and video stitching," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 263–270. 35

[86] M. Solh and G. AlRegib, "Miqm: A novel multi-view images quality measure," in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*. IEEE, 2009, pp. 186–191. 35, 142, 143

[87] H. Qureshi, M. Khan, R. Hafiz, Y. Cho, and J. Cha, "Quantitative quality assessment of stitched panoramic images," *IET Image Processing*, vol. 6, no. 9, pp. 1348–1358, 2012. 35, 142

[88] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2. IEEE, 2005, pp. ii–573. 36, 65, 66

[89] A. C. Brooks, X. Zhao, and T. N. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," *IEEE Transactions on image processing*, vol. 17, no. 8, pp. 1261–1273, 2008. 36, 65, 66

[90] F. Battisti and P. Le Callet, "Quality Assessment in the context of FTV: challenges, first answers and open issues," *IEEE COMSOC MMTC Communications - Frontiers*, vol. 11, no. 2, pp. 22–26, Mar. 2016. 37

[91] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux, P. T. Kovács, and V. K. Adhikarla, "Subjective evaluation of Super Multi-View compressed contents on high-end light-field 3D displays," *Signal Processing: Image Communication*, vol. 39, pp. 369–385, Nov. 2015. 38

[92] O. Stankiewicz, K. Wegner, T. Senoh, G. Lafruit, V. Baroncini, and M. Tanimoto, "Revised summary of call for evidence on free-viewpoint television: Super-multiview and free navigation," *ISO/IEC JTC1/SC29/WG11 MPEG2016/N16523*, Oct. 2016. 38

[93] P. Carballeira, J. Gutiérrez, F. Morán, J. Cabrera, and N. García, "Subjective evaluation of super multiview video in consumer 3d displays," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, Costa Navarino, Greece, May 2015, pp. 1–6. 38, 39, 103, 193

[94] P. Carballeira, J. Gutiérrez, F. Morán, J. Cabrera, F. Jaureguizar, and N. García, "Multiview perceptual disparity model for super multiview video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 113–124, 2017. 38, 103, 193

[95] R. Recio, P. Carballeira, J. Gutierrez, and N. Garcia, "Subjective Assessment of Super Multiview Video with Coding Artifacts," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 868–871, Jun. 2017. 38

[96] A. T. Hinds, D. Doyen, and P. Carballeira, "Toward the realization of six degrees-of-freedom with compressed light fields," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, China, Jul. 2017, pp. 1171–1176. 38, 101

[97] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "Visual Quality Assessment of Synthesized Views in the Context of 3D-TV," in *3D-TV System with Depth-Image-Based Rendering*. New York, NY: Springer New York, 2013, pp. 439–473. 38

[98] "Irccyn ivc dibr database website," ftp://ftp.ivc.polytech.univnantes.fr/IRCCyN_IVC_DIBR_Images/. 38, 63, 67, 77, 99

[99] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3d video systems," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–11, 2009. 38, 159, 160, 194

[100] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering," pp. 1809–1812, 2010. 38, 159, 160, 194

[101] "3d-htm," http://hevc.hhi.fraunhofer.de/. 39

[102] "Jm," http://iphome.hhi.de/suehring/tml/. 39

[103] "Hm," http://hevc.hhi.fraunhofer.de/. 39

[104] "Kakadu," http://www.kakadusoftware.com/. 39

[105] J. Gautier, O. Le Meur, and C. Guillemot, "Efficient depth map compression based on lossless edge coding and diffusion," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012, pp. 81–84. 39

[106] F. Pasteau, C. Strauss, M. Babel, O. Déforges, and L. Bédat, "Adaptive color decorrelation for predictive image codecs," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 1100–1104. 39

[107] E. Bosc, "Compression of multi-view-plus-depth (mvd) data: from perceived quality analysis to mvd coding tools designing," Ph.D. dissertation, INSA de Rennes, 2012. 39

[108] I. Tutorial, "Objective perceptual assessment of video quality: Full reference television," *ITU-T Telecommunication Standardization Bureau*, 2004. 41

[109] I. Recommendation, "1401,"methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," rec," *Union Std*, 2012. 41

[110] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, pp. 1–6. 41, 42, 43, 44, 112, 191

[111] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data?" in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. Ieee, 2016, pp. 1–6. 41, 42, 43, 44, 112, 141, 191

[112] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: new methods from atis/t1a1," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 101–107, 2004. 43

[113] G. Barnard, "A new test for 2× 2 tables," *Nature*, vol. 156, p. 177, 1945. 43

[114] J. LI and P. L. CALLET, "Improving the discriminability of standard subjective quality assessment methods: a case study." in *QoMEX*, 2018. 45

[115] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Data Compression Conference (DCC), 2017*. IEEE, 2017, pp. 52–61. 45

[116] D. M. Watson, T. Hartley, and T. J. Andrews, "Patterns of response to visual scenes are linked to the low-level properties of the image," *NeuroImage*, vol. 99, pp. 402–410, 2014. 53

[117] M. R. Greene and A. Oliva, "The briefest of glances: The time course of natural scene understanding," *Psychological Science*, vol. 20, no. 4, pp. 464–472, 2009. 53

[118] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive psychology*, vol. 9, no. 3, pp. 353–383, 1977. 53

[119] J. A. Movshon and E. P. Simoncelli, "Representation of naturalistic image structure in the primate visual cortex," in *Cold Spring Harbor symposia on quantitative biology*, vol. 79. Cold Spring Harbor Laboratory Press, 2014, pp. 115–122. 54, 60

[120] L. W. Renninger and J. Malik, "When is scene identification just texture recognition?" *Vision research*, vol. 44, no. 19, pp. 2301–2311, 2004. 54

[121] J. R. Bergen and B. Julesz, "Rapid discrimination of visual patterns," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 5, pp. 857–863, 1983. 54

[122] B. Julesz, "Texton gradients: The texton theory revisited," *Biological cybernetics*, vol. 54, no. 4-5, pp. 245–251, 1986. 54

[123] J. Aloimonos, "Shape from texture," *Biological cybernetics*, vol. 58, no. 5, pp. 345–360, 1988. 54

[124] J. R. Kender, "Shape from texture: An aggregation transform that maps a class of textures into surface orientation," in *Proceedings of the 6th international joint conference on Artificial intelligence-Volume 1*. Morgan Kaufmann Publishers Inc., 1979, pp. 475–480. 54

[125] C. Li and A. C. Bovik, "Three-component weighted structural similarity index," in *Proc. sPIE*, vol. 7242, 2009, pp. 1–9. 54

[126] L. Sharan, C. Liu, R. Rosenholtz, and E. H. Adelson, "Recognizing materials using perceptually inspired features," *International journal of computer vision*, vol. 103, no. 3, pp. 348–371, 2013. 57, 58, 60, 61, 62

[127] L. Xu, W. Lin, L. Ma, Y. Zhang, Y. Fang, K. N. Ngan, S. Li, and Y. Yan, "Free-energy principle inspired video quality metric and its use in video coding," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 590–602, 2016. 58

[128] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962. 60

[129] ——, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968. 60

[130] S. L. Brincat and C. E. Connor, "Underlying principles of visual shape selectivity in posterior inferotemporal cortex," *Nature neuroscience*, vol. 7, no. 8, p. 880, 2004. 60

[131] N. J. Priebe and D. Ferster, "Mechanisms of neuronal computation in mammalian visual cortex," *Neuron*, vol. 75, no. 2, pp. 194–208, 2012. 60

[132] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex." *Nature neuroscience*, vol. 2, no. 11, 1999. 60

[133] S. Bae, S. Paris, and F. Durand, "Two-scale tone management for photographic look," in *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 637–645. 60

[134] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *International journal of computer vision*, vol. 81, no. 1, pp. 24–52, 2009. 61

[135] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *ACM transactions on graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 257–266. 61

[136] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1270–1281, 2008. 61

[137] I. Biederman and G. Ju, "Surface versus edge-based determinants of visual recognition," *Cognitive psychology*, vol. 20, no. 1, pp. 38–64, 1988. 61, 62

[138] J. De Winter and J. Wagemans, "Contour-based object identification and segmentation: Stimuli, norms and data, and software tools," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 4, pp. 604–624, 2004. 61

[139] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893. 62

[140] Y. Zhai, D. L. Neuhoff, and T. N. Pappas, "Local radius index-a new texture similarity feature," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 1434–1438. 62

[141] S. Golestaneh and L. J. Karam, "Reduced-reference quality assessment based on the entropy of dwt coefficients of locally weighted gradient magnitudes," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5293–5303, 2016. 64, 65

[142] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010. 64, 65

[143] D. Sandic-Stankovic, D. Kukolj, and P. Le Callet, "Dibr synthesized image quality assessment based on morphological pyramids," in *2015 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON).* IEEE, 2015, pp. 1–4. 67, 77, 111

[144] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Dibr synthesized image quality assessment based on morphological wavelets," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on.* IEEE, 2015, pp. 1–6. 67, 74, 77, 99, 109, 111

[145] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008. 72, 73, 109

[146] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012. 72

[147] ——, "Slic superpixels," Tech. Rep., 2010. 72, 74, 149

[148] W. Mio, A. Srivastava, and S. Joshi, "On shape of plane elastic curves," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 307–324, 2007. 75, 76

[149] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, 2011. 75, 76

[150] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013. 79

[151] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 500–513, 2011. 80

[152] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference.* BMVA Press, 2009, pp. 124–1. 80

[153] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8. 80

[154] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision.* Springer, 2006, pp. 428–441. 80

[155] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 54, 2013. 82, 135, 156

[156] M. Narwaria, "Toward better statistical validation of machine learning-based multimedia quality estimators," *IEEE Transactions on Broadcasting*, 2018. 82

[157] E. Siahaan, A. Hanjalic, and J. A. Redi, "Augmenting blind image quality assessment using image semantics," in *Multimedia (ISM), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 307–312. 82, 135

[158] J. Kubilius, J. Wagemans, and H. P. O. de Beeck, "Encoding of configural regularity in the human visual system," *Journal of Vision*, vol. 14, no. 9, pp. 11–11, 2014. 91

[159] F. Attneave, "Some informational aspects of visual perception." *Psychological review*, vol. 61, no. 3, p. 183, 1954. 91

[160] Z. Kourtzi and C. E. Connor, "Neural representations for object perception: structure, category, and adaptive coding," *Annual review of neuroscience*, vol. 34, pp. 45–67, 2011. 91

[161] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt, "A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization." *Psychological bulletin*, vol. 138, no. 6, p. 1172, 2012. 91

[162] G. Rhodes and L. Jeffery, "Adaptive norm-based coding of facial identity," *Vision research*, vol. 46, no. 18, pp. 2977–2987, 2006. 91

[163] D. A. Leopold, A. J. O'Toole, T. Vetter, and V. Blanz, "Prototype-referenced shape encoding revealed by high-level aftereffects," *Nature neuroscience*, vol. 4, no. 1, p. 89, 2001. 91

[164] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3158–3165. 95, 98, 109

[165] J. N. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on*. IEEE, 2009, pp. 819–822. 96, 97

[166] A. A. Goshtasby, *2-D and 3-D image registration: for medical, remote sensing, and industrial applications*. John Wiley & Sons, 2005. 97

[167] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009. 98

[168] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8. 98

[169] VQEG, "Report on the Validation of Video Quality Models for High Definition Video Content," Jun. 2010. 102, 105, 111

[170] G. Lafruit, K. Wegner, and M. Tanimoto, "Call for evidence on free-viewpoint television: Super-multiview and free navigation," *MPEG N15348, Warsaw*, 2015. 103, 104

[171] "Nagoya university sequences," http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/. 104

[172] K. Wegner and O. Stankiewicz, "Ders software manual," *ISO/IEC JTC1/SC29/WG11 M*, vol. 34302. 104

[173] G. Lafruit, K. Wegner, T. Grajek, T. Senoh, P. Kovács, P. Goorts, L. Jorissen, B. Ceulemans, P. C. Lopez, S. G. Lobo *et al.*, "Ftv software framework," *MPEG N15349, Warsaw*, 2015. 104

[174] U. Engelke and P. Le Callet, "Perceived interest and overt visual attention in natural images," *Signal Processing: Image Communication*, vol. 39, pp. 386–404, 2015. 105

[175] ITU, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," *ITU-T Recommendation P.913*, 2014. 105

[176] ——, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT.500*, 2012. 105

[177] S. Ling and P. Le Callet, "Image quality assessment for free viewpoint video based on mid-level contours feature," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on.* IEEE, 2017, pp. 79–84. 113, 171

[178] A. Zheng, G. Cheung, and D. Florencio, "Context tree-based image contour coding using a geometric prior," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 574–589, 2016. 115, 116, 117, 118, 119, 120, 121

[179] H. Freeman, "Application of the generalized chain coding scheme to map data processing," 1978. 115, 118

[180] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order markov models," *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 385–421, 2012. 115

[181] J. Rissanen, "A universal data compression system," *IEEE Transactions on information theory*, vol. 29, no. 5, pp. 656–664, 1983. 115

[182] C. C. Lu and J. G. Dunham, "Highly efficient coding schemes for contour lines based on chain code representations," *IEEE Transactions on Communications*, vol. 39, no. 10, pp. 1511–1514, 1991. 115

[183] I. Daribo, D. Florencio, and G. Cheung, "Arbitrarily shaped motion prediction for depth video compression using arithmetic edge coding." *Image Processing IEEE Transactions on*, vol. 23, no. 11, pp. 4696–708, 2014. 117

[184] S. W. Golomb, "Run-length encodings," *IEEE Transactions on Information Theory*, vol. 12, no. 3, pp. 399–401, 1966. 121

[185] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision.* Springer, 2014, pp. 818–833. 129

[186] L. Deng and J. Chen, "Sequence classification using the high-level features extracted from deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.* IEEE, 2014, pp. 6844–6848. 130

[187] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," 2015. 130

[188] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in neural information processing systems*, 2013, pp. 2553–2561. 130

[189] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current opinion in neurobiology*, vol. 14, no. 4, pp. 481–487, 2004. 131

[190] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010. 131

[191] A. Ahar, A. Barri, and P. Schelkens, "From sparse coding significance to perceptual quality: A new approach for image quality assessment," *IEEE Transactions on Image Processing*, 2017. 131

[192] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2528–2535. 132, 133

[193] M. Šorel and F. Šroubek, "Fast convolutional sparse coding using matrix inversion lemma," *Digital Signal Processing*, vol. 55, pp. 44–51, 2016. 132, 133, 134, 137, 138, 193

[194] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 391–398. 133

[195] R. Song, H. Ko, and C. Kuo, "Mcl-3d: A database for stereoscopic image quality assessment using 2d-image-plus-depth source," *arXiv preprint arXiv:1405.1403*, 2014. 134

[196] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," WASHINGTON UNIV SEATTLE DEPT OF ELECTRICAL ENGINEERING, Tech. Rep., 2011. 141

[197] J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving dlt," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2339–2346. 141

[198] F. Zhang and F. Liu, "Parallax-tolerant image stitching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3262–3269. 141

[199] J. Gao, S. J. Kim, and M. S. Brown, "Constructing image panoramas using dual-homography warping," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 49–56. 141

[200] L. Zhang, L. Zhang, X. Mou, D. Zhang *et al.*, "Fsim: a feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. 142

[201] L. Zhang and H. Li, "Sr-sim: A fast and high performance iqa index based on spectral residual," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 1473–1476. 142

[202] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740. 143

[203] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 145

[204] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," *arXiv preprint arXiv:1610.04490*, 2016. 145

[205] W. Wang, Q. Huang, S. You, C. Yang, and U. Neumann, "Shape inpainting using 3d generative adversarial network and recurrent convolutional networks," *arXiv preprint arXiv:1711.06375*, 2017. 145

[206] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493. 145, 146, 147, 150

[207] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544. 145, 146, 147, 149, 150, 157

[208] W. Zhang, B. Ni, Y. Yan, J. Xu, and X. Yang, "Depth structure preserving scene image generation," *arXiv preprint arXiv:1706.00212*, 2017. 145

[209] S. Liu, J. Pan, and M.-H. Yang, "Learning recursive filters for low-level vision via a hybrid neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 560–576. 146, 147

[210] J. S. Ren, L. Xu, Q. Yan, and W. Sun, "Shepard convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 901–909. 146

[211] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in neural information processing systems*, 2012, pp. 341–349. 146

[212] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on image processing*, vol. 17, no. 1, pp. 53–69, 2008. 146

[213] Y. J. Jung, H. Sohn, S.-i. Lee, Y. M. Ro, and H. W. Park, "Quantitative measurement of binocular color fusion limit for non-spectral colors," *Optics express*, vol. 19, no. 8, pp. 7325–7338, 2011. 149

[214] J. Li, "Methods for assessment and prediction of qoe, preference and visual discomfort in multimedia application with focus on s-3dtv," Ph.D. dissertation, Université de Nantes, 2013. 149

[215] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015. 150

[216] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017. 150

[217] D. Kinga and J. B. Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015. 150

[218] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 150, 157

[219] G. Silva and P. Rodriguez, "Efficient convolutional dictionary learning using partial update fast iterative shrinkage-thresholding algorithm." 169

[220] Y. Rai, A. Aldahdooh, S. Ling, M. Barkowsky, and P. Le Callet, "Effect of content features on short-term video quality in the visual periphery," in *Multimedia Signal Processing (MMSP), 2016 IEEE 18th International Workshop on*. IEEE, 2016, pp. 1–6. 169, 171

[221] S. Ling, P. L. Callet, and Z. Yu, "The role of structure and textural information in image utility and quality assessment tasks," *Electronic Imaging*, vol. 2018, no. 14, pp. 1–13, 2018. 171

[222] S. Ling and P. Le Callet, "Image quality assessment for dibr synthesized views using elastic metric," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1157–1163. 171

[223] P. L. C. Ling, Suiyi and C. Gene, "Quality assessment for synthesized view based on variable-length context tree," in *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*. IEEE, 2017. 171

[224] S. Ling and P. Le Callet, "How to learn the effect of non-uniform distortion on perceived visual quality? case study using convolutional sparse coding for quality assessment of synthesized views," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 286–290. 171

[225] J. Li, R. Mantiuk, J. Wang, S. Ling, and P. Le Callet, "Hybrid-mst: A hybrid active sampling strategy for pairwise preference aggregation," in *Advances in neural information processing systems*, 2018, pp. 3475–3485. 171

[226] S. Ling, J. Li, P. Le Callet, and J. Wang, "Perceptual representations of structural information in images: application to quality assessment of synthesized view in ftv scenario," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1735–1739. 171

[227] S. Ling, J. Gutiérrez, K. Gu, and P. Le Callet, "Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 204–216, 2019. 171

[228] Y. Zhou, L. Li, S. Ling, and P. Le Callet, "Quality assessment for view synthesis using low-level and mid-level structural representation," *Signal Processing: Image Communication*, vol. 74, pp. 309–321, 2019. 171

[229] S. Ling, J. Li, Z. Che, X. Min, G. Zhai, and P. L. Callet, "Quality assessment of free-viewpoint videos by quantifying the elastic changes of multi-scale motion trajectories," *arXiv preprint arXiv:1903.12107*, 2019. 171

[230] S. Ling, J. Li, J. Wang, and P. L. Callet, "Gans-nqm: A generative adversarial networks based no reference quality assessment metric for rgb-d synthesized views," *arXiv preprint arXiv:1903.12088*, 2019. 171

List of Tables

List of Figures

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Suiyi LING

Représentations perceptuelles de l'information structurelle et géométrique des images : approches bio inspirées et par apprentissage machine

Application à la qualité visuelle de médias immersifs

Perceptual representations of structural and geometric information in images: bio-inspired and machine learning approaches

Application to visual quality assessment of immersive media

Résumé

Ce travail vise à mieux évaluer la qualité perceptuelle des images contenant des distorsions structurelles et géométriques notamment dans le contexte de médias immersifs. Nous proposons et explorons un cadre algorithmique hiérarchique de la perception visuelle. Inspiré par le système visuel humain, nous investiguons plusieurs niveaux de représentations des images : bas niveau (caractéristiques élémentaires comme les segments), niveau intermédiaire (motif complexe, encodage de contours), haut niveau (abstraction et reconnaissance des données visuelles). La première partie du manuscrit traite des représentations bas niveau pour la structure et texture. U n modèle basé filtre bilatéral est d'abord introduit pour qualifier les rôles respectifs de l'information texturelle et structurelle dans diverses tâches d'évaluation (utilité, qualité. . .). Une mesure de qualité d'image/video est proposée pour quantifier les déformations de structure spatiales et temporelles perçues en utilisant une métrique dite élastique. La seconde partie du mémoire explore les représentations de niveaux intermédiaires. Un modèle basé « schetch token » et un autre basé sur codage d'un arbre de contexte sont présentés pour évaluer la qualité perçue. La troisième partie traite des représentations haut niveau. Deux approches d'apprentissage machine sont proposées pour apprendre ces représentations : une basée sur un technique de convolutional sparse coding, l'autre sur des réseaux profonds de type generative adversarial network. Au long du manuscrit, plusieurs expériences sont menées sur différentes bases de données pour plusieurs applications (FTV, visualisation multivues, images panoramiques 360. . .) ainsi que des études utilisateurs.

Mots clés

évaluation de la qualité visuelle, représentation perceptuelle et cognitive, TV à point de vue libre, apprentissage profond, convolutional sparse coding, réseaux "generative adversarial network", réalité virtuelle, systèmes multi vues

Abstract

This work aims to better evaluate the perceptual quality of image/video that contains structural and geometric related distortions in the context of immersive multimedia. We propose and explore a hierarchical framework of visual perception for image/video. Inspired by representation mechanism of the visual system, low-level (elementary visual features, e.g. edges), mid-level (intermediate visual patterns, e.g. codebook of edges), and higher-level (abstraction of visual input, e.g. category of distorted edges) image/video representations are investigated for quality assessment. The first part of this thesis addresses the low-level structure and texture related representations. A bilateral filter-based model is first introduced to qualify the respective role of structure and texture information in various assessment tasks (utility, quality . . .). An image quality/video quality measure is proposed to quantify structure deformation spatially and temporally using new elastic metric. The second part explores mid-level structure related representations. A sketch-token based model and a context tree based model are presented in this part for the image and video quality evaluation. The third part explores higher-level structure related representations. Two machine learning approaches are proposed to learn higher-level representation: a convolutional sparse coding based and a generative adversarial network. Along the thesis, experiments an user studies have been conducted on different databases for different applications where special structure related distortions are observed (FTV, multi-view rendering, omni directional imaging . . .) .

Key Words

Visual quality assessment, Perceptual and cognitive representation, Free view-point TV, Deep learning, Convolutional sparse coding, Generative adversarial network, Virtual reality, Multi-view systems