Dense correspondence estimation, image set compression and image retargeting quality assessment

Zhang, Yabin

2018

Zhang, Y. (2018). Dense correspondence estimation, image set compression and image retargeting quality assessment. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/75890>

<https://doi.org/10.32657/10356/75890>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

DENSE CORRESPONDENCE ESTIMATION, IMAGE SET COMPRESSION AND IMAGE RETARGETING QUALITY ASSESSMENT

A thesis submitted
for a degree of Doctor of Philosophy
by

YABIN ZHANG

School of Computer Science and Engineering
Nanyang Technological University

July 2018

Abstract

Establishing dense correspondence among images is one fundamental topic in image set compression, image retargeting, computer vision and other image manipulation tasks. As image set compression for cloud-based storage begins to attract more attentions and the increasing variety of display devices has imposed the demand for image adaption to different sizes and corresponding objective quality assessment metrics, existing dense corresponding techniques such as image alignment and optical flow show limited ability to clarify the complicated image relationships. In this thesis, we first study and design specialized dense correspondence estimation techniques, and employ them to reduce inter-image redundancy for image set compression and model the retargeting modifications for image retargeting quality assessment.

To reduce inter-image redundancy in image set for cloud-based storage, we develop a novel dense correspondence based prediction approach to model the consistent pixel-to-pixel relationship between images via the fast-convergent random search. While dense correspondence provides effective local similarity estimation, it may suffer from large side information expense when it cannot be parameterized effectively. By introducing consistency check constraints, we obtain the smooth and consistent dense correspondence as the inter-image prediction results, so we can efficiently reconstruct a new reference image by geometric transformations and luminance adjustments for the HEVC inter-prediction based video-like encoding. Because of the reconstruction in local units, our approach is more robust to the complex local variations compared with the global estimation based methods.

To better understand the image retargeting process and perform the evaluation for different image retargeting operators, we provide a unified interpretation of image retargeting with the resampling grid generation and forward resampling. We demonstrate that the geometric change estimation is an efficient way to clarify the dense correspondence relationship among images. We formulate the geometric change estimation as a backward registration problem via Markov random field and provide an effective and practical solution.

For image retargeting quality assessment applications, backward registration can provide reliable dense geometric change estimation, which links the properties between the original and retargeted images and makes the measurement of the corresponding quality degradation more efficiently. Under the guidance of the estimated dense correspondence, we develop an effective aspect ratio similarity metric by exploiting the local block changes to evaluate the visual quality of retargeted images. Furthermore, we point out that the existing low-level feature based methods are limited due to the visual importance dependency and the lack of high-level visual distortion modelling for the high-level image retargeting quality assessment task. To address these problems, we combine the dense correspondence estimation with different level feature detection methods and develop three effective general-purpose features to model various kinds of quality degradation factors in the retargeted image. Finally, we propose a multiple-level feature based framework to predict the perceptual retargeted image quality.

Acknowledgments

First and foremost, I would like to thank my supervisor, Dr. Weisi Lin, for his great guidance and support during four years of my Ph.D. study. Under his supervision, I have learned not only the cutting-edge research on image coding and quality assessment, but also his inspiring and down-to-earth working attitude. His guidance and wisdom always give me confidence, polish my research works and encourage me to explore interesting new things in my research field. Besides my supervisor, I also would like to thank Dr. Jianfei Cai for his valuable guidance, continuous encouragement as well as the care about my daily life in NTU.

I am greatly indebted to Long Xu, Jinjian Wu, Xinfeng Zhang, Yuming Fang, Huan Yang, Yuan Yuan, Lu Dong, Qiaohong Li, Sharath Chandra Guntuku, Prakhya Sai Manoj, Wentao Cheng, Tao Xi, Sheng Yang, Yupeng Cheng, Zhuo Chen, Naushad Ansari and other group members, for their friendship and insightful discussions. We have pleasant collaborations and/or great joy and fun together in the past four and a half years. Thanks to my roommates Tao Guo, Xingli Wang and Tianyi Zhang, and I would also like to thank my lab colleagues, Leida Li, Guanyu Gao, Junlin Hu, Jinpeng Wang, Jianxiong Yin, and Weide Liu. I wish to express my thanks to Rapid-Rich Object Search (ROSE) Lab and Media & Interactive Computing (MML) Lab for providing me with all the facilities for research. Thanks to them, I live a wonderful life here. Last but not the least, I sincerely thank my parents, my elder sister, Liya, and other family members from the bottom of my heart for their ceaseless support.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	vii
List of Tables	xii
List of Abbreviations	xii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Objective and Scope	6
1.3 Summary of Contributions	7
1.4 Thesis Organization	8
2 Literature Review	10
2.1 Dense Correspondence Estimation	10
2.2 Image Set Compression	15
2.3 Image Retargeting Quality Assessment	17
3 Image Set Compression	23
3.1 The Proposed Method	23
3.1.1 Dense correspondence based prediction	25
3.1.2 New reference reconstruction	30
3.2 Experimental Results	33
3.3 Conclusion	36

4	Backward Registration for Image Retargeting	37
4.1	Motivation to Model Image Retargeting Process	37
4.1.1	Challenge to evaluate the retargeted image	37
4.1.2	Difference between IQA and IRQA	38
4.2	Backward Registration Estimation	41
4.2.1	A unified interpretation of image retargeting	41
4.2.2	Problem formulation	44
4.2.3	Problem approximation	46
4.2.4	Objective energy minimization	47
4.2.5	The proposed hybrid descriptor	48
4.3	Forward and Backward Registration Comparison	48
4.4	Experimental Study for Backward Registration	52
4.4.1	Validation databases with the groundtruth	52
4.4.2	Backward registration accuracy	54
4.5	Conclusion	55
5	Aspect Ratio Similarity for IRQA	57
5.1	Aspect Ratio Similarity Definition	57
5.2	Aspect Ratio Similarity Analysis	59
5.3	Experimental Study for Aspect Ratio Similarity	64
5.3.1	Image retargeting databases	64
5.3.2	Performance evaluation on MIT RetargetMe database	67
5.3.3	Performance evaluation on CUHK database	68
5.4	Conclusion	70
6	Multiple Level Feature Measure	71
6.1	Problem Statement	71
6.2	Framework of the MLF Measure	72
6.3	Aspect Ratio Similarity Feature Extension	74
6.3.1	Brief introduction	74
6.3.2	Tradeoff between information loss/visual distortions	74
6.4	Edge Group Similarity Feature	75

6.4.1	Edge group representation and matching	79
6.4.2	Edge group similarity	81
6.5	Face Block Similarity Feature	83
6.6	Machine Learning based Multiple-Level Features Fusion	86
6.7	Experimental Study	88
6.7.1	Performance evaluation on MIT database	88
6.7.2	Performance evaluation on CUHK database	91
6.7.3	Parameter study	92
6.7.4	Feature analysis	94
6.7.5	Cross database performance evaluation for MLF	94
6.8	Conclusion	96
7	Summary and Future Work	97
7.1	Summary	97
7.2	Future Directions	99
7.2.1	Automatic image cropping	99
7.2.2	Image aesthetic assessment	100
	Appendix - List of Publications	101
	References	102

List of Figures

1.1	Overall structure of the thesis.	2
1.2	Illustration of the typical seam-carving and scale-and-stretch image retargeting operators. (a) Original image, ‘butterfly’. (b) Detected optimal seams. (c) Image retargeted by seam-carving retargeting operator. (d) Mesh grid on the original image. (e) Mesh grid on the retargeted image. (f) Image retargeted by scale-and-stretch retargeting operator.	4
3.1	Framework of image pair compression with dense correspondence based prediction.	24
3.2	Example of the encoding for a pair of images.	26
3.3	Dense correspondence estimation for the reference image (a) and current image (b). (c) and (d) are the visualized flow map without consistency checks and with consistency checks, respectively.	30
3.4	Dense correspondence based prediction and new reference reconstruction. (Note that the matched pixels in (a) and (b) are partially shown every 32 pixels.) . . .	31
3.5	Image set compression test image pairs.	34
3.6	Coding performance comparisons. (“Intra” and “Inter” curves are the HEVC Intra-prediction coding and Inter-prediction coding, and “SIFT-based multi-model” [1].)	35
4.1	Illustration of how the image content is modified and the visual quality change during image retargeting process.	38
4.2	Comparisons of <i>geometric</i> and <i>intensity</i> changes (Part I). The SSIM scores for regenerated images are 1.0000 (CR), 0.9078 (SV), 0.9322 (MOP) and 0.9366 (SC).	39

4.3	Comparisons of <i>geometric</i> and <i>intensity</i> changes (Part II). The SSIM scores for regenerated images are 0.9006 (SCL), 0.9270 (SM), 0.8965 (SNS) and 0.8993 (WARP).	40
4.4	Typical grid models for image retargeting, where the blue grid L is the resampling grid and the yellow region $\mathcal{L}_c$ is the resampling space.	42
4.5	Grid models for the image enlargement retargeting.	43
4.6	Markov random field (MRF) model for dense correspondence estimation, where the red node p represent the observed pixel in the retargeted image, and the blue node l is its resampling location (x, y) from the resampling grid. The subscripts $\{L, T, B, R\}$ denote the four-connected neighborhood.	45
4.7	Backward registration results on ‘Brasserie_L_Aficion’ image set (Part I). . . .	49
4.8	Backward registration results on ‘Brasserie_L_Aficion’ image set (Part II). . . .	50
4.9	Comparison of <i>forward</i> and <i>backward</i> registration by measuring information loss.	51
4.10	Validation databases. (a) Original image ‘ArtRoom’. (b) Image retargeted by SC with 25% width reduction. (c) The ground-truth for (b): a binary map indicating the removed pixels (black regions are the removed seams). (d) The image retargeted by SNS with 50% width reduction. (e) The approximated ground-truth map for (d).	53
5.1	(a) Original image, ‘ArtRoom’. (b) Visualized backward registration results, where each pixel from the retargeted image is distributed onto its resampling location in the original image. (c) Image retargeted by SC. (d) The geometric change from the regular blocks in the original image (left) to the retargeted blocks in the retargeted image (right), the block along the lime route is preserved in high quality while the block along the red route is squeezed seriously.	58
5.2	3D plot for aspect ratio similarity	60
5.3	ARS evaluation on ‘ArtRoom’ set (Part I). (a) Retargeted images. (b) ARS maps. (c) Importance weighted ARS maps.	61
5.4	ARS evaluation on ‘ArtRoom’ set (Part II). (a) Retargeted images. (b) ARS maps. (c) Importance weighted ARS maps.	62

5.5	A toy example. (a) Original image (512×384). (b) Image (192×384) retargeted by cropping out the middle half of the original image and scaling the right half of the rest by 50% width reduction. Original image is divided into 64×64 non-overlapped blocks as shown in (c) and the corresponding retargeted blocks are illustrated in (d), where B_1 is removed totally, B_2 is well-preserved as B_4 , and B_3 is scaled to B_5	63
5.6	The influence of block approximation	64
5.7	Original images from MIT RetargetMe database.	65
5.8	Original images from CUHK database.	66
5.9	The scatter plots of subjective rating against objective scores on CUHK database.	69
6.1	The block diagram of the proposed multiple-level feature based method. In the first stage, the backward registration is used to estimate the pixel-level correspondence between the original and retargeted images; in the second stage, we generate the multiple-level features to model different aspects of the quality degradation; in the last stage, a learned feature fusion model is used to predict the perceptual quality of the retargeted image.	73
6.2	‘Brasserie_L_Aficion’ image set (Part I). Note that (d) is clustered on the initial edge map of the original image individually and unused in following EGS feature generation.	75
6.3	‘Brasserie_L_Aficion’ image set (Part II). From the top to the bottom in each column in (e) CR, (f) SV, and (g) MOP, there are the image retargeted by the corresponding operator, the visualized backward registration result, the calculated ARS map (brightness indicates the quality score), the importance weighted ARS map, the EGS map for the original image, and the EGS map for the retargeted images, respectively. The edge group’s quality is indicated by the colorbar on the right side.	76

6.4	‘Brasserie.L_Aficion’ image set (Part III). From the top to the bottom in each column in (e) SC, (f) SCL, and (g) SM, there are the image retargeted by the corresponding operator, the visualized backward registration result, the calculated ARS map (brightness indicates the quality score), the importance weighted ARS map, the EGS map for the original image, and the EGS map for the retargeted images, respectively. The edge group’s quality is indicated by the colorbar on the right side.	77
6.5	‘Brasserie.L_Aficion’ image set (Part IV). From the top to the bottom in each column in (k) SNS, and (l) WARP, there are the image retargeted by the corresponding operator, the visualized backward registration result, the calculated ARS map (brightness indicates the quality score), the importance weighted ARS map, the EGS map for the original image, and the EGS map for the retargeted images, respectively. The edge group’s quality is indicated by the colorbar on the right side.	78
6.6	Influence of α from ARS on MIT and CUHK databases. The prediction accuracies are evaluated using KRCC on MIT database, and PLCC and SRCC on CUHK database.	79
6.7	The illustration of Chamfer matching for different transformed shapes. (a) Original image. (b) The shape contour of the giraffe (white). (c~f) the query shapes (red) with aspect ratio or scale transformation applied on the original shape. The matching distances for (c~f) are 148.5, 759.4, 84.3, and 217.0, respectively. The distance variations show its limitation for the object search.	81
6.8	Comparisons 1 of the EGS feature Q_{EG} and ARS. (a) The original ‘DKNYgirl’ image. (d) Corresponding visual importance map. (b) Image retargeted by MOP. (e) Image retargeted by SC. (c) EGS map for (b). (f) EGS map for (e).	83
6.9	Comparisons 2 of the EGS feature Q_{EG} and ARS. (a) The original ‘Woman’ image. (d) Corresponding visual importance map. (b) Image retargeted by SC. (e) Image retargeted by SCL. (c) EGS map for (b). (f) EGS map for (e).	84
6.10	Comparisons 3 of the EGS feature Q_{EG} and ARS. (a) The original ‘child’ image. (d) Corresponding visual importance map. (b) Image retargeted by SC. (e) Image retargeted by SCL. (c) EGS map for (b). (f) EGS map for (e).	85

6.11	Examples of (a) ‘child’ face images. (b~i) are images retargeted by CR, SV, MOP, SC, SCL, SM, SNS, and WARP, respectively. As indicated by the red boxes, the facial regions are first detected in the original image and the corresponding facial regions are identified in retargeted images using the pixel-level correspondence estimated in the first stage.	86
6.12	Examples of (a) ‘girls’ face images. (b~i) are images retargeted by CR, SV, MOP, SC, SCL, SM, SNS, and WARP, respectively. As indicated by the red boxes, the facial regions are first detected in the original image and the corresponding facial regions are identified in retargeted images using the pixel-level correspondence estimated in the first stage.	87
6.13	The individual KRCC results of ARS and MLF measure for the 37 image sets on MIT database. For each image set (including eight retargeted images), the KRCC is calculated as the rank correlation between the ranking of subjective and objective scores.	89
6.14	<i>Rank-n</i> accuracy of different IRQA measures. The rank prediction by MLF shows better overall prediction accuracy.	91
6.15	Influence of α from MLF measure on MIT and CUHK databases.	93
6.16	Influence of β from MLF measure on MIT and CUHK databases.	94

List of Tables

4.1	Performance evaluation on validation databases	55
5.1	Benchmark databases for IRQA evaluation	64
5.2	Performance evaluation of different metrics on MIT database.	68
5.3	Performance evaluation of different metrics on CUHK database	69
6.1	Performance comparison of different IRQA measures on MIT database.	88
6.2	Performance comparison of different measure on MIT database by two kinds of KRCC3. The first one (left) considers the pairs <i>s.t.</i> $(rank_s(i) \leq 3 \vee rank_s(j) \leq 3) \wedge (rank_o(i) \leq 3 \vee rank_o(j) \leq 3)$, while the second one (right) considers the pairs <i>s.t.</i> $rank_s(i) \leq 3 \vee rank_s(j) \leq 3$	90
6.3	Performance of different quality measures on CUHK database	92
6.4	Investigation of individual features on CUHK database	93
6.5	Feature analysis on MIT database	95
6.6	Feature analysis on CUHK database	95

List of Abbreviations

AR	Aspect Ratio
ARS	Aspect Ratio Similarity
AS	Absolute Size
BDS	Bi-Directional Similarity
BDW	Bi-Directional Warping
BMC	Block based Motion Compensation
BME	Block based Motion Estimation
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
CL	Color Layout
CR	Cropping
EGS	Edge Group Similarity
EH	Edge Histogram
EMD	Earth Mover's Distance
ENER	Energy-based Deformation
FBS	Face Block Similarity
FSIM	Feature Similarity Index
GLS	Global and Local Similarity
GOP	Group of Picture
GPM	General PatchMatch
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IQA	Image Quality Assessment
IRQA	Image Retargeting Quality Assessment
KRCC	Kendall Rank-Order Correlation Coefficient
LCC	Linear Correlation Coefficient
LOOCV	Leave-One-Out Cross-Validation
MAE	Mean Absolute Error

MLF	Multiple Level Feature based Measure
MOS	Mean Opinion Score
MRF	Markov Random Field
MST	Minimal Spanning Tree
NMS	Non-Maximum Suppression
NNF	Nearest Neighbor Field
OR	Outlier Ratio
PGDIL	Perceptual Geometric Distortion and Information Loss
PHOW	Pyramid Histogram of Visual Words
PLCC	Pearson Linear Correlation Coefficient
PM	PatchMatch
QP	Quantization Parameter
RANSAC	Random Sample Consensus
RMSE	Root Mean Squared Error
SC	Seam-Carving
SCL	Uniform Scaling
SCSC	optimized Seam-Carving and Scale
SI	Side Information
SIFT	Scale Invariant Feature Transform
SM	Shift-Map
SNS	Scale-and-Stretch
SRCC	Spearman Rand-Order Correlation Coefficient
SSIM	Structural Similarity Index
SV	Streaming Video
SVM	Support Vector Machine
SVR	Support Vector Regressor
WARP	Non-homogeneous Warping

Chapter 1

Introduction

In this chapter, we provide the background and motivations of image set compression and image retargeting quality assessment in Section 1.1. Next, we present the overview of our objective and research scope in Section 1.2, and summarize the main contributions of our studies on image set compression and image retargeting quality assessment in Section 1.3. The organization of this thesis is provided in Section 1.4.

1.1 Background and Motivation

Establishing dense correspondence among images is one fundamental topic in image set compression, image retargeting, computer vision and other image manipulation tasks. As image set compression for cloud-based storage begins to attract more attentions and the increasing trend of display devices has imposed the demand for image adaption to different resolutions and corresponding objective image retargeting quality assessment metrics, existing dense corresponding techniques such as image alignment and optical flow show limited ability to clarify the complicated image relationships. In this thesis, we first study and design specialized dense correspondence estimation techniques, and employ them to reduce inter-image redundancy for image set compression and model the retargeting modifications for image retargeting quality assessment.

In Figure 1.1, we show the overall structure of the thesis and the connections among the major works. We are eager to model the inter-image similarity for image set compression, and extract effective features to predict the perceptual image quality for image retargeting quality assessment. As existing dense correspondence estimation techniques show limited ability to meet the requirements, we develop new dense inter-image prediction approach and backward registration method to clarify the image dense correspondence relationships for image set compression and image retargeting quality assessment, respectively. Based on the dense inter-image estimation, a video-like image coding method is developed to reduce inter-image redundancy for image set compression. For image retargeting quality assessment, we utilize the backward registration results to identify the property changes between the original and retargeted images, thus develop the aspect ratio similarity metric and multiple-level feature based measure to predict perceptual retargeted image quality.

Figure 1.1: Overall structure of the thesis.

In the following, we will introduce the background for image set compression and objective image retargeting quality assessment, and show the motivation to exploit dense correspondence estimation for these applications at the same time.

For the first application, image set compression for the cloud-based storage begins to attract more and more attentions in the past few years. Due to the existence of inter-image redundan-

cy among the clustered similar images, it is feasible to compress the image set at higher ratio than encoding them using the individual coding approaches, including JPEG [2] or HEVC intra image coding [3]. At the beginning the samples for image set compression are mostly collected from the specific fields such as medical and satellite imagery. Usually, the images are well-aligned and inter-image variations exist in the pixel-to-pixel sense. To reduce the inter-image redundancy, the widely used methods are the prediction encoding of the image differences by directly subtracting the images with their references. The problems are mostly concerned with the anchor reference image generation and similar images clustering [4, 5]. When it comes to more complex wild image sets, e.g. personal photo albums, the direct image subtraction results may be even larger than the original ones. Therefore, the subsequent work mostly adopted the video-like encoding way to reduce the inter-image redundancy using video coding techniques. Since the motion estimation in video coding has limited ability to inter-image similarity among the complex wild images, it is necessary to develop inter-image similarity estimation techniques to facilitate the image set redundancy reduction. In this thesis, we mainly focus on how to effectively estimate the dense correspondence to model the inter-image relationship, thus generate the good anchor reference image for the video-like encoding.

The second application is image retargeting quality assessment (IRQA). The increasing variety of widely used display devices has imposed the demand for image adaptation to different sizes and aspect ratios. In addition, designers would like to edit images into different sizes with specific purposes, and the automatic adaption of images is favored. In the past few years, a number of content-aware image retargeting methods have been developed for image resizing [6]. The image retargeting operators can be broadly categorized into two types: discrete and continuous approaches [7], based on whether they treat images as discrete pixels or continuous signals. The typical discrete approaches mainly include manual Cropping (CR), Seam-Carving (SC) [8], and Shift-Map (SM) [9], while uniform Scaling in one dimension (SCL), non-homogeneous Warping (WARP) [10], Streaming Video (SV) [11] and Scale-and-Stretch

(SNS) [12] are representative continuous methods. Both kinds of content-aware approaches remove pixels or warp the image to the targeted resolution according to the visual importance of image content. The typical SC and SNS operators are shown in Figure 1.2, where the red lines are the calculated seams in SC and the purple mesh grid is used to manipulate image content during the warping process in SNS. As we can see in Figure 1.2, the seams and mesh grid tend to remove or shrink unimportant regions.

Figure 1.2: Illustration of the typical seam-carving and scale-and-stretch image retargeting operators. (a) Original image, ‘butterfly’. (b) Detected optimal seams. (c) Image retargeted by seam-carving retargeting operator. (d) Mesh grid on the original image. (e) Mesh grid on the retargeted image. (f) Image retargeted by scale-and-stretch retargeting operator.

The principle of the content-aware image retargeting is to preserve visually important content and structure (i.e. to reduce information loss), and at the same time, to limit the visual distortion in the retargeted image [6, 7, 13]. CR and SCL are two conventional retargeting methods based on geometric constraints without any consideration of image content. In images retargeted by CR, there is merely information loss occurring, while in images generated by SCL,

visual distortions due to squeezing or stretching degrade the image quality. Because the image size reduction during retargeting is inevitable, most content-aware retargeting operators try to remove or shrink the unimportant content, and thus achieve better overall performance by balancing the tradeoff between information loss and visual distortion. Content-aware retargeting operators [8, 11, 12, 10, 14, 9] have been developed in the last decade and show impressive results by preserving the important content with minimal distortion. However, there is still no single retargeting operator that can work well for every image, thus it is meaningful to develop effective IRQA measures to further advance image retargeting techniques.

The most reliable way to estimate the perceptual quality of the retargeted image is the subjective test, since human visual system (HVS) is usually the ultimate receiver in retargeting applications. While subjective evaluation is generally cumbersome, non-automatic and expensive, objective IRQA methods [15, 13, 16, 17] are more favorable to predict the retargeted image quality. It is worth mentioning that there has been a significant progress in the image quality assessment (IQA) studies [18, 19, 20] in the past few decades. In traditional (IQA) studies, the pristine and distorted images are usually assumed to be well-aligned and the change of images can be obtained by the direct subtraction easily, while in IRQA the content change during image retargeting process is much more complicated due to arbitrary kinds of artificial modifications. In our opinion, the essential difference is that in IQA the quality is influenced by the intensity change while in IRQA it is the geometric change, which can be illustrated by the manipulation of seams or mesh grid in Figure 1.2. Most of IQA metrics cannot be applied to IRQA directly, and even when we manage to exploit these metrics for IRQA successfully, it is hard to achieve satisfactory prediction performance.

The complicated relationship between the original and retargeted images, a.k.a, how the original image is modified into the retargeted image, makes the quality evaluation for the retargeted image extremely challenging. Given the original image and its multiple retargeted images, normally it is unknown about the details how the retargeted images are generated for

IRQA tasks. Considering that these details play an important role in the retargeted image generation, it is non-trivial to clarify the relationship between images by modelling the undergoing retargeting process explicitly. Although existing IRQA works [15, 13, 16] have provided preliminary studies to estimate relationship between the original and retargeted image, there is still a lack of effective modeling of the image retargeting process and correspondence estimation methods to reveal artificial modifications in the retargeted image. Research efforts are required to interpret image retargeting process and develop effective dense correspondence estimation methods to reveal the relationship between the images.

There is still no single retargeting operator that can work well for every image, thus it is meaningful for use to design new effective and reliable features to measure the quality change in the retargeted image. There are no regular/fixed patterns or constraints for the retargeting modifications and new retargeting operators will be developed based on different criteria over time as well. The visual distortions are arbitrary and content-dependent, such as the twist-ed line/structure, deformed face and broken symmetry pattern. Existing objective methods [16, 13] using low-level features usually pool the local quality scores with visual importance map, which lack the high-level features to depict these characteristics and measure the corresponding quality degradation evoked by image retargeting. The visual importance dependence also makes these methods tend to overlook visual distortions in unimportant regions. To better address IRQA problems, we need to develop more comprehensive features with the semantic high-level understanding of the image.

1.2 Objective and Scope

The objective of this thesis is to exploit the dense correspondence among images, and develop effective prediction approach for image set compression and objective image retargeting quality assessment models. In particular, we explore the following problems in this thesis:

- to capture and reduce the inter-image redundancy in image set.

- to model and reveal the artificial modifications during image retargeting process.
- to evaluate the perceptual quality of retargeted images.
- to address the visual importance dependence problem in existing IRQA methods.

1.3 Summary of Contributions

Firstly, we present a study on image set compression to reduce inter-image redundancy. We develop a novel dense correspondence based prediction approach to estimate the consistent pixel-to-pixel relationship between images via the fast-convergent random search. Dense correspondence can provide more effective local similarity estimation between images, but may suffer from large side information expense if it cannot be parameterized effectively. By introducing consistency check constraints, we obtain the smooth and consistent dense correspondence as the inter-image prediction results. Next, we reconstruct a new reference image by geometric transformations and luminance adjustments for the subsequent HEVC inter-prediction based video-like encoding. Because of the reconstruction in local units, our approach is more robust to the complex local variations compared with the global estimation based methods. Experimental results show that the proposed method achieves better performance compared with relevant state-of-the-art methods, especially when the local variations are dominant.

Secondly, to better understand the image retargeting process and perform the evaluation for different image retargeting operators, we provide a unified interpretation of image retargeting with the resampling grid generation and forward resampling. We demonstrate that the dense geometric change estimation is an efficient way to clarify the dense correspondence relationship between the original and retargeted images. We formulate the geometric change estimation as a backward registration problem via Markov random field and provide an effective and practical solution. For image retargeting quality assessment applications, the dense backward registration results can provide reliable geometric change guidance, which link the

properties between the original and retargeted images and make the measurement of the corresponding quality degradation more directly and efficiently. Experimental results on two validation databases show that the proposed backward registration method achieves high overall estimation accuracy, qualitatively and quantitatively.

Thirdly, by exploiting the geometric change estimated by the backward registration approach, we develop an effective aspect similarity ratio metric by exploiting the local block changes to evaluate the visual retargeted image quality. Compared with other existing IRQA metrics on the publicly available MIT RetargetMe and CUHK databases, the proposed ARS yields statistically better results in the prediction accuracy. Furthermore, we point out that existing low-level feature based methods are limited due to the visual importance dependency and the lack of high-level visual distortion modelling for the high-level IRQA task. To address these problems, we combine the dense correspondence estimation with different level feature detection methods and develop three effective general-purpose features to model various kinds of quality degradation factors in the retargeted image. Finally, we propose a multiple-level feature (MLF) based framework to predict the perceptual retargeted image quality. Experimental results show that the proposed MLF measure achieves better and more reliable quality prediction performance compared with other existing methods.

1.4 Thesis Organization

The thesis consists of seven chapters and each chapter is summarized as follows. Chapter 1 gives a brief introduction to the motivation and background of this thesis, and highlights the major contents and contributions. Chapter 2 is the literature review to get a broad view of the current research status of related topics. We mainly introduce the image set compression, image retargeting, and image retargeting quality assessment. Chapter 3 proposes a novel dense correspondence based prediction approach for image set compression. Chapter 4 provides a unified interpretation of image retargeting process. We show the formulation, approximation,

and solution for the dense correspondence estimation problem in IRQA. Chapter 5 proposes an effective aspect ratio similarity metric by exploiting the local block changes to evaluate the visual quality of retargeted images. Chapter 6 develops three effective general-purpose features to model various kinds of quality degradation factors in the retargeted image, and proposes a multiple-level feature based framework to predict the perceptual retargeted image quality. We address the limitation of existing low-level feature based methods due to the visual importance dependency and the lack of high-level visual distortion modelling. Chapter 7 concludes the thesis with a summary of the aforementioned research works and provides the possible research directions in the future.

Chapter 2

Literature Review

Our works are inextricably connected with dense correspondence for the estimation of inter-image redundancy in image set and geometric change in image retargeting. Therefore, we first review the related dense correspondence estimation techniques in Section 2.1. Then, in Section 2.2, we review the related works on image set compression. Finally, we review the literature about image retargeting, and image retargeting quality assessment in Section 2.3.

2.1 Dense Correspondence Estimation

Establishing dense correspondence among images has been a fundamental task in many computer vision and image processing applications, including image retrieval [21], object recognition [22] and tracking [23]. When computing the dense correspondence among images, the image alignment, optical flow and nearest neighbor field (NNF), aim to solve different problems but share certain similar philosophy. Here we introduce the related works about image alignment, optical flow, and NNF, respectively.

Image alignment : Image alignment is a traditional topic in computer vision, which aligns images from the same scenes in different views for images stitching, panorama images, and video stabilization etc. [24]. Here we focus on the dense correspondence related Direct method (pixel-to-pixel matching), which normally warps one image relative to the other based on certain motion model (2D motion models refer to the translation model, affine model, homograph

model, or 3D motion models) and find the optimal motion parameters by checking how much the pixels between two images agree. The translation model is the simplest motion model where we shift one image relative to the other, and search the optimal translation parameters. Given the template image $I_0(x)$, with pixel locations $\{p_i = (x_i, y_i)\}$, we need to find the corresponding pixels in image $I_1(x)$. If we use the sum of square differences as the error metric, it will be a least square minimization problem to find the optimal translation parameter w as

$$E(w) = \sum_i [I_1(p_i + w) - I_0(p_i)], \quad (\text{Eq. 2.1})$$

where $w = (u, v)$ is the constant translation parameter for the entire image, and the parameter can be $w(p) = (u(p), v(p))$ for other more complex motion models, e.g. affine and homograph model.

The simple brute-force solution to solve this error cost function is to do a full search over certain range of shift with either integer or sub-pixel steps, which is the popular Block-based Motion Estimation (BME) in the video coding for the predicted frames, where we search the accurate motion vector in the range e.g. 16×16 pixels. To obtain better sub-pixel estimation, the famous gradient-based Lucas-Kanade method [25] is commonly used to refine the obtained $w = (u, v)$. A Taylor Series expansion of Eq. 2.1 is shown as

$$\begin{aligned} E(w) &= \sum_i [I_1(p_i + w + \Delta w) - I_0(p_i)] \\ &\approx \sum_i [I_1(p_i + w) + (I_x(p_i), I_y(p_i)) \cdot \Delta w - I_0(p_i)] \end{aligned} \quad (\text{Eq. 2.2})$$

$$= \sum_i [(I_x(p_i), I_y(p_i)) \cdot \Delta w + I_t(p_i)], \quad (\text{Eq. 2.3})$$

where $I_t(p_i) = I_1(p_i + w) - I_0(p_i)$. The least square problem can be minimized by solving

$$\begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}. \quad (\text{Eq. 2.4})$$

Optical flow: Optical flow is the estimation of apparent velocities of movement of brightness patterns in an image. The optical flow estimation [26] has a long history and most techniques

exploit two constraints: brightness constancy and spatial smoothness, which are introduced by the Horn-Schunk [27] algorithm. The brightness constancy is the assumption that the intensity values of the image objects will not change over time as

$$f(x + u, y + v, t + 1) = f(x, y, t), \quad (\text{Eq. 2.5})$$

where the displacement vectors (u, v) are called optical flow.

Optical flow cannot be computed individually for each pixel, because there is only one independent measurement at one point while the flow vector has two components. To solve the ill-posed problem, the Lucas-Kanade [25] algorithm can be used to compute the optical flow by assuming the optical flow vector is constant within some neighborhood of size ρ , while the Horn-Schunk [27] assumes the apparent velocity of the brightness pattern to be smooth, which is constrained by a regularization term in the corresponding energy function. Commonly, the optical flow field is expected to be dense where each pixel is assigned one flow vector. When the Lucas-Kanade algorithm is applied, the local neighborhood ρ is assigned only one constant flow vector for the block and to obtain the dense flow field, the results must be interpolated to generate more flow vectors but the accuracy will decrease relatively [28].

The energy functions of most optical flow techniques have data term and smoothness term, e.g. the famous Horn-Schunk algorithm [27, 29], and it computes the dense flow field by introducing the smoothness term in the global energy function by

$$E_{HS}(w) = \sum (|(I_x, I_y) \cdot (u, v) + I_t|^2 + \lambda(|\Delta u|^2 + |\Delta v|^2)), \quad (\text{Eq. 2.6})$$

where the latter term is the *smoothness* term and λ is the smoothness weight. The larger λ enforces stronger penalization on discontinuity of the adjacent flow vectors, thus leads to smoother flow fields.

Nearest neighbor field (NNF) : Given two image I_0 and I_1 , we densely sample image patches, treat each patch as a feature point corresponding to one pixel in a high-dimensional space, and

for each patch in I_0 we find the nearest neighbor in I_1 based on certain distance metric. It is called Nearest-Neighbor Field (NNF) problem [30, 31]. If we use the centroid pixel p to represent the patch P , the NNF problem can be formulated as the following energy function,

$$E(w) = \sum \|I_1(P(p_i + w_i)) - I_0(P(p_i))\|^2, \quad (\text{Eq. 2.7})$$

where p_i is the pixel with index i , and $I(P(p_i))$ is the corresponding neighborhood patch as the feature vector. By minimizing the energy function, we can obtain the nearest neighbor pixel in I_1 denoted by the relative displacement vector w_i for the pixel p_i in I_0 .

The patch sampling problem is first derived from image editing tasks such as texture synthesis. It can also be modeled in Markov random field (MRF) framework, where the pixel value given its neighborhood image patch is totally independent of the rest part of the image. The image patches (e.g. 7×7) are densely sampled in an overlapped way at multiple scales of the image. The sampled patches can be regarded as the feature descriptors of the corresponding pixels (e.g. the center or left-top pixel in the patch). The NNF estimation for each pixel is able to synthesize the image texture and structures by exploiting the probability distribution of the pixel in the given context image, and also applied to other image processing applications such as video completion, summarization, image in-painting, image denoising and so on. The computation cost for NNF can be very expensive, and the brutal method to approach the problem has computation cost $O(mM^2)$ for the image with size M and patch size of m pixels. It means that as the image size increases, the computation expense will become intolerable. Various fast NNF techniques have been proposed, including the tree-structure based method (e.g. KD-trees) [32], and the recently proposed PatchMatch [30] method, which boost the computation speed for interactive image editing applications.

Dense correspondence among images : In the view of the dense pixel-to-pixel correspondence among images, the aforementioned studies share similar properties. They all have the similar data terms which aim to find the closest pixel correspondence with certain distance metric, while the smoothness terms (neighborhood constraints) vary in different research fields.

Energy equations for image alignment, optical flow and NNF problem can be reformulated comparatively as

$$E_{alignment}(w) = \sum \|F_1(p + w) - F_0(p)\|_2^2, \quad (\text{Eq. 2.8})$$

$$E_{optical_flow}(w) = \sum \|F_1(p + w) - F_0(p)\|_2^2 + \lambda(|\nabla u|^2 + |\nabla v|_2^2), \quad (\text{Eq. 2.9})$$

$$E_{NNF}(w) = \sum_i \|F_1(p_i + w_i) - F_0(p_i)\|_2^2, \quad (\text{Eq. 2.10})$$

where $F(p)$ is either the pixel value or the neighborhood patch based the feature vector to exploit the image similarity.

Firstly, the differences between the energy functions depend on their application purposes: in image alignment, the displacement vectors $w = (u, v)$ are generated from the same motion model with fixed parameters for the entire image; in optical flow estimation, dense displacement vector (flow vector) is calculated for each pixel, and adjacent displacement vectors obey the smoothness constraints for practical solutions; in NNF problems, the estimation merely searches the most similar pixel without any explicit neighborhood constrains.

Secondly, these objective energy functions can be viewed as the log likelihood of the posterior distribution of an MRF [33], which consist of the data energy term and the smoothness energy term: in image alignment, there is only the data term since the motion parameters for all the pixels are assumed to be the identical, which is kind of the strongest smoothness constraints; in optical flow, there are both data and smoothness terms. The data term by itself is not sufficient to solve the problem and local smoothness constraint is applied, which is consistent with the physical world; in NNF, there is no explicit smoothness constraint since the main target is to find the nearest neighbor in the high dimensional feature space constructed from the images. (The search process in NNF is based on the local patches or feature descriptors, and the incorporation of local context information can enforce certain weak smoothness constraints inexplicitly.)

Finally, the differences among these research fields are mostly due to their specific application requirements. In image alignment, the camera motion relationship among the images can be estimated by various motion models for the entire image, including the translational and affine models; in optical flow, we aim to find the apparent motion of the brightness pattern projected from 3D real physical world; in NNF problem, the nearest neighbor of the pixel reveals the stochastic distribution of the pixel in the context of image synthesis or denoising applications. Similar to the Lucas-Kanade algorithm, it can be applied to the entire image to refine the alignment precision, and also can be applied to small image blocks to obtain the less dense optical flow by assuming flow vector to be locally smooth or approximately constant within the blocks [26]. Most existing optical flow techniques can only establish the small displacement around the initial value, and large displacement motion estimation is a challenging problem. NNF techniques can be leveraged to improve large displacement flow estimation by introducing nearest neighbor pixels as better initial estimation candidates [34].

With the above analysis, we can draw the conclusion that the principles under dense correspondence should be designed specially for targeted applications by varying their objective energy functions. In this thesis, we mainly study how to design appropriate dense correspondence estimation techniques to reduce the inter-image redundancy for image set compression and model the retargeting changes for image retargeting quality assessment.

2.2 Image Set Compression

Image set compression for the cloud-based storage has become more attractive in recent years. Image/video compression is possible because of the data redundancy therein. The individual coding approaches, including JPEG [2] and HEVC intra coding [3], aim to reduce the intra-image redundancy. Due to the existence of inter-image redundancy in the image set, it is feasible to encode these images at the even higher compression ratio.

At the beginning, the concerned samples for image set compression are mostly collected from the specific fields such as medical and satellite imagery. In one single image database, there can be a considerable amount of images with the same subjects. In the radiology such as X-ray examination, the images are usually taken with rigorous standards on the positions, orientations and parameters of the imaging device. These images are usually well-aligned and the inter-image variations normally exist in the pixel-to-pixel sense. To reduce the inter-image redundancy, the widely used methods are the prediction encoding of the image differences by directly subtracting the images with their references. The research problems are mostly concerned with the anchor reference image generation and similar images clustering [4, 5]. Unfortunately, such methods cannot be applied to more complex wild image sets such as personal photo albums, where the direct image subtraction results may be even larger than the original ones.

Therefore, subsequent works mostly adopt the video-like encoding way to reduce the inter-image redundancy in image set. Au *et al.* [35] proposed the inter prediction in video coding with global and local motion compensation in the given similar image set. In their scheme, the given image is encoded with one anchor image. The global motion between the images is estimated with the affine model, then the anchor image is aligned to the given image during the global motion estimation, next it is the local block-based motion compensation and residual signal encoding. Normally the global motion estimation is applicable for the global motion caused by camera motion, so the chosen image sets are still very similar to the video sequence. Zou *et al.* [36] mainly focused on the encoding structure of the image set compression. They clustered the similar images as a group of picture (GOP) encoded by High Efficiency Video Coding (HEVC) in a $I - P - P - P$ structure, and proposed Minimal Spanning Tree (MST) encoding structure with depth control to achieve better coding gain for the entire set and fast random access as well. They formulated the images into video-like sequences and compressed them using video inter-prediction coding. However, this method can only be applied to image

sets with strong temporal correlation or small viewpoint changes, and its performance on image sets with large variations are quite limited. In other image sets with larger variations these methods can be limited. To deal with the wild image sets with larger viewpoint differences, Shi *et al.* [1] first proposed a Feature-based Minimum Spanning Tree (FMST) to generate the encoding structure of the image sets. To reduce the inter-image redundancy, they proposed a feature based three-step inter-image prediction framework for inter-image coding. Given a pair of images, they performed the sparse scale-invariant feature transform (SIFT) feature matching between the images and estimated the multiple geometric deformation models via Graph-Cut algorithm. With the multiple geometric deformed reference image, the photometric relation were established and adjusted. Finally, the generated multiple reference images were used to predict the current image by the block-based motion estimation in HEVC. With the global geometric deformation and photometric transformation, multiple new reference images were generated from the original reference to better predict the current image. In this way, the Block-based Motion Estimation and Compensation (BME & BMC) in HEVC can be used to generate better prediction from the multiple reference images. However, the SIFT-based global photometric transformation is not always effective as the feature keypoints are generally sparse and not sufficient enough for reliable luminance adjustment. Furthermore, the global geometric deformation is also not always effective to deal with local variations.

2.3 Image Retargeting Quality Assessment

Image retargeting : The basic scaling retargeting operator (SCL) usually uses the simple scaling uniformly along one direction and bi-cubic interpolation for image retargeting. Another basic cropping retargeting operator (CR) adopts the cropping window in the target sizes, where the optimal window can be chosen manually or automatically based on the saliency detection or composition evaluation.

Many different content-aware image retargeting operators have been proposed, which can be mainly classified as discrete or continuous approaches [7]. The discrete approaches, including seam-carving (SC) [8] and shift-map (SM) [9], remove pixels or patches judiciously, while the continuous approaches, including non-homogeneous warping (WARP) [10], scale-and-stretch (SNS) [12] and streaming video (SV) [11], seek an optimal mapping/warping to adapt an image to the target size with certain content protection constraints.

In SC [8] operator, a seam is defined as one optimal 8-connected path from top to bottom (vertical seam) or left to right (horizontal seam), which is a contiguous chain of pixels as the vertical seams illustrated in Figure 1.2 (b). The optimal seams are the paths with the least total energy, and they can be computed using dynamic programming. Their chosen energy map, or so-called visual importance map, is the simple L_1 norm of the image intensity gradient. SC operator can repeatedly remove or duplicate the seams passing through the least important image regions for the shrinkage or enlargement of images. SM [9] operator characterizes the image retargeting process using a shift-map, which describes the geometric shift of each pixel in the retargeted image from the original image. They adopts the graph-cut method to optimize the graph labelling problem. Instead of removing individual seam repeatedly, SM operator reformulates the entire image to the target size at a time in the discrete manner. However, this method was experimented with setting of different shift map directions, saliency maps and boundary conditions, and the best retargeting result for each image need to be chosen manually.

WARP [10] operator squeezes the image non-homogeneously with a one-dimensional warping function. Its importance map is the L_2 norm of the image intensity gradient along with face detection results, and it is used to weight the warping objective energy function, which helps to preserve the important content and shrink the less important regions. When WARP operator is applied to video retargeting applications, it utilizes additional motion information to improve the temporal channel consistency. Similarly, SNS [12] operator is a warping based retargeting method by computing the optimal local scaling factors. The adopted strategy is to allow

the important region to scale uniformly and homogenous regions to be distorted. The quad deformation term and grid line bending energy term are utilized in the objective function for the mesh-based image warping method. The quad deformation term forces the important quad to the uniform scaling, and the grid bending term aims to minimize the bending of grid lines in the mesh grid. The chosen alternative optimization method provides a quite efficient solution for the linear system. In addition, their designed importance map combines both the gradient and saliency detection, which is less sensitive to the disturbance detected in the saliency map. SV [11] is another continuous warping based operator, which works with both automatic and interactively user-defined features. Their special importance map is a combination of saliency detection, edge detection, scene cut detection, and optional high-level user-marking constraints. In the objective function, there are additional terms to take into account the user defined constraints for the object position and line preservation, and the optimization method for the warping energy function is derived from SNS and WARP operators.

As discussed in [6] the main objectives shared by most image retargeting operators are to preserve important content and internal structure, as well as prevent visual artifacts. Most retargeting operators work well on images with removable regions such as sky and water, where the information loss can be easily controlled. However, visual distortions are inevitable for images with dense information or global patterns when we minimize the information loss. It becomes challenging to achieve an appropriate tradeoff between information loss and visual distortion in the retargeted image.

Image retargeting quality assessment : There has been a significant progress in the image quality assessment (IQA) studies [18, 37] in the past few decades. However, the IQA metrics, including the successful structural similarity index (SSIM) [37] and the feature similarity (FSIM) index [38], cannot be applied to IRQA directly because they require the reference and distorted images to be of the same resolution.

Although a number of image retargeting methods have been proposed for image resizing during recent years, the visual quality evaluation of retargeted images is still a challenging task. To measure the quality of retargeted images, Rubinstein *et al.* [6] conducted a comparative study of different retargeting operators, and also provided six objective quality metrics: Bidirectional Similarity (BDS), Bidirectional Warping (BDW), Edge Histogram (EH), Color Layout (CL), SIFT flow [39] and Earth-Mover's Distance (EMD) [40]. The BDS is a patch based similarity metric formulated with a bidirectional mapping between two images. For patch in one image, the matched patch is searched in the other image and vice versa. The mean distance of these patches is summarized as the BDS distance. The BDW is one metric similar to BDS, but the mapping between two images is constrained to be monotonic. The EH and CL are two low-level metrics in moving picture experts group 7 (MPEG-7) standard, where EH uses the histogram of the spatial edge distribution as the image representation and the L_1 distance of the two histograms as the image distance, while similarly, CL is an image descriptor which captures the spatial distribution of colors in the image. SIFT flow is a recently proposed across-scene image alignment algorithm and EMD is a dissimilarity metric between two distributions defined as the minimal cost to transform one distribution into the other. Among them EMD and SIFT flow show better agreement with subjective rankings, and the reason might be that their constrained matching appears to be able to model the retargeting modification on images more accurately.

In the early IRQA work [41], Ren *et al.* developed an automatic image retargeting quality method in the principle of user perception based criteria of important content emphasis, global information coverage and visual distortion reduction. In [15], Liu *et al.* exploited global structure and local correspondence to measure the quality of retargeted images. By traversing from coarse scales to fine scales, the metric extracted the global geometric structure and established the local pixel correspondence in a top-down manner. Finally, based on the correspondence, the retargeted image is evaluated with a saliency weighted similarity metric. Ma *et al.* [42]

examined the quality prediction performance of different existing shape descriptors such as PHOW [43], GIST [44], MPEG-7 descriptors [45], and EMD [40]. Fang *et al.* [16] proposed an algorithm (IR-SSIM) to generate Structural Similarity (SSIM) [37] map to measure the quality of preserved structural information in retargeted images. They first utilized the SIFT flow to explore the dense correspondence between the images, and then generated a SSIM map to measure the amount of the preserved structural information in the retargeted image. They exploited both bottom-up and top-down image saliency estimation as the visual importance map for retargeting. The spatial pooling of the visual importance map with the SSIM map provided the overall evaluation of the retargeted image.

In [46], Zhang *et al.* investigated the features from three determining factors for objective quality of experience assessment on retargeted images: global structural distortions (G), local region distortions (L) and loss of salient information (S), and proposed the GLS metric by fusion with regression methods. Liu *et al.* [47] also utilized machine learning to fuse different spatial estimator and frequency domain quality estimator, and achieved appealing quality prediction performance. In [13], Hsu *et al.* obtained the dense correspondence between original and retargeted images with SIFT flow technique as well. By measuring the local variance of the SIFT flow field between the images, the perceptual geometric distortions are measured, and by measuring the saliency loss during warping the original saliency map to the target resolutions, the perceptual information loss is measured. An adaptive fusion method is proposed to fuse perceptual geometric distortion and information loss as the overall evaluation of the retargeted image. The relationship between the images are still not fully investigated and the correspondence estimation has limitations on finding the retargeting modification. Recently, Chen *et al.* [48] leveraged the relative quality difference and proposed a general regression neural network (GRNN) model based two-step learning method to predict the ranking of retargeted image.

Visual importance map : Visual importance map is necessary for content-aware image retargeting to identify the important regions, and even for the identical retargeting operator, different

visual importance maps may lead to quite different retargeted results. The common visual importance maps adopted by different retargeting operators, are edge maps (L_1 or L_2 norm of the gradient), Itti's saliency map [49] and eye gaze measurement [50]. To achieve better performance, some works have further refined the importance map using the multiplication of the edge map and saliency map [12], and the importance filtering [51], where the saliency map is filtered with the guidance of the image itself to preserve structure consistency. There are also some saliency detection methods [52, 53] developed specifically for image retargeting, and these visual importance maps play an important role in most existing IRQA works.

Chapter 3

Image Set Compression

In this chapter, we propose a novel dense correspondence based prediction approach to reduce the inter-image redundancy in image set. Different from existing related methods, we manage to utilize the dense correspondence to predict and parameterize the inter-image relationship and reconstruct a new reference for video-like encoding. This chapter is organized as follows. Section 3.1 provides the technical details about dense correspondence based prediction approach and new reference reconstruction. Section 3.2 reports the experimental results and Section 3.3 concludes this study.

3.1 The Proposed Method

The image set compression system can be separated into two parts: image set encoding structure organization and image pair compression. Image set encoding structure represents the similarity relationship among images. The optimal encoding structure searches the most similar image pairs and generates the best dependence order for the pair-wise encoding. The similarity estimation of encoding structure strongly relies on the chosen image pair encoding approach. As we can find in [36, 54], different encoding structure generation methods have been chosen, because only the similarity metric with appropriate image pair compression approach can achieve the optimal encoding performance for the entire set. In this chapter, we

mainly focus on the image pair compression approach about how to better exploit the inter-image redundancy.

Figure 3.1: Framework of image pair compression with dense correspondence based prediction.

The framework of our proposed image pair compression approach is shown in Figure 3.1. The purpose is to encode the current image I_c with less bitrate using the prediction from the reference image I_r . The dense correspondence relationship between I_c and I_r is estimated first. The advantage of dense correspondence estimation is that it provides more effective local similarity information between images and is closely correlated with coding purpose, but it may suffer from extremely large side information (SI) burden if the dense correspondence relationship cannot be parameterized effectively. Existing dense matching approaches usually produce inconsistent dense prediction, which will counteract the coding gain obtained by the inter-prediction to a great extent. For example, the SIFT flow [39] algorithm is designed to establish dense correspondences between images across scenes. Each pixel will be assigned a match in the reference image under the similarity metric using SIFT descriptor and the objective energy function in SIFT flow allows the relative large discontinuous matching, which leads to the inconsistent matching results for coding purpose. Therefore, the dense correspondence based prediction is constrained to be consistent in our method, which becomes appropriate as

SI with efficient parametrization. After the correspondence estimation, we partition the reference image into regular non-overlapped units (e.g. 256×256 blocks), so that we can calculate the transformation parameters for each local unit individually via the widely used RANSAC [55] algorithm.

Then one new reference I_{rec} is reconstructed by geometric transformations and luminance adjustments in the selected units for the subsequent BME & BMC in the second video-like encoding stage. In the second stage, the reconstructed reference I_{rec} is transmitted to the decoded picture buffer for BME & BMC in HEVC. The reference I_{rec} works as a better candidate for inter-image prediction. Thus the current image I_c is encoded with the prediction from the reconstructed reference in the same way as video inter-prediction coding and formulated into the bitrate along with the necessary parameters as SI. To decode I_c , the new reference I_{rec} will be reconstructed again from the reference image I_r and works as the prediction reference for I_c . An example of the encoding process for the current image is shown in Figure 3.2. In the step 1, we find the dense correspondence prediction between I_c and I_r and apply to I_r to obtain the reconstructed image I_{rec} ; in the step 2, we adopt the HEVC inter-prediction to encode I_c with prediction from I_{rec} in I-P structure.

3.1.1 Dense correspondence based prediction

The sparse feature-based methods estimate the global geometric deformation between images based keypoint feature descriptors, e.g. as SIFT [56], using geometric models including affine and homograph transformation. Since the transformation among real-world images are mostly non-rigid, it is hard to use one single global geometric transformation to model the inter-image relationship in the wild image sets. Therefore, multiple geometric transformations are required for different local regions in the images, and we need to reconstruct multiple new references accordingly for the following video-like encoding in image set compression. The increasing number of reconstructed reference frames leads to higher video-like encoding complexity and

Figure 3.2: Example of the encoding for a pair of images.

if we restrict the number of reconstruction frames, certain local image contents may be not modelled effectively, especially when the local variations are significant. While the sparse feature-based methods are limited to the global geometric transformations, dense correspondence estimation is able to provide much more detailed pixel-to-pixel relationship in local regions, which makes it feasible to construct better new reference frame.

To capture the local image content similarity among images, we are eager to develop the dense correspondence estimation based approach both in the range of the entire image and in the translations, rotations, and scales search channels. The conventional translational block-based motion estimation in video coding is usually limited to a relative small local search window due to the computational complexity. Although various fast BME algorithms have been developed in the last decade to boost the search speed, the computational expense will become intolerable when its search window is extended to the entire image. It's also impractical to directly extend the search channels of BME algorithms to translations, rotations, and scales

at the same time.

If we extend the search range of BME algorithm to the entire image, it is actually a nearest neighbor field search problem. To enhance the search efficiency and seek a feasible solution, we leverage a fast-convergent random search algorithm [57] by utilizing the natural image coherence. The key observation is that random search can somehow find some good matching results accidentally, and the natural image coherence allows us to propagate these good matching results quickly to the neighboring regions. The alternative iterations of random search and neighborhood propagation make the global search process converging fast with effective approximate NNF matching results.

Here, we use the left-top pixels of the blocks to denote the corresponding blocks. Compared with non-overlapped block partition scheme in BME, the natural coherence is gained in the overlapped blocks division scheme, which plays an important role in the fast-convergent solution for the NNF search problem. The dense correspondence estimation problem in the range of entire reference image with search channels of translations, rotations, and scales can be formulated into the following objective function,

$$E = \sum_{ij} MSE(I_c(C_{ij}), I_r(R_{ij})), \quad (\text{Eq. 3.1})$$

where

$$R_{ij} = A \cdot C_{ij}, \quad (\text{Eq. 3.2})$$

$$A = \begin{bmatrix} scale \cdot \cos(\theta) & \sin(\theta) & X \\ -\sin(\theta) & scale \cdot \cos(\theta) & Y \\ 0 & 0 & 1 \end{bmatrix}, \quad (\text{Eq. 3.3})$$

E is the overall error energy, MSE is the mean squared error between two blocks, C_{ij} and R_{ij} are the block coordinates at (i, j) in the current image I_c and the reference image I_r , m and n are the row and column indices for the overlapped block in I_c , and A is the affine matrix with four parameters $(X, Y, \theta, scale)$ representing the geometric transformation relationship.

The generalized PatchMatch (GPM) [57] provides a practical, fast-convergent implementation with iterations of neighborhood propagation and random search. However, the neighborhood propagated searching results will be distributed in the whole range of I_r with high possibility and also not consistent due to its greedy collaboratively gradient-descent search. Different from the Markov random field based backward registration algorithm, which we will introduce in Chapter 4, there is no smoothness term in the energy function to regularize the optimization around the noisy or discontinuous regions. It is hard to achieve the neighborhood consistency with only the inexplicit natural coherence in the data term, which is close to the search for each pixel independently. It will lead to extremely large SI burden if the dense correspondence relationship cannot be parameterized efficiently. To further improve the consistency of dense correspondence results, we adopt Yoav Hachohen's [58] solution to compromise the greedy minimization process of E to achieve consistent matching results by applying consistency checks on the random search method as the following Algorithm 1 and explanations.

Algorithm 1: General PatchMatch (GPM) with consistency checks

Result: Consistent constrained dense correspondence map

```
1 GPM initialization;
2 for Parameter range from coarse to fine do
3   GPM convergent searching results;
4   Neighbor pixel consistency check;
5   Local random region consistency check;
6   Small regions elimination;
7   Narrow parameter range  $(X, Y, \theta, scale)$ ;
8 end
```

While in the GPM algorithm the natural coherence among overlapped blocks works as weak constraints for the neighborhood propagation and has no guarantee for the consistent results, the additional consistency checks are targeted to remove inconsistent mapping results in neighboring pixels and local regions. The rest consistent results are preserved as the initial results

for the next round of GPM estimation. Hereby we introduce the neighbor pixel consistency check in detail. First, the adjacent matches are defined as consistent if their nearest neighbor field transformations are similar. The underlying reason for the neighbor consistency checks is that although it is hard to determine whether the matches are accurate or not individually, the several consistent adjacent matches are more likely to be reliable compared with any individual match. We use the L_2 -norm of the difference between a pair of matches as the computed consistency error. If the consistency error for the match pair is below the given threshold, we generate the connected components in the four-connected neighborhood of the image. In this way we can gradually obtain the desired connected region where all the adjacent matches are consistent. In the consistent region Z obtained by neighbor pixel consistency checks, all the adjacent pixels are consistent, but the pair of matches further apart might not be. To avoid the checking for every pair of matches in region Z , we only examine a random subset of pairs in Z . In practice, we sample $\sqrt{|Z|}$ pairs of matches for neighbor pixel checking. In this local random region consistency check, we only accept the consistent region when the ratio of the consistent pairs to the total number of pairs is above 0.5. Considering that the small consistent regions is likely be coherent by chance, in the small regions elimination stage, we eliminate the small regions with the number of pixels smaller than 500.

By narrowing down the searching parameter ranges $(X, Y, \theta, scale)$ gradually, the following round of GPM always produces more consistent results in these regions. At each finer parameter range, we adapt the search range using the parameters of the matched transformation in the reliable consistent regions obtained in the previous steps. At last we keep the final consistent results as the prediction results for the new reference reconstruction. In Figure 3.3, the estimated dense correspondence maps without and with consistency check are shown, where the 2-dimensional flow vectors are visualized in the color space. After multiple iterations, only the consistent dense correspondence are preserved according to the refinement constraint. It can be observed that in the texture and structure image regions the correspondences are

well preserved, while the correspondences in the textureless regions are mostly wiped out by the adopted consistency check, because their matching results are relatively random and inconsistent. In addition, since the coding gain for the textureless or smooth regions is usually insignificant, the consistency check leads to quite limited side effects when we abandon the inter-prediction for these regions. More dense correspondence estimation results are shown in Figure 3.4 (a) and (b).

Figure 3.3: Dense correspondence estimation for the reference image (a) and current image (b). (c) and (d) are the visualized flow map without consistency checks and with consistency checks, respectively.

3.1.2 New reference reconstruction

We utilize the regular block division scheme to parameterize the consistent prediction results in the local region. By partitioning the matched pixels of the current image I_c into regular non-overlapped units, the correlation in each unit pair is approximated with homograph trans-

(a) Matched points in I_c

(b) Corresponding points in I_r

(c) One selected unit in I_c

(d) Corresponding unit in I_r

(e) Current image I_c

(f) Reconstructed reference I_{rec}

Figure 3.4: Dense correspondence based prediction and new reference reconstruction. (Note that the matched pixels in (a) and (b) are partially shown every 32 pixels.)

formation, and the luminance adjustment is employed to balance the luminance variations in each unit as well. The new reference reconstruction process is shown in Figure 3.4. The new reference I_{rec} has the same resolution as the current image I_c and the coordinates are partitioned into regular non-overlapped $N \times N$ units from the left-top. To balance the tradeoff between the redundancy reduction and SI expense, we choose the multiples of the maximal coding unit size in HEVC as the basic unit size. In our experimental setting, the multiplier is 4 and the unit size is 256×256 . Units with the low matched points percentage are normally unreliable and not worthy to further reconstruct for prediction. We choose the unit selection threshold as $\tau = 0.2$ empirically. In addition, we employ the general RANSAC algorithm [55] to find the homograph transform H for the approximation of the geometric relationship in each of selected units in Eq. 3.4.

$$U_r = U_c \cdot H, \quad (\text{Eq. 3.4})$$

where

$$H = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}, \quad (\text{Eq. 3.5})$$

and u_c and u_r are the matched point coordinates in each unit from I_c and I_r . The homograph transform H with nine parameters is used to approximate the inter-unit relationship between the image pair by capturing the major deformation as shown in Figure 3.4 (c) and (d). Based on the geometric correlation parameterization, the geometric transformation and luminance adjustment in each unit are performed to reconstruct new reference I_{rec} as Eq. 3.6.

$$I_{rec}(U_c) = \alpha \cdot I_r(U_c \cdot H) + \beta, \quad (\text{Eq. 3.6})$$

where

$$\alpha = \sigma(I_c(U_c)) / \sigma(I_r(U_c \cdot H)), \quad (\text{Eq. 3.7})$$

$$\beta = \mu(I_c(U_c)) - \alpha \cdot (I_r(U_c \cdot H)), \quad (\text{Eq. 3.8})$$

I_c and I_r are the current and reference images, and I_{rec} is the reconstructed reference image. U_c are coordinates of pixels in the selected unit in I_c and $U_c \cdot H$ are their corresponding coordinates in I_r , μ and σ are the mean and variance operations, α and β are the luminance scale and offset parameters. In the experimental settings, we perform the adjustment only for luminance channel because the luminance component is more important in HEVC with 4:2:0 color space sampling. Next, the new reference I_{rec} is reconstructed in the selected transformed units and the rest uncovered regions, including the regions with low number of matched points and the regions that are not matched by the dense correspondence estimation as well, are directly filled with contents from I_r as shown in Figure 3.4 (f). At last, we utilize the HEVC inter-prediction codec to encode I_c with the prediction from reference I_{rec} . The SI consists of selected unit coordinates, homograph matrices, and luminance adjustment parameters. The matrices are quantized and rounded with the appropriate constant quantization table. The selected unit coordinates and luminance adjustment parameters are also quantized and round properly. The quantized parameters are compressed with arithmetic coding and stored along with video-like encoded bitrate.

3.2 Experimental Results

To conduct the fair experimental comparisons, we re-implement the SIFT-based multi-model method proposed in [1]. We utilize the HEVC HM15.0 for both proposed method and the SIFT-based multi-modal method, and choose the Quantization Parameter (QP) setting of 22, 27, 32 and 37. The experiments have been conducted on four image sets: ‘Mail room’ set, ‘Notre Dame’ set, ‘Wadham college’ set provided in [1] and ‘Mount Rushmore’ set collected from Google Images as shown in Figure 3.5. There are four pair of test images, and their resolutions are ‘1072x600’, ‘1024x768’, ‘1024x768’, and ‘1600x1064’, respectively.

The image pair encoding results are presented in Figure 3.6. The SIFT-based multi-model method saves 31.57%, 51.76%, 38.63%, and 9.05% over the standard HEVC inter-prediction

Figure 3.5: Image set compression test image pairs.

(Top row: image pairs from ‘Mail room’ and ‘Wadham College’;
Bottom row: image pairs from ‘Notre Dame’ and ‘Mount Rushmore’)

coding in the four image pairs using BD-PSNR, while our approach achieves 43.63%, 55.60%, 34.20%, and 18.73% bitrate saving. We can observe that our approach achieves better coding performance for the image pairs in Figure 3.6 (a) and (b). In Figure 3.6 (c), the performance is comparable as the SIFT-based multi-model method. The reason is possibly that SIFT-based method is more advantageous for the global alignment of images captured in the different point view, while the block-based local reconstruction in the proposed method may result in some discontinuous content around the block border, which influences the prediction performance of the new reconstructed frame. It is obvious that the inter-image coding improvement is content-dependent. Although in Figure 3.6 (d) the images collected from Google Images are not as similar as other three sets, the improvement by our approach is still significant and demonstrates that our local luminance adjustments are more robust compared with the feature-based global photometric transformation when the local variations are dominant and luminance relationship is complicated.

Compared with the state-of-the-art methods, our approach achieves better performance and is robust to complex luminance variations, especially when the local variations are dominant.

Figure 3.6: Coding performance comparisons. (“Intra” and “Inter” curves are the HEVC Intra-prediction coding and Inter-prediction coding, and “SIFT-based multi-model” [1].)

Our dense correspondence based prediction approach has close correlation with BME & BMC in video coding, and it has the potential to more accurately estimate the encoding improvement for the entire image set encoding structure generation.

3.3 Conclusion

As a response to the new requirement and challenge in big data and cloud-based storage, we have proposed a novel dense correspondence based prediction approach for image set compression, which achieves better performance compared with relevant state-of-the-art methods especially when the local variations are dominant. The dense correspondence is first estimated to predict the pixel-to-pixel relation via fast-convergent random search, and with the parameterization of the prediction results, we reconstruct a new reference image by geometric transformations and luminance adjustments for the subsequent video-like encoding. Because of the reconstruction in local units, our approach is more robust to the complex local variations compared with the global estimation based methods.

Chapter 4

Backward Registration for Image Retargeting

This chapter is organized as follows. Section 4.1 provides our motivation to model image retargeting process. Section 4.2 formulates the problem and proposes an effective backward registration method for geometric change estimation. Section 4.3 highlights the difference between forward and backward registration. Section 4.4 reports the experimental results and Section 4.5 concludes the study.

4.1 Motivation to Model Image Retargeting Process

4.1.1 Challenge to evaluate the retargeted image

The visual quality evaluation of retargeted images is a semantic high-level task. As the analysis in [6, 13, 16], the objective of image retargeting is to preserve the important image content while avoiding the introduction of the visual distortion. However, due to the complicated relationship between the original and retargeted images, it is difficult to measure the information loss and visual distortion directly.

Similar to the study [6], it is crucial to capture the retargeting modification that the image has undergone without specifying the retargeting operator. As shown in Figure 4.1, it is helpful for the image retargeting quality assessment to acquire the details about how the butterfly and flowers are changed. We suggest that before the quality evaluation it is necessary to

approximate the retargeting modification, so that the image content in the original and retargeted images can be linked to find the corresponding quality degradations.

Figure 4.1: Illustration of how the image content is modified and the visual quality change during image retargeting process.

4.1.2 Difference between IQA and IRQA

In the traditional IQA, the reference and distorted images are normally assumed to be well-aligned and the pixel-to-pixel relationship is self-evident. The common image distortions, including Gaussian noise, blur, and JPEG compression artifacts, are mostly the *intensity* change, and the substantial difference can be obtained by the direct subtraction of two images. In IRQA, it is unlikely to obtain the undergone artificial retargeting modification on images simply by the subtraction. Before the quality evaluation, it becomes necessary to find the relationship between the original and retargeted image in distinctly different resolutions.

When we investigate the retargeted image generation process, we classify the artificial modifications into *intensity* change and *geometric* change¹. The *intensity* change mostly happens in the continuous retargeting operators such as SCL and WARP when the subpixel interpolation is involved. The subpixel sampling may create the intensity fluctuations such as blockiness and aliasing artifacts [7].

¹The geometric change describes how the retargeted image is resampled from the original image with the resampling grid (see details in subsection 4.2.1).

Figure 4.2: Comparisons of *geometric* and *intensity* changes (Part I). The SSIM scores for regenerated images are 1.0000 (CR), 0.9078 (SV), 0.9322 (MOP) and 0.9366 (SC).

The *geometric* change refers to the artificial modification on the image grid to achieve overall resolution alternation. Note that the *geometric* change discussed here is different from

Figure 4.3: Comparisons of *geometric* and *intensity* changes (Part II). The SSIM scores for regenerated images are 0.9006 (SCL), 0.9270 (SM), 0.8965 (SNS) and 0.8993 (WARP).

the projection change in the real physical world, where the *geometric* deformation is caused by viewpoint disparity [24].

In Figures 4.2 and 4.3, we investigate the influence of the *intensity* change and *geometric* change on the process of the retargeted image generation. It includes four groups of retargeted images, the regenerated images, and error maps. The regenerated images are built from original images with the estimated *geometric* change, so there may be the intensity errors between the retargeted and regenerated images as shown in Figures 4.2 (c) and 4.3 (c). The SSIM [37] quality scores for regenerated images are 1.0000 (CR), 0.9078 (SV), 0.9322 (MOP), 0.9366 (SC), 0.9006 (SCL), 0.9270 (SM), 0.8965 (SNS) and 0.8993 (WARP). It shows that the regenerated images in Figures 4.2 (b) and 4.3 (b) are generally in high quality compared with reference images from Figures 4.2 (a) and 4.3 (a). As we can see, with the estimated geometric change it is feasible to regenerate the nearly identical retargeted image from the original image. It indicates that the geometric change plays an important role in the generation of the retargeted image, and the estimation of the geometric change is an efficient way to discover the relationship between the given images. This is the reason that we adopt the geometric change as the guidance for the image retargeting quality assessment.

4.2 Backward Registration Estimation

4.2.1 A unified interpretation of image retargeting

To estimate the geometric change during the image retargeting, it is necessary to have in-depth understanding of different retargeting operators. Here we formulate different kinds of retargeting operators in a unified interpretation framework.

We use four representative retargeting operators to introduce the unified interpretation of image retargeting, as shown in Figure 4.4. In each model, the top part is the regular retargeted image grid (red nodes), while the bottom parts are the regular original image grid (gray nodes) and the resampling grid in different shapes (blue nodes). The image retargeting is interpreted as the resampling grid generation and forward sampling. The resampling grids are generated in different shapes according to the retargeting operator and the green arrows denote the forward

Figure 4.4: Typical grid models for image retargeting, where the blue grid $\mathbf{L}$ is the resampling grid and the yellow region $\mathcal{L}_c$ is the resampling space.

sampling from original image by directly pixel copy or interpolation at subpixel locations, to construct retargeted images in regular grids.

The CR operator is to manually choose an optimal rectangle window and directly copy the image content within the window. Instead of the fixed cropping window, the content-aware discrete methods generalize CR by determining pixels for removal in a seam-wise or pixel-wise way. The reserved pixels are collected together to make up the retargeted image in regular grids. The SCL operator resizes the image in one direction uniformly. It inflates the resampling grid of target resolution onto the original image size to cover all the image content and resamples

Figure 4.5: Grid models for the image enlargement retargeting.

the image based on the new grid via the interpolation methods, e.g. bi-cubic interpolation. The content-aware continuous methods can be regarded as extensions of SCL, which manipulate the resampling grid consistently in a more sophisticated way by certain energy minimization process.

With these grid models, we are able to interpret the image retargeting with two steps: resampling grid generation and forward resampling. The resampling grids for discrete methods are special cases of the grids for the continuous methods at the sub-pixel level, thus we can adopt the subpixel level resampling grid to model different retargeting operators. Accordingly, if the resampling grid coincides with the original image grid, a.k.a. at pixel level, the forward resampling process is the direct pixel copy; otherwise, the retargeted image pixels are resampled at the sub-pixel location with the adopted interpolation method.

In addition, we generalize the grid model for image enlargement cases as shown in Figure 4.5. In Figure 4.5 (a), the image enlarged by seam-carving is obtained by inserting new seams.

As illustrated by the grid model, the original image (3×4) is enlarged to the retargeted image (5×4) by the insertion of two new seams, where we search two optimal vertical 8-connected seams (lowest energy) and duplicate these seams. The bold blue seams are the computed optimal seams and the new inserted green seams are duplicated from bold red seams (The resampling locations for bold red seams and green seams are actually same, but we show them separately for better visualization). For the continuous image retargeting operators, the resampling grid is denser than the original image grid, thus the original image is enlarged by interpolation of the resampling grid. As shown in Figure 4.5 (b), the regular grid (5×4) is denser than the original grid, based on which we interpolate the original image to obtain the scaled image. Furthermore, the density of the resampling grid can be content-aware non-uniform and complicated for the more sophisticated retargeting operators.

Obviously, we are able to conclude that the retargeted image is almost determined once the resampling grid is generated. The resampling grid represents the geometric change of the image during the retargeting process and is the strong evidence about the retargeting modification details. As we can observe, the part with dense resampling grid reserves more image content while the zero-density parts can remove the corresponding image region. The sophisticated retargeting operators usually try to preserve visually important content through dense sampling and shrink or even remove other visual unimportant content through sparse or zero sampling. On the other hand, some retargeting operators may destroy the regularity of the resampling grid in the important regions, e.g. foreground objects, people, and faces, which would cause the obvious annoying visual artifacts.

4.2.2 Problem formulation

According to the unified interpretation above, we need to estimate the resampling grid to reveal the geometric change for IRQA. Compared with the forward resampling, to reveal the resampling grid is a reverse problem, and here we formulate it as a *backward registration*

Figure 4.6: Markov random field (MRF) model for dense correspondence estimation, where the red node p represent the observed pixel in the retargeted image, and the blue node l is its resampling location (x, y) from the resampling grid. The subscripts $\{L, T, B, R\}$ denote the four-connected neighborhood.

problem with Markov Random Field (MRF). Given the original and retargeted image pair $I = \{I_{org}, I_{ret}\}$, the retargeted image grid is chosen as the graph for labelling. We denote $\mathcal{P}$ as the set of pixels in I_{ret} , $\mathbf{L}$ as the resampling grid, and $\mathcal{L}_c$ as the continuous resampling space corresponding to the I_{org} , as shown in Figure 4.6. We assume that $p \in \mathcal{P}$ is the observed pixel from the retargeted image and in grid $\mathbf{L}$, each node $l_p = (x_p, y_p) \in \mathcal{L}_c$ is the unknown resampling location for pixel p , where the coordinates are $(x_p, y_p) \in \mathbb{R}$.

Without loss of generality, there are two assumptions about the resampling grid: the uniqueness assumption and ordering assumption. The uniqueness assumption is that each pixel in I_{ret} has one unique resampling point in I_{org} , which is a one-to-one mapping between the pixels in I_{ret} and the resampling grid $\mathbf{L}$. The uniqueness assumption is used to exclude the forward registration as we will discuss in Section 4.3. The resampling grid $\mathbf{L}$ following the ordering assumption should be monotonic without any fold, which is reasonable for qualified image retargeting operators, otherwise the image resampled based on the resampling grid would be

incomplete. The grid nodes $(x_i, y_j) \in \mathbf{L}$ satisfy the constraints *s.t.* $x_i < x_{i+1}$ and $y_j < y_{j+1}$, where $i, j \in \mathbb{N}$. It indicates the ordering of the resampling location for retargeted image should be same to the pixel ordering in the original image.

We model the resampling grid with the MRF and the backward registration problem can be formulated as the labelling problem for retargeted image pixels in the energy minimization framework as

$$E = E_d + \lambda E_s. \quad (\text{Eq. 4.1})$$

The data term E_d describes the likelihood that pixel p is resampled from the location l_p in the original image based on the similarity distance of them using

$$E_d = \sum_p d_p(f(p), f(l_p)), \quad (\text{Eq. 4.2})$$

where d_p is the distance between the feature $f(p)$ for p and the feature $f(l_p)$ at location (x_p, y_p) may be involved with subpixel interpolations when $(x_p, y_p) \notin \mathbb{N}$.

The smoothness term E_s corresponds to the ordering assumption in the modelling assumption

$$E_s = \sum_{(p,q) \in \mathcal{N}} V(l_p, l_q), \quad (\text{Eq. 4.3})$$

where $\mathcal{N}$ is the local neighborhood in the four-connected retargeted image grid. $V(l_p, l_q)$ measures the cost of assigning two resampling locations l_p and l_q to the neighboring pixels. The smoothness term is used to achieve the smooth monotonic resampling grid, while the well-designed data term contributes to the accurate backward registration results.

4.2.3 Problem approximation

There are several algorithms such as graph-cut and loopy belief propagation to solve the energy minimization problem [33]. To solve the problem, we first consider the label space for the resampling grid. When $\mathcal{L}_c$ is the continuous resampling space, the space can be very large

(say $10^5 - 10^8$) even when we quantize it at the quarter subpixel precision. It means that the complexity of the data term D_s is extremely high. Moreover, the subpixel precision $f(l_p)$ in the I_{org} is always related to the interpolation, which is also time-consuming.

A practical way is to approximate the model by replacing the continuous space $\mathcal{L}_c$ with the pixel-level discrete space $\mathcal{L}_d$. With the consideration that the subpixel error of the resampling grid is quite small compared with the space $\mathcal{L}$, the tradeoff between the subpixel accuracy and fast approximation is worthy as long as the designed IRQA metric is robust to the backward registration errors.

Let $p \in \mathcal{P}$ be the pixel from the retargeted image and in grid $\mathbf{L}$ each node $l_p = (x_p, y_p) \in \mathcal{L}_d$ be the hidden resampling location for pixel p , where the coordinates are $(x_p, y_p) \in \mathbb{N}$.

4.2.4 Objective energy minimization

The objective function Eq. 4.1 is defined as:

$$E_d = \sum_p \|f(p) - f(l_p)\|_1 \quad (\text{Eq. 4.4})$$

$$E_s = \sum_{(p,q) \in \epsilon} \min(\alpha |x_p - x_q|, d) + \min(\alpha |y_p - y_q|, d) \quad (\text{Eq. 4.5})$$

where the truncated L_1 -norm distance metric $\min(|*|, d)$ is used in the smoothness term to make the objective function robust to outliers and large discontinuities with d as the threshold and the ϵ is the four-connected neighborhood around the pixel p in the retargeted image grid. In the smoothness term, according to the image coherence, we expect that the resampling location $l_p = (x_p, y_p)$ of pixel p is similar to the resampling location $l_q = (x_q, y_q)$ of pixel q in the four-connected neighborhood of p . We adopt energy function minimization implementation in [39], which is a dual-layer loopy belief propagation based algorithm and utilize a coarse-to-fine scheme to speed up the optimization. The implementation parameters are the default setting $\alpha = 2$, $d = 40$, $\lambda = 1$, $nIterations = 60$ and the adaptive pyramid level $nlevels = \lceil \log_2(MAX(W, H)/10) \rceil$ for the original image in $W \times H$ resolution, where $\lceil \cdot \rceil$ is the ceiling operation.

4.2.5 The proposed hybrid descriptor

The smoothness term mainly guarantees a smooth monotonic resampling grid, while the feature descriptor choice in the data term is important to obtain accurate backward registration results. To get an accurate backward registration results, we develop a hybrid feature descriptor by using CIE-Lab intensity component, dense SIFT descriptor and the relative position component

$$f_p = [\text{Lab}_{1 \times 3} \text{ SIFT}_{1 \times 128} \text{ Pos}_{1 \times 2}] \quad (\text{Eq. 4.6})$$

The intensity component is the pixel intensity in CIE-Lab color space. Since the intensity change due to subpixel interpolation is minor, the intensity value is reliable for the likelihood estimation. However, it is not distinctive enough when there are similar pixels in the neighborhood. Recently the dense SIFT descriptor has shown promising performance for matching across scene [39], where the SIFT descriptor is the histogram of the gradient information in the local neighborhood. With the dense SIFT component, the included local neighbor information helps to formulate a distinctive descriptor. With these two components, the feature descriptor might still be ineffective at the smooth or textureless regions. Thus, we also adopt the relative positions of the pixels as the third component. For each image, the pixel coordinates are normalized into the range of $[-1, 1]$. The relative position component complies with the ordering assumption for the resampling grid and has the tendency to inflate the retargeted image grid onto the original size.

4.3 Forward and Backward Registration Comparison

The backward registration results are shown in Figures 4.7 and 4.8. In each group, the first row is the retargeted image and the second row is visualized backward registration result by distributing retargeted image pixels onto their resampling locations. We can find that the revealed resampling grids are quite appealing for different retargeting operators. The accuracy evaluation for the backward registration solution is shown in the following section.

Figure 4.7: Backward registration results on ‘Brasserie_L_Aficion’ image set (Part I).

The contrary way is the forward registration and obviously we can formulate the model with the original image grid as the grid graph and the retargeted image as the resampling space. The objective function and solution can still be the same. One may wonder whether the forward

Figure 4.8: Backward registration results on ‘Brasserie_L_Aficion’ image set (Part II).

registration can reveal the geometric change similarly, although it is obviously inconsistent with our unified resampling interpretation and the results are not the resampling grid as well.

In Figure 4.9, we provide a comparison sample to show the drawbacks of the forward regis-

Figure 4.9: Comparison of *forward* and *backward* registration by measuring information loss.

tration. To better visualize the difference, we choose the CR operator with the consideration of its apparent cropping window. The information loss measurement is perceptually intuitive and also suitable as the measurement for comparison. We create a toy visual importance map as Figure 4.9 (c) with horizontally different colors. In Figure 4.9 (e), it is the visualized forward registration results by reconstruction with the retargeted image pixels. There are three parts to be noted: the red rectangle is the chosen cropping window, the gray regions are the pixels

which have no appropriate matches in the retargeted image, and the region within the green box represents the pixels that find the wrong matches in the retargeted image. In this case, the data term cannot find any suitable match. The remaining effective smoothness term for these pixels is reversed and we can see the matches in the yellow dash box are mirrored compared with the correct matches on their right side.

To measure the information loss based on the forward registration results, one feasible way is to warp the visual importance map to the retargeted image size and measure importance loss amount compared with the original one. In Figure 4.9, (d) and (f) are two warped visual importance maps in the scan-order from top-left to right-bottom, and top-right to left-bottom, respectively. There are obvious holes in the warped importance maps. The reason is that the mismatches for the removed part override the correct matches due to the many-to-one mapping. This will seriously disturb the information loss measurement. In contrast, the backward registration satisfied with the uniqueness assumption can avoid the unnecessary matches. Similarly, in the bi-directional registration, there is the same problem in the forward part and only the backward part is useful. In Figure 4.9 (g) and (h), the backward registration results reveal the cropping window correctly and based on the estimation we can measure information loss in a reliable way.

4.4 Experimental Study for Backward Registration

4.4.1 Validation databases with the groundtruth

To evaluate the accuracy of the geometric change estimation by backward registration, it is necessary to provide the databases with the retargeting modification as the ground-truth information. However, due to limited number of retargeting operators that provide the source codes, here we only choose the Seam-Carving [59] and Scale-and-Stretch (SNS) [12] to build the validation databases as shown in Figure 4.10. To obtain the ground-truth information, we implement the SNS in MATLAB and the code is public available at [yabinzhangjohn.github.io](https://github.com/yabinzhangjohn).

Figure 4.10: Validation databases. (a) Original image ‘ArtRoom’. (b) Image retargeted by SC with 25% width reduction. (c) The ground-truth for (b): a binary map indicating the removed pixels (black regions are the removed seams). (d) The image retargeted by SNS with 50% width reduction. (e) The approximated ground-truth map for (d).

We use source images from both MIT and CUHK databases. In MIT part, we use the 37 source images from MIT database [60] and retarget the images into similar target resolutions. In CUHK part, we choose the 57 source images and build all the images by retargeting each image with 50% width reduction.

In Figure 4.10 (c), the black regions represent the removed seams by SC. With the binary map we can reconstruct the identical retargeted image by removing the pixels in black regions. In Figure 4.10 (e) it is the approximated ground-truth map for SNS. Since the ground-truth information for the continuous image retargeting operators is the resampling grid, we transform by brute force the resampling grid into the binary map by approximating each resampling point to its nearest-neighbor pixel for the visualization.

4.4.2 Backward registration accuracy

The performance evaluation of the backward registration results on the validation databases is shown in Table 4.1. The performances of different kinds of descriptors: SIFT, Lab, SIFT+Lab and the proposed descriptor, are evaluated on the SC and SNS validation sets. According to the proposed unified resampling framework, the backward registration accuracy can be effectively evaluated by measuring the similarity between ground-truth and the estimated resampling grids. Here we adopt the Mean Absolute Error (MAE) as the main accuracy measure:

$$\text{MAE} = \frac{1}{N} \sum_i^N \left\| (x_i, y_i) - (x'_i, y'_i) \right\|_1 \quad (\text{Eq. 4.7})$$

where (x_i, y_i) is the ground-truth resampling point and (x'_i, y'_i) is the estimated resampling point; N is the total number of pixels in the retargeted image.

The discrete SC can be regarded as a special case of the continuous retargeting operators. For images retargeted by SC, the accuracy can be measured by whether we have find the removed pixels correctly, so we utilize the Recall (Rc), Precision (Pr), and the overlapped ratio R_o as the additional accuracy measurement. Here, R_o refers to the percentage of pixels in retargeted image with same resampling locations as other pixels, and in the ideal situation R_o should be equal to zero. We extend these metrics to evaluate the accuracy on images retargeted by SNS. To make it feasible to measure Rc , Pr and R_o , we take the brutally approximated binary maps instead of the ground-truth resampling grids for SNS set.

In Table 4.1 the experimental results show that our descriptor achieves the best overall performance. The MAE results show that our descriptor can estimate the resampling grid more accurately. For the Rc , Pr , and R_o on SC validation set, the results show the hybrid descriptor is more advantageous. On SNS validation set, the Lab descriptor shows better Recall and Precision performance. The reason is that the Lab descriptor, without neighborhood information, makes the backward registration a too greedy search strategy for the warping cases with serious deformed resampling grids. As we can see, the MAE for Lab descriptor is still

Table 4.1: Performance evaluation on validation databases

Set	Descriptor	MIT part				CUHK part			
		R_c	Pr	R_o	MAE	R_c	Pr	R_o	MAE
SC	SIFT	0.7831	0.8023	0.034	14.231	0.6277	0.6769	0.077	38.009
	Lab	0.8045	0.8175	0.022	11.383	0.6767	0.6996	0.034	20.914
	SIFT+Lab	0.8164	0.8322	0.024	7.977	0.7075	0.7367	0.040	15.333
	Our hybrid descriptor	0.8257	0.8336	0.014	4.228	0.7304	0.7410	0.014	6.425
SNS	SIFT	0.6306	0.6524	-0.043	25.856	0.4838	0.5279	-0.090	60.271
	Lab	0.7785	0.7806	-0.003	8.816	0.6542	0.6621	-0.013	22.605
	SIFT+Lab	0.6905	0.6952	-0.008	7.058	0.5695	0.5825	-0.023	16.192
	Our hybrid descriptor	0.6947	0.6949	-0.001	3.723	0.5806	0.5812	-0.001	7.888

quite large. Since the images in CUHK part are all retargeted with 50% width reductions, where the discontinuities and deformations in retargeted images are more serious, the overall performance is relatively lower compared with the MIT part. Although the single SIFT descriptor or Lab descriptor is capable to reveal the resampling grid, the proposed hybrid descriptor shows significantly better performance. The overall performance of the backward registration is satisfactory, which is consistent with the results in Figures 4.7 and 4.8, and with its guidance different IRQA metrics can be further developed.

4.5 Conclusion

It is usually unknown that how the image is retargeted into a target resolution for multiple image retargeting operators, and the complicated image relationship impedes the effective perceptual quality evaluation for the retargeted image. To better understand the image retargeting process, we have first provided a unified interpretation of image retargeting with the resam-

pling grid generation and forward resampling. We show that the geometric change estimation is an efficient way to clarify the relationship between the original and retargeted images. We formulated the geometric change estimation as a backward registration problem via Markov random field and provided an effective and practical solution. For image retargeting quality assessment applications, backward registration can provide reliable geometric change guidance, with which we can link the properties between the original and retargeted images and measure the corresponding quality degradation more efficiently. Experimental results on two validation databases show that the proposed backward registration method achieve high overall estimation accuracy qualitatively and quantitatively.

However, as a dense correspondence estimation method, the accuracy of the proposed backward registration approach may suffer from the severe content changes in the discontinuous/twisted or the almost removed regions. It is partly caused by the fact that there is insufficient information to effectively constraint the dense correspondence relationship in these regions for the ill-posed problem. Although we somehow alleviate the negative influence with the adopted strategies, new designed IRQA methods should be robust to the local unsteadiness in the severely changed regions when they are based on the backward registration approach.

Chapter 5

Aspect Ratio Similarity for IRQA

This chapter is organized as follows. Section 5.1 provides the formulation of aspect ratio similarity metric and Section 5.2 gives additional analysis of the metric. Section 5.3 provides the experimental study for aspect ratio similarity and Section 5.4 concludes this study.

5.1 Aspect Ratio Similarity Definition

As discussed in the previous chapter, the geometric change can effectively clarify the dense correspondence relationship between the original and retargeted images. The well-estimated geometric change captures the detail information about how images are retargeted. If we partition the original image into regular blocks, with its guidance we can establish the retargeted blocks in the retargeted images and then estimate whether they are well-preserved, removed or distorted. To effectively evaluate the visual quality of the retargeted image, here we design a novel aspect ratio similarity (ARS) metric by measuring the local block changes between the original and retargeted images.

The purpose of image retargeting is actually the Aspect Ratio (AR) alteration for the entire image. The global AR change is achieved with the local AR changes collaboratively across the image. The AR changes of the local blocks indicate how they are preserved in the retargeted image, based on which we can measure their information loss and visual distortions. Therefore, the visual quality of retargeted images can be assessed by measuring the local AR changes.

Figure 5.1: (a) Original image, ‘ArtRoom’. (b) Visualized backward registration results, where each pixel from the retargeted image is distributed onto its resampling location in the original image. (c) Image retargeted by SC. (d) The geometric change from the regular blocks in the original image (left) to the retargeted blocks in the retargeted image (right), the block along the lime route is preserved in high quality while the block along the red route is squeezed seriously.

As shown in Figure 5.1, the original image is divided into $N \times N$ regular blocks and the corresponding retargeted blocks are established in retargeted image under the geometric change guidance. More results are shown as the second columns in Figures 4.7 and 4.8. We calculate the maximal width w_{ret} and maximal height h_{ret} of each retargeted block. The height and width change ratios can be denoted as $r_w = w_{ret}/N$ and $r_h = h_{ret}/N$, and the mean ratio $u_r = (r_w + r_h)/2$ denotes the absolute block size changes. The similarity score of the block pair is formulated as follows:

$$S = \left[\frac{2 \cdot r_w \cdot r_h + C}{r_w^2 + r_h^2 + C} \right] \cdot \left[e^{-\alpha(u_r - 1)^2} \right] \quad (\text{Eq. 5.1})$$

where C is a small positive constant to increase the stability by avoiding the division by zero and the α is the parameter larger than zero to balance the information loss and visual distortions by adjusting the information loss penalty degree.

5.2 Aspect Ratio Similarity Analysis

In Eq. 5.1, the first aspect ratio (AR) term measures the aspect ratio similarity between the block pairs. Since the aspect ratio loses the absolute size change information, in the second absolute size (AS) term we utilize the Gaussian function of the mean ratio of u_r to take into account the absolute size change influence. When S is close to 1, the block content in the original image is generally preserved in high quality, while when S is close or equal to zero, it indicates that the retargeted block is suffering from serious information loss or visual distortions. In Figure 5.2, we provide a 3D plot for the aspect ratio similarity, where the magenta dashed line denotes the points with $r_w = r_h$ and the cyan dashed lines denote the points with $r_w = 1$ or $r_h = 1$. The absolute size term changes along the direction indicated by the magenta dashed line and the aspect ratio term changes along its perpendicular direction. As we can observe in the plot, ARS always prefers the well-preserved blocks with $r_w = 1$ and $r_h = 1$. The aspect-ratio-preserved but squeezed blocks and squeeze-and-stretch blocks, mainly suffer from information loss and visual distortion, respectively. In addition, we can also observe that the formulation of ARS can deal with both the shrinkage and enlargement of the blocks.

To predict the visual quality of the entire retargeted image, we utilize the saliency detection method [52] specifically designed for image retargeting to calculate the saliency map as our visual importance map. The visual quality score ARS for the retargeted image is defined as Eq. 5.2 by pooling the similarity score of each block with the corresponding visual importance weight.

$$ARS = \sum_m \sum_n S_{mn} \cdot V_{mn} / \sum_m \sum_n V_{mn} \quad (\text{Eq. 5.2})$$

Figure 5.2: 3D plot for aspect ratio similarity

where S_{mn} represents similarity score S for block (m, n) , V_{mn} denotes the sum of the visual importance value within the block (m, n) , where the visual importance map is partitioned regularly in the same way as the original image; m and n are the block indices. The ARS evaluation on ‘ArtRoom’ set is illustrated in Figures 5.3 and 5.4, where we can observe that the local deformation are captured by the designed ARS and the visual importance pooling makes the measurement content-aware.

To obtain more intuitive understanding of ARS metric, we provide one toy example to show the quality measurement for different types of block changes in Figure 5.5. The original toy image (a) is retargeted to image (b) via the combination of CR and SCL. We investigate the typical well-preserved, totally removed, and horizontal uniformly scaled blocks. The block B_4 is well-preserved from the original image, and thus both AR and AS terms are equal to 1. As we all know, if one image is uniformly down-scaled in both width and height, there should be little visual artifact. When we uniformly down-scale the local blocks, AR term is always equal to 1 and AS term takes account of the overall size changes for information loss. When it comes to the extreme case, e.g. totally removed block B_1 , AR term is still equal to 1 while AS term

Figure 5.3: ARS evaluation on ‘ArtRoom’ set (Part I). (a) Retargeted images. (b) ARS maps. (c) Importance weighted ARS maps.

reduces to the minimal $e^{-\alpha}$. For the horizontal scaling case from B_3 (64×64) to B_5 (64×32), AR term becomes to 0.8 and AS term reduces to $e^{-0.06\alpha}$. We can see that AR and AS terms are used to evaluate the information loss and visual distortion for each block collaboratively.

Another problem we need to address is the block approximation in the retargeted image. We observe that the retargeted blocks may have irregular shapes and it is difficult to measure

Figure 5.4: ARS evaluation on ‘ArtRoom’ set (Part II). (a) Retargeted images. (b) ARS maps. (c) Importance weighted ARS maps.

the ratio change of the various kinds of shapes themselves. The bounding box approximation of the irregular retargeted shape provides an effective solution to capture the local changes by calculating the dimension change. When one block and the neighboring blocks are preserved in good shape, they will all be assigned high ARS scores. If one single block is deformed to an irregular shape, it will lead to the quality drop in either of the block itself and its adjacent

Figure 5.5: A toy example. (a) Original image (512×384). (b) Image (192×384) retargeted by cropping out the middle half of the original image and scaling the right half of the rest by 50% width reduction. Original image is divided into 64×64 non-overlapped blocks as shown in (c) and the corresponding retargeted blocks are illustrated in (d), where B_1 is removed totally, B_2 is well-preserved as B_4 , and B_3 is scaled to B_5 .

block or both of them. The illustration example is shown in Figure 5.6. The middle rectangle block in the first row is sheared into a red parallelogram in the second row. The width ratio of its bounding box is $r_w = 2$ and the height ratio is $r_h = 1$, but its neighboring two trapezoid blocks still have the same ratio changes as $r_h = 1$ and $r_h = 1$. Although the bounding boxes of its two neighboring trapezoid blocks are unaware of the deformation of their embedded shapes, the bounding box of the red parallelogram can still capture the major deformation in the local region. From this example, we can see that with the aid of neighboring block interaction, ARS using the bounding box approximation is still feasible to capture the image content deformation in the local neighborhood.

Figure 5.6: The influence of block approximation

5.3 Experimental Study for Aspect Ratio Similarity

5.3.1 Image retargeting databases

In the experiments, we adopt two widely used public databases: MIT RetargetMe [60] and CUHK databases [61, 62] to evaluate the performance of the proposed objective quality metric, in terms of the correlation between the objective scores and the subjective scores provided in the databases. The detailed information of these databases are summarized in Table 5.1 and their chosen sources images are shown in Figures 5.7 and 5.8.

Table 5.1: Benchmark databases for IRQA evaluation

Database	MIT	CUHK
Source Image No.	37	57
Retargeted Image No.	296	171
Retargeted Image -25% No.	184	46
Retargeted Image -50% No.	112	125
Retargeting Operator No.	8	10
Subject No.	210	64
Subjective Score Type	Pair-wise	MOS

Figure 5.7: Original images from MIT RetargetMe database.

In MIT database, there are 37 source images as shown in Figure 5.7 and their retargeted images are generated by eight retargeting operators, which are CR, SCL, SC [8], MOP [14], SM [9], SNS [12], SV [11], and WARP [10] with 25% (23P) or 50% (14P) of width or height reduction. There are six major image attributes provided for better insights: *Line/Edge*, *Face/People*, *Foreground Objects*, *Texture*, *Geometric Structures* and *Symmetry*, and each image may own more than one attribute. The subjective tests are conducted in the way of the paired comparison, where subjects choose the better one from two retargeted images shown in juxtaposition. The database provides the corresponding numbers of times that the retargeted image is favored over others as the subjective scores. The correlations between the objective and subjective scores are measured by the Kendall Rank Correlation Coefficient (KRCC) [6]:

$$\text{KRCC} = \frac{n_c - n_d}{0.5n(n-1)}, \quad (\text{Eq. 5.3})$$

where n is the length of the ranking (here $n = 8$), n_c is the number of concordant pairs and n_d is the number of discordant pairs from all the pairs.

The CUHK database contains 171 retargeted images generated from 57 natural source images as shown in Figure 5.8. Compared with MIT database, two other retargeting opera-

Figure 5.8: Original images from CUHK database.

tors, optimized seam-carving and scale (SCSC) [63] and energy-based deformation (ENER) [64], are included as well. Each original image can be retargeted by either 25% or 50% of the width or height reduction. Different from the pair-wise comparison in MIT database, the subjective tests employ 5 category quality scales as “Bad”, “Poor”, “Fair”, “Good”, and “Excellent”, and generate the Mean Opinion Score (MOS) for each retargeted image, which serves for the correlation evaluation between subjective MOSs and the objective scores. It is similar to the subjective testing in the traditional IQA databases [18, 37, 65].

In CUHK database, four commonly used evaluation metrics are employed to evaluate the relationship of the objective quality metric scores and the provided MOSs. The first one is the Pearson Linear Correlation Coefficient (LCC) between MOS and the objective scores after nonlinear regression. The nonlinear regression is the mapping function suggested by Sheikh *et*

al. [66]:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2(x-\beta_3)}} \right) + \beta_4 x + \beta_5. \quad (\text{Eq. 5.4})$$

The second metric is the Spearman Rank-order Correlation Coefficient (SRCC), which can measure the prediction monotonicity of the objective IRQA metric scores. The third metric is the Root Mean Squared Error (RMSE) between MOS and the objective scores after nonlinear regression. The fourth metric, the Outlier Ratio (OR) [67] is the ratio of the false objective score number to the total score number. The false score means that the score lies outside the interval $[\text{MOS}-2\sigma, \text{MOS}+2\sigma]$ after the nonlinear regression, where σ is the corresponding standard deviation. As we can see, larger LCC and SRCC values indicate that the objective metric scores correlate better with the subjective MOSs. In other words, the performance of the objective metric is better. The other RMSE and OR metrics correspond to the correlation errors, thus the smaller values indicate the better objective quality metrics.

5.3.2 Performance evaluation on MIT RetargetMe database

To demonstrate the effectiveness of the proposed ARS metric, we present rank correlation results on the benchmark MIT RetargetMe database [60] in Table 5.2. The proposed ARS is compared with BDS [68], EH [45], SIFT flow [39], EMD [40], CSim [15], IR-SSIM [16] and PGDIL [13]. We have adopted the same saliency detection method [52] as the IR-SSIM and PGDIL.

We give the mean and standard deviation values of the rank correlations as well as p -value in Table 5.2. The significance test using p -value is used to examine the distribution of the KRCC coefficients. In the MIT database, we can estimate the distribution of KRCC for $N = 8$ by considering all rank correlation of all possible permutations of eight elements, and Rubinsein *et al.* [6] found that it has a normal distribution with zero-mean and $\sigma = 0.2887$. We use the χ^2 test the null hypothesis that the observed KRCC coefficients are randomly sampled from the normal distribution with zero-mean and $\sigma = 0.2887$. A small p -value (typically ≤ 0.05)

Table 5.2: Performance evaluation of different metrics on MIT database.

Metric	mean KRCC in each subset						Total		
	<i>Line Edge</i>	<i>Faces People</i>	<i>Foreground Objects</i>	<i>Texture</i>	<i>Geometric Structure</i>	<i>Symmetry</i>	mean KRCC	std KRCC	<i>p-val</i>
BDS	0.040	0.190	0.167	0.060	-0.004	-0.012	0.083	0.268	0.017
EH	0.043	-0.076	-0.079	-0.060	0.103	0.298	0.004	0.334	0.641
SIFTflow	0.097	0.252	0.218	0.161	0.085	0.071	0.145	0.262	0.031
EMD	0.220	0.262	0.226	0.107	0.237	0.500	0.251	0.272	1e-5
CSim	0.097	0.290	0.293	0.161	0.053	0.150	0.164	0.263	0.028
IR-SSIM	0.309	0.452	0.377	0.321	0.313	0.333	0.363	0.271	1e-3
PGDIL	0.431	0.390	0.389	0.286	0.438	0.523	0.415	0.296	6e-10
ARS	0.463	0.519	0.444	0.330	0.505	0.464	0.452	0.283	1e-11

indicates strong evidence against the null hypothesis, so we can reject the null hypothesis. From this table, we can see that the proposed ARS can obtain statistically better performance than the state-of-the-art methods and in image subsets labelled by different attributes except *Symmetry*, the proposed metric achieves the best correlation results among the compared methods. The relative lower performance in *Symmetry* attribute subset shows the limitation of ARS in the measurement of the global and semantic structures distortions.

5.3.3 Performance evaluation on CUHK database

In Table 5.3, we compare the performance of the proposed ARS with BDS [68], EH [45], SIFT flow [39], EMD [40], CSim [15], GLS [46] and PGDIL [13] on CUHK database. The performance is measured by LCC, SRCC, RMSE and OR. As shown in Table 5.3, the proposed ARS performs consistently better than other metrics significantly.

In Figure 5.9, we show the scatter plots of subjective scores against objective scores predicted by PGDIL and ARS on CUHK database. The curves in Figure 5.9 were fitted by the nonlinear regression using Eq. 5.4. The scatter distributions are used to show the correlation between the MOS and objective prediction. The logistic fitted curve shows the overall trend of

Table 5.3: Performance evaluation of different metrics on CUHK database

Metric	LCC	SRCC	RMSE	OR
BDS	0.2896	0.2887	12.922	0.2164
EH	0.3422	0.3288	12.686	0.2047
SIFT flow	0.3141	0.2899	12.817	0.1462
EMD	0.2760	0.2904	12.977	0.1696
CSim	0.4374	0.4662	12.141	0.1520
GLS	0.4622	0.4760	10.932	0.1345
PGDIL	0.5403	0.5409	11.361	0.1520
Proposed ARS	0.6835	0.6693	9.855	0.0702

Figure 5.9: The scatter plots of subjective rating against objective scores on CUHK database.

the objective scores regarding the MOS. In general, the good objective prediction scores should be scattered around the fitted curve closely, and the monotonic fitted curves are preferred as well. In Figure 5.9, the PGDIL prediction scores around the straight line are more scattered compared with the ARS prediction scores around the fitted curve, which objective scores predicted by ARS are more correlated with subjective scores compared with the state-of-the-art PGDIL scores.

5.4 Conclusion

Under the guidance of the geometric change estimated by the backward registration approach proposed in Section 4, we have developed an effective ARS metric by exploiting the local block changes to evaluate the visual quality of retargeted images. Compared with other existing IRQA metrics on the publicly available MIT and CUHK databases, the proposed ARS yields statistically better results in the prediction accuracy. The proposed ARS metric shows the success by measuring the local block changes under the geometric change guidance. Even though ARS metric is designed based on the local low-level features and it is infeasible to explicitly evaluate the specific attribute degradations such as the broken line and violated symmetry, it outperforms other existing previous methods. Some recent studies have already investigated the influence of specific image attributes, including *Line*, *Symmetry* and *Aesthetics*, both for image retargeting and its quality assessment [11, 69, 70, 71, 72]. In the following Chapter 6, we will explore additional detection techniques for image attributes such as *Line/Edge* and *Face*. With the estimated geometric change, we can further establish more attribute correspondence between the original and retargeted images. The explicit evaluation of corresponding attribute quality degradation can help develop better IRQA measures.

Chapter 6

Multiple Level Feature Measure

This chapter is organized as follows. Section 6.1 provides the problem statement and Section 6.2 provides the framework of the proposed MLF measure. Section 6.3 extends the ARS feature. Sections 6.4 and 6.5 present the proposed edge group similarity and face block similarity features. Section 6.6 gives the details about the learning based feature fusion and Section 6.7 provides the experimental study and Section 6.8 concludes this study.

6.1 Problem Statement

The visual importance map is the key to achieve accurate quality prediction performance in existing IRQA studies. It ranks image pixels according to their visual importance, so that the information loss and visual distortion in the important regions are prevented as much as possible. The low-level feature based quality measures [16, 13, 17] highly depend on the visual importance map to achieve the content-aware purpose. Here we show the limitation of the visual importance dependence and the necessity to develop effective features to model the high-level quality degradation aspects.

Although there are many successful visual attention models over the past two decades, the visual importance map for IRQA is not always reliable, or even misleading sometimes. As we can see in Figure 6.9, the importance map identifies the vase as important region but ignores the woman in the bottom left corner, which leads to inconsistent prediction results with human

perception. The low-level feature based measures become vulnerable due to its high visual importance dependence. Although one possible way is to develop better importance map, it will become another very challenging problem. Furthermore, even the ideal importance map has quite limited ability to avoid the visual distortion, e.g. the deformation of lines and structures. The importance map only identifies the important region that should be intact during image retargeting process, but it cannot guarantee the intactness after retargeting, especially when the percentage of important region is large or the important structure distributes over the entire image. This is also the reason that additional constraints, e.g. straight line preservation, are adopted in works [11, 71] to suppress the structure distortions.

Generally we think that visual importance map is effective to measure information loss by estimating the preservation of the important region. On the other hand, it may be inappropriate by only focusing on the distortions in the important region, because the deformations in the unimportant regions with dense information such as shape/structure patterns can also degrade the image quality in a significant way. Currently the existing low-level features are ineffective to capture the annoying artifacts such as face and structure deformations, which are mostly the semantic high-level distortions. As we are seeking an optimal tradeoff between information loss and visual distortion, it is necessary to accurately measure both of them to develop an effective IRQA measure. Therefore, in this chapter, we aim to develop multiple-level features to model the visual quality degradation in the retargeted image better.

6.2 Framework of the MLF Measure

The framework of the proposed measure is shown in Figure 6.1. Due to the complicated relationship between the original and retargeted images, we adopt the proposed backward registration method to estimate the pixel-level correspondence in the first stage. With the estimated correspondence, in the second stage we generate the multiple-level features to model different kinds of property changes during image retargeting. ARS captures the information

loss and visual distortion of local blocks at the same time. EGS is designed to include the shape/structure related visual distortion by estimating the edge spatial arrangement change. FBS penalizes the sensitive facial region deformation in a top-down manner. In the last stage, a fusion model of multiple-level features is learned to predict the retargeted image quality.

Figure 6.1: The block diagram of the proposed multiple-level feature based method. In the first stage, the backward registration is used to estimate the pixel-level correspondence between the original and retargeted images; in the second stage, we generate the multiple-level features to model different aspects of the quality degradation; in the last stage, a learned feature fusion model is used to predict the perceptual quality of the retargeted image.

In general, each of the features is designed to measure the quality change of certain properties such as shape/structure and human face. The property is first detected or labelled in either the original or retargeted image, and the corresponding property is matched in the other image based on the pixel-level correspondence revealed in the first stage. After the property matching procedure, it is feasible to measure the corresponding quality degradation by estimating to what extent the concerned properties change during the retargeting process. While the pixel-level correspondence matching error in the first stage will possibly disturb the matching and the property changes are complicated in different retargeted images, the designed features are in the form of relative large blocks or edge group representations, to achieve the robust property matching and reliable feature generation.

6.3 Aspect Ratio Similarity Feature Extension

We still adopt the low-level feature ARS due to its simplicity and effectiveness. After a brief introduction, we present an in-depth analysis on ARS feature.

6.3.1 Brief introduction

As shown in Figure 6.1, the original image is partitioned into regular blocks (e.g. 16×16) indicated by the red lines, then the corresponding modified blocks are matched in the retargeted image based on the estimated pixel-level correspondence. The retargeted blocks usually have irregular shapes and we use their bounding boxes as an approximation to capture the local block deformation. For each block pair, the similarity is calculated by Eq. 5.1 and the similarity score for the entire image is given by Eq. 5.2. More ARS maps and importance weighted ARS maps are shown as the third and fourth rows in Figures 6.2~ 6.5. We can see that ARS relies on the visual importance map to achieve the content-aware quality measurement.

6.3.2 Tradeoff between information loss/visual distortions

In Eq. 5.2, the first aspect ratio change term mainly corresponds to the visual distortion occurred in images retargeted by the continuous retargeting operators such as SCL and SNS, while the other absolute size change term complements the information loss missed by the aspect ratio term, which avoids that images retargeted by CR are always favored. Parameter α adjusts the relative importance between the aspect ratio change and the absolute size change terms. Larger α indicates more information loss penalty. It controls the tradeoff between the information loss and visual distortion. The performance of ARS with different α values on the MIT [60] and CUHK [61] databases is shown in Figure 6.6. On both databases α has the similar influence on the overall performance. The prediction accuracy increases from 0 to 0.3 and decreases from 0.3 to 1. In the previous chapter, the parameter setting is fixed as $\alpha = 0.3$ for both databases. As we try to develop multiple-level features to measure the information loss

Figure 6.2: ‘Brasserie L Aficion’ image set (Part I). Note that (d) is clustered on the initial edge map of the original image individually and unused in following EGS feature generation.

and visual distortion in different aspects, the tradeoff between information loss and visual distortion will change obviously when ARS works as one of the features. Since we introduce more features in the following to address visual distortion, we adopt ARS feature Q_{AR} at $\alpha = 0.7$ with more information loss penalty to achieve a new appropriate balance between information loss and visual distortion.

6.4 Edge Group Similarity Feature

In the retargeted image, the shape/structure deformation artifacts are common distortions that seriously degrade the viewing experience of users, even when they occur in less important regions. The low-level features have limited abilities to model these visual distortions. Due to their visual importance dependence, the distortions in the unimportant region are likely to be

Figure 6.3: ‘Brasserie_L_Aficion’ image set (Part II). From the top to the bottom in each column in (e) CR, (f) SV, and (g) MOP, there are the image retargeted by the corresponding operator, the visualized backward registration result, the calculated ARS map (brightness indicates the quality score), the importance weighted ARS map, the EGS map for the original image, and the EGS map for the retargeted images, respectively. The edge group’s quality is indicated by the colorbar on the right side.

Figure 6.4: ‘Brasserie.L_Aficion’ image set (Part III). From the top to the bottom in each column in (e) SC, (f) SCL, and (g) SM, there are the image retargeted by the corresponding operator, the visualized backward registration result, the calculated ARS map (brightness indicates the quality score), the importance weighted ARS map, the EGS map for the original image, and the EGS map for the retargeted images, respectively. The edge group’s quality is indicated by the colorbar on the right side.

Figure 6.5: ‘Brasserie_L_Aficion’ image set (Part IV). From the top to the bottom in each column in (k) SNS, and (l) WARP, there are the image retargeted by the corresponding operator, the visualized backward registration result, the calculated ARS map (brightness indicates the quality score), the importance weighted ARS map, the EGS map for the original image, and the EGS map for the retargeted images, respectively. The edge group’s quality is indicated by the colorbar on the right side.

Figure 6.6: Influence of α from ARS on MIT and CUHK databases. The prediction accuracies are evaluated using KRCC on MIT database, and PLCC and SRCC on CUHK database.

ignored as well. We are eager to develop higher level features to capture the shape/structure deformation distortions directly. As humans can perceive the object using different kinds of cues such as color, texture, and shape information, the edge/contour features provide effective information for some object-level tasks, including object detection and recognition [73, 74, 75, 76]. Hence, we are intuitively inspired to use the sparse edge groups based shape/structure representation to model the related mid-level information, and accordingly develop the EGS feature to measure the shape/structure related distortion.

6.4.1 Edge group representation and matching

For the gradient edge information, we are more interested in the gradient spatial arrangement for the shape/structure information rather than the gradient width or magnitude. Given an image we utilize the structured edge detector [77] to generate the initial edge map. The non-maximum suppression (NMS) [78] and edge group clustering procedures [75] are performed to obtain sparse edge group representation as shown in Figure 6.1 and Figure 6.2 (c) and (d), where neighboring edge groups are marked in different colors. The NMS is an edge thinning technique to help reduce the redundant gradient information and preserve the

sparse spatial arrangement of the edge response, which is more suitable as the representation of shape/structure.

In the resultant edge map, each edge pixel has the orientation along the edge response. To facilitate the subsequent similarity measurement, it is preferred to cluster the sparse edge pixels into edge groups as the basic evaluation unit. The edge pixels are clustered greedily along the 8-connected neighboring edge pixels until the accumulative orientation difference summation is larger than the specified threshold $\pi/2$, and the small edge group will be merged with the neighboring edge groups further following the setting in [75].

The designed EGS is to measure the similarity between edge groups from the original and retargeted images, but the separate greedy edge group clustering, which clusters edge groups for the retargeted image and the original image individually, cannot guarantee that the obtained two edge groups are corresponding to the exact same content. This will lead to the unreliable similarity estimation between the edge group pair. Therefore, it is necessary to utilize the pixel-level correspondence revealed in the first stage to match the edge groups.

As EGS feature measures visual distortion in the retargeted image, the main problem is to measure the shape deformation of the edge groups preserved in the retargeted image. Therefore, we first cluster the edge groups in the retargeted image and then match their corresponding edge groups in the original image based on the pixel-level correspondence. To be specific, each edge group $EG' = \{e'_j\}$ in the retargeted image is mapped to the original image to search the corresponding edge group $EG = \{e_i\}$ therein. Since the backward registration estimation is not exactly accurate, we allow the edge pixel searching in the local neighborhood to refine the matching results using the edge information itself. As shown in Figure 6.1, the neighboring edge groups are marked in different colors and the edge group pairs are generally matched in high accuracy.

6.4.2 Edge group similarity

The Chamfer matching proposed by Barrow *et al.* [74] is a popular technique used for contour alignment, object detection and human pose estimation [79, 80]. For the purpose of shape based object searching, it cannot effectively handle the shape variations when the scale, rotation and aspect ratio changes as shown in Figure 6.7, and is also unreliable when the background clutter exists.

Figure 6.7: The illustration of Chamfer matching for different transformed shapes. (a) Original image. (b) The shape contour of the giraffe (white). (c~f) the query shapes (red) with aspect ratio or scale transformation applied on the original shape. The matching distances for (c~f) are 148.5, 759.4, 84.3, and 217.0, respectively. The distance variations show its limitation for the object search.

Since our edge group representations are always in the clean sparse form as shown in Figure 6.1 and Figure 6.2 (c) and (d), the background clutter influence is negligible. After the edge group representation and matching in the previous part, the Chamfer matching, however, can be used to estimate how the shape/structure represented by the edge group deviates from the

original one. At the current stage, we regard all the shape/structure deformations as the perceivable visual distortions, so the Chamfer matching works as an appropriate distance metric to estimate the similarity between the edge group pair.

Here is the Chamfer distance to calculate the similarity between the edge group pair. Let $EG_k = \{e_i\}$ and $EG'_k = \{e'_j\}$ be the k th pair of edge groups in original and retargeted images respectively. In Eq. 6.1, Chamfer distance d_{CM} between EG_k and EG'_k is the average distance of each edge pixel in $e'_j \in EG'_k$ and its nearest match in EG_k , where $L = |EG'_k|$ is the length of edge group EG'_k .

$$d_{CM}(EG_k, EG'_k) = \frac{1}{L} \sum_{e'_j \in EG'_k} \min_{e_i \in EG_k} |e_i - e'_j| \quad (\text{Eq. 6.1})$$

We treat each edge group equally, and EGS for the retargeted image is calculated by Eq. 6.2, where N_2 is the total number of the edge group pairs. The feature score is transformed into the range of 0 to 1 by an exponential operator,

$$Q_{EG} = e^{-\beta \cdot \sqrt{\frac{1}{N_2} \sum_{k=1}^{N_2} d_{CM}(EG_k, EG'_k)}}, \quad (\text{Eq. 6.2})$$

where β controls the distribution of transformed scores. As the parameter study shown in Figure 6.16, different β settings in the range of $[0.1, 1]$ achieve similar performance and in the experiments, we choose the setting $\beta = 0.2$.

More EGS maps for the original and retargeted images are shown in the fifth and sixth rows in Figures 6.2~ 6.5, and the edge group quality is indicated by the colorbar on the right side. EGS maps for the original and retargeted images are actually equivalent, but we present both of them to show their matching relationship as well. Since EGS is formulated to measure the shape/structure related distortion, the images retargeted by CR are always in high quality. The comparisons between EGS feature and ARS are shown in Figures 6.8~ 6.10. From the results, we can see that the low-level feature ARS, which relies on the visual importance, tends to ignore the distortion in the unimportant region, which leads to the prediction results

Figure 6.8: Comparisons 1 of the EGS feature Q_{EG} and ARS. (a) The original ‘DKNYgirl’ image. (d) Corresponding visual importance map. (b) Image retargeted by MOP. (e) Image retargeted by SC. (c) EGS map for (b). (f) EGS map for (e).

inconsistent with the subjective votes in these cases. Even though EGS only measures the visual distortion, the predictions Q_{EG} by the EGS feature correlate better with the human preference than ARS, when there is no obvious content removal as shown in Figures 6.8~6.10.

6.5 Face Block Similarity Feature

Face is an important element in a large number of images, and usually attracts more attentions. The study [6] shows that viewers show consistently high sensitivity to face deformation. Therefore, it will be beneficial to develop specific high-level features to capture the deformation in facial region.

Votes = 32 v.s. 12 ARS = 0.889 v.s. 0.947 $Q_{EG} = 0.891$ v.s. 0.836

Figure 6.9: Comparisons 2 of the EGS feature Q_{EG} and ARS. (a) The original ‘Woman’ image. (d) Corresponding visual importance map. (b) Image retargeted by SC. (e) Image retargeted by SCL. (c) EGS map for (b). (f) EGS map for (e).

As shown in Figure 6.1, FBS feature is designed to explicitly address the annoying deformation in the facial region indicated by the red boxes. Beyond the local distortions measured by ARS, FBS feature calculates the facial region deformation at the face block scale and measures the visual distortions with explicit semantic meaning. To obtain a satisfied face detection rate, we adopt the Face++ toolkit [81] to detect human faces in the original image, and the retargeted faces are established using the bounding box based on the pixel-level correspondence. In Figures 6.11 and 6.12, we show more examples of the detected faces in the original and their matched face blocks in retargeted images. The deformation between each face block pair is

Figure 6.10: Comparisons 3 of the EGS feature Q_{EG} and ARS. (a) The original ‘child’ image. (d) Corresponding visual importance map. (b) Image retargeted by SC. (e) Image retargeted by SCL. (c) EGS map for (b). (f) EGS map for (e).

calculated with Eq. 5.2 denoted as $s_{ar}(n)$ and FBS score is given by

$$Q_{FB} = \begin{cases} \frac{1}{N_3} \sum_{n=1}^{N_3} s_{ar}(n) & N_3 > 0 \\ 1 & N_3 = 0 \end{cases}, \quad (\text{Eq. 6.3})$$

where N_3 is the total number of detected faces.

Since human faces appear in a portion of natural images, for the general purpose, only the face images will be penalized by FBS and other images will be assigned with the high quality score $Q_{FB} = 1$. The similarity scores for the multiple faces are averaged within one image. The novel FBS feature, capturing the sensitive facial region deformation, is nontrivial compared to previous works, where the face detection is only used to refine the visual importance map [16, 51] or classify the images into binary categories [47].

Figure 6.11: Examples of (a) ‘child’ face images. (b~i) are images retargeted by CR, SV, MOP, SC, SCL, SM, SNS, and WARP, respectively. As indicated by the red boxes, the facial regions are first detected in the original image and the corresponding facial regions are identified in retargeted images using the pixel-level correspondence estimated in the first stage.

6.6 Machine Learning based Multiple-Level Features Fusion

A feature fusion model is learned to map feature scores into quality indices, providing a final measure of the retargeted image quality. In our implementation, a support vector machine (SVM) regression (SVR) [82] is adopted for learning on MIT RetargetMe [6] and CUHK [61] databases. SVR has been widely applied to IQA problems [83, 84, 85]. We use the LIBSVM package [86] to implement SVR on CUHK database with radial basis function (RBF)

Figure 6.12: Examples of (a) ‘girls’ face images. (b~i) are images retargeted by CR, SV, MOP, SC, SCL, SM, SNS, and WARP, respectively. As indicated by the red boxes, the facial regions are first detected in the original image and the corresponding facial regions are identified in retargeted images using the pixel-level correspondence estimated in the first stage.

kernel. Different from the mean opinion score (MOS) provided on CUHK database, pair-wise comparison votes are provided on MIT database, where only rankings among retargeted images are reliable. The direct regression using SVR is not efficient. By this consideration, we utilize the SVM^{rank} implementation [87] with RBF kernel for ordinal regression on MIT database instead.

6.7 Experimental Study

In this section, we present a series of performance evaluation of the proposed MLF on the two public MIT RetargetMe [6] and CUHK [61] databases, which we have already introduced in subsection 5.3.1.

6.7.1 Performance evaluation on MIT database

Table 6.1: Performance comparison of different IRQA measures on MIT database.

Methods	mean KRCC for each subset						Total		
	<i>Line Edges</i>	<i>Faces People</i>	<i>Foreground Objects</i>	<i>Texture</i>	<i>Geometric Structure</i>	<i>Symmetry</i>	mean KRCC	std KRCC	<i>p-val</i>
BDS [68]	0.040	0.190	0.167	0.060	-0.004	-0.012	0.083	0.268	0.017
EH [45]	0.043	-0.076	-0.079	-0.060	0.103	0.298	0.004	0.334	0.641
SIFT flow [39]	0.097	0.252	0.218	0.161	0.085	0.071	0.145	0.262	0.031
EMD [40]	0.220	0.262	0.226	0.205	0.237	0.500	0.251	0.272	1e-5
CSim [15]	0.097	0.290	0.293	0.161	0.053	0.150	0.164	0.263	0.028
PGDIL [13]	0.431	0.390	0.389	0.286	0.438	0.523	0.415	0.296	6e-10
ARS	0.463	0.519	0.444	0.330	0.505	0.464	0.452	0.283	1e-11
MLF	0.486	0.605	0.544	0.384	0.536	0.536	0.512	0.251	1e-14

Following the previous work [72] we apply leave-one-out cross-validation (LOOCV) on MIT database with SVM^{rank} [87] using RBF kernel for the feature fusion. In each image set, we treat the original image as a query and the subjective votes of the 8 retargeted images as their quality ranking. There are 37-folds in total and in each fold we use one set for testing and the rest 36 sets for training. After LOOCV, each set is tested once exactly, and then their results are evaluated accordingly. When using the SVM^{rank} to learn the regression models, the parameters (C, γ) are set using grid search in the range of $[2^{-8}, 2^8]$ with a $2^{0.8}$ step and the optimal (C, γ) were found to be $(2^{4.8}, 2^{3.2})$ on MIT database.

The rank correlation results on MIT database [60] are shown in Table 6.1. We compare the proposed MLF measure with BDS [68], EH [45], SIFT flow [39], EMD [40], CSim [15],

PGDIL [13] and ARS. We present the mean and standard deviation of KRCC for each IRQA measure, as well as the mean KRCC for each image subset with certain attributes. From the results, we can see that our proposed measure outperforms other measures in each image subset as well as on the overall database. In Figure 6.13, we compare the individual KRCCs of ARS and MLF for the 37 image sets from MIT database. Compared with ARS, the proposed MLF shows more stable and reliable quality prediction accuracy.

Figure 6.13: The individual KRCC results of ARS and MLF measure for the 37 image sets on MIT database. For each image set (including eight retargeted images), the KRCC is calculated as the rank correlation between the ranking of subjective and objective scores.

We also follow the *top 3 ranking* KRCC evaluation strategy adopted in [6, 13], which we denote it as KRCC3 here. For the $N = 8$ length ranking, all 28 ranking pairs are compared for the complete rank correlation calculation, while the KRCC3 evaluation *s.t.* $(rank_s(i) \leq 3 \vee rank_s(j) \leq 3) \wedge (rank_o(i) \leq 3 \vee rank_o(j) \leq 3)$, focuses more on the prediction accuracy for the top ranking images. As the constraint shows, there are the *top 3 ranking* selections for both subjective ranking and objective ranking. With the *top 3 ranking* selection for the subjective ranking, we compare the constant 18 high ranking pairs. The additional *top 3 ranking* selection for the objective ranking will further reduce it to a uncertain smaller number. For example, the average number of the compared pairs for EMD is 13.30 and its standard deviation is 2.47. It indicates that we are comparing a varying number of pairs for the objective measures and it may

be biased towards certain measures. To exclude this possible influence factor, we also compare the KRCC3 performance only with the *top 3 ranking* selection for the subjective ranking. As shown in Table 6.2, we can observe that our proposed MLF measure shows the comparable or better prediction performance.

Furthermore, to examine the ability to select the best subject-rated retargeted image, we measure the *rank-n* accuracy for different IRQA measures. Given the eight retargeted images for each original image, the *rank-n* accuracy is the percentage of the objective scores where the subjective-rated best one is within their *top n* positions. In Figure 6.14, we present the *rank-1*, *rank-2*, *rank-3*, and *rank-4* accuracies for MLF and other measures on MIT database. We can see that the proposed MLF measure shows better overall prediction accuracy than other measures.

Table 6.2: Performance comparison of different measure on MIT database by two kinds of KRCC3. The first one (left) considers the pairs *s.t.* $(rank_s(i) \leq 3 \vee rank_s(j) \leq 3) \wedge (rank_o(i) \leq 3 \vee rank_o(j) \leq 3)$, while the second one (right) considers the pairs *s.t.* $rank_s(i) \leq 3 \vee rank_s(j) \leq 3$.

Methods	<i>top 3 subjective ranking</i> <i>top 3 objective ranking</i>		<i>top 3 subjective ranking</i>	
	mean	std	mean	std
	KRCC3	KRCC3	KRCC3	KRCC3
BDS [68]	0.108	0.532	0.132	0.372
EH [45]	-0.071	0.593	0.009	0.413
SIFT flow [39]	0.298	0.483	0.255	0.348
EMD [40]	0.326	0.496	0.339	0.442
CSim [15]	0.277	0.467	0.240	0.305
PGDIL [13]	0.533	0.383	0.532	0.280
ARS	0.450	0.469	0.520	0.347
MLF	0.518	0.459	0.589	0.310

Figure 6.14: *Rank-n* accuracy of different IRQA measures. The rank prediction by MLF shows better overall prediction accuracy.

6.7.2 Performance evaluation on CUHK database

On CUHK database MLF measure is evaluated using SVR with 5-fold cross-validation following IQA method BRISQUE [84]. We divide the database into two randomly chosen subsets, 80% training and 20% testing. There will be no overlap between train and test subsets. We repeat the random train-test procedure 1000 times and report the median performance of these 1000 iterations. When using the SVR to learn the regression models on CUHK database, the parameters (C, γ) are set using grid search in the range of $[2^{-8}, 2^8]$ with a $2^{0.8}$ step and the optimal (C, γ) were found to be $(2^8, 2^{-3.2})$.

In Table 6.3, we compare the performance of the proposed MLF measure with BDS [68], EH [45], SIFT flow [39], EMD [40], CSim [15], GLS [46], PGDIL [13] and ARS on CUHK database using PLCC, SRCC, RMSE and OR. As the results show, MLF achieves significantly better prediction performance compared with other existing measures.

In Table 6.4 we also compare the MLF features individually with other shape descriptor features evaluated in [42], including Pyramid Histogram of Visual Words (PHOW) [43] and

Table 6.3: Performance of different quality measures on CUHK database

Methods	PLCC	SRCC	RMSE	OR
BDS [68]	0.2896	0.2887	12.922	0.2164
EH [45]	0.3422	0.3288	12.686	0.2047
SIFT flow [39]	0.3141	0.2899	12.817	0.1462
EMD [40]	0.2760	0.2904	12.977	0.1696
CSim [15]	0.4374	0.4662	12.141	0.1520
GLS [46]	0.4622	0.4760	10.932	0.1345
PGDIL [13]	0.5403	0.5409	11.361	0.1520
ARS	0.6835	0.6693	9.855	0.0702
MLF	0.7577	0.7383	8.525	0.0294

GIST [44]. We can see that all our features show competitive performance compared with other existing shape features. When the Q_{FB} works as an individual feature, it is evaluated among the 33 detected face images from the 171 retargeted images. It should be clear that FBS feature can predict the quality of face images rather effectively and it is worthy to develop the specific high-level features for these sensitive attributes.

6.7.3 Parameter study

We provide the sensitivity studies for parameters α and β from the proposed MLF measure on MIT and CUHK databases in Figures 6.15 and 6.16. Since the major part of information loss is captured by ARS feature, α plays an important role in the tradeoff between information loss and visual distortion similar to that in ARS. The influence of α in Figure 6.15 is similar to that in Figure 6.6, but the peak shifts from 0.3 to 0.7. Visual distortions measured by EGS and FBS features have broken the balance in the case where only ARS is adopted, and a larger α with more information loss penalty helps to achieve the new optimal balance.

We have also observed that the variation tendency of the overall performance on CUHK database is relatively less obvious compared with that on MIT database. In Figure 6.16, we

Table 6.4: Investigation of individual features on CUHK database

Features	PLCC	SRCC	RMSE	OR
SCD	0.1508	0.1792	13.347	0.2164
CSD	0.1520	0.1688	32.731	0.5322
CLD	0.1033	0.0850	13.429	0.2398
HTD	0.0829	0.0890	35.151	0.5673
EHD	0.3031	0.2729	12.866	0.2047
EMD	0.2760	0.2904	12.977	0.1696
PHOW	0.3706	0.2308	12.540	0.2222
GIST	0.5443	0.5114	11.326	0.1579
Q_{AR}	0.6752	0.6587	9.960	0.0760
Q_{EG}	0.5706	0.5617	11.087	0.1287
Q_{FB}	0.8052	0.6984	8.717	0.0909

Figure 6.15: Influence of α from MLF measure on MIT and CUHK databases.

show the influence of β from EGS feature on MLF measure. The overall performance on both databases is relatively stable with small variations when β is larger than 0.1. In general, the performance of MLF measure is insensitive to the change of β in the range of $[0.1, 1]$, but still influenced by α with regard to the tradeoff between information loss and visual distortion.

Figure 6.16: Influence of β from MLF measure on MIT and CUHK databases.

6.7.4 Feature analysis

To further investigate the significance of different features, in Tables 6.5 and 6.6 we conduct the performance evaluation using different feature combinations on MIT and CUHK databases, respectively. In combination #3*, only face images are focused, since other images are assigned with the same scores. From the tables, we can see that the combination #7 with all features outperforms other feature combinations with noticeable margins on both databases. In combination #6, EGS and FBS features mainly measure visual distortion and without effective information loss measurement, their performance has decreased obviously compared with #7. ARS feature measures information loss and visual distortion at the same time, and is able to work as a baseline part individually. Based on the comparisons of #4, #7 and #5, #7, we can observe that EGS and FBS features both are able to further improve the overall performance effectively.

6.7.5 Cross database performance evaluation for MLF

In order to check the generalization ability of MLF, we conduct the cross database performance evaluation by training on one entire database and testing on the other one. First, we use SVR to learn the feature fusion model on the entire CUHK database, and the mean KRCC and its

Table 6.5: Feature analysis on MIT database

No.	Feature combination			mean KRCC for each Attribute						Total	
	Q_{AR}	Q_{EG}	Q_{FB}	<i>Line Edges</i>	<i>Faces People</i>	<i>Foreground Objects</i>	<i>Texture</i>	<i>Geometric Structure</i>	<i>Symmetry</i>	mean KRCC	std KRCC
1	✓			0.369	0.424	0.401	0.313	0.424	0.536	0.396	0.244
2		✓		0.217	0.257	0.270	0.268	0.250	0.024	0.222	0.251
3*			✓	-	0.464	-	-	-	-	0.464	-
4	✓	✓		0.463	0.576	0.524	0.357	0.522	0.512	0.492	0.249
5	✓		✓	0.383	0.462	0.425	0.339	0.424	0.536	0.411	0.249
6		✓	✓	0.229	0.310	0.306	0.304	0.250	0.024	0.243	0.243
7	✓	✓	✓	0.486	0.605	0.544	0.384	0.536	0.536	0.512	0.251

Table 6.6: Feature analysis on CUHK database

No.	Feature combination			PLCC	SRCC	RMSE	OR
	Q_{AR}	Q_{EG}	Q_{FB}				
1	✓			0.6752	0.6587	9.960	0.0760
2		✓		0.6179	0.5725	10.616	0.1228
3*			✓	0.8052	0.6984	8.717	0.0909
4	✓	✓		0.7314	0.7089	9.029	0.0588
5	✓		✓	0.7015	0.6900	9.393	0.0588
6		✓	✓	0.5898	0.5527	10.642	0.1176
7	✓	✓	✓	0.7577	0.7383	8.525	0.0294

standard deviation tested on MIT database are 0.469 and 0.256, respectively. Next, we perform the learning on MIT database using SVM^{rank} . The PLCC, SRCC, RMSE and OR tested on CUHK database are 0.730, 0.713, 9.230, and 0.047, respectively. Compared with Tables 6.1 and 6.3, we can observe that MLF's cross database performance is inferior to the performance when training and testing data come from the same database but still better than that of other existing quality measures. It should be clear that MLF performs better in terms of correlation with human perception and has good generalization ability across databases as well.

6.8 Conclusion

For the high-level IRQA task, we pointed out that the existing low-level feature based methods are limited due to the visual importance dependency and the lack of high-level visual distortion modelling. To address these problems, we combined the dense correspondence estimation with different level feature detection methods and developed three effective general-purpose features to model various kinds of quality degradation factors in the retargeted image. Then we proposed a multiple-level feature based framework to predict the perceptual retargeted image quality. On the public MIT and CUHK databases, the proposed MLF measure shows better and more reliable quality prediction performance compared with other existing methods. While the proposed method addresses the related problems and shows promising performance, it is still quite necessary to develop more effective features as well as the correspondence estimation and visual importance map algorithms for better IRQA measures in the future work.

Chapter 7

Summary and Future Work

7.1 Summary

In this thesis, we studied specialized dense correspondence estimation techniques, and employed them to reduce inter-image redundancy for image set compression and model the retargeting modifications for image retargeting quality assessment. We developed one dense correspondence based prediction approach for the inter-image redundancy reduction for image set compression, one backward registration method to clarify the dense correspondence relationship for image retargeting, and two objective image retargeting quality assessment measures.

We presented a study on image set compression for cloud-based storage. We developed a novel dense correspondence based prediction approach to estimate the consistent pixel-to-pixel relationship between images via the fast-convergent random search. Dense correspondence can provide more effective local similarity estimation between images, but may suffer from large side information expense when it cannot be parameterized effectively. By introducing consistency check constraints, we obtained the smooth and consistent dense correspondence as the inter-image prediction results with efficient parameterization. Next, we reconstructed a new reference image by geometric transformations and luminance adjustments for the subsequent HEVC inter-prediction based video-like encoding. Because of the reconstruction in local units, our approach is more robust to the complex local variations compared with the global estimation based methods. Experimental results show that the proposed method achieves better per-

formance compared with relevant state-of-the-art methods, especially when the local variations are dominant.

For the comparative evaluation of multiple image retargeting operators, the complicated relationship between the original and retargeted images impedes the effective perceptual quality evaluation of the retargeted image. To better understand the image retargeting process, we have first provided a unified interpretation of image retargeting with the resampling grid generation and forward resampling. We showed that the geometric change estimation is an efficient way to clarify the dense correspondence relationship between the original and retargeted images. We formulated the geometric change estimation as a backward registration problem via Markov random field and provided an effective and practical solution. For image retargeting quality assessment applications, backward registration can provide reliable geometric change guidance, which link the properties between the original and retargeted images and make the measurement of the corresponding quality degradation more directly and efficiently. Experimental results on two validation databases show that the proposed backward registration method achieves high overall estimation accuracy qualitatively and quantitatively.

Under the guidance of the geometric change estimated by the backward registration approach, we have developed an effective ARS metric by exploiting the local block changes to evaluate the visual retargeted image quality. Even though ARS is designed based on the local low-level features and it is infeasible to explicitly evaluate the specific attribute degradations such as the broken line and violated symmetry, it outperforms other existing previous methods.

Furthermore, we pointed out that existing low-level feature based methods are limited due to the visual importance dependency and the lack of high-level visual distortion modelling for the high-level IRQA task. To address these problems, we combined the dense correspondence estimation with different level feature detection methods and developed three effective general-purpose features to model various kinds of quality degradation factors in the retargeted image. Then we proposed a multiple-level feature based framework to predict the perceptual retargeted

image quality. On the public MIT and CUHK databases, the proposed MLF measure shows better and more reliable quality prediction performance compared with other existing methods.

7.2 Future Directions

7.2.1 Automatic image cropping

In this thesis, the concerned image retargeting is to preserve the original image content in terms of maximal fidelity. The chosen original images are usually professional ones with high perceptual quality, and the manual cropped image has no higher perceptual quality than the original one. However, photos in the real world are not always in good compositions, especially for amateur photographers without professional skills. In these scenarios, without fidelity constraint, image cropping with the consideration of aesthetic and composition, becomes another research problem, which will improve the perceptual quality by removing the superfluous scene content. As manual cropping remains tedious and requires experience, automatic image cropping are in demand for applications such as thumbnail generation and view finding.

Automatic image cropping generally aims to find an optimal subview in terms of its aesthetics or composition. Traditional automatic cropping techniques can be classified as attention-based and aesthetic-based methods [88]. The success of deep learning motives us to learn automatic cropping directly from images. To efficient build sufficient training samples, recent work [89] suggests that the professional photo and its crops are qualified as pairs of positive and negative training samples, because professional photographs are normally edited in excellent composition and the cropping would break their good composition. The large training database can be collected in a cheaper way, which makes it feasible to perform the ranking learning using convolutional neural networks. The learned networks would suffer from searching in the huge amount of subview candidates. It is feasible to choose several random seed tiny windows, gradually search the better larger views around the seed, so we can obtain the optimal subview inside the image when the growing process converges. It is worth investigating the

deep integration of ranking neural network learning and automatic cropping window finding scheme to develop more effective and efficient automatic image cropping methods.

7.2.2 Image aesthetic assessment

While it is difficult to design effective handcraft features to model the quality changes for complicated and high-level IRQA tasks, the machine learning based methods are possible to extract better features. However, there are no sufficient training samples on existing benchmark IRQA databases, since even the largest RetargetMe database taking quite a lot of human power only covers 37×8 images.

Recent image aesthetic assessment studies have shown significant progress on the large-scale aesthetic databases. Image aesthetic assessment aims to predict the perceived beauty of images objectively. Compared to IRQA it has quite similar purpose on the understanding and appreciation of images, so there is an opportunity to explore the deep aesthetic features and transfer the domain knowledge to IRQA. Existing convolutional neural network based works [90, 91, 92] adopt the data-driven approach with the training samples scaling from hundreds of images to millions, and show promising performance on the binary aesthetic quality prediction by judging whether one image is perceptually high-quality or low-quality. The binary aesthetic classification is far from enough for us to investigate and interpret the perceptual quality or the aesthetic value of one image. It would be an interesting problem to extend it to fine-scale rating and derive the effective aesthetic feature for specific applications such as image retargeting, image composition and image enhancement. In addition, on the large-scale AVA database [93], except the ground-truth quality labels, most images are provided with additional aesthetic annotation and semantic annotations. We notice that textual tags and competition topics are usually closely correlated with the image content. There is still a lack of research effort on this problem. It is possible to model these semantic content influence factors to further improve image aesthetic evaluation.

Publication List

- [1] **Yabin Zhang**, Weisi Lin, Qiaohong Li, Wentao Cheng, Xinfeng Zhang, Multiple-Level Feature-Based Measure for Retargeted Image Quality. *IEEE Trans. Image Processing* 27(1): 451-463 (2018). (**TIP**)
- [2] **Yabin Zhang**, Yuming Fang, Weisi Lin, Xinfeng Zhang, Leida Li, Backward Registration-Based Aspect Ratio Similarity for Image Retargeting Quality Assessment. *IEEE Trans. Image Processing* 25(9): 4286-4297 (2016). (**TIP**)
- [3] **Yabin Zhang**, Weisi Lin, Xinfeng Zhang, Yuming Fang, Leida Li, Aspect Ratio Similarity (ARS) for image retargeting quality assessment. *ICASSP 2016*: 1080-1084.
- [4] **Yabin Zhang**, Weisi Lin, Jianfei Cai, Dense correspondence based prediction for image set compression. *ICASSP 2015*: 1240-1244.
- [5] Xinfeng Zhang, **Yabin Zhang**, Weisi Lin, Siwei Ma, Wen Gao, An inter-image redundancy measure for image set compression. *ISCAS 2015*: 1274-1277.
- [6] Xinfeng Zhang, Weisi Lin, **Yabin Zhang**, Shiqi Wang, Siwei Ma, Lingyu Duan, Wen Gao, Rate-Distortion Optimized Sparse Coding with Ordered Dictionary for Image Set Compression. *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*

References

- [1] Zhongbo Shi, Xiaoyan Sun, and Feng Wu. Photo album compression for cloud storage using local features. *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 4(1):17–28, March 2014.
- [2] Gregory K Wallace. The JPEG still picture compression standard. *Consumer Electronics, IEEE Transactions on*, 38(1):xviii–xxxiv, 1992.
- [3] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the H. 264/AVC video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(7):560–576, 2003.
- [4] Kosmas Karadimitriou and John M Tyler. The centroid method for compressing sets of similar images. *Pattern Recognition Letters*, 19(7):585–593, 1998.
- [5] Anthony Schmieder, Howard Cheng, and Xiaobo Li. A study of clustering algorithms and validity for lossy image set compression. In *IPCV*, pages 501–506, 2009.
- [6] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. A comparative study of image retargeting. *ACM Trans. Graph.*, 29(6):160, 2010.
- [7] Ariel Shamir and Olga Sorkine. Visual media retargeting. In *SIGGRAPH ASIA, Yokohama, Japan, December 16-19, 2009, Courses Proceedings*.
- [8] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. *ACM Trans. Graph.*, 27(3), 2008.
- [9] Yael Pritch, Eitam Kav-Venaki, and Shmuel Peleg. Shift-map image editing. In *IEEE 12th International Conference on Computer Vision, ICCV, Kyoto, Japan, September 27 - October 4, 2009*, pages 151–158.

REFERENCES

- [10] Lior Wolf, Moshe Guttman, and Daniel Cohen-Or. Non-homogeneous content-driven video-retargeting. In *IEEE 11th International Conference on Computer Vision, ICCV, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–6.
- [11] Philipp Krähenbühl, Manuel Lang, Alexander Hornung, and Markus H. Gross. A system for retargeting of streaming video. *ACM Trans. Graph.*, 28(5), 2009.
- [12] Yu-Shuen Wang, Chiew-Lan Tai, Olga Sorkine, and Tong-Yee Lee. Optimized scale-and-stretch for image resizing. *ACM Trans. Graph.*, 27(5):118, 2008.
- [13] Chih-Chung Hsu, Chia-Wen Lin, Yuming Fang, and Weisi Lin. Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss. *J. Sel. Topics Signal Processing*, 8(3):377–389, 2014.
- [14] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Multi-operator media retargeting. *ACM Trans. Graph.*, 28(3), 2009.
- [15] Yong-Jin Liu, Xi Luo, Yuming Xuan, Wenfeng Chen, and Xiaolan Fu. Image retargeting quality assessment. *Comput. Graph. Forum*, 30(2):583–592, 2011.
- [16] Yuming Fang, Kai Zeng, Zhou Wang, Weisi Lin, Zhijun Fang, and Chia-Wen Lin. Objective quality assessment for image retargeting based on structural similarity. *IEEE J. Emerg. Sel. Topics Circuits Syst.*, 4(1):95–105, 2014.
- [17] Yabin Zhang, Yuming Fang, Weisi Lin, Xinfeng Zhang, and Leida Li. Backward registration-based aspect ratio similarity for image retargeting quality assessment. *IEEE Trans. Image Processing*, 25(9):4286–4297, 2016.
- [18] Weisi Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *J. Visual Communication and Image Representation*, 22(4):297–312, 2011.
- [19] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1):98–117, Jan 2009.
- [20] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

REFERENCES

- [21] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *CVPR (2)*, pages 506–513, 2004.
- [22] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.
- [23] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4), 2006.
- [24] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1), 2006.
- [25] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81), Vancouver, BC, Canada, August 1981*, pages 674–679, 1981.
- [26] John L. Barron, David J. Fleet, and Steven S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.
- [27] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1-3):185–203, 1981.
- [28] Ben Galvin, Brendan McCane, Kevin Novins, David Mason, and Steven Mills. Recovering motion fields: An evaluation of eight optical flow algorithms. In *Proceedings of the British Machine Vision Conference 1998, BMVC 1998, Southampton, UK, 1998*, pages 1–10, 1998.
- [29] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3):211–231, 2005.
- [30] Connelly Barnes, Dan B. Goldman, Eli Shechtman, and Adam Finkelstein. The patchmatch randomized matching algorithm for image manipulation. *Commun. ACM*, 54(11):103–110, 2011.

REFERENCES

- [31] Neeraj Kumar, Li Zhang, and Shree K. Nayar. What is a good nearest neighbors algorithm for finding similar patches in images? In *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II*, pages 364–378, 2008.
- [32] Chanop Silpa-Anan and Richard Hartley. Optimised kd-trees for fast image descriptor matching. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.
- [33] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall F. Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):1068–1080, 2008.
- [34] Zhuoyuan Chen, Hailin Jin, Zhe Lin, Scott Cohen, and Ying Wu. Large displacement optical flow from nearest neighbor fields. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2443–2450, 2013.
- [35] Oscar Au, Sijin Li, Ruobing Zou, Wei Dai, and Lin Sun. Digital photo album compression based on global motion compensation and intra/inter prediction. In *Audio, Language and Image Processing (ICALIP), 2012 International Conference on*, pages 84–90. IEEE, 2012.
- [36] Ruobing Zou, Oscar C Au, Guyue Zhou, Wei Dai, Wei Hu, and Pengfei Wan. Personal photo album compression and management. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 1428–1431. IEEE, 2013.
- [37] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [38] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Processing*, 20(8):2378–2386, 2011.

- [39] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011.
- [40] Ofir Pele and Michael Werman. Fast and robust earth mover’s distances. In *IEEE 12th International Conference on Computer Vision, ICCV, Kyoto, Japan, September 27 - October 4, 2009*, pages 460–467.
- [41] Tongwei Ren and Gangshan Wu. Automatic image retargeting evaluation based on user perception. In *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*, pages 1569–1572.
- [42] Lin Ma, Long Xu, Huanqiang Zeng, King Nghi Ngan, and Chenwei Deng. How does the shape descriptor measure the perceptual quality of the retargeting image? In *IEEE International Conference on Multimedia and Expo Workshops, Chengdu, China, July 14-18, 2014*, pages 1–6.
- [43] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Image classification using random forests and ferns. In *ICCV*, pages 1–8. IEEE Computer Society, 2007.
- [44] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [45] B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, and Akio Yamada. Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):703–715, 2001.
- [46] Jiangyang Zhang and C. C. Jay Kuo. An objective quality of experience (QoE) assessment index for retargeted images. In *Proceedings of the ACM International Conference on Multimedia, MM ’14, Orlando, FL, USA, November 03 - 07, 2014*, pages 257–266, 2014.
- [47] Anmin Liu, Weisi Lin, Hai Chen, and Philipp Zhang. Image retargeting quality assessment based on support vector regression. *Sig. Proc.: Image Comm.*, 39:444–456, 2015.
- [48] Yang Chen, Yong-Jin Liu, and Yu-kun Lai. Learning to rank retargeted images. In *CVPR*, page 39944002. IEEE Computer Society, 2017.

REFERENCES

- [49] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.
- [50] Douglas DeCarlo and Anthony Santella. Stylization and abstraction of photographs. *ACM Trans. Graph.*, 21(3):769–776, 2002.
- [51] Yuanyuan Ding, Jing Xiao, and Jingyi Yu. Importance filtering for image retargeting. In *CVPR*, pages 89–96. IEEE Computer Society, 2011.
- [52] Yuming Fang, Zhenzhong Chen, Weisi Lin, and Chia-Wen Lin. Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Transactions on Image Processing*, 21(9):3888–3901, 2012.
- [53] Radhakrishna Achanta and Sabine Süsstrunk. Saliency detection for content-aware image resizing. In *ICIP*, pages 1005–1008. IEEE, 2009.
- [54] Zhongbo Shi, Xiaoyan Sun, and Feng Wu. Feature-based image set compression. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.
- [55] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [56] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [57] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized patchmatch correspondence algorithm. In *Computer Vision–ECCV 2010*, pages 29–43. Springer, 2010.
- [58] Yoav HaCohen, Eli Shechtman, Dan B Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. In *ACM Transactions on Graphics (TOG)*, volume 30, page 70. ACM, 2011.

REFERENCES

- [59] Seam Carving code. Available: <http://people.csail.mit.edu/mrub/code>.
- [60] RetargetMe Benchmark. [Online]. Available: <http://people.csail.mit.edu/mrub/retargetme>.
- [61] Lin Ma, Weisi Lin, Chenwei Deng, and King Ngi Ngan. Image retargeting quality assessment: A study of subjective scores and objective metrics. *J. Sel. Topics Signal Processing*, 6(6):626–639, 2012.
- [62] Image Retargeting Subjective Quality Database. [Online]. Available: <http://ivp.ee.cuhk.edu.hk/projects/demo/retargeting/index.html>.
- [63] Weiming Dong, Ning Zhou, Jean-Claude Paul, and Xiaopeng Zhang. Optimized image resizing using seam carving and scaling. *ACM Trans. Graph.*, 28(5):125:1–125:10, 2009.
- [64] Zachary Karni, D. Freedman, and Craig Gotsman. Energy-based image deformation. *Comput. Graph. Forum*, 28(5):1257–1268, 2009.
- [65] Nikolay N. Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir V. Lukin, Karen O. Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: peculiarities, results and perspectives. *Sig. Proc.: Image Comm.*, 30:57–77, 2015.
- [66] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.
- [67] Zhou Wang, Ligang Lu, and Alan C. Bovik. Video quality assessment based on structural distortion measurement. *Sig. Proc.: Image Comm.*, 19(2):121–132, 2004.
- [68] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *IEEE Computer Society Conference on CVPR, 24-26 June 2008, Anchorage, Alaska, USA*.
- [69] Huisi Wu, Yu-Shuen Wang, Kun-Chuan Feng, Tien-Tsin Wong, Tong-Yee Lee, and Pheng-Ann Heng. Resizing by symmetry-summarization. *ACM Trans. Graph.*, 29(6):159, 2010.

REFERENCES

- [70] Fang-Lue Zhang, Miao Wang, and Shi-Min Hu. Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Transactions on Multimedia*, 15(7):1480–1490, 2013.
- [71] Che-Han Chang and Yung-Yu Chuang. A line-structure-preserving approach to image resizing. In *IEEE CVPR, Providence, RI, USA, June 16-21, 2012*, pages 1075–1082.
- [72] Yun Liang, Yong-Jin Liu, and Diego Gutierrez. Objective quality prediction of image retargeting algorithms. *IEEE Trans. Vis. Comput. Graph.*, 23(2):1099–1110, 2017.
- [73] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44:1–44:10, 2012.
- [74] Harry G. Barrow, Jay M. Tenenbaum, Robert C. Bolles, and Helen C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, pages 659–663. William Kaufmann, 1977.
- [75] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 391–405. Springer, 2014.
- [76] Serge J. Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.
- [77] Piotr Dollár and C. Lawrence Zitnick. Structured forests for fast edge detection. In *ICCV*, pages 1841–1848. IEEE Computer Society, 2013.
- [78] John Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.
- [79] Ming-Yu Liu, Oncel Tuzel, Ashok Veeraraghavan, and Rama Chellappa. Fast directional chamfer matching. In *CVPR*, pages 1696–1703. IEEE Computer Society, 2010.
- [80] Arasanathan Thayananthan, Bjoern Stenger, Philip H. S. Torr, and Roberto Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR*, pages 127–133. IEEE Computer Society, 2003.

REFERENCES

- [81] Megvii Inc. Face++ research toolkit. www.faceplusplus.com, December 2013.
- [82] Bernhard Schölkopf, Alexander J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.
- [83] Manish Narwaria and Weisi Lin. SVD-based quality metric for image and video using machine learning. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 42(2):347–364, 2012.
- [84] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Processing*, 21(12):4695–4708, 2012.
- [85] Manish Narwaria and Weisi Lin. Objective image quality assessment based on support vector regression. *IEEE Trans. Neural Networks*, 21(3):515–519, 2010.
- [86] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [87] Thorsten Joachims. Training linear SVMs in linear time. In *KDD*, pages 217–226. ACM, 2006.
- [88] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *WACV*, pages 226–234. IEEE, 2017.
- [89] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *ACM Multimedia*, pages 37–45. ACM, 2017.
- [90] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, pages 497–506. IEEE Computer Society, 2016.
- [91] Xin Lu, Zhe Lin, Xiaohui Shen, Radomír Mech, and James Zijun Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, pages 990–998. IEEE Computer Society, 2015.

REFERENCES

- [92] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Process. Mag.*, 34(4):80–106, 2017.
- [93] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415. IEEE Computer Society, 2012.