ADVANCED VIDEO PROCESSING TECHNIQUES IN VIDEO TRANSMISSION SYSTEMS

By

ZHENG YUAN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2013

I dedicate this dissertation to my loving mother Xiaoli Xu, father Minxu Yuan, and other family members, Guihua Wang, Ting Yue, Lei Mu, Minxia Yuan, Guangyan Xu, Bingkun Yuan, Yulian Zhao and Zhenwu Xu for your unlimited support and love.

ACKNOWLEDGMENTS

First of all, I would like to thank my PhD advisor Dr. Dapeng Oliver Wu. With his broad knowledge of my research areas, Dr. Wu always provided me with strategic advises and helped me to shape my research courses very wisely. He devoted a lot of time and effort in exchanging and developing ideas with me about the scientific research. His erudite while easy-going style sets me a role model of how to conduct good quality research in the future.

I would like to thank Dr. Xiaolin Li, Dr. Tao Li and Dr.Shigang Chen for serving on my committee. They have given me important comments and suggestions on both my proposal and defense so that I know how to make improvement precisely. I appreciate their time and effort in reading my dissertation and attending my oral defense.

I would like to thank my colleges and classmates, Dr. Taoran Lu, for helping me in many aspects when I first came to the US. I enjoyed our academic collaboration truly; Dr. Zhifeng Chen, for guiding me with video streaming research topic; Dr. Qian Chen, Dr. Zongrui Ding and Dr. Yuejia He, for your valuable suggestions and support; Dr. Jun Xu, Dr. Lei Yang, Dr. Bing Han for your opportunities. A special thanks to our departmental graduate coordinator Ms. Shannon Chillingworth for her always prompt and precise reply to my questions. This makes my graduate academic life reassured.

I sincerely thank Dr. Yu Huang, Dr. Zhenyu Wu, Dr. Yuwen He, Dr. Jun Xin and Dr. Dongqing Zhang for your mentoring during my internship in the industry. The hand-on experiences learned from you are significant to me. Special thanks are also extended to Dr. Miao Zhao from Stony Brook University and Dr. Tianyi Xu from University of Delaware, for your constructive advice during our discussion on my PhD research.

Finally, I thank my friend Chong Pang, Chao Li, Yang Lu, Steve Hardy, Huanghuang Li etc. in Gainesville and Lei Xu, Xiaohua Qian, Haiquan Wang and Ke Wang back in China. Because of them, my life during my PhD studies is colorful!

I always cherish your support!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ADVANCED VIDEO PROCESSING TECHNIQUES IN VIDEO TRANSMISSION SYSTEMS

By

Zheng Yuan

August 2013

Chair: Dapeng Wu
Major: Electrical and Computer Engineering

In this dissertation, we address four processing techniques in a video transmission system. First, we propose a homogenous video retargeting approach to dynamically adapt the video from a large resolution to a small one, keeping the interesting contents without introducing artifacts. We then design a video summarization system to summarize a long video into a shorter version, allowing viewers to access content more efficiently. The experiments of the proposed retargeting and summarization approaches show performance gain over existing solutions. For the events of video packet loss during transmission, we propose a perceptual quality model that mimics the human response scores after watching the impaired videos. Based on discovering the visual psychological factors that directly relate to the human scores, the perceptual quality model achieves better correlation with the true human response. Finally, we address the perceptual quality based rate-distortion optimization problem in the encoder design. We propose a piecewise approximation method to find out a rate distortion (perceptual) model, based on which the best rate-distortion trade-off is achieved.

CHAPTER 1
INTRODUCTION

1.1 Problem Statement

The fast development of the multimedia compression and network technologies enables a fast proliferation of video transmission applications, such as multimedia streaming (Video on Demand), Video Live Broadcasting, Video Conferencing and Video Conversional Applications. Although the detailed implementation of these systems may vary, the common architecture of a transmission system for video media typically consists of a video acquisition and encoding modules on the sender side, a transmission channel with certain capacity in the middle, and the video decoding and display modules on the receiver side. Fig. 1-1 shows the architecture of a video transmission system.

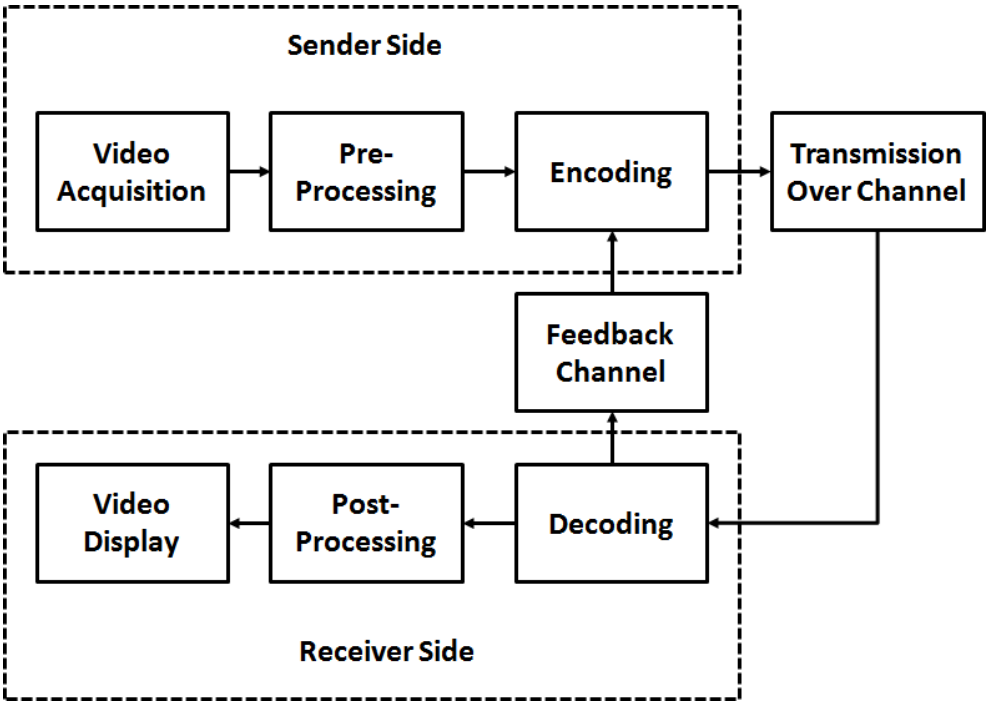

Figure 1-1. The architecture of a video transmission system.

Video acquisition component is a camera module to collect the pixel values of a scene. Encoding component compresses the pixel values into a much smaller version to facilitate the transmission. In between, there may be preprocessing module to adjust the video content for efficient encoding. Then the compressed video is packetized

into media packets for transmission. During the transmission, a channel with transport protocols delivers video packets. In the receiver side, a decoder depacketizes and decompresses the video packets into pixel values. Most often artifacts due to compress and channel error events can happen. In these situations, the decoder need try its best to decode the possible impaired bitstream and a post-processing unit is usually involved to mitigate the artifacts. The display component displays the final content to viewers.

Two key values of designing a video transmission system are content adaptability and visual quality awareness design. The main purpose of content adaptability is to distribute multimedia to maximum reachable audiences. Since the implementation of each component can vary across a very large spectrum, adapting the output of a particular component into an eligible or even optimal input of the next component can bring a “seamless connection” between the two. The ultimate goal of adaptability is the ubiquitous access of media, in which a component can handle the output of any implementation of its logically connected component. The main motivation of visual quality awareness design is to use human visual perception simulated index to measure how well of a component of a video transmission system behaves and thus make strategies to optimize the component. After all, it is up to humans to evaluate the video transmission services, therefore, visual quality awareness design may exploit human visual quality characteristics directly to enhance the performance of a transmission system, e.g. a source encoder.

1.2 The Scope of the Dissertation

This dissertation considers both the two values: for content adaptability, the author proposes two techniques video retargeting and video summarization, which adapt the resolution and frame rate of video source in the video acquisition component to the ones in the video display component. Video retargeting is a technique that intelligently adapts the original video to the resolution of the display. This technique can enable the viewer with small screen devices more effectively to attend the region-of-interest of the original

at much less cost of artifacts. It can also be widely used in video transcoding area, where an encoded bitstream is needed to be converted to another bitstream matching channel or receiver-side change. Parallelingly, Video summarization is a technique that adapts the original video to the frame rate to other value preferred by the recipient. It can help viewers to efficiently access the content of a Video-on-Demand system and also be used in video transcoding area. For video surveillance, which is a special live broadcasting application, video summarization may automatically detect the unusual activities in the video and thus only notify the recipient upon detected.

For visual quality awareness design, in the receiver side the author proposes a perceptual quality metric to measure the decoded video quality. In the sender side, the author addresses the essential problem (the trade-off between bitrate and distortion) of encoder design in the perceptual context and proposes a perceptual quality based encoder optimization method accordingly.

In a video transmission system, channel sometimes fails to deliver video packets from the sender to the recipient due to reasons such as reduced bitrate and network congestion. In these error events, although the decoder will still make best effort to recover or conceal the corrupted video, it is not uncommon that the artifacts still can be observed. In order to monitor the video quality, an automatic quality assessment method based on artifact edge map and face detection is proposed from the decoder side.

In the encoder side, we aim to achieve a bitstream in lowest bitrate possible given a certain distortion level. This requires the knowledge of the model of distortion with bitrate. When the distortion is measure by perceptual quality metric, such model may not be easily derived from the perspective of information theory as the distortion measure is in a different nature. The author instead proposes an experiment based sampling method with curve fitting technique to learn the model and thus achieve the best rate distortion trade-off.

1.3 The Organization of the Dissertation

The remaining of this dissertation is organized as follows. Chapter 2 and Chapter 3 introduce the proposed video retargeting and video summarization technique, respectively. Chapter 4 proposes the perceptual quality model to assess the video quality during packet loss event and Chapter 5 is about the perceptual rate distortion optimization in encoder design. Chapter 6 concludes the work of this dissertation.

CHAPTER 2
VIDEO RETARGETING

2.1 Background

Thanks to the enriched communication network resources and efficient compression techniques, a diversity of mobile devices and streaming terminals gain significant shares in multimedia access over years. While they are designed in unique resolutions and aspect ratios, most media sources, when created, generally follow the standard formats (e.g. resolution 1920×1080, 720×576). As the proxy to transfer video media across platforms, video retargeting techniques automatically adapt source video contents to fit for the display size of target devices (generally from large to small). While they attempt to preserve most Visual Interestingness (VI) for each individual frame as required in the image retargeting task, they also demand the temporal consistencies among adjacent frames to generate visually agreeable retargeting results.

As video retargeting seeks to preserve the VI efficiently, it is sensible to understand how the VI distributes over pixels spatially and temporally. Attention modeling, as a self-contained research topic, provides such distribution information by mimicking the visual stimulus from each pixel of video frames. In literature, many methods are proposed, including Ref. [2, 8–11] and also [12] where multi-modalities are considered. Several aliases of visual interestingness include visual importance [13], visual energy [14], saliency [9] [2].

2.1.1 Previous Methods

Current video retargeting techniques are conceptually classified into two major categories: the homogeneous vs. heterogeneous approaches. The homogeneous methodology formulates video retargeting as searching for a sequence of rigid retargeting windows on each original frame followed by homogeneously resizing the contents within the window into the target display. Although the homogeneous methodology may sacrifice contents outside the retargeting window, this scheme allows

a systematic regulation of retargeted pixels. Spatially, the selected pixels are treated equally to avoid the geometric distortion, which is especially meaningful when the content contains the well-defined or viewer familiar objects such as human face or architectures. Temporally, the retargeting consistency can be easily achieved by merely imposing constraints on the window parameters across neighboring frames.

Given that viewers exercise non-uniform attention response to stimulus from different pixels shown by visual psychophysical studies, Liu et al.'s auto-pan-scan [15] and Hua et al.'s search window method [16] look for a window to move dynamically to secure the most visually interesting regions for each individual frame. Regarding the arisen consistency issue, they both utilize curve fitting to smooth the window parameters. Although this procedure alleviates the shaky artifacts to some extent when visually interesting areas locate closely over frames, it can not guarantee consistent retargeted views in presence of rapid and irregular content changes. Instead, Deselaers et al. [17] considers the visual interestingness preservation and retargeting consistency together as the retargeting scores and traces back the best window sequence that maximizes the accumulation scores. This approach for the first time models the retargeting consistency issue as a build-in consideration and thus ensures the retargeted views consistent. Nevertheless, this setup of maximization of accumulation score may not quite suit for the VI preservation: without knowing a pre-recognized destination that specifies the retargeting window to track the updated visual interesting areas, the calculated window sequence is anchored near its initial position due to the encouragement of an over-inertia transition among adjacent frames, which we name as "consistency clamping" artifact. This artifact is visible especially in long video whilst its complex background change requires the retargeting window to update to the current position. Also, if there is "saliency inaccuracy" in the initial frame, the identified parameters of the first RW is impaired. The impairment will propagate into

all the following frames due to clamping. For long video retargeting, the propagation is even longer.

On the other hand, the heterogenous methodology does not perform “hard” selection of a chunk of pixels continuously aligned in a rectangle window, but rather takes each pixel individually and imposes “soft” manipulation of potentially every pixel. This flexibility makes the heterogenous methodology the focus of media retargeting in academia for many years, which develops primarily in two tracks. The first track focuses on seam carving: [14] shrinks original image by eliminating pixels aligning in seams (continuous curves) in priority of the one with less visual energy. The same idea extends to video retargeting [18] by cutting 2D seam manifolds therein from a 3D video frame volume. Recently, Matthias et al. [19] impose the discontinuity constraint on seam structures to alleviate the cutting of featured objects. [20] adopts multi-operator scheme to choose seam carving when appropriate. The other track is warping based approaches: they do not explicitly remove pixels away but rather morphologically squeeze them to various extents proportional to their visual importance. Wolf et al. [13] initializes this idea by formulating the mapping from original pixel to the retargeted correspondence as a sparse linear system of equations. Wang et al. [21] then introduce the “scale-and-stretch” method to warp image with the matched local scaling factors. The authors [22] then expand the same philosophy to video retargeting by incorporating motion-aware constraints. In practice, Krahenbuhl et al. [23] implement a unique system based on warping for streaming applications.

In essence, the flexible pixel rearrangement of the heterogenous methodology avoids explicit content sacrifice, suggesting somewhat latent preference to a globally retargeted view. They produce excellent results in natural images and in scenarios when aspect ratio change is significant. However, individual manipulations of pixels also demand for such a large number of pixel parameters to be jointly optimized that there are always some pixels are not coordinated well spatially or temporally. Therefore, it is

common to observe the resultant deformation and/or inconsistency, which can be quite noticeable or even disastrous when the content involves well-defined objects.

2.1.2 Our Approach

Motivated by the difference between homogeneous and heterogenous methodologies, our approach takes into account the major design considerations: a) content preservation, b) the temporal retargeting consistency and c) the prevention of deformation. We ask a big question: from the perspective of observers, what are the real priorities among these considerations to perform a compelling video retargeting?

For content preservation vs. non-deformation, since they are both the in-frame considerations, it is insightful to explore the conclusions from image retargeting evaluations. Recently, Rubinstein et al. [1] created a benchmark image database and performed comprehensive statistical analysis on the human response to several state-of-the-art techniques from both homogeneous and heterogenous methodologies. They found that “viewers consistently demonstrate high sensitivity to deformation”, and “in many cases users prefer sacrificing content over inserting deformation to the media” [1]. Then for the temporal retargeting consistency that stands out in video retargeting, it plays an even more important role: even a mild inconsistency among adjacent frames would lead to the annoying flickering or jittering, which may fatigue human eyes quickly; whereas, content sacrifice may be less tangible without presence of original video since viewers are capable of recovering the content via imagination.

1) Our approach enables a user-specified retargeting scale, making the trade-offs between the visual content preservation and retargeting consistency appropriate for each individual viewer. Comparing heterogenous approaches that assertively impose the pro-global-view retargeting or the traditional homogeneous methods that impose the pro-local-view retargeting, our refined approach is closer to the real viewer aesthetic fact. This user-specific practice is inspired and justified by the study of human response to the retargeting scale.

Figure 2-1. Left: original image. Right top: retargeted image by pro-global view. Right bottom: retargeted image by pro-local view. Photos courtesy of RetargetMe database by MIT [1] for the image retargeting benchmark.

2) Our video retargeting system is capable of processing long-duration video with generic contents, not limited to the realm of short video clips as in many existing work. Unlike short clips that last only one scene, long videos contains many scene changes and the length of a scene can be very long or very short. The proposed system studies the temporal retargeting consistency comprehensively, which to our best knowledge is the first time to discuss this issue on such an elaborate level. Considering that frames at different temporal locations require different consistency, our retargeting system adapts the trade-off structure to different frame types, aiming at striving for the most freedom for saliency preservation.

On the whole, this framework bridges the miscellaneous retargeting considerations in practice with a structuralized analysis of retargeting objectives and endows the generic video retargeting problem with elegant mathematical formulations.

3) As the substantial algorithm of the video retargeting system, we formulate the retargeting as an optimization problem, where the variables to solve are the sequential positions of the retargeting windows over a subshot. Regarding the objective function, we propose the volume retargeting cost metric to systematically considers the retargeting consistency and VI preservation together.

We further represent the optimization into a graph context and then prove the equivalency of the optimization as searching for the path with minimal total cost on the graph. The solution is obtained in dynamic programming fashion. It is encouraging that the solution may extend to the other measures of visual interestingness preservation and consistency refined in the future.

4) For the attention modeling, we propose an innovative computational model with non-linear fusion of spatial and motion channels. The proposed model with independent channels captures the distinct mechanism of human perception to luminance, chrominance and motion stimulus, avoiding the shape twist of salient entities due to joint processing of intertwined spatial-temporal data in many attention models. Also, the non-linear fusion scheme takes advantage of the properties of the computed visual interestingness distribution from the two channels and strives for the detection of meaningful entities, which may subsequently make the suggested interesting objects more likely be preserved in the retargeting window.

This chapter is organized as follows. Section 2.2 describes the statistical study of human response to retargeting scales. Section 2.3 presents the proposed video retargeting system architecture. Section 2.4 describes a visual information loss metric to measure interestingness preservation and Section 2.5 proposes a non-linear fusion based attention model. Section 2.6 proposes a volume retargeting cost metric and the corresponding graph representation with the dynamic programming solution. Section 2.7 presents a method to choose a unified scale for a shot. Section 2.8 shows our experimental results.

2.2 Global or Local? — Statistical Study on Human Response to Retargeting Scale

This section studies how viewers evaluate retargeting scales. Ideologically, most heterogeneous approaches follow the hypothesis that a global scale is preferred in the retargeting task while most current homogeneous approaches tend to retarget content in

a local scale (they may also be pro-global scale if the aspect ratio change is not drastic). Considering both the merits and weaknesses with the two methodologies, we inquire the validity of the hypotheses by examining whether there really exists a consensus of perception bias towards a particular scale, either global or local. This inquiry is inspired by the subjectivity of human aesthetics, the randomness of the image content, the retargeting purpose and many other non-objective factors, which probably suggest there is no consensus of preference to make one scale dominate the other and thus both hypotheses are opposed. Note that although preserving the original image on a global scale is intuitively justified among many viewers and then is enforced by most existing works, a significant number of people alternatively consider it would not be strictly necessary if at the cost of geometric distortions or inconsistency over adjacent frames. They prefer a nice local retargeting that enhances the emphasis on the object of interest. For example, in Fig. 2-1, the top right image is by purely scaling the original image globally without any content removal and the bottom right image is obtained by first cropping a local region and then scaling to fit the target display, but with a smaller scaling factor. As expected, many viewers we surveyed claimed the top is better for its complete content preservation; however, other viewers argued that it is reasonable to cut off the boundary regions as the green twigs there are not visually important and not even intact in the original frame, but the cropped retargeting renders the sunflower with finer resolution as a considering perception advantage of the bottom image. To clear the disagreement, we conduct a statistical study http://www.mcn.ece.ufl.edu/public/ZhengYuan/statistical_study_retargeting_scale.html to test which hypothesis of retargeting scale is true, aiming at the potential computational measure for the supported hypothesis. Or otherwise neither can override the other if suggested by the study, we may leave the freedom of choosing a retargeting scale to the individual viewer if neither hypothesis can be statistically supported to dominate the other.

Figure 2-2. Up left to bottom right: ArtRoom, MusicSound, DKNYGirl, Redtree, Trees, Football, Sunflower, Architecture, Butterfly, CanalHouse, Perissa, Child, Fish, GreekWine, Fatem. Each original image is in one of the three common aspect ratios (4:3, 3:2 and 16:9) and is retargeted to all three aspect ratios. They are represented as thumbnails due to the conflict between various sizes and the space limitation. Our survey website provides the retargeting results in their authentic sizes. Photos courtesy of RetargetMe database by MIT [1] for the image retargeting benchmark.

We devise the statistical experiment as follows: given 15 different images that cover most popular topics in photography, we retarget each image with two distinct scales, which represent pro-global-view[1] and pro-local-view strategies, respectively. We also incorporate target aspect ratio as a variable to our statistical study: each one of the 15 images is in one of three common display aspect ratios $3:2$, $4:3$ and $16:9$; for each image, we retarget it into all three aspect ratios with two retargeting scales. Then we collect the response scores from 60 viewers to evaluate the retargeted image. The response scores are rated according to the individual aesthetic standard, ranging from 1 to 10 with the increment as 1. Our objective is to determine if there is a meaningful gap of the response scores between the two retargeting strategies.

[1] Here we use two methods to implement the global-scale retargeting, the global cropping-and-scaling and the heterogeneous method in [21]. The former introduces no shape deformation while its modest cropping may remove a little border areas and the latter keeps all contents with least shape deformation. We assign the higher score from the two methods as the score for global retargeting to measure its best perception performance whereas a single method is hardly to be deformation-free and keeping all contents as well.

Statistically speaking, given the grouped score samples $X_{ik} = [x_{ik1}, \dots, x_{ikj}, \dots, x_{ikn}]$, $i = 1:60, k = 1:2, n = 15 \times 3 = 45$, where i is the viewer index, j is the image index, k is the strategy/group index and n is the number of retargeting processes, we want to infer if the subjective scores suggest retargeting equivalence between two groups. Retargeting equivalence is statistically defined as the norm of the group mean difference $\Delta\mu = \bar{X}_1 - \bar{X}_2$ is bounded by the subspace $\mathcal{H} = \{\Delta\mu : -1 < \Delta\mu_j < 1, \forall j = 1:n\}$. (Note that two ratings with difference less than one is considered the equivalent as the discretization of the score rating is 1) with a high probability $1 - \alpha$. Its two complement spaces are $\mathcal{H}^* = \{\Delta\mu : 1 < \Delta\mu_j < \infty, \forall j = 1:n\}$ and $\mathcal{H}^\circ = \{\Delta\mu : -\infty < \Delta\mu_j < -1, \forall j = 1:n\}$. $\mathcal{H}^*$ represent the space where viewers generally considers local scale is better and $\mathcal{H}^\circ$ refers to the space where global scale is better.

$$P(\Delta\mu \in \hat{\mathcal{H}}) > 1 - \alpha \tag{2–1}$$

We utilize confidence interval estimation to solve the problem. Assume the collected response scores X_{ik} of each group follows a multivariate Gaussian distribution with mean μ_k and variance-covariance matrix Σ.

$$X_{ik} \sim N(\mu_k, \Sigma) \quad \forall i = 1:n, k = 1:2 \tag{2–2}$$

Hence, the difference of two group means $\Delta\mu$ also follows Gaussian distribution with mean $\mu_1 - \mu_2$ and covariance Σ,

$$\Delta\mu \sim N(\mu_1 - \mu_2, \Sigma) \tag{2–3}$$

Given the confidence level $1 - \alpha$, we estimate the corresponding interval $\hat{\mathcal{H}}$ where $\Delta\mu$ probably lies in according to the distribution in Eq. (2–3). If $\hat{\mathcal{H}}$ is a subspace of $\mathcal{H}$, we considered the two strategies are retargeting equivalent. If $\hat{\mathcal{H}}$ is a subspace of $\mathcal{H}^*$, it suggests that viewers prefer local retargeting. If $\hat{\mathcal{H}}$ is a subspace of $\mathcal{H}^\circ$, it is highly likely

that viewers prefer global retargeting. Otherwise, no significant preference bias exists. Table 2-1 describes the 95% confidence interval of the group mean difference for each dimension/image, using T^2 confidence interval analysis.

As indicated in Table 2-1, nine retargeting processes marked with $^\circ$ result in the confidence intervals in $\mathcal{H}^\circ$, so they suggest the preference to global retargeting. Meanwhile, another six retargeting processes marked with * has the confidence intervals in $\mathcal{H}^*$, suggesting local retargeting is preferred. For the rest 30 retargeting processes, their intervals are either retargeting scale equivalence or no significant preference bias. Therefore, if we take the retargeting generically, there is no consensus on the preference to some particular scale. This conclusion suggests that in the retargeting task, one does not have to preemptively preserve a global view or local view. Based on this inference, our proposed system endows the freedom of choosing the global view or local view or some other scales in between to individual users according to their own aesthetic preferences and needs. This strategy maximizes the retargeting performance by allowing the greatest flexibility. The corresponding interface of a scale optimizer is designed in a try-and-error fashion to facilitate viewers to determine a more customized viewing scale.

2.3 System Design

This section discusses the design methodology elaborately. We propose a practical retargeting system that adopts the homogeneous methodology, with retargeting process equivalent to searching for the scale and position of a proper retargeting window for each frame. In our refined homogeneous approach, we let viewers determine their aesthetically pleasing scale, hence the corresponding removal of the periphery region based on the customized scale should be considered reasonable for the individual viewer. Particularly, the homogeneous methodology allows a retargeting window (RW) to contain any region, even to the full inclusion.

Table 2-1. Confidence intervals by T^2 estimation with level $1-\alpha = 95\%$. Courtesy of RetargetMe database by MIT [1] for the image retargeting benchmark.

Images	3:2		4:3		16:9	
	Lbound	Ubound	Lbound	Ubound	Lbound	Ubound
ArtRoom	-1.00	-0.02	-0.78	0.23	-2.81	-1.64°
DKNYGirl	-0.40	0.58	-0.26	0.57	-0.15	1.08
Child	-0.79	0.86	0.78	1.64*	-0.42	0.55
Butterfly	-0.09	0.59	1.12	1.85*	1.11	1.60*
Greenwine	-1.95	-1.10°	-2.51	-1.70°	-1.05	-0.41
CanalHouse	1.04	2.71*	-0.80	0.69	-0.37	0.49
Sunflower	-0.28	0.85	-0.28	0.69	-0.89	0.11
Fatem	-2.59	-1.66°	-1.47	0.61	-1.19	0.59
Fish	0.10	0.90	1.55	2.36*	-0.52	0.84
Perissa	-3.42	-2.03°	-2.03	-0.98°	-2.22	-0.94°
MusicSound	0.17	1.16	-0.07	0.79	0.84	2.05*
Football	-3.36	-2.67°	-1.94	-1.16°	-1.38	-0.47
Trees	-1.08	0.79	-0.99	1.19	0.03	0.94
Architecture	-1.05	0.52	-0.89	0.01	-0.87	0.34
Redtree	0.28	1.32	-0.30	0.55	-0.11	0.72

2.3.1 System Architecture

Fig. 5-1 describes the system architecture and explains the work flow of how to find the scale and location for the RW of a frame. The system consists of six major components: a) Shot Detection b) Saliency Calculation c) Scale Optimization d) Visual Information Analysis e) Boundary-Frame Retargeting and f) Inner-Frame Retargeting. Shot detection module divides a long video into visually coherent units for subquential independent retargeting. Saliency detection module implements the attention model we proposed in Section 2.5 to quantize the interestingness of each pixel on the frame. Scale optimization module evaluates an entire shot and then determines a single optimal scale of the RWs for the shot. Visual-Info Analysis module transforms the saliency distribution into the potential visual information loss value incurred by the RW at all possible locations. Boundary-frame Retargeting module searches the best location of the RW for the boundary frames. Finally, the inner-frame retargeting module takes the inner frames altogether and determines the dynamic trace of the RWs over them.

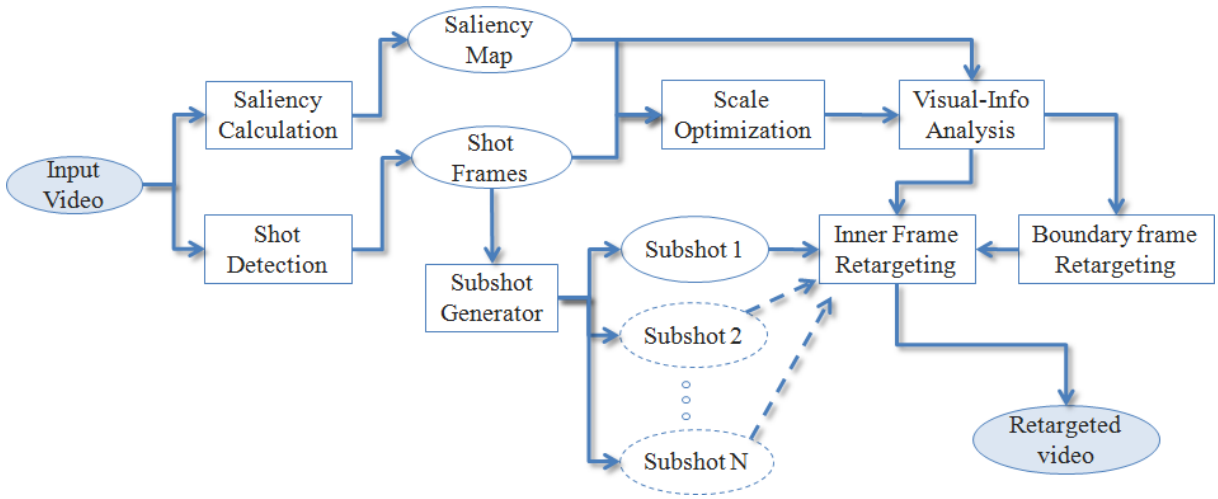

Figure 2-3. Rectangle block: system modules. Elliptical block: video or image frame.

As a new frame coming in, the shot detection module compares the statistics of its luminance and chrominance with previous frames and decides if a shot is initiated [24][25]. At the same time, the saliency calculator takes the incoming frame and generates a saliency distribution map in the same frame size. Then all frames in the detected shot with their saliency maps are streamed into the scale optimizer to find a unified viewing scale for the RWs of the entire shot. As the size of the RW is aesthetically determined with the optimized scale, the Visual-Info Analyzer computes the potential loss due to the cropping and scaling incurred by the RWs at all possible locations for each frame. Finally, the optimal locations of the RWs are searched in two cases: for the frames at the boundary of a subshot (the subshot generator chops a shot into several even length subshots), Boundary-Frame Retargeting module finds the RW by only minimizing the visual-info loss of the individual frame; for other within-subshot frames, Inner-Frame Retargeting module considers the sequential frames jointly and determines a smooth trace of RWs.

2.3.2 Design Principle

The proposed system is designed in such a way in order to comprehensively address the visual consistency, rather than treating it as a trivial post processing

procedure. It is sensible to notice that the frames at different temporal locations raise distinct extent of consistency requirements. Utilizing this non-uniformity, we exercise customized retargeting schemes to specialize the trade-off between the required consistency and visual information preservation. This flexibility commits to the most exploitable visual-info preservation as long as the consistency requirement is satisfied. The consistency considerations are summarized as follows,

1) If a frame is asserted as the boundary of a shot, normally a clear cut is inserted before the next frame. The instantaneous content switching in the original video is prevalently used and it looks natural to viewers to symbolize the start of a new episode. Hence, no visual consistency concerns are required in the retargeted video either. For the same reason, a shot can be processed independently as the retargeting unit.

2) When a frame is within a shot, any abrupt RW change over two adjacent frames would result in a very noticeable jitter against the similar visual contents among adjacent frames. However, for these subshot boundary frames, we may only consider the VI preservation because the consistency requirement pertaining to them will be solved by their two immediate neighbors in 3). This strategy actually keeps an intermittent but timely update of where salient entities in the original frame go, rendering a retargeted video that is very likely to contain the interesting parts. This update mechanism avoids the “clamping” and prevents the possible inaccurate retargeting due to saliency inaccuracy from propagating to the next updated frame.

3) When a frame belongs to the inner-subshot frames, the two competitive objectives should be jointly considered. In order to obtain a visual-info-keeping but also smooth trace of RWs linking the optimal locations of the two subshot boundary frames, we stack the inner-subshot frames together as a volume and minimize the total visual-information loss of the volume under the constraint that any RW transits over adjacent frames with curvature less than a bound.

In addition, we fixed one unified scale of RWs for the entire shot provided that human perception is susceptible to scale variation of RWs over the frames with similar contents, known as the “flicker” artifact. As for the locations, on the contrary, certain amount of translation is tolerable by human perception. Actually, it is the translation of neighboring RWs that make the tracking of interesting content possible. The real challenge is how to configure the allowed translation carefully to smoothly catch up with the movement of the interesting region of the original video.

2.4 Visual Information Loss

As the aforementioned fact, our retargeting system aims at the best possible preservation of visual interestingness and also permits the retargeting consistency. This section discusses the measure of the former and the mathematical quantification of the visual-info loss caused by a retargeting window with candidate parameters. The measure is implemented in the Visual-Info Analysis module in the system.

The visual-info loss closely relates to the attention model in Section 2.5 as the pixels manipulated by the RW shed distinct interestingness. It comes from two sources: 1) the cropping loss: the retargeting system removes the periphery areas outside the RW so that viewers cannot attend the contents therein. 2) the scaling loss: the cropped content is further downsampled with scale s to be exactly in the retargeting size, which degrades the original frame into the coarser resolution.

Suppose the attention distribution $\phi(x, y)$ corresponding a frame is available, the cropping loss is then measured as the accumulated interestingness of those pixels discarded.

$$L_c = 1 - \sum_{(x,y)\in W} \phi(x, y) \tag{2–4}$$

Here, $\phi(x, y)$ is normalized saliency map such that $\sum_{(x,y)} \phi(x, y) = 1$, W is the retargeting window.

The measure of scaling loss is more challenging as the original frame cannot be compared directly with the retargeted counterpart due to their distinct sizes. [16]

compares the original frame with the one after applying a low-pass Gaussian filter, which presumably substitutes the real retargeted frame. However, this heuristic does not exactly match the real operation in the retargeting scheme — downsampling.
In our measure, we generate a synthesized frame which contains exact the same visual information as the retargeted frame but in the same size of the RW still. To guarantee the visual-info equivalency of the synthesized frame with the retargeted frame,we upsample the synthetic frame by mapping each pixel in the retargeted frame repetitively to be the successive $1/s \leq 1$ pixels in the synthetic frame. When $1/s$ is not an integer, bilinear interpolation is adopted. The scaling loss is then defined as the square difference between the synthesized frame and the content of original frame within the RW.

$$L_s = \sum_{(x,y)\in W} (F(x,y) - \hat{F}(x,y))^2 \tag{2–5}$$

where $\hat{F} = upsizing(g * downsizing(F, s), s)$. The visual information loss is thus the combination of the two sources,

$$L(x,y) = (1-\lambda)L_c(x,y) + \lambda L_s(x,y) \tag{2–6}$$

where λ is a factor to balance the importance of content completeness and resolution, adjustable to user's preference. Given a λ, we may find the optimal RW parameter $(\hat{x}, \hat{y}, \hat{s})$ by minimizing the visual information loss measure,

$$P(\hat{x}, \hat{y}, \hat{s}) = \arg\min_{x,y,s} L(x, y, s \cdot W_t, s \cdot H_t) \tag{2–7}$$

where W_s, W_t, H_s, H_t are the width and height of the source and target frames, respectively. Note the search range of (x, y, s) is constrained by $0 \leq x \leq W - s \cdot Wt$, $0 \leq y \leq H - s \cdot Ht$. Therefore, for the subshot boundary frames, the RW is finalized by

$\hat{x}, \hat{y}, \hat{s}$ and the retargeted frame is generated immediately by zooming the RW out by $\hat{s}$ times.

Note that the scale of RW is fixed within a shot (see Section 2.7). Therefore, for subshot boundary frames, we only search along (x, y) to minimize the visual information loss measure.

2.5 The Non-linear Fusion Based Attention Modeling

Attention modeling computes the meaningful saliency distribution for the evaluation of how much a retargeting window preserves the visual interestingness. Among a diversity of modeling methods, Guo et al. [8] take the advantage of the Quaternion Fourier Transform [26] to extract saliency from frequency domain, which is a principled approach free of the sensitivity of laborious parameter tuning. Meanwhile, [8] proves that it is in fact the phase spectrum of an image that captures its most saliency feature. Hence, it is proper to substitute the entire spectrum residue (Hou et al. [2] show the spectrum residue is an efficient way to detect saliency) with only the phase spectrum for the succinct purpose. In this dissertation, we inherit the merit of the Phase Quaternion Fourier Transform (PQFT) to detect spatial saliency but compute temporal saliency separately. We argue, it is not justifiable as in [16] to mix the three spatial images (one luminance and two chrominance) together with one temporal (motion) image into the quaternion and derive the saliency from the interweaved spatiotemporal data. Since human perceive spatial and temporal stimulus by distinct psychophysical mechanisms, treating spatial and temporal channels jointly with a unified transform would produce a saliency model that twists the actual spatiotemporal interaction. Instead, we first simulate the spatial saliency with PQFT and the temporal saliency with the local motion intensity. Then we fuse them nonlinearly to mimic the human responses to various spatial and temporal saliency distribution scenarios.

2.5.1 Spatial Saliency with Phase Quaternion Fourier Transform

Spatial saliency, to a large amount, originates from the stimulus of those pixels with high contrast in luminance and chrominance against their neighborhoods. The high contrasts are largely mathematically captured by the phase spectrum while Quaternion Fourier transform provides a nice principled approach to calculate the phase spectrum of a color image. In YC_bC_r color space, a video frame $I(x, y, t)$ is represented as three independent scalar images, $Y(x, y, t)$, $C_b(x, y, t)$ and $C_r(x, y, t)$, where x, y are the location of a discrete pixel on the frame, t is the id number of the frame in temporal order, Y is the luminance component and C_b and C_r are two chrominance components. Quaternion, generally a hypercomplex number, is $q = a + b\mu_1 + c\mu_2 + d\mu_3$, where, a, b, c, d are numbers in real value, and

$$\mu_i{}^2 = -1, \mu_i \perp \mu_j, i \neq j, \mu_1 \times \mu_2 = \mu_3 \tag{2–8}$$

Reorganize q in the symplectic form as the combination of two complex numbers,

$$q = f_1 + f_2\mu_2, f_1 = a + b\mu_1, f_2 = c + d\mu_1 \tag{2–9}$$

Then the quaternion Fourier Transform can be performed by two standard fast Fourier transforms (F_1 and F_2), where,

$$Q(u, v, t) = F_1(u, v, t) + F_2(u, v, t)\mu_2 \tag{2–10}$$

where $F_i(u, v, t)$ is the Fourier Transform of $f_i(x, y, t)$.

Assume $a = 0$, and substitute b, c, d with Y, C_b, C_r respectively, we may represent the video frame $I(x, y, t)$ as a pure quaternion frame $q(x, y, t)$,

$$q(x, y, t) = Y(x, y, t)\mu_1 + C_b(x, y, t)\mu_2 + C_r(x, y, t)\mu_3 \tag{2–11}$$

Then apply Eq. (2–10) to calculate the Quaternion Transform $Q_l(u, v, t)$ of the quaternion frame $q(x, y, t)$ and then its phase spectrum $P_l(u, v, t)$ is derived by dividing Q_l over its norm $|Q_l|$, $P_l = Q_l/|Q_l|$. Finally, we take the inverse quaternion Fourier transform for $P_l(u, v, t)$

to get the phase quaternion image $q_p(x, y, t)$. Finally the spatial saliency $\phi_s(x, y, t)$ is obtained by smoothing out the squared L_2 norm of $q_p(x, y, t)$ with a two-dimensional Gaussian smoothing filter g.

$$\phi_s = g * ||q_p(x, y, t)||_2 \tag{2–12}$$

2.5.2 Temporal Saliency with Local Motion Intensity

Generally, the disparity of temporally adjacent pixels comes from both the camera motion and local object motion. While camera motion applies to every pixel on the image globally, the local motion embodies the actual contrast of a pixel against its neighborhood. Hence, the local motion reflects the temporal saliency. We first use Kanade-Lucas-Tomasi (KLT) tracker to obtain a set of matched good feature points [27] of one video frame and its neighboring frame and then estimate the global motion parameters [28] with affine model. The location motion is extracted after removing the global motion from temporal disparity.

Denote the disparity of a KLT feature point (x_{t-1}, y_{t-1}) at previous frame I_{t-1} matched with $x = (x_t, y_t)$ in current frame I_t as $d = (d_x, d_y)^T$. Then the disparity can be approximated by a six-parameter affine global motion model d = Dx+t, where t is the translation component $t = (t_x, t_y)^T$ and D is a 2×2 rotation matrix. Represent the affine model in form of the matched feature points, $x_{t-1} = Ax_t + t$, where $A = E + D$ and E is a 2×2 identity matrix. The motion parameters in t and D can be estimated by minimizing the total neighborhood dissimilarity of all the matched features.

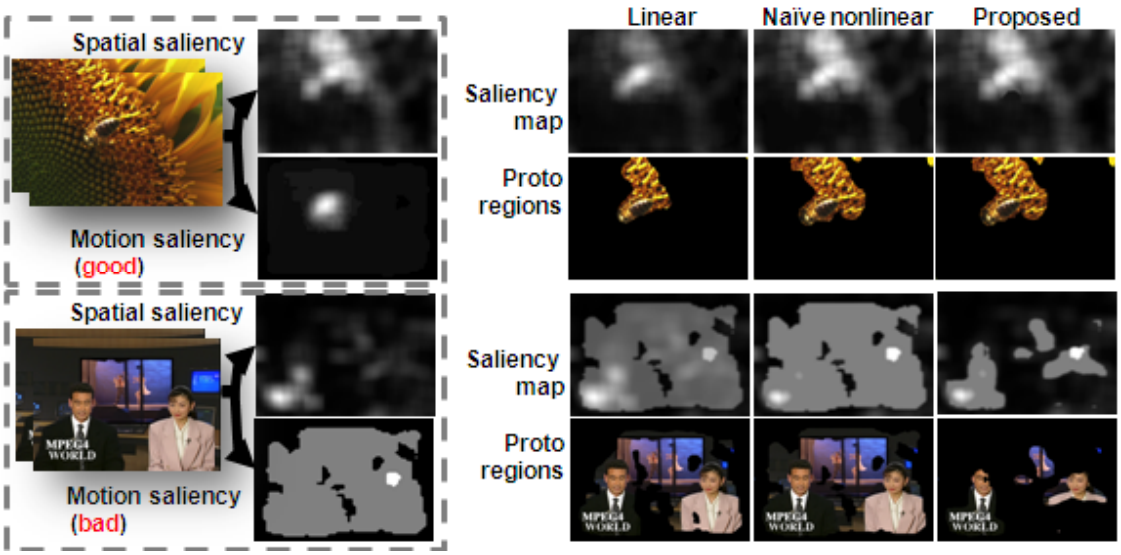

Figure 2-4. Comparison of linear combination, naive MAX operation and proposed approach when global motion is correct or wrong. Photos courtesy of the video trace library [3] for YUV video sequences.

$$\{\hat{A}, \hat{t}\} = \arg\min_{A,t} \int_W (I_t(\boldsymbol{Ax}+\boldsymbol{t}) - I_{t-1}(\boldsymbol{x}))^2 d\boldsymbol{x} \tag{2–13}$$

where W denotes a 8×8 neighborhood of feature points. We adopt the Least Median Squares approach to estimate the affine parameters robustly [28]. We generate the global motion predicted frame by warping the current frame $I_t(x, y, t)$ with the estimated parameter $\hat{A}$ and $\hat{t}$. The absolute difference (after smoothing) of the predicted frame with the previous frame $I_{t-1}(x, y, t)$ reflects the intensity of local motion or the temporal saliency,

$$\phi_m = g(\boldsymbol{x}) * |I_{t-1}(\boldsymbol{x}) - I_t(\hat{\boldsymbol{A}}^{-1}[\boldsymbol{x} - \hat{\boldsymbol{t}}])| \tag{2–14}$$

2.5.3 Nonlinear Fusion of Spatial and Temporal Saliency

The actual strategies adopted by human attention when fusing spatial and temporal saliency components are rather complex, depending on particular distributions of spatial and temporal salient areas. It is noticeable that humans are likely to attend video

frames on pixel clusters, which may suggest the existence of a meaningful entity, rather than attracted to a solitary pixel. In a huge variety of videos, salient entities almost universally express high spatial saliency value since most videos are captured, by either professionals or amateurs, to promote a certain foreground target. For example, when director shoots a TV show, the actor or actress at the center of focus is often depicted with unique characteristics to distinguish from the background or others. Furthermore, an entity also demonstrates high motion saliency if it happens to move. However, since the entity may also not move in a number of scenarios, motion saliency does not carry as discriminative power as spatial saliency. Hence, aiming at a fused saliency map featuring the high confidence and stability to suggest the existence of an entity, we make the spatial saliency as the primary cue and the motion saliency as secondary.

On the other hand, we observe that the detected spatially salient area (obtained by thresholding the spatial saliency map) is generally continuous and concentrated since the smoothing procedure in Eq. (2–12) is suited for the detection of high spatial contrast region, rather than an individual pixel. This trait nicely stands out the existence of an underlying salient entity. In order to further enhance the entity from other spatial salient areas, we resort to the correlation of the motion saliency map and increase the saliency value of the spatially salient areas by the amount proportional to the corresponding motion saliency value. The reason why this measure works is that both the spatial and motion saliency are driven by the same entities. Hence their intersection area suggests the probable existence of real salient entity. For example, the bee in Fig. 2-4 represents a salient entity that viewers usually attend. As indicated in the spatial saliency map, the high saliency area indeed covers where the bee locates but also admits other areas that also considered spatially salient. In this case, the calculated motion saliency map concentrates and suggests where the bee is. So the motion saliency is able to improve the conspicuousness of the bee by increasing the saliency value of the bee area of the spatial saliency map.

However, since motion saliency is performed on pixel level, the detected salient area can be rather dispersed as the pixels in scattered locations may contribute competitive local motions so that none of them can distinguish. In this case, the motion saliency map cannot suggest reliable and dominant salient entities, although the related pixels are indeed with high local motion. Fortunately, the spatial saliency map are not affected and thus can be utilized as a filter to confine the existence of the salient entity within the spatially saliency area and the pixels in such areas are still allowed to be enhanced by their motion saliency value. For example, in Fig. 2-4, the motion salient areas spread over the entire image, capturing none of the two anchors (the salient entities). However, the spatial saliency map successfully concentrates on the two anchors. Therefore, the final saliency map may use the spatial salient area to filter the motion saliency and only increase the saliency of pixels covering the anchors. Based on the analysis above, we devise the following nonlinear fusion scheme,

$$\phi(x, y, t) = \max\{\phi_s(x, y, t), \mathcal{M} \cap \phi_m(x, y, t)\} \tag{2–15}$$
$$\mathcal{M} = \{(x, y, t) : \phi_s(x, y, t) \geq \varepsilon\}$$

The intersection of $\mathcal{M}$ and motion saliency map simulates the filtering of spatial salient areas over temporal saliency. Within the intersection, the max operator use motion saliency value to intensify the areas where both two cues agree. This operator utilizes the characteristics of two saliency maps and produces a saliency map that emphasizes on salient entity.

2.6 Joint Considerations of Retargeting Consistency and Interestingness Preservation

This section discusses the more general case: the retargeting of the inner-subshot frames. Here we not only seek to avoid the visual information loss but also ensure the retargeting consistency over adjacent frames. In essence, the two objectives are

conflicting trade-offs. On one hand, if we find RWs by merely minimizing intra-frame visual loss for every frame, the resultant RW indeed always tracks the most updated information-rich areas. However, those areas do not necessarily move with consistent patterns and so with the searched RWs. Eventually we may end up with a video contaminated with annoying jitters. On the other hand, if we only desire the absolute retargeting consistency, the positions of a RW during the entire subshot should be fixed as otherwise any non-static RW would introduce an artificial global motion into the retargeted video. Nevertheless, the static RW in this situation is unable to track the dynamic visual information rich areas and preserve them.

Keep in mind that to permit retargeting consistency, it is impossible for the RW of each frame individually to attain its best position with local minimal visual-info loss. Hence, the joint consideration of the two objectives requires us to treat the entire inner-subshot frames together. We thus propose the volume retargeting cost metric in Eq. (2–16) to evaluate the retargeting of a whole subshot,

$$L_v = \sum_{t=1}^{N} L(x_t, y_t) + \omega \sum_{t=1}^{N} D(x_t, y_t, x_{t-1}, y_{t-1}) \tag{2–16}$$

where $D(x_t, y_t, x_{t-1}, y_{t-1}) = |x_t - x_{t-1}, y_t - y_{t-1}|_2$ is the differential of the RW trace at the frame t, measuring the retargeting inconsistency therein. L is the visual interestingness loss with the same interpretation as in Eq. (2–6). ω is the trade-off factor of the two objectives and N is the total number of inner frames in a subshot.

The volume retargeting cost metric features the total visual-info loss plus total retargeting inconsistency and emphasizes searching for a dynamic RW trace for the entire subshot. When those temporally adjacent frames are stacked together, the minimization of the volume metric explores a configuration of the RW positions with low total cost, forcing the information loss and retargeting inconsistency regarding each individual frame low as well. This metric is a relaxation of individual visual interestingness preservation when mingled with the retargeting consistency concern.

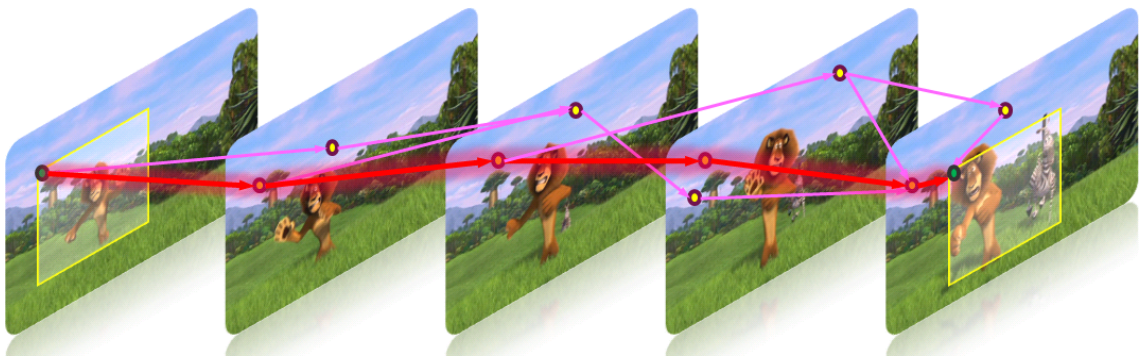

Figure 2-5. Each frame corresponds to a layer. Green: source and destination vertexes. Yellow: candidate vertex for each frame. Red: the path with least cost to denote optimized dynamic trace. Photos courtesy of the movie Madagascar by DreamWorks Animation [29].

Furthermore, in order to guarantee the retargeting consistency, we explicitly add a constraint that the norm of the differential of the RW trace at each frame should be less than a value. Therefore, the searching for the best trace of RWs is formulated as the following optimization problem,

$$\{\hat{x}_t, \hat{y}_t\}_{t=1}^{N} = \arg\min_{\{x_t, y_t\}_{t=1}^{N}} L_v(x_t, y_t) \tag{2–17}$$
$$s.t. \quad D(x_t, y_t, x_{t-1}, y_{t-1}) \leq \varepsilon$$

where ε is a psychophysical threshold below which human attention can tolerate view inconsistency.

2.6.1 Graph Representation of the Optimization Problem

The solution to the optimization problem in Eq. (2–17) is not trivial due to two aspects. 1) the arguments in the optimization are the trace of the RWs; they are the sequential positions of RWs in high dimension. 2) The objective function may be non-linear and even non-convex due to the nature of the computed saliency distribution. Thus, regular analytical or computation methods may not work in this situation; however, we observe it is meaningful to represent the optimization in a graph and explore the solution in that context.

As depicted in Fig. 2-5, we construct a graph with vertexes across N layers. Each layer symbolizes an inner frame from 1 to n within a subshot and each vertex on one layer represents a possible position of the RW to be decided. We assign a cost value for each vertex as the visual information loss incurred by the RW at the corresponding position. Then for every pair of vertexes on the adjacent layers, if the norm of differentiate with them is less than the bound ε, we establish an edge to link them together, suggesting a possible transition of the RW over adjacent frames. This establishment ensures the retargeting consistency constraint is satisfied. Then we also assign each edge a cost value as the norm of the differentiate with the two vertexes on its two ends. Specifically, the source and destination vertexes are the positions of the RWs on the two boundary frames of the subshot, which are obtained by minimizing visual information loss only. In this graph, any path from source to destination vertexes would be a possible trace of the RWs over the subshot. We define the cost of a path as the total cost of the vertexes and edges on it. Evidently, the cost of a path denotes the volume retargeting cost metric in Eq. (2–17) and the solution to the constrained optimization problem is the path with the minimal cost.

2.6.2 The Dynamic Programming Solution

We propose a dynamic programming method to find the path with minimal cost. Suppose the optimal path from the source vertex s to the jth vertex v_i^j on the ith layer, is $s \rightarrow v_i^j$, the question is how to find out the optimal path to the all the vertexes on the next layer. For the kth vertex v_{i+1}^k on layer $i+1$, denote the set of vertexes on layer i that has an edge to link v_{i+1}^k as $\mathcal{V}$. Then the option to find the best path to v_{i+1}^k is to augment the optimal paths up to every vertex in $\mathcal{V}$ to v_{i+1}^k and choose the one with minimal updated cost. Therefore, the recursive format of the objective function in Eq. (2–17) is as follows,

$$L_v(s \to v_{i+1}^k) = \min_{v_i^j \in \mathcal{V}} \{L_v(s \to v_i^j) + \omega \cdot D(v_i^j, v_{i+1}^k)\} + L(v_{i+1}^k) \tag{2–18}$$

where $s = v_1$ and $L_v(s \to v_1) = 0$. $L_v(s \to v_i^j)$ denotes minimized volume retargeting cost up to frame i assuming v_i^j is the destination or equivalently the shortest path from source vertex s of frame 1 to the jth vertex of frame i. $L_v(s \to v_{i+1}^k)$ is the shortest path up to the kth vertex of frame $i+1$, $D(v_i^j, v_{i+1}^k)$ denotes the cost of edge connecting the jth vertex of frame i to the kth vertex of frame $i+1$ and $L(v_i^j, v_{i+1}^k)$ is the cost of the kth vertex of frame $i+1$.

Notice that both the source s and destination d are predetermined by minimizing Eq. (2–6) for boundary frames. Starting from s, we update the best path up to the vertexes on each layer from 1 to N. In the end, for all the vertexes on layer N, we find out the one through which the best path leading to d,

$$v_N^k = \arg\min_{v_N^j} L_v(s \to v_N^j) + \omega \cdot D(v_N^j, d) \tag{2–19}$$

Then the best path leading to v_N^k is the final solution to the optimization in Eq. (2–17).

2.7 Optimal Selection of Scale in a Shot

As mentioned before, a unified scale is chosen for the entire shot. It determines the actual size of the retargeting window and reflects the aesthetic preference of a particular viewer. In the minimization of the information loss function in Eq. (2–6), the chosen scale depends on the cropping-scaling trade-off factor λ. Our proposed system allows viewers to initialize it, giving the system a general idea of generating a local or global retargeting view or somewhere in between.

Based on the initial λ, we find the optimal scale by making the partial derivative of the visual-info loss function with regard to the scale equal to zero. We perform this

operation on each frames in the shot and average the obtained scales as the unified scale of the shot.

The size of the RW definitely affects how quickly the RW responds to the dynamic content change. It is enlightened to think the size of a RW as gears of the manual transmission in a car and the transition rate of RWs between adjacent frames as the tachometer and the change of dynamic content as the road condition. We desire to maintain the transition rate within a pleasing level for consistency concerns. Meanwhile, the featured contents are supposed to be tracked and preserved no matter how quickly it changes.

Just as choosing a higher gear for a high speed if road condition permits, when the salient entities move rapidly, it is sensible to choose a larger RW size to satisfy the aforementioned two desires. On the contrary, when the salient entities moves low, we may alter to a smaller RW to save the resolution.

Therefore, give the initial λ which settles the aesthetic preference, we tune it afterwards to better suit for the content change. We use the velocity the RW transits (predicted by the initial λ) to estimate how fast salient entities move. Then based on the velocity estimate, we adjust the weight λ in order to obtain a more suitable scale, which then resize the RW to track salient entities more wisely.

$$\lambda' = \frac{\lambda}{1 + e^{-(\frac{1}{N} \cdot \sum_{i=1}^{N} \frac{(v_i - v_{i-1})^2}{\beta_i^2} - v_\alpha)}} \tag{2–20}$$

where $\frac{1}{N} \cdot \sum_{i=1}^{N} \frac{(v_i - v_{i-1})^2}{\beta_i^2}$ is the velocity estimate of the moving RW, N is the total number of frames in the shot. β_i is the maximum distance a RW can move from v_{i-1} at frame $i - 1$. v_α denotes a reference velocity which human finds pleasing. Given the updated weight λ', a new optimal scale average is calculated for the shot.

2.8 Experimental Results

We design three groups of experiments 1) spatial saliency modeling 2) video saliency modeling and 3) video retargeting to demonstrate the effectiveness and efficiency of the proposed attention modeling method and the proposed retargeting system. For evaluating convenience, the testing video sequences are clips in 1 to 2 minutes length. They cover many genres, have multiple scenes and complex backgrounds and sufficiently demonstrate the validness of our retargeting system. Also, in order to test the performance of our method on long videos, we retargeted two 10 minute movies the *big buck bunny* and *elephant dream* to variant resolutions and aspect ratios. They are full movies from Open Movie projects without license issue. In all three experiments, we compare our schemes with the representative existing approaches. Our schemes are implemented in C++ and Matlab with open sources libraries OpenCV (http://opencv.willowgarage.com/wiki/), FFTW3 (http://www.fftw.org/) and KLT (http://www.ces.clemson.edu/~stb/klt/).

Since the attention modeling and the entire video retargeting are highly subjective tasks, we summarize the experimental results in two fashions, including the image snapshots and online videos to provide readers the visual comprehension of the proposed schemes and also the subjective tests of the viewer perception scores.

2.8.1 Spatial Saliency Modeling

2.8.1.1 Proto-region detection results on spatial saliency map

Fig. 2-6 shows the attention modeling results (saliency maps) generated by human, saliency toolbox (STB, http://www.saliencytoolbox.net/), context-aware saliency (CS) [30], Histogram-based Contrast (HC) [31], Region-based Contrast (RC) [31] and our attention modeling algorithm respectively. We use a collection of images in multiple resolutions and aspect ratios for the experiments. Beside the resultant saliency maps, we also illustrate the so called “proto-regions” [32], which are found by thresholding the

normalized saliency maps with the same threshold [2], to show the contours of salient entities in contents, as shown in the red-circled regions on original images in Fig. 2-6.

Note that in rigid video retargeting, it is the shapes of the detected salient objects and whether they are accurately distinguished from the background that matter most because the salient objects, as integrated entities, are intended to be kept in the retargeting window. A good saliency map here does not only capture all the pixels of the salient objects, but also avoids to capture too much details from the background, which weakens the emphasis of saliency objects. Therefore, we use the similarity of the "proto-regions" with the ground-truth "attended objects" (generated by us as viewers in Col. 2 of Fig. 2-6) to measure the effectiveness of each method.

Comparing the saliency maps of STB, our algorithm successfully detects more pixels on the salient objects and thus our "proto-regions" have much more similarity of the shape with the ground-truth. For example, for image 4 which shows two children playing at beach with a sailing boat in the sea, our algorithm successfully extracts the regions of children and the boat as salient objects while STB is only able to capture the line between the sea and the sky. As for HC and RC, their methods produce a saliency map that emphasizes regions, which is suitable for image segmentation (e.g. image 1 and image 7 of Fig. 9). However, we observe occasionally the objects of interest, although wholly extracted as regions, are not very "salient" comparing with other regions. For example, for HC, the house region in image 3 is considered less salient then the flower region in both the two methods and the saliency of the old lady's face is overshadowed by the wall nearby; for RC, the children in image 4 are less salient than the sea surface. For the comparison with CS, both CS and our algorithm detect all the pixels of the salient objects. Indeed, CS captures more details in both the salient objects and the background, which potentially counterbalances the importance of the salient objects. In image 2, CS results show both the house and the ground are captured, resulting in a "proto-region" that includes many unnecessary background regions.

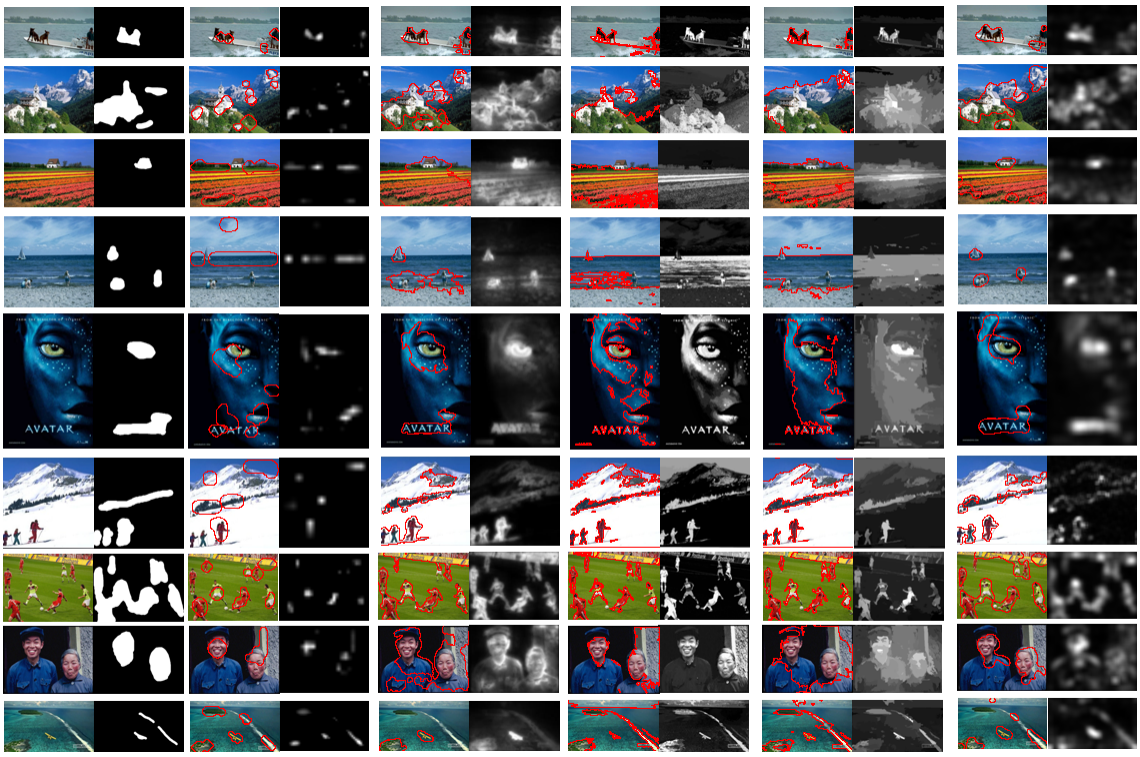

Figure 2-6. Col.1: original image. Col.2: human labeled salient regions. Col.3: proto-regions detected by STB. Col.4: saliency map by STB. Col.5: proto-regions detected by CS. Col.6: saliency map by CS. Col. 7: proto-regions by HC Col. 8: saliency map by HC. Col. 9: proto-regions by RC. Col. 10: proto-regions by RC. Col. 11: proto-regions detected by our method. Col.12: saliency map of our method. Photos courtesy of the image database by Xiaodi Hou [2].

2.8.1.2 Subjective test of the attention modeling

We conduct a subjective test to collect the scores ranging 1 to 5 graded by 60 participants (http://www.mcn.ece.ufl.edu/public/ZhengYuan/spatial_saliency_comparison.html). Then we conduct confidence interval analysis to calculate where the mean scores of each method lies in with a 95% probability.

The results of confidence interval analysis are summarized in Table 2-2 and Fig. 2-7 is the bar chart. It is shown that for all images except image 7 and image 9, they exhibit a consistency that the confidence intervals of STB do not overlap with ours. Since our method has a higher mean score than STB, it suggests that ours method outperforms STB. Comparing with HC and RC, except for image 2 and

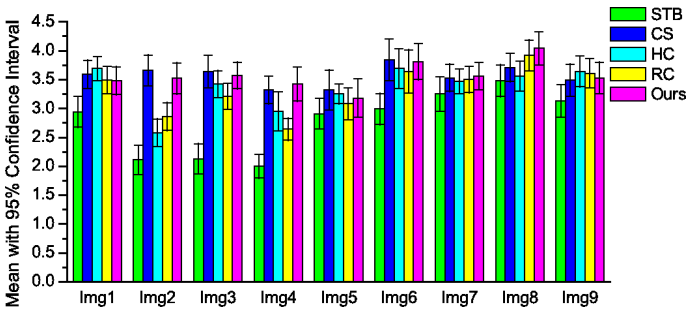

Figure 2-7. Statistics for saliency maps by STB, CS, HC, RC and Our method.

image 4, the confidence intervals of our method overlap with the two methods but generally occupy slightly higher ranges for other images. The possible reason is that the saliency of the objects of interests by HC and RC may be diluted by the color variations nearby. For the comparison with CS, it seems that for most images, the confidence intervals overlap. Generally, in terms of the grading on the similarity of detected salient objects with the ground truth, the performances of ours and the CS method are similar. Sometimes, the two methods still demonstrates a potential of performance gap. In image 5, the confidence interval of CS is higher than ours. However in image 4, our confidence interval is higher. The possible reason is that in the saliency map by CS, too much details on the two children are extracted. Thus the same threshold level which perfectly distinguishes the shape of the sailing boat results in two oversized children. This weakened salient object due to over-extracted details also explains the score comparison on image 8.

2.8.1.3 Computational complexity

We test the time complexity of the five methods on the operating system as Windows 7, 64bits and hardware as CPU 2.67GHz and memory 4GB. The time complexity is also in Table 2-2. Indicated by the computation time, our method is faster than all the other four methods, especially our computational efficiency is 1000 times of that of SC, thus is suitable for real time video retargeting system.

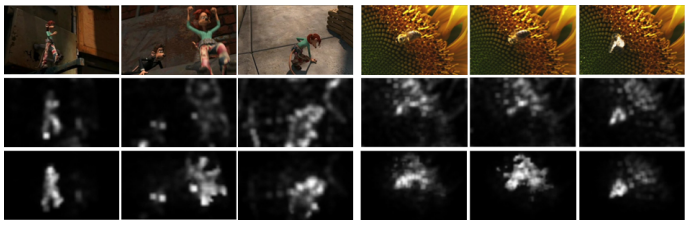

Figure 2-8. The first and fourth rows: original video sequences Rat and Sunflower. The second and fifth rows: the saliency modeling results of baseline-PQFT. The third and sixth rows: the saliency modeling results of our method. Photos courtesy of the video trace library [3].

2.8.2 Attention Modeling Comparison in Video Retargeting

2.8.2.1 Saliency video comparison

Here we focus on the attention modeling unit and compare the proposed non-linear fuse modeling with the baseline-PQFT [8]. We present the saliency comparisons on six videos, Chicken, Rat, News, Sink, Jets and Sunflower. The comparisons are in form of live videos with baseline-PQFT side by side. Please visit http://www.mcn.ece.ufl.edu/public/ZhengYuan/saliency_comparison.html to watch them. Due to the space limitation, Fig. 2-8 presents the snapshots of two representative saliency video results (*rat* and *sunflower*).

As both the online videos and Fig. 2-8 show, the salient objects in Rat detected by our method has more resembling shape and gesture with the Rat (the salient object in rat) than that of Baseline-PQFT. The same case happens in the bee in Sunflower. The reason is that in our non-linear fusion modeling, the fusion of two channels emphasizes the potential existence of salient objects rather than individual salient pixel, thus results in better shape of the detected salient objects. Baseline-PQFT feeds the spatial and motion data directly into the PQFT calculation. Since the values are different in nature, the shape of detect salient objects may be twisted.

2.8.2.2 Subjective test

We also carry out the subjective test for the saliency modeling comparison. Like the spatial saliency evaluation, we collect the scores from 60 participants and perform the

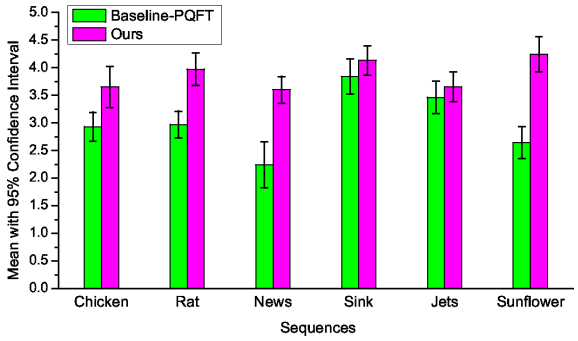

Figure 2-9. Green: saliency Baseline-PQFT. Purple: ours.

confidence interval analysis of the mean scores for the two saliency modeling methods. Table. 2-3 presents the results of the estimated interval where the two mean scores lie in for each video and Fig. 2-9 shows the bar chart. From Fig. 2-9, we can see for Chicken, Rat, News and Sunflower, the interval by our method does not overlap with that of Baseline-PQFT and our interval is higher. For Jets and Sink, although our confidence intervals have a small overlap with Baseline-PQFT, they still occupy a higher range. It suggests that participants generally consider our saliency modeling is better than Baseline-PQFT in terms of salient object detection.

2.8.3 The Comparison of Video Retargeting Approaches

2.8.3.1 Video and image snapshot comparison

We present video targeting comparison on 6 videos, Barnyard, FashionShow, HearMe, Madagascar, Rat and Soccer. They are in 1 to 2 minutes length with multiple scene changes and complex backgrounds. We perform video retargeting with our retargeting system and by two previous homogeneous methods: Single Frame Smoothing (SFS) [15] [16] and Back Tracing (BT) [17]. Each original video has an aspect ratio between $\frac{1}{2}$ and $\frac{3}{4}$ with width more than 600 pixels. The retargeted output sizes are 320×240 and 480×240, so in the retargeting process, the videos are both squeezed and stretched.

We demonstrate the video comparison results on http://www.mcn.ece.ufl.edu/public/ZhengYuan/video_retargeting_comparison.html. From all the videos, we may

generally observe that SFS suffers from jittering, which causes uncomfortable feelings of viewers. Back Tracing is mostly acceptable, however, the retargeted video is not always able to preserve regions of interest of the original video. In comparison, our method throughout preserves salient region as a frame goes further and avoids jitter effects as well[2] .

Due to the space limitation, Fig. 2-10 captures the snapshots of two representative video results, Madagascar (retarget to 320×240, aspect ratio squeezed) and the Fashionshow (retarget to 480×240, aspect ratio stretched). For the authentic performance comparison, readers may still refer to our website for observation.

In the results of SFS, although the lion and the zebra are preserved completely, the retargeting window shifts back and forth frequently, which suggests huge jitter effects in the retargeted video. Regarding BT, from the 1st to the 3rd frames, the retargeting window includes complete zebra, however, as frame goes to the 4th and the 5th, it is left behind by the zebra due to the fast motion of the latter. So most parts of zebra is lost in the retargeted video. In contrast, our result yields a visually consistent retargeting window trace to preserve zebra completely. For the *fashionshow*, BT results in a retargeting window that does not include the model's face as she moves across frames; In comparison, for the corresponding frames, our results alleviate the consistency clamping artifact and mostly keep the model's face inside. For SFS retargeting, it still has shaky artifact, viewers may perceive that through online video results.

2.8.3.2 Subjective test

Since it is obvious from the retargeted videos that both BT and our approach are better than SFS, we only need to evaluate BT and our approach quantitively. For each

[2] Occasionally it may include modest unnatural camera motion, which can be alleviated by increasing ω in Eq. (2–17)

Figure 2-10. Two snapshots of video retargeting comparison among SFS, BT and our approach. The first and fourth rows: single frame search and smoothing. The second and fifth rows: back tracing. The third and sixth rows: our proposed approach. Photos courtesy of the movie Madagascar by DreamWorks Animation [29] and the runway model video by Michael Kors at YouTube [33].

retargeting process, we collect the subjective scores ranging 1 to 5 from 60 participants, as in the previous tests.

Based on the collected scores, we perform confidence interval analysis to compare two methods according to where their true mean scores lie. Table 2-4 summarizes the confidence intervals for each video being retargeted to 320×240 and 480×240. The bar chart in Fig. 2-11 illustrates the relative locations of the mean scores. From Table 2-4 and Fig. 2-11, it seems for all the retargeting process where the aspect ratio is stretched, the confidence intervals by our method are higher than those of BT, although sometimes they overlap with BT with a small percentage. For the retargeting process where the aspect ratio is squeezed, the confidence intervals of our method indeed overlap with

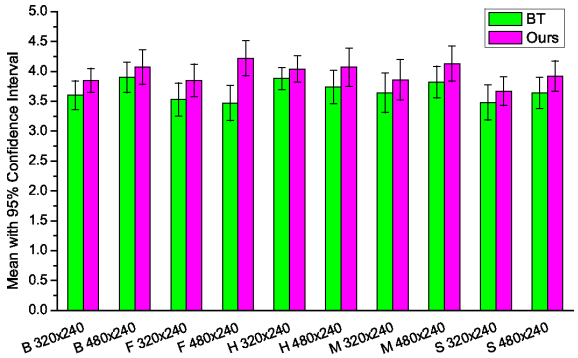

Figure 2-11. Statistical analysis for video retargeting approaches. Green: Back tracing. Purple: Ours.

those of BT; however, it seems that they still occupy higher ranges. This subjective test results suggest that our method is generally better than BT, especially in the case where the aspect ratio is stretched.

Table 2-2. Confidence interval analysis for subjective evaluation on spatial saliency modeling $\alpha = 0.05$. Courtesy of image database by Xiaodi Hou [2].

		L Bound	Mean	U Bound	Time
Img1	STB	2.68	2.94	3.21	97.0 ms
	CS	3.34	3.59	3.83	129608 ms
	HC	3.49	3.69	3.90	5323 ms
	RC	3.26	3.49	3.73	35501 ms
	Ours	3.24	3.48	3.72	10.9 ms
Img2	STB	1.86	2.11	2.36	103.2 ms
	CS	3.39	3.66	3.92	74972 ms
	HC	2.34	2.58	2.82	2069 ms
	RC	2.62	2.86	3.10	25971 ms
	Ours	3.25	3.52	3.79	12.1 ms
Img3	STB	1.87	2.13	2.39	95.7 ms
	CS	3.36	3.64	3.92	105808 ms
	HC	3.19	3.42	3.66	1076 ms
	RC	2.99	3.21	3.44	20531 ms
	Ours	3.34	3.57	3.80	10.4 ms
Img4	STB	1.79	2.00	2.20	79.8 ms
	CS	3.08	3.32	3.56	96148 ms
	HC	2.61	2.95	3.29	385 ms
	RC	2.45	2.64	2.83	1831 ms
	Ours	3.12	3.42	3.71	14.0 ms
Img5	STB	2.65	2.91	3.17	100.4 ms
	CS	2.97	3.32	3.66	37638 ms
	HC	3.08	3.25	3.42	1445 ms
	RC	2.81	3.08	3.36	7529 ms
	Ours	2.85	3.18	3.51	73.3 ms
Img6	STB	2.72	2.99	3.26	78.8 ms
	CS	3.48	3.84	4.20	106181 ms
	HC	3.35	3.69	4.04	428 ms
	RC	3.27	3.64	4.02	2215 ms
	Ours	3.50	3.81	4.12	9.8 ms
Img7	STB	2.95	3.25	3.54	98.3 ms
	CS	3.29	3.53	3.76	172269 ms
	HC	3.26	3.47	3.68	5561 ms
	RC	3.28	3.50	3.73	220497 ms
	Ours	3.32	3.56	3.79	12.4 ms
Img8	STB	3.20	3.48	3.75	94.6 ms
	CS	3.46	3.71	3.95	130905 ms
	HC	3.31	3.56	3.82	765 ms
	RC	3.66	3.92	4.19	5170 ms
	Ours	3.75	4.04	4.32	8.5 ms
Img9	STB	2.85	3.13	3.41	100.0 ms
	CS	3.21	3.49	3.76	49442 ms
	HC	3.38	3.64	3.91	5757 ms
	RC	3.36	3.61	3.87	44585 ms
	Ours	3.25	3.53	3.79	11.7 ms

Table 2-3. Confidence interval analysis for subjective evaluation on video saliency modeling $\alpha = 0.05$. Courtesy of the video trace library [3].

		L Bound	Mean	U Bound
Chicken	Baseline-PQFT	2.54	2.93	3.06
	Ours	3.28	3.65	4.02
Rat	Baseline-PQFT	2.73	2.97	3.21
	Ours	3.67	3.97	4.26
News	Baseline-PQFT	1.82	2.24	2.66
	Ours	3.36	3.60	3.84
Sink	Baseline-PQFT	3.52	3.84	4.16
	Ours	3.86	4.13	4.39
Jets	Baseline-PQFT	3.16	3.46	3.75
	Ours	3.38	3.65	3.92
Sunflower	Baseline-PQFT	2.35	2.64	2.93
	Ours	3.92	4.24	4.56

Table 2-4. Confidence interval analysis for video retargeting $\alpha = 0.05$. Courtesy of the video trace library [3].

		BT			Ours		
Videos	RetargetSize	lB	M	uB	lB	M	uB
Barn	320x240	3.36	3.60	3.84	3.65	3.85	4.05
	480x240	3.64	3.90	4.15	3.78	4.07	4.36
Fashion	320x240	3.25	3.53	3.80	3.58	3.85	4.12
	480x240	3.17	3.47	3.76	3.92	4.22	4.51
Hearme	320x240	3.69	3.88	4.06	3.82	4.04	4.26
	480x240	3.46	3.74	4.02	3.75	4.07	4.39
Madaga	320x240	3.31	3.64	3.97	3.52	3.86	4.20
	480x240	3.55	3.82	4.08	3.83	4.13	4.42
Soccer	320x240	3.18	3.48	3.77	3.43	3.67	3.91
	480x240	3.34	3.64	3.86	3.66	3.92	4.17

CHAPTER 3
VIDEO SUMMARIZATION

3.1 Background

The fundamental purpose of video summarization is to epitomize a long video into a succinct synopsis, which allows viewers to quickly grasp the general idea of original video. The resultant summary provides a compact representation of the original content structure. Although brief, a good summary preserves all necessary hallmarks of the original video and viewers are sufficiently able to recover original content through reasoning and imagination. To-date, there are excellent surveys on video summarization, abstraction and skimming [34, 35]. These papers cover many detailed approaches with one common strategy: being formulated as an optimization problem, it selects a subset of video units (either static frames or dynamic shot clips) from all possible units in the original video such that they maximize some metric function of the summary quality. Based on the cognitive level (signal and semantic) where a metric function lies, we categorize current summarization techniques into three types. Type I utilizes signal level features to measure the difference of a video summary from its original. Various implementations include the motion trajectory curve [36], visual redundancy [37], visual centroid [38], inter-frame mutual information [39], similarity graph [40] and summarized PSNR [41]. All these metrics are manipulations of pure intensities and in essence measure the visual diversity contained in a summary. Hence the maximization leads to the summary with most content diversity, but not necessarily the one that presents most important clues that enhance viewers' understanding.

Type II characterizes with high level semantic analysis, in which semantic events with explicit meanings are detected and the resultant semantic structure is utilized to guide the summarization. Generally, semantics are defined explicitly by some offline database, which annotates its entries with meaningful tags. Through supervised learning from labeled data, various methods in this category detects events with explicit

meanings. Typical implementations include the recognition of the “emotional dialogue and violent action”

Type III lies in the intermediate level, seeking entities with implicit meanings. The philosophy is that implicit semantical entities also suffice viewers to understand and recover original plot while avoiding the heuristic attempts for explicit semantic recognition. Some researchers in [42–45] assume the semantics are implicitly expressed by popular human perception models [46, 47] and they yield summaries with most salient video units. Unfortunately, salient features do not necessarily mean semantic distinguishable as they basically measure how interesting of a video while the interesting part may be an important clue for understanding or may be not.

In this dissertation, we propose an intermediate level approach to explore the implicit semantics of original video. We pursue a self-explanatory video summary through discovering and preserving “concepts”. Concepts symbolize abstract clues on which viewers base to recover the original plot thread; although on the intermediate level, they are patterns learned from the original video and represent implicit semantic meanings, rather than in an explicit and rigorous manner but with less generalization. The motivation of concept is intuitive: emulating the human cognitive process, naturally a list of key patterned hints such as characters, settings, actions and their orders are needed in the short summary for viewers to stitch these hints logically and use imagination to fill the omitted part. The concepts correspondingly encode the semantic meanings of patterned hints. Specifically, we extract visual features and use spectral clustering to discover “concepts” and consider the repetition of shot segments which instantiate the same concept as summarization redundancy. We further analyze the criteria of a good summary and formulate the summarization as integer programming problem. A ranking based solution is given.

The main contributions of this chapter are therefore:

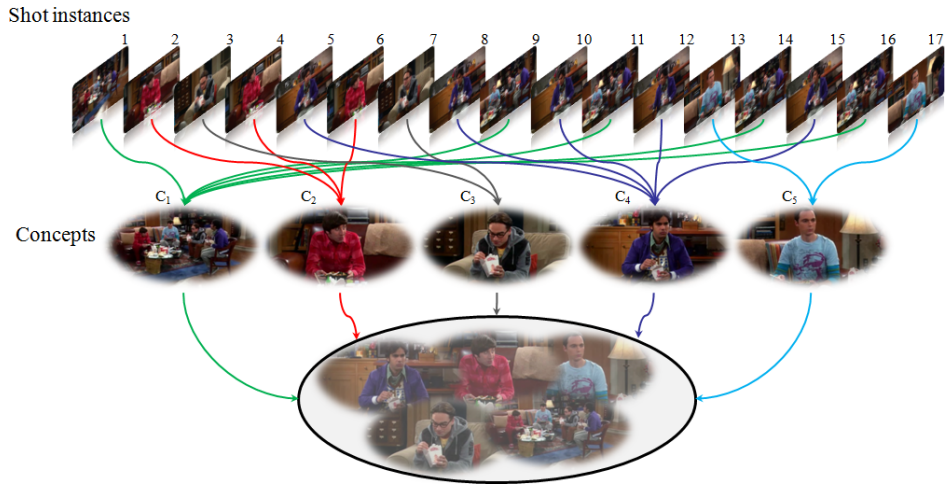

Figure 3-1. The top row: shot instances. The middle row: concepts. The bottom row: an understandable summary by presenting concepts. Photos courtesy of the TV show Big Bang Theory by CBS Broadcasting, Inc [4].

(i) A in-depth analysis on the philosophy of video summarization and answered the key questions: what are the semantically important elements and redundances; and how to keep the semantics in the summary.

(ii) Proposed a clustering based method to discover the semantic structure of the video as concepts and instances.

(iii) Proposed a ranking based method for summary generation, which is a scalable framework that adapts the detail preservation to the summary length.

3.2 The Summarization Philosophy

In order to capture the video semantics, the first question is how the semantics are expressed in the video. Observing many meaningful videos, we find out the major semantics embedded are a plot outline these videos try to express. Generally, the outline can be abstracted in a sequence of “entity combinations” (e.g. in Fig. 3.1, the four persons, as four entities, are combined together to have a conversation). Each “entity combination” forms a particular concept that exhibits certain semantic implications. Here, concept differs from the “cast list” [48] in that the constituent entities are not necessarily “faces” or “actors” but may include any entity that is distinguishing

to characterize the concept. Also, the combination are not a simple list of entities, but emphasize their interaction: through certain postures and actions of the entities, the semantic is vividly implied within the concept.

These concepts shape the course of the plot, thus are the skeleton of the video. Furthermore, each concept is materialized by multiple video shots (we call them instances of a concept). Each shot instance, as the muscle of the video, elaborates the concept from a specific perspective. (e.g. in Fig. 3.1, the concept of having conversation is expressed by shot 1 to shot 17. Some instances depict one person while others depict another, together they convey four persons are having a conversation.) However, since the instances for the portrait the same concept, they are destined to include redundancies in the summarization sense.

Therefore, the concepts can be seen as a parsimonious abstraction of video semantics, which suits our summarization purpose very well: by eliminating unnecessary instances with semantic redundancies but keeping the concepts intact in the summary, we shorten the long video without compressing the semantics. The preservation of a concept may be fulfilled by keeping the instances with more distinguishable characteristic to express the concept in higher priorities. From viewer's perspective, if viewers are informed with the semantic skeleton, they are able to refill a full body of video outline using reasoning: they witness the concepts and those entities within and concatenate them into a plot thread based on experience or even imagination, thus understand the video. For example, In Fig. 3-1, when all concepts in the middle row are available in summary, viewers are capable to realize the four people are having a conversation. Therefore, the video summarization consists of two phases: exploring the inherent concept-instance structure followed by preserving those prioritized instances in the summary.

3.3 Recognize the Concepts

In our context, discovering concepts is an unsupervised learning problem. We first divide a full video into multiple shots based on signal level change detection and regard each shot as a sample. Then we use an unsupervised learning method to cluster the samples into several classes. Each class of samples, collectively, expresses a concept. For feature extraction, the concept defined as “entity combination” implies the samples within the same concept/class share the similar interaction of entities. Thus, a feature vector capturing the constituents of entities of a shot is required, preferably with the robustness to variations of entities due to postures, scale and locations.

For that purpose, we propose the bag-of-words (BoW) model as the feature in this task. BoW model [49] was initially utilized in natural language processing to represent the structure of a document. It regards a document as a collection of certain words from a reference dictionary but ignores their order. In our context, “Bag” is a shot instance regarded as a container and “Words” are the visual entities inside. The BoW model provides the macro look of semantic ingredients within an instance and emphasizes the conceptual similarity among instances, considering entity attendance only. In Fig. 3-2, a far view of rose and butterfly and a close-up look of the same entities should both imply the same semantic implications, despite the rose and butterfly may appear in different locations or scales. Evidently, the BoW feature graciously expresses the order-irrelevant property and measures the two images in similar values.

The next question is how to construct the BoW feature for each shot. For the representation of “words”, ideally it should feature a perfectly extracted entity. However, due to the limitations of current object detection techniques, we alternatively use the SIFT feature points [50] within the detected salient objects [47] to represent an entity/words. Parsing throughout the original video, all occurring “words” constitute a “dictionary”, which describes a space of all entities occurring in the video. In order to distinguish those significant entities, we compress the “words” into “codewords”, which

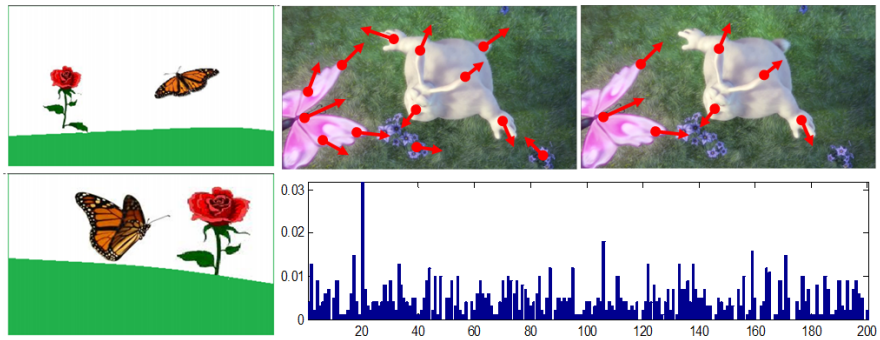

Figure 3-2. Left: two semantic similar frames. Right top: words and codewords. Bottom: a BoW vector. Photos courtesy of the animation Big Buck Bunny by the Blender institute [5] .

convey characteristic entities after dimension reduction. For each sample/shot, we count the number of occurrence of each "codeword" as its final BoW feature vector. Fig. 3-2 shows the original "words" for a frame, the compressed "codewords" and also the bag-of-words feature. Finally, we exploit spectral clustering [51] to cluster shots into several classes, each of which corresponds to a concept.

3.4 The Summarization Methodology

3.4.1 Criteria of a Good Video Summary

The most significant criterion of a good summary is to preserves all original concepts. If there is absence of any concept in the summary, viewers may be misled by fractional clues and fantasize a totally deviated plot.

Also, concepts should be rendered with balance, i.e. in the summary there are equal or comparable occurrences of instances for each concept. As every concept in the summary are semantically compact and decisive, ignoring or overemphasizing anyone due to imbalanced rendering may mislead the viewers to recover a plot deviating from the original.

The third criterion is that each concept is preferably rendered by the most interesting instances. Then for a given length, the summary not only preserve the semantics but also triggers viewers most interests to recover the plot.

3.4.2 Constrained Integer Programming

Given a clear semantic structure, we model video summarization as an integer programming problem. Let c_{ij} be the binary indicator that denotes if we retain an instance in the summary or not, s_{ij} is the jth instance for concept i. We want to find a combination of instances indicated by $\hat{c} = \{\hat{c}_i\}$, which aim to deliver all concept with most interestingness, or equivalently, maximizes the total attention saliency for each concept.

$$\hat{c}_i = \arg_{c_i} \max \sum_{j=1}^{n_j} c_{ij}\xi(s_{ij}) \forall i = 1 \cdots N \tag{3–1}$$

$$s.t. \sum_{i=1}^{N} \sum_{j=1}^{n_j} c_{ij} \leq r \tag{3–2}$$

$$\min\left(\lfloor \frac{r}{N} \rfloor, n_i\right) \leq k_i = \sum_{j=1}^{n_i} c_{ij} \leq \min\left(\lceil \frac{r}{N} \rceil, n_i\right) \tag{3–3}$$

$$\forall i \in 1, \ldots, N$$

$$c \in \{c_{ij} | \arg\max_{c_{ij}} \sum_{i=1}^{N} \sum_{j=1}^{n_i} c_{ij} l_i\} \tag{3–4}$$

where $\xi(s_{ij})$ is the saliency of s_{ij}, n_i is the number of instances for the ith concept in the original video and k_i is the number of instance of concept i in the summary. N is the total number of concepts in the original video. r is the maximum number of instances we can retain in the summarized video. l_i is the importance rank of concept i, which is dependent on the number of frames for this concept pattern, based on the common sense that director may naturally distribute more time to portrait a more important concept.

Constraint I in Eq. 3–2 comes from the summary length limit. In our approach, we ceil the maximum number of frames in a detected a shot below some constant number, thus r is almost proportional to the predefined summarization ratio. Constraint II in Eq. 3–3 is imposed by the concept completeness and balance criterion. Given a summarization ratio, we always try to deliver all concepts using commeasurable number of shot instances.

Constraint III in Eq. 3–4 deals with critical situations when r decreases to a small value so that the summary could not keep all concepts with even one shot instance for each or in the situation that a summary length does not allow all concepts to be delivered with absolutely equal number of instances. Here we give priority to concepts with larger importance to be kept in the summary.

Considering constrain II in Eq. 3–3, it is required the number of instances for each concept should be almost the same, if achievable, no matter how long the summary is. Therefore, we construct the summary in a bottom-up fashion: starting from an empty summary, every time we allow only one instance from one concept to be admitted in the summary and continue this process until the resultant summary length reach its limit. Therefore, no matter when the process terminates, we may guarantee that the numbers of instances for different concepts differs only by one (except the time that all instances are recruited in a long summary).

The scalability of the summary is straightforward: if the summary length is short, we keep a modest number of most distinguishable instances for every concept. If the summary length is longer, we may add more instances for every concept, rendering a more detailed summary. This trait to adapt detail preservation in the summary to variable length suits very well for the needs of ubiquitous multimedia terminals that accepts different levels of summaries.

In the algorithm, which instance will be included in the summary is the key question. An instance is indexed by the rank of concept it belongs to and also its rank within the concept. As aforementioned assumption, we rank the concepts according to its length in the original video. As for the ranking within a concept, we rank the instances of a concept according to its saliency value. Thus, the admitted instances for a concept contain the highest interestingness accumulation for the given summary length, fulfilling the objective function in Eq. 3–1. Here, based on human attention modeling in [47], we define the visual saliency $\xi^v(s_i)$ of a shot s_i is the average saliency over all frames t

within the shot.

$$\xi^{v}(s_i) = \frac{1}{|s_i|}\sum_{t=1}^{|s_i|}\pi(t) \tag{3–5}$$

3.5 Experiment Results

3.5.1 An Example to Illustrate Our Algorithm

In this section, we evaluate the performance of the proposed concept recognition method and also the overall video summarization approach, with two summarization ratio 50% and 20% used. We implement the proposed approach in C code. The evaluation video sequence is the 10-minute Big Buck Bunny, which comes from the Open Movie project without license issue. Fig. 3-3 shows the recognized concepts. Since the original video is too long to visualize, we illustrate the frames that are sampled every 15 seconds from the original video. In Fig. 3-3, each concept consists of the frames in the same border color. It shows that the frames that express the similar semantics are clustered into the same concept indeed. For example, the frames in red box all depict the bunny is playing with the butterfly; the frames in green box all show that the bunny is standing by himself. Also, we visualize three frames with no border color as they are the outliers of clustering. This result can be explained by their negligible occurrences in the video: they are too trivial to express a concept. For the number of concepts (the complexity of clustering model), we enumerate 5, 10, 15, 20 and 25 as the candidates to measure the clustering error and use Occam's razor rule to finalize it as 10. The clustering result indicates our concept recognition method works well.

Fig. 3-4 shows the summary results of 50% and 20% ratio, respectively. Note that the 20% summary is a subset of the 50% summary. This result suggests that the summary can be packaged into multiple layers, with each layer as the summary difference for two adjacent length scales. This fact suits for the scalability in the multimedia system, since a terminal can depend on its own needs or situation to

Figure 3-3. The images with the same color belong to the same concept. Photos courtesy of the animation Big Buck Bunny by the Blender institute [5].

Figure 3-4. Top: the 50% summary. Bottom: the 20% summary. The concepts are complete in each summary. Photos courtesy of the animation Big Buck Bunny by the Blender institute [5].

receive how many layers without changing of the summarization process. Also, both the two summary contain complete concepts and thus preserve semantics effectively. It is natural for the 50% summary to include more details since the summary length allows.

3.5.2 Subjective Evaluation

We also carry out a subjective test of the human response to the generated summaries. We adopt two metrics informativeness and enjoyability proposed in [42] to quantify the quality of the summary under different summarization ratios. Enjoyability reflects user's satisfactory of viewing experience. Informativeness accesses the capability of maintaining content coverage while reducing redundancy.

Table 3-1. The statistics of scores of four video clips. Video courtesy of [4–6].
E: enjoyability. I: informativeness.

		Our approach		Scalable approach [52]	
		Mean	Var.	Mean	Var
LoR	E	64.05	18.21	62.88	19.82
50%	I	66.08	18.71	65.50	18.67
BBT	E	63.28	20.31	61.07	21.79
30%	I	69.82	18.56	66.49	19.60
BBB	E	61.25	21.86	62.30	20.32
20%	I	68.00	16.90	67.50	17.26
CNN	E	55.06	19.59	53.92	20.19
10%	I	60.87	18.58	59.49	17.15

The subjective test is set up as follows. First, in case of viewing tiredness of participants, we carefully picked four test videos: a 7-minute movie clip in "lord of the rings" (LoR) from the MUSCLE movie database (http://poseidon.csd.auth.gr/EN/MUSCLE_moviedb/index.php), a 6-minute TV clip in "the big bang theory" (BBT) [4], a 10-minute animation clip in "the big buck bunny" (BBB) [5] and a 6-minute news clip in "CNN student news" (CNN) [6]. Then, we chose summarization ratios 50%, 30%, 20% and 10% respectively to consider both the ordinary summarization cases and the extreme cases and for LoR to test. We use both our approach and the scalable approach in [52] to generate the summary.

During the test, each participant was asked to give each summary enjoyability and informativeness scores in percentage ranging from $0\% - 100\%$. We collected scores submitted by 60 participants and calculated the means and variances of the scores for the two approaches, as indicated in Table 3-1. From the scores to our approach, it shows that by reducing 50% to 90% of the video content, the enjoyability and informativeness only drops less than 50%, which is an encouraging result. Also, compared with the scalable approach, our scores in all informativeness and most enjoyability cases are slightly better. This results suggests our approach may preserve more semantic information.

CHAPTER 4
PERCEPTUAL QUALITY ASSESSMENT

4.1 Background

Due to the tremendous advance of video compression and network communication technologies, video media in much larger resolutions (1280x720, 1920x1080) gains popularity from content providers and is increasingly supported in the Video-on-Demand, Live broadcasting and video telephony applications [53, 54]. When wired or wireless networks occasionally fail to deliver media packets, the received media may be implicated by the packet loss events. Different from the traditional video media, each of the video packets in large resolution cases may impair a portion of frame instead of dropping the whole frame. However, the impairment may propagate through a sequence of frames followed; the degradation becomes much more noticeable as is in sharp contrast with the adjacent unaffected regions. Thus, in current HD video transmission, the degradation of a packet loss is significantly different from the conventional artifact like frame freeze and jitter [55]; instead, the packet loss event results in a unique "glittering block contamination" when the received content is displayed and the contamination aggravates along with the object motion trace in the video media. Fig. 4-1 shows the glittering artifact propagating a sequence of frames along skater motion traces.

In order to better predict and thus control the degradation resulted from packet loss, an automatic subjective quality measurement (perceptual quality) mechanism dedicated to packet loss related degradations is required. The perceptual quality brings the benefit of quantifying the end-to-end distortion of a video transmission system and guiding it to configure optimistically. Also, the perceptual quality may help content operators to assess the quality of their provided service in terms of human visual satisfaction, which potentially differentiates the various service levels and thus helps to generate revenues more efficiently.

Figure 4-1. The glittering blocks originate from transmitted packet loss in the first frame. As the frames followed depend on the first frame, the packet loss artifacts aggravates along the motion trace, appearing as glittering block contamination. Photos courtesy of the video trace library [3] for YUV video sequences.

Quality metric techniques can be divided into full reference [56], reduced reference [57] and no reference [58] methods according to how much information of the reference video frame available when assessing the received and decoded video. Full reference methods assume that all pixel of the original video are available; reduced reference methods assume no available full pixel information but some sparse parameters/features extracted from the original video are accessible; no reference methods assume that there is no original video information at all. From another taxonomy point of view, quality metric techniques can be categorized into objective, perceptual and subjective quality methods based on the level of distortion responses.

Objective quality measurement techniques such as Peak Signal Noise Ratio (PSNR) and Sum of the Square Error (SSE) are signal level metrics. It is very easy and light to compute them as well as their derivatives. Thus many video transmission systems use them as distortion metrics to optimistically adapt the encoding and decoding to the varying channel conditions. However, it is proved in many literatures that the objective quality measurements do not necessarily correlate with the human

visual response well [59]. On the other hand, subjective quality refers to a numeric that qualifies the true human response to video degradation. Although it is the true value of human response, the tedious requirement for labor and constraints for laboratory-like environment prohibits it from the usage of real-time and large scale quality assessment. In between, the perceptual quality [60–63] is a computational model for quality measure which aims at mimicking the human score response. It enjoys the strength of a close relation to human scores as well as the capability to be conveniently applied into video transmission systems due to its automatic nature.

The general methodology for perceptual quality metric is to extract factors from either the received bit stream or the decoded pictures and combine them using a model to predict the quality value. For the factor extraction, many techniques [7][64] utilize compression parameters such as frame type, motion vector strength and packet loss parameters such as the length of loss burst and the loss location. These methods provide the opportunities to estimate the quality on network level; however, since these factors do not directly shape the degradation, there is always a cognitive gap between them and the human response. To tackle this problem, these methods also define a model such as generalized linear model, support vector machine, decision tree to map the factors to the human response score and train the model from some packet loss patterns vs. the human response score samples. Generally, the model parameters are sorely obtained by training while sometimes the model form is also trained (e.g. decision tree). In fact, the mapping from factors to the human score involves a complicated error propagation process and a psychological response process. Therefore, the highly non-linear nature of the mapping may not be properly matched by the pure data-driven model trained with a limited number of samples generated by a limited number of packet loss patterns. As a result, the model may be sensitive to the variation of packet loss patterns and encoding parameters.

In this dissertation, we propose a perceptual quality model that addresses the human cognitive response to the video packet loss degradation. The model is a non-reference quality metric since it is designed for a video transmission system where no information of the original video is available. Aiming at the first-hand psychologically sensible factors that are related to human grading score, the authors conduct a survey that focuses on discovering concrete the psychological effects when human viewers perceive the corresponding artifacts. Then the proposed model quantifies the two discovered factors, obvious face deformation and glittering block edge strength, using advanced image processing techniques. The combination of the two factors is learned from training samples (video clips with degradation vs. human response scores) with linear regression technique. Since the model exploits direct psychological factors, which are cognitively explainable, the model itself could be represented by a simpler form and the training process is much more at ease, as oppose to the case where indirect transmission and video stream layer factors are used and thus a much more complicated model is expected. Also, our model is independent with most pre-decoding factors; thus when the actual transmission system with different encoding parameters and transmission packet loss events from the ones used for model training, the proposed model is more robust.

4.2 The Construction of Training Sample Database

In order to learn the perceptual quality model to predict human response, video footage vs. human scores samples for training are required. A compelling sample database here should span a variety of video contents, contain a wide distribution of the hypothetical reference circuits (the encoding, transmission and decoding chain) that generate the artifacts and also have humans scores cover fairly over the entire score range.

4.2.1 The Weakness of Publicly Available Database

Currently, the publicly available human score databases include the Video Quality Expert Group (VQEG) HDTV data [65], LIVE Video Quality Database and LIVE Mobile Video Quality Database [66]. The first two databases aim at a variety of video samples with common artifacts such as blurring, noise, compression loss and transmission loss. Since they are databases for all-inclusive perceptual quality models, they do not narrow down to the specific artifacts related to packet loss during transmission and hence not provide enough samples with various hypothetical reference circuits configuration. In contrast, LIVE Mobile database is a dedicated sample set for the artifacts in mobile applications and it contains video samples generated by various encoding and transmission parameters. However, the database seems to focus on the mapping of encoding and transmission factors, rather than the psychological factors, to the subjective quality value. Therefore they do not provide information about the psychological effects of human response to the artifacts, which is instead our proposed methodology.

4.2.2 Video Sample Generation

In this dissertation, the author focuses on the modeling of perceptual quality of the artifacts resulted from packet loss and provide a new sample database that does not only contain the Hypothetical Reference Circuit (HRC) parameters vs. human scores but also the information about how human viewers' psychological response to the artifacts and reasons why they give a certain score level. The author attempts to generate packet loss related artifacts as complete as possible so that the samples in the dataset include a wide variety of encoding and transmission parameters.

In this database, the video footage are generated by varying the encoding and transmission parameters. The original videos contain five sequences crew, ice, harbor, crowdrun, parkjoy from VQEG website [67]. The resolutions are 4cif (704x576, used in DVD) for the first three and 720p (1280x720, used in HDTV) for the last two. Each of the

five sequences has duration around 10s and carries a unique motion and texture form, including high-foreground-low-background motion, low-foreground-high-background motion, human faces, natural scene and human crowd. Since we focus on the artifacts related to packet loss rather than other artifacts such as compression loss, we encode each video sample with a fairly high bit rate 6Mbps. We choose an encoding structure IPPP with Group of Pictures length as 25 to guarantee the artifact is able to propagate and is noticeable to the viewers. Also the coding tools in H.264 AVC standard such as Arbitrary Slice Order and Flexible Macro Block are designed to counteract the loss events: for bit streams encoded using different slice groups (partition of a frame), the same loss events may result in significantly different artifacts. Therefore we apply different slice group parameters to encode the same video, leading to a diverse form of artifacts in packet loss events. As regards to the setup of transmission, we packetize each bit stream with conformation to H.264 Annex B standard (each slice contains in one packet and the delimiter is 0x000001) and control that the packet loss events happen after the 3rd second to prevent viewers from neglecting artifacts prematurely. We also assume the loss pattern is a burst loss as it is a frequent case in current mobile applications. By varying the number of lost packets and also the location of the first packet loss, we generate a collection of 99 corrupted bit streams. The corrupted bit streams are then decoded by JM 16.0 using the error concealment method (motion copy) and we obtained all the videos with artifacts. Table 4-1 summarizes the distribution of generated video samples with their characteristics and the HRC used.

4.2.3 Score Collection and Survey Conduction

In the score collection phase, we adopt the recommendation of VQEG HDTV-1 test plan [68]. We select 26 viewers to rate 99 video samples with packet-loss related artifacts. The viewer candidates consist of the following categories, based on their background related to video processing expertise, video coding scientists, Electrical and Computer Engineering graduate students, general engineering students, photography

Table 4-1. The distribution and properties of the generated video clips with packet loss artifacts. I-interlace, C-chessboard, R-Region of Interest. Courtesy of the video trace library [3].

Video source	Sample size	Video size	Motion property	Slice partition	Packet loss length	Error concealment
Crew	27	CIF	High fg low bg	I, C, R	3, 5, 7	Motion copy
Ice	18	CIF	High fg low bg	I, R	3, 5, 7	Motion copy
Harbor	18	CIF	Low fg low bg	I, R	3, 5, 7	Motion copy
Crowdrun	18	720p	High bg high fg	I, C	3, 5, 7	Motion copy
Parkjoy	18	720p	High bg low fg	I, C	3, 5, 7	Motion copy

hobbyists, animation artists and amateurs. The viewers have original/corrected vision acuity of at least 20/20. Each participant is presented with many video samples as yuv documents, displayed by YUV Viewer running in DELL desktops with 1280x720 resolution. The distance of the students facing toward the computer screen lies in between 40 - 60 inches. Considering that the video samples with the same content may easily make viewers feel fatigue, we divide the 99 ratings into four sessions, with each session less than 30 minutes. Viewers are allowed to watch the same sample back and forth until they reach a confident rating. In order to counteract the users' adjustment to certain loss patterns and thus the possible unnecessary forgiveness to later video samples, we randomize the order of the video sample for each viewer. In the rating process, the five original video is also presented without noticing the viewers. The viewers are asked to give one score using Absolute Category Rating Scale (ACRS) ranging from 1 to 5, according to the psychological satisfaction on the perceived artifacts. Also, in order to discover the psychological factors, the viewers are asked to choose at least 20 video samples to answer the questions in Table 4-2. The ratings of the 20 chosen video samples are preferred to span equally over the score range since we aim at psychological factors explainable for all scores. After the score collection, we subtract each rating score from the score of its original video to calculate the Difference Score. Then for each video sample, we calculate the mean value of the Difference Score over all participants as the final Difference Mean Opinion Score (DMOS).

Table 4-2. The questions for the video samples chosen by each viewer

ID	Questions
1	Do you think you have a good reason to give such a score?
2	Why do you think it is such a score?
3	What did you observe?

4.2.4 Analysis of the Survey Result

After manually parsing the answers attached to the chosen video samples, we summarize that 97% of participants explain their ratings using reasons that can be attributed to the following two aspects. They are “how human face is distorted” and “the strength of glittering blocks”. Therefore, we conclude that the two aspects are the dominant psychological factors of the human response on the video packet loss related artifacts. In Sec. 4.3 and Sec. 4.4 , we will elaborate the image processing techniques we use to describe the two factors.

4.3 Glittering Artifact Detection

This section addresses the measurement of the first dominant psychological factor “the strength of glittering blocks”. Fig. 4-2 is a typical video frame that is impaired by packet loss related artifacts. The areas indicated by the red arrows show the exemplar appearance of the “glittering blocks”. As shown in Fig. 4-2, the areas have “crystal” shapes and edges and also demonstrate significant contrast with the neighboring unaffected regions. Also, the colors of the areas appear to be artificial, imposing a distinguishing discontinuity over the “natural” unaffected regions. Therefore, we propose an edge-map detection method to capture the discontinuity of the areas.

Note that the reason why the “glittering blocks” have such “regular” appearance lies in the block-based encoding in compression standards such as H.264 as well as the fact that a lost packet induces integer number of blocks lost on the decoded frame. When a packet lost during transmission, the corresponding slice data containing multiple 16x16 Macroblocks (MB) is lost; therefore the boundary lines between the artifacts and the unaffected regions follow the 16x16 lattices of the frame. The decoder tries to handle the

Figure 4-2. An exemplar frame in video *crew* degraded with glittering blocks. Photos courtesy of the video trace library [3] for YUV video sequences.

lost MBs gracefully using error concealment techniques, including copying the content from the referred MBs of the lost MB along the motion vectors. However, due to the erroneous motion estimation, encoding mode and the MB residue, the referred MBs are not necessarily consistent with the lost MB adjacencies. Therefore, it is still possible for humans to observe the artifacts.

4.3.1 Edge Map Generation

Based on the analysis, we first apply a high pass filter over the 16x16 lattice of the decoded frame to emphasize the possible locations that artifacts can occur. The high pass filter can be implemented based on Eq. 4–1, and the edge map $E(x, y, i)$ is calculated by applying the filter onto r, g, b color plane of the original frame $I(x, y, i)$ respectively.

$$E(x,y,i) = \begin{cases} 0 & \text{for} \quad \lfloor \frac{x}{16} \rfloor, \lfloor \frac{y}{16} \rfloor \neq 0 \\ \frac{1}{2} \sum_{v=y-1}^{y+1} \frac{1}{2}^{|v-y|} \cdot |\triangle_x^1 I(x,v,i)| \cdot (1+e^{-\alpha}) & \text{for} \quad \lfloor \frac{x}{16} \rfloor = 0, \lfloor \frac{y}{16} \rfloor \neq 0 \\ \frac{1}{2} \sum_{u=x-1}^{x+1} \frac{1}{2}^{|u-x|} \cdot |\triangle_y^1 I(u,y,i)| \cdot (1+e^{-\beta}) & \text{for} \quad \lfloor \frac{x}{16} \rfloor \neq 0, \lfloor \frac{y}{16} \rfloor = 0 \end{cases} \tag{4–1}$$

where $\triangle_x I(x,v,i) = I(x,v,i) - I(x-1,v,i)$ and $\triangle_y I(u,y,i) = I(u,y,i) - I(u,y-1,i)$. $\alpha = \min |\triangle_x I(x-1,v,i)|, |\triangle_x I(x+1,v,i)|$ and $\beta = \min |\triangle_y I(u,y-1,i)|, |\triangle_y I(u,y+1,i)|$. (x,y) and (u,v) are locations of an edge pixel and i is the index of the color plane. The final edge value for a pixel is the maximum edge value of the three color plane.

$$E(x,y) = \max E(x,y,r), E(x,y,g), E(x,y,b) \tag{4–2}$$

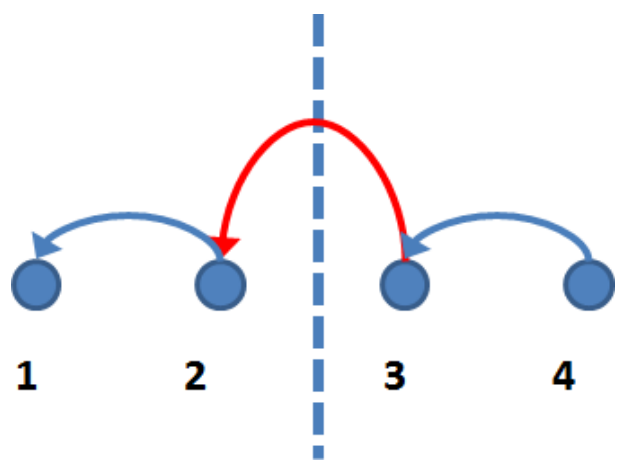

Figure 4-3. Artifact along MB boundary with two pixels in each side.

Fig. 4-3 shows the rationale of the Eq. 4–1. The vertical dash line is a vertical boundary in the 16x16 lattice. Pixel 3 is right on the lattice so pixel 1 and 2 belongs to one MB while pixel 3 and pixel 4 belong to another. On a typical artifact that occurs across the two MBs, the pixel value would discontinue from pixel 2 to pixel 3, indicated by the red curve, whereas the pixels that belong to the same MB tend to have similar pixel value, indicated by the blue curves. Thus we use the absolute difference value of

the two pixels as the block edge strength indicator. The indicator is further amplified if at least one pair of pixels on the same MB demonstrates similar pixel value. Also, to make the indicator more robust to noise, we average the edge along the MB boundary direction. In this edge location selection step, we aim at selecting all possible edge locations; so we use the maximum value of the three color planes as the final edge value. Note that the artifacts can only happens at the MBs that are lost or affected (their referred MBs are lost), we mask the edge detection on the affected regions. Fig. 4-4 is the Edge location selection result of Fig. 4-3.

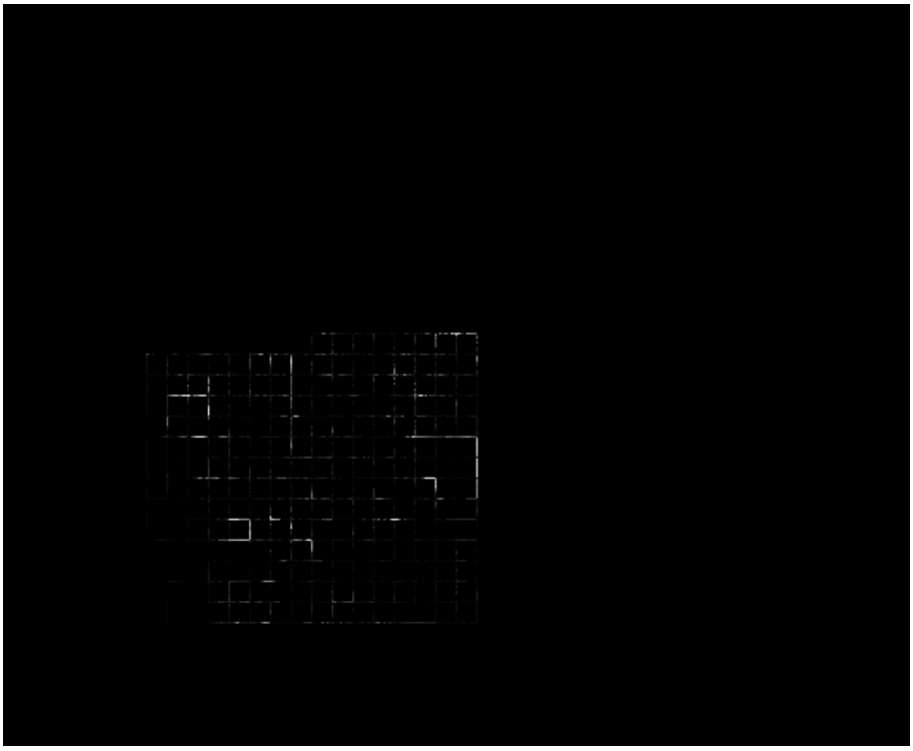

Figure 4-4. The result of edge location selection.

4.3.2 Structure Detection

Note that the output of the edge map is pixel-wise edge pixel value while the artifacts are attended in a block-wise fashion by the viewers. Fig. 4-5 is a close-up view of the artifact region in Fig. 4-2 and its detected edge value from the last step. The edge strength of artifact against its adjacency, highlighted by the red circle, varies along

the MB boundary, which is best reflected by variation of the edge strength in the right edge-map. However, humans tend to psychologically regard that the artifacts occurring along a MB boundary as atomic, instead of treating each edge pixel individually. Therefore, structuralizing the edge pixel value into structure-wise representation may improve the measure the "glittering block" factor. We first binarize the edge map with a threshold θ. Then for each boundary (horizontal or vertical) of the 16x16 lattice, if the number of edge pixels whose value is above θ is more than 4, we mark the all half of the 16 boundary pixels as edge pixels; otherwise remove them from edge pixels. In this manner, the detected edge map is 8-pixel-structure-wise.

Figure 4-5. Block-wise perception vs. pixel-wise detection. Photos courtesy of the video trace library [3] for YUV video sequences.

4.3.3 False Intersection Suppression

When we choose θ conservatively in case of missing edge locations, we unnecessarily include some false edge structures. The most possible case is edge intersections in Fig. 4-4. Normally, a "crystal" shape "glittering block" appears to be convex so only the contour is desired in edge map instead of the internal intersections due to block-based edge detection. Although the scenario where edge intersection is indeed desired can occur, its possibility is very low. Thus we suppress the edge intersection as a less valid edge structure. Fig. 4-6 is a representation of the scenarios of the artifact distribution and their corresponding edge structure to be detected. Each square denotes an MB and

is coded using a color. The artifacts happen on boundaries of MBs in different colors. When considering four squares together, the first two scenarios happen most often and the third happens at times but rarely the fourth. The possible reason for the rare case is that only when the area is original filled with highly complicated texture at the same time that the error concealment method refers to very different regions to paint the missing regions that make all the four MBs considerably different.

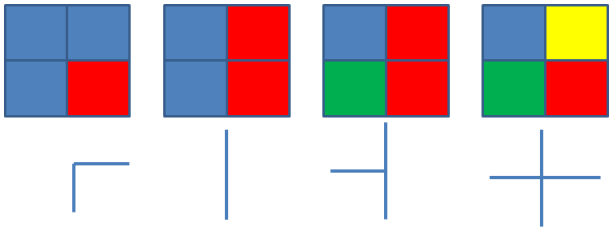

Figure 4-6. Artifact layout scenarios with their desired edge structures to detect.

Thus, for an intersection detected by the edge location selection step, we suppress the intersection to reduce into the second scenario in Fig. 4-6. We measure the edge strength in the horizontal direction and the vertical direction as the number of pixels above θ in Sec. 4.3.1. If the strength of one direction is more than the other by one and a half times, we remove the weaker direction from the edge map. Fig. 4-7 shows that edge map after structure detection and false Intersection Suppression of Fig. 4-2. We use the number of the detected edges in the edge map as the metric of the "glittering block" factor.

4.4 Face Deformation Estimation

The other psychological factor that dominates the human response to the packet loss related artifacts is "how human faces are deformed in the decoded frame". It is suggested in the experiments in Sec. 4.2.4 that human viewers show little tolerance against the deformation of human face regions, as opposed to the forgiveness to other non-face regions. This situation is best explained by Fig. 4-8. The two images are two decoded frames with artifacts; they correspond to the same original frame but go through different hypothetical reference circuits. The left image is resulted from a

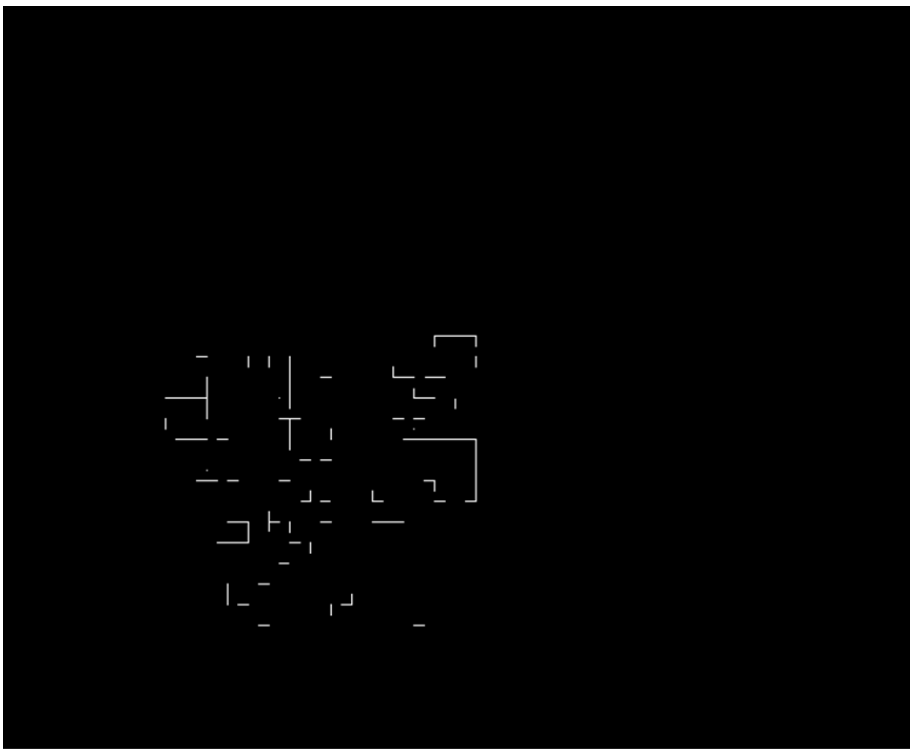

Figure 4-7. The result of structural detection followed by false intersection suppression.

HRC with 3 lost packets and the right image has a HRC with 5 lost packets. For the appearance of the artifacts, the contrast of the affected region against its adjacency, which is highlighted by the red circle in each of the two images, is much sharper in the right than the left. However, the MOS of the right image is 3.6 while the one of the left image is 2.1. The survey participants explain their ratings for the left image are based on “the human face deformation can be acutely perceived”. Therefore, face region deformation should be treated independently from other “glittering blocks”.

The mechanism of a deformed human face region is that the MBs or partitions within the region are copied from external frames with different displacements/motion vectors. While the decoder makes the best effort to choose a suitable referred region for each MB/partition, the copied contents do not necessarily match, especially when the encoding modes of the face region are complex, such as inter 4x4, intra 4x4. Fig. 4-9 shows a deformed face and its two sub regions with different motion vectors to the

Figure 4-8. Visually perceived annoyingness vs. the contrast of the artifact against adjacency. Photos courtesy of the video trace library [3] for YUV video sequences.

referred frame region for copying. As a result, the lower sub region has a larger motion vector than the upper region, overriding the middle area of the face region.

Figure 4-9. A deformed face region and its two sub regions lost and copied from inconsistent referred areas. Photos courtesy of the video trace library [3] for YUV video sequences.

We propose a motion-copy vector disorder metric to estimate possible face deformation. First, we use Viola-Jones face detector [69] to locate a human face region. Then we calculate the deformation value according to Eq. 4–3,

$$D(r) = -\sum_{sr \in r} p(mv(sr)) \cdot \log p(mv(sr)) \tag{4–3}$$

where $D(r)$ is the estimated value of face deformation. r is the detected face region, consisting of sub regions indexed by sr. $mv(sr)$ is the displacement/motion vector of the sub region sr from its referred region copied. $p(mv(sr))$ is the distribution of the motion vectors within the region r. We quantize motion vectors into discrete intervals and approximate the distribution of mv using the histogram of the mv intervals. The entropy like estimator in Eq. 4–3 measures how the motion vectors are different from each other, thus denotes the likelihood of the face deformation.

4.5 Model Training to Map Psychological Factors to Perceptive Scores

Since the two factors we extract are closely related to the human response score, we use a simple linear model to link them to the MOS. The parameters of the model are obtained by Linear Regression method.

$$y = X\beta + \epsilon \tag{4–4}$$

As shown in Eq. 4–4, linear regression assumes that the response y is the inter product of the factor vector X and coefficient vector β plus a noise term ϵ. The regression technique tries to solve β by minimizing the error between the predicted response and its actual value.

$$\hat{\beta} = \arg\min_{\beta} \sum_{(x_i, y_i) \in trainingset} \| y_i - X_i\beta \| + \lambda \sum_{(x_i, y_i) \in testingset} \| y_i - X_i\beta \| \tag{4–5}$$

Eq. 4–5 is the optimization used in the regression technique with cross validation. The first term on the right side is the fitting error of the training set and the second term

is the prediction error of the testing set. λ is the trade-off between the two terms. In cross validation, we split the samples of the database into two non-overlapping subsets, training set and testing set, and jointly minimize the error to avoid model overfitting. In fact, we adopt 10-folded cross validation: the samples are randomly split according to the ratio 9:1 (training set vs. testing set) and this process is repeated 10 times until getting the final value of β.

4.6 Experiments

We use 27 video samples as the testing test. The video samples with packet loss artifact are generated from 5 video sequence crew, ice, harbor, crowdrun and parkjoy, using the same HRCs as in Sec. 4.2.2. Then the human score collection process also follows the routine in Sec. 4.2.3.

After obtaining the glittering block edge and human face deformation estimator for the image in each video sample, we use software Weka 3.6.8 [70] to perform the linear regression. The trained model to predict human score is in Eq. 4–6,

$$score = -0.0193 \times E - 0.4982 \times D + 3.7578 \tag{4–6}$$

where E is the measure of "glittering blocks" in Sec. 4.3 and D is the measure of "human face deformation" in Sec. 4.4.

We also compare with the bitstream/QoS factor based the perceptual model [7]. The available bitstream/QoS factor packet loss length, initial packet loss location and bitstream slice group and decision tree model are used to predict the perceptual quality scores.

Fig. 4-10 shows the scatter plot of the predicted quality scores vs. actual human scores by the method in [7] and our proposed method. As shown in the figure, the score predicted by the proposed method is more close to the actual score with less variance.

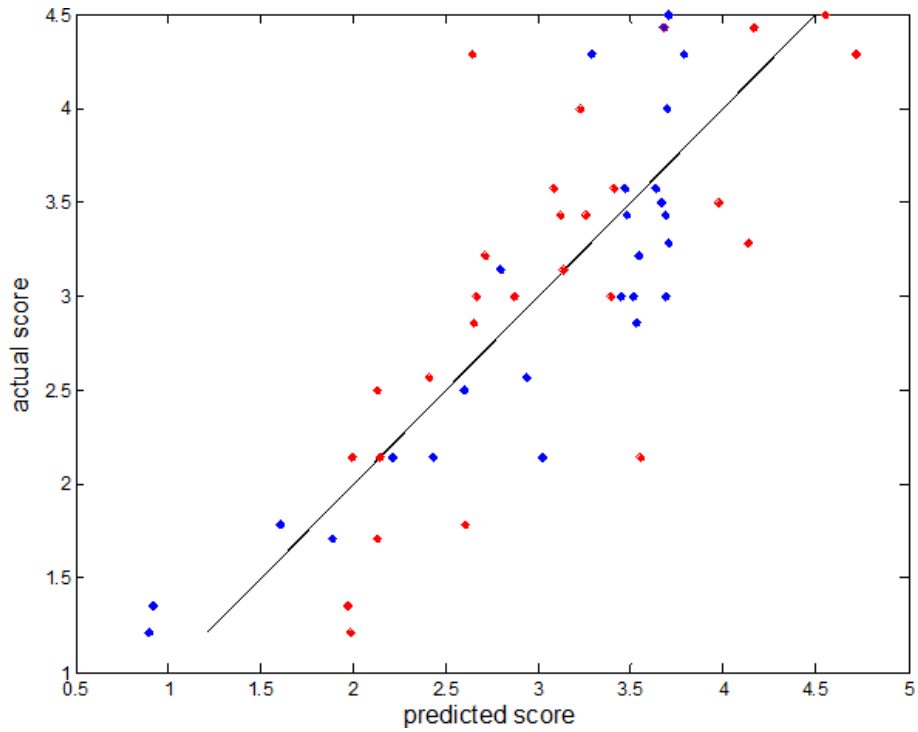

Figure 4-10. Red: the QoS/bitstream factor based approach. Blue: the proposed approach.

We also use Pearson's coefficient 4–7, Spearman Rank Order Correlation 4–8 and the Relative Root Mean Square Error 4–9 to compare the two methods numerically. Table 4-3 gives the results.

Table 4-3. The comparison of the perceptual quality model in [7] and our proposed perceptual model

Model metric	Pearson correlation r	Spearman rank correlation ρ	Relative RMSE
QoS/bitstream[7] factor based	0.7628	0.7768	12.18%
Our proposed	0.8559	0.8104	8.13%

The Pearson coefficient is defined as,

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4–7}$$

where X_i, Y_i are predicted score and actual scores and $\bar{X}$, $\bar{Y}$ are their means.

The Spearman Rank Correlation is defined as,

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{4–8}$$

where x_i, y_i are the orders of X_i, Y_i in Eq. 4–7.

The Relative Root Mean Square Error (RRMSE) is defined as in Eq. 4–9,

$$RRMSE = \frac{\sqrt{\sum_{(X_i, Y_i) \in testingset} (Y_i - X_i)^2}}{score\_dynamic\_range} \tag{4–9}$$

where X_i, Y_i are predicted score and actual score.

Thus, based on Pearson and Spearman coefficients, our model is more correlated with the actual human score than the bitstream/QoS factor based method. Based on RRMSE, our proposed perceptual model increases the prediction accuracy by 4%.

CHAPTER 5
PERCEPTUAL QUALITY BASED VIDEO ENCODER RATE DISTORTION OPTIMIZATION

5.1 Background

In the past 20 years of video compression standard evolution, each generation, from H.261 in the early 90s to the current HEVC, aims at creating encoding tools that can achieve the best rate and distortion trade-off, suited for the computational capacity at that time. The question to find the best trade-off between rate and distortion is defined as rate-distortion optimization (RDO). The core aspects in RDO are finding the best encoding mode and choosing the quantization parameter (QP) for each block.

The encoding modes include the prediction modes and also the partition modes, which essentially define how to predict the current block from its spatially or temporally neighboring blocks. The quantization parameter, critical in encoding the residue of a block subtracted by its prediction, indexes the step size used to quantize the residue into discrete signal, which can be represented by limited number of bits. Ideally mode decision and choosing QP should be jointly optimized, however, [71] reported that choosing the best mode depends on a known QP since representing mode-related information needs bit budgets, where QP is used to predict bit consumption. Also, choosing a QP also requires the best mode be settled since the best mode produces the least block residual for QP to exploit most. Thus, it is a chicken-egg problem. To make the encoding implementable, in current encoder configuration, a QP is given before finding the best mode [72][73][74]. Then based on the best mode, a QP can be reevaluated and possibly adjusted for a second pass encoding [75][76]. Therefore, the RDO can be simplified into such mathematical form,

$$\begin{aligned} \hat{mode} &= \arg\min_{mode} D(mode|QP) \\ s.t. R(mode|QP) &\leq R_c \end{aligned} \tag{5–1}$$

where $D(mode|QP)$ is the distortion of the encoded block from its uncompressed version, $R(mode|QP)$ is the bit rate consumed for encoding the block. R_c is total bitrate budget for this block. This constrained optimization problem can be reformatted into an unconstrained optimization problem using Lagrange method [77],

$$\min_{mode} J = D(mode|QP) + \lambda R(mode|QP) \tag{5–2}$$

where $\bigtriangledown D(mode|QP) = -\lambda \bigtriangledown R(mode|QP), \lambda > 0$.

$$\lambda = -\frac{\partial D(QP)}{\partial R(QP)} \tag{5–3}$$

When λ is available, the encoder loops over all the modes [78] and calculate the corresponding objective value in Eq. 5–2 and use the mode that results in the least objective function value. Note that to accelerate the optimization, many solutions are proposed, including mode early termination [79], mode scope reduction by machine learning [80].

Thus, the core question in RDO is to choose a suitable λ for Eq. 5–2, which requires the knowledge of the analytical form of D in terms of R. Throughout the encoder generations, Mean Square Error (MSE) is conventionally used as the distortion measure for its simplicity. In literature, many R-D models [81][71] are proposed to model the relations of distortion, as MSE, and rate. Based on the R-D models, the negative derivative with respect of R is assigned to λ as in Eq. 5–3. Specifically, in the H.264/AVC coding configuration, λ is suggested as $0.85 \times 2^{\frac{QP-12}{3}}$.

Meanwhile, it is reported that MSE does not necessarily correlate with human visual characteristic (HVC) very well [82]. Thus, many perceptual quality based distortion metric are proposed [83][84][85], each of which claims certain HVC are captured and incorporated into the proposed metric. Therefore, if an encoder employs perceptual quality metric instead of MSE into its RDO framework, it is estimated that a better RD trade-off can be achieved. The intuition is that perceptual quality metric distinguishes the

certain distortion aspects that human visual system is most sensitive of, thus RDO can arrange bit budgets more wisely to accommodate the aspects, while signal level metric such as MSE does not consider. In this chapter, we use the most popular perceptual metric SSIM [85] in our RDO framework,

$$SSIM(x, y) = l(x, y) \times c(x, y) \times s(x, y) \tag{5–4}$$

When perceptual quality distortion metric is used in RDO, it is intuitive to replace the D in terms of MSE with the perceptual quality metric in Eq. 5–4. However, the challenge is that the Lagrange multiplier λ should be changed accordingly. In literature, two types of methodology are proposed to predict the perceptual RDO λ. The first one is experiment-based approach [86][87]. The authors in [86] first perform many encoding processes with MSE and perceptual quality based RDO, respectively. Based on the visualization of the RD samples resulted from the encoding, they assume that the RD curve by the perceptual quality based RDO parallels with that by the MSE based RDO and adopt λ predicted by the MSE RDO into the perceptual quality based RDO. However, the assumption is not usually the case, as shown in Section 5.3 . The other type of approach is information-theory based approach [88][89]. The authors represent the perceptual quality (SSIM) in terms of QP by the physics of the perceptual quality metric, together with the rate in terms of QP and eventually derive the relation of D(SSIM) with rate. However, due to many perceptual quality metric are in non-parametric fashion [7][90], it is unrealistic to relate them with QP in an analytical fashion. Also, in the derivation, many assumptions based on large number of samples are used and thus the resultant asymptotical conclusion (the derived RD model) may have a gap with a concrete encoding process.

5.2 Perceptual RDO framework by Piecewise Linear Approximation

It is desirable to have a comprehensive method to model the perceptual quality based distortion with the rate. Preferably this method is not to be limited to a certain

perceptual quality metric. The Lagrange multiplier λ then can be derived as the negative derivative of the RD model. In this dissertation, a modeling method for perceptual quality metric with respect to rate is proposed. Also, the author takes SSIM as an example, to perform the modeling, where approximation is used for easy implementation.

This dissertation proposes a framework for perceptual quality metric based RDO in video encoding. The proposed method, as a type of experiment based approach, produces a RD model associated with a perceptual quality distortion metric. The RD model is based on the best achievable RD trade-off and thus can give the best RD performance. The authors first collect RD samples by running RDO with a set of Lagrange multiplier λ, where λ is enumerated in a way that its best value (to be found) is included. Then based on the visualization of RD samples on the RD plane, the envelope curve that encloses all the RD samples is recognized as the best achievable RD curve for the perceptual quality metric. Finally the envelope/desired RD model is fitted using piecewise lines. The framework includes five modules, RD sampling, local RD curve fitting and piecewise envelope generation, as shown in Fig. 5-1.

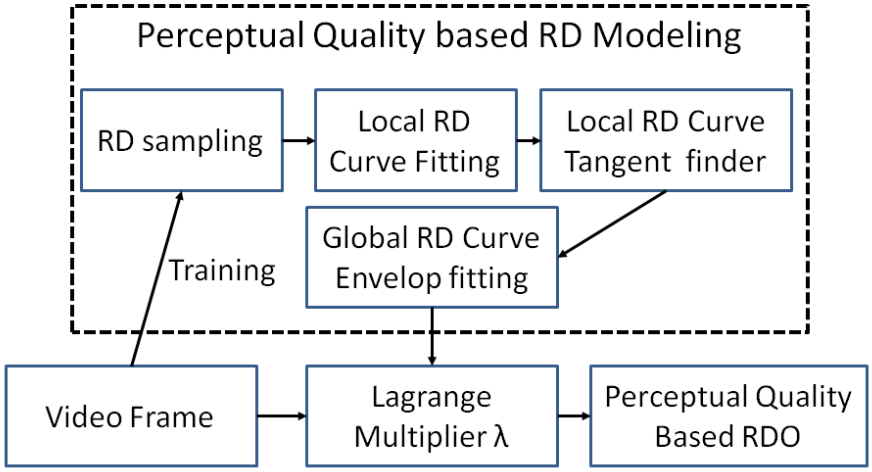

Figure 5-1. The block diagram of the perceptual quality based RDO system.

In the RDO framework, a video frame is first input into the perceptual RD modeling unit for training purpose. After the RD model is learned, the best λ is derived as the

negative derivative of D with respect to R. Then the RDO processes for the following frames are able to be initiated. Since the video characteristics tend to have high correlations over a period of time, the following frames can use the learned RD model to choose their own desired Lagrange multiplier, without additional computational load.

For the perceptual quality metric SSIM, this dissertation gives an exemplar RD modeling solution. In the RD sampling section, the MSE based RDO framework is utilized. Its λ is scaled to $\lambda\prime$, the value that matches SSIM dynamic range and $\lambda\prime$ is used as the Lagrange multiplier in the SSIM based RDO framework. Since the two RDO give comparable encoding results and both the results are visually good, the best λ_{ssim} is presumably in the neighborhood of $\lambda\prime$. Also, as shown in Eq. 5–2, running the RDO process for an encoding unit depends on a given QP. Therefore, RD sample points for the same QP but different λ_{ssim} form a local RD curve, representing the local RD behavior under the particular QP. For the family of local RD curves over different QP, they span an envelope that closely enclose them, which is apparently the RD bound of the perceptual quality metric. In the local RD curve fitting module, each local RD curve is fitted using a quadratic curve (circle for SSIM) with least square regression. Then for every two neighboring RD curves, a common tangent line is derived in the Piecewise Envelope Generation module to capture the gradient of the RD envelope at that location. Note that all the tangent line segments form a piece-wise approximation of the RD envelope.

5.3 RD Sampling

Notice that the conventional MSE based RDO is able to produce visually pleasant compression (blue line in Fig. 5-2). Thus its RDO framework can be used as a starting point for finding the best λ_{ssim} for the perceptual quality based RDO. We propose a λ rescaling method to utilize the MSE based RDO. For SSIM based RDO, we rescale the λ of the MSE RDO into a value that matches the dynamic range of D_{SSIM}. Eq. 5–5 shows the two RDO frameworks. After encoding blocks using the existing MSE-RDO, we

get the statistics of the average MSE metric and SSIM-distortion metric of a block. Their ratio is applied to scale the SSIM-RDO Lagrange multiplier.

$$\begin{aligned} &\min_{mode} D_{MSE}(mode|QP) + \lambda_{MSE} R(mode|QP) \\ &\min_{mode} D_{SSIM}(mode|QP) + \lambda_{SSIM} R(mode|QP) \end{aligned} \tag{5–5}$$

where $\lambda_{ssim} = \frac{\bar{D}_{ssim}}{\bar{D}_{mse}} \times \lambda_{mse}$

In this configuration, each mode produces similar rate distortion trade-off for two RDO frameworks and thus similar modes should be chosen for both RDO frameworks. Therefore, the λ rescaling method has comparable RD performance with MSE based RDO and also produces visually pleasant compression (purple line in Fig. 5-2). This fact suggests that the best λ_{ssim} for SSIM based RDO should be in the neighborhood of the rescaled λ_{mse}.

In order to include the best λ_{ssim}, we vary the λ_{ssim} in Eq. 5–5 with offset interval from -30% - 200% and perform the perceptual quality based RDO. For each QP, multiple RD sample points are generated [91], each one (black sample points in Fig. 5-2) corresponds to offsetting λ_{ssim}. They compose a local RD curve that describes the RD behavior of a given QP. Also, over different QP, a family of local RD curves is generated to reflect the global RD behavior. Its envelope (red line in Fig. 5-2) on the lower left side describes the best achievable RD behavior since each RD points in the interior region is at least worse than two RD points (its horizontal and vertical projections) on the boundary. As shown in Fig. 5-3, an interior RD point R_s has the same distortion with R_l on the envelope, but needs larger bitrates and also has the same birate with R_b on the envelope, but results in larger distortion. Therefore, the envelope corresponds to the desired RD model. As shown in Fig. 5-3, the best RD curve based on perceptual quality metric RDO does not necessarily parallel with that of the MSE based RDO [86].

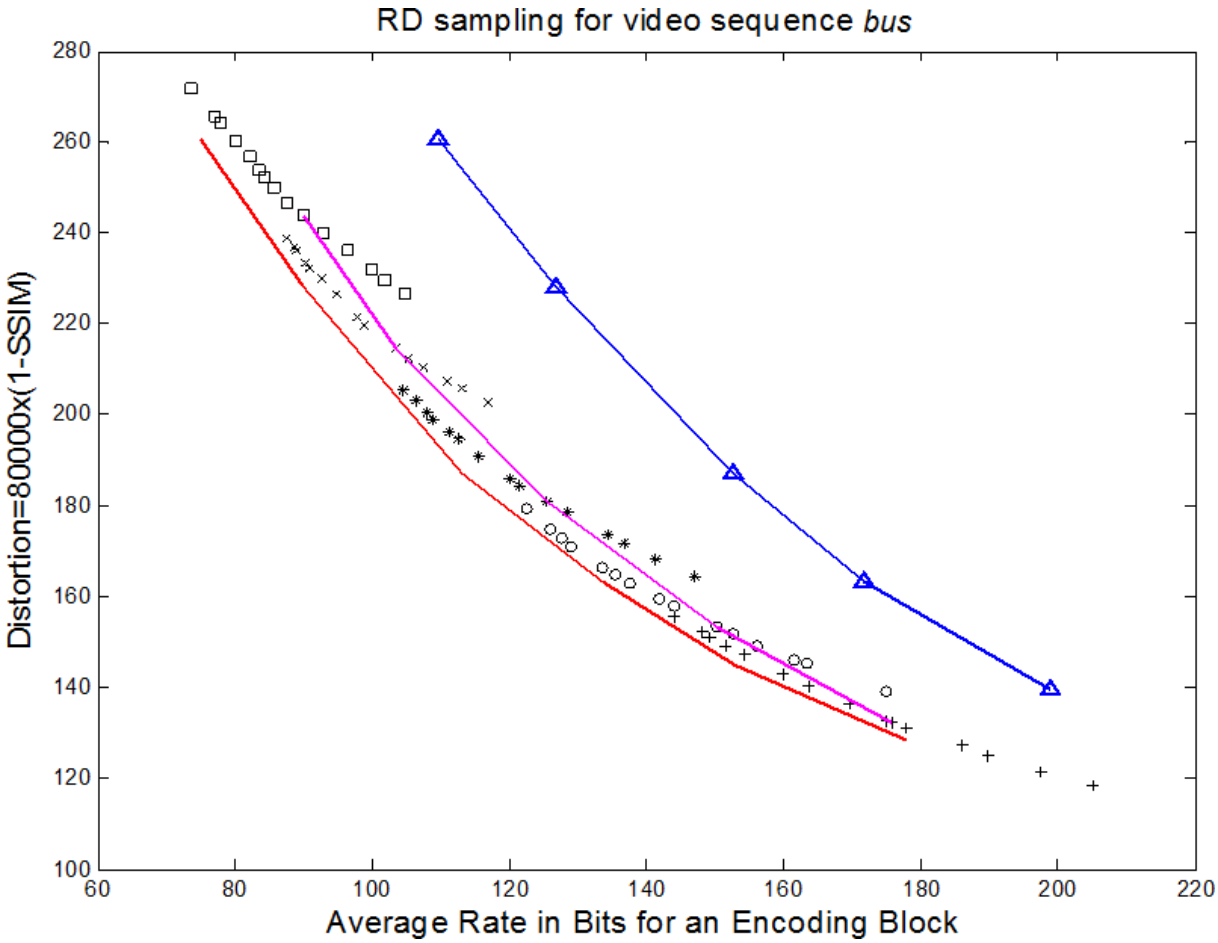

Figure 5-2. Black points in the same marker: RD samples for a given QP but varying λ. Black points in different marker: RD samples for different QP. Blue line: RD curve resulted from MSE based RDO. Purple line: RD curve resulted from perceptual quality based RDO with scaling λ associated with MSE-RDO. Red line: RD curve that is the bound of best achievable perceptual quality based RD trade-off, enclosing all RD points.

5.4 Local RD Curve Fitting

Since the desired RD envelope is spanned by the family of local curves, we need to obtain the analytic form for each local RD curve before fitting the envelope out.

Based on the characteristics of RDO, we can infer some geometric properties about the local curve, as shown in Fig. 5-4,

1) It is monotonously decreasing. When choosing a larger λ_{ssim}, the RDO in Eq. 5–2 tolerates distortion but penalize bitrate, resulting a RD point with less bitrate but larger distortion. So if $R_1 > R_2$, $D_1 < D_2$.

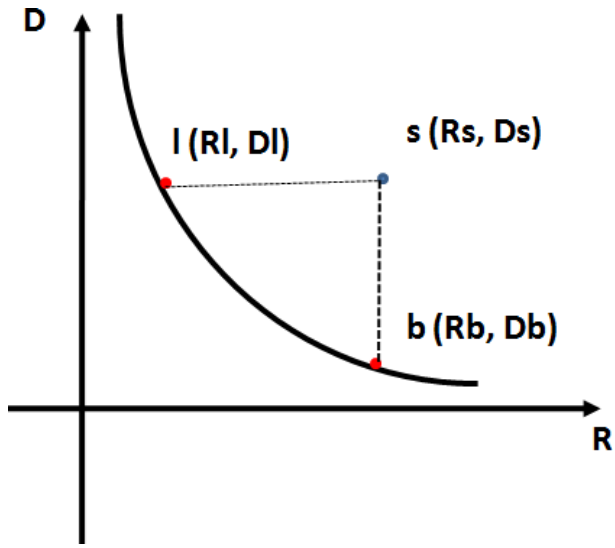

Figure 5-3. The RD performance comparison of an interior RD point with two RD points on the RD envelope.

2) The local curve starts from its lower right end and goes to its upper left end as λ_{ssim} starts from 0 and increases.

3) The curve is convex. When the sampled $\lambda_{ssim} = 0$, the RDO only considers distortion. When R reaches a value, the distortion reaches its bound (quantization always loses some information and there are distortion always) and cannot decrease. So the lower right end of the local RD curve approximates a horizontal line. When the sampled λ_{ssim} is large enough, the RDO framework only consider bitrates. When distortion increases and reaches a value, the rates consumed cannot be reduced anymore since there will be bits used to represent the video signal. Therefore, the upper left end of the local curve approximates a vertical line. To connect the upper-left end to the lower-right end using a smooth line, a convex curve is required.

Fig. 5-5 shows the RD samples by running multiple SSIM-based RDO on video sequence Bus and Mobile. The sample points with the same marker belong to the same QP but varying λ_{ssim}. The samples points with different markers correspond to different QP. Based on the visualization of the RD samples, we use a quadratic model (circle is the simplest) to fit the local curve (RD samples with the same marker). The functional

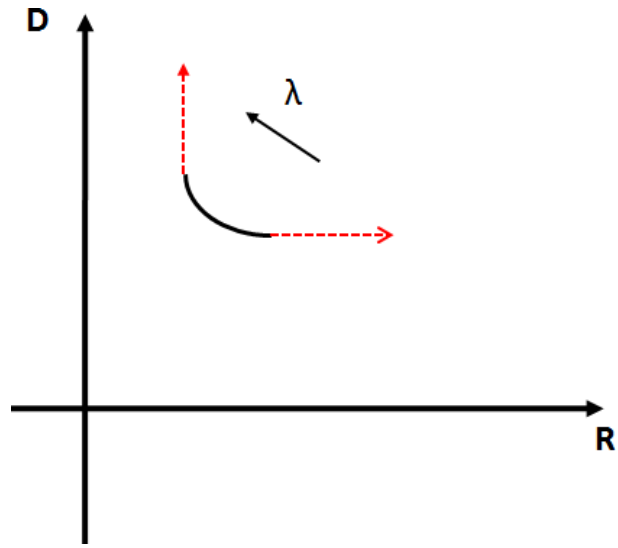

Figure 5-4. The geometric characteristics of a local RD curve.

form of a circle is as in Eq. 5–6, where (R, D) are available RD samples, c, d and e are coefficients of the quadratic curve to fit. We perform least square regression to solve the coefficients as in Eq. 5–7.

$$f_c(R, D|c, d, e) = R^2 + D^2 + c \times R + d \times D + e \tag{5–6}$$

$$\begin{pmatrix} \cdots & \cdots & \cdots \\ R_i & D_i & 1 \\ \cdots & \cdots & \cdots \end{pmatrix} \begin{pmatrix} c \\ d \\ e \end{pmatrix} = \begin{pmatrix} \cdots \\ -R_i^2 - D_i^2 \\ \cdots \end{pmatrix} \tag{5–7}$$

$$\Longleftrightarrow \mathbf{Ax} = \mathbf{b} \Longleftrightarrow \mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

Fig. 5-6 shows the performance of the circle curve (the blue line) fitted from the local RD samples (The black markers).

5.5 Piecewise Envelope Generation

We propose a piece-wise approximation method to fit the global RD envelope. The idea is that the global RD envelope can be approximated by a family of piece-wise line

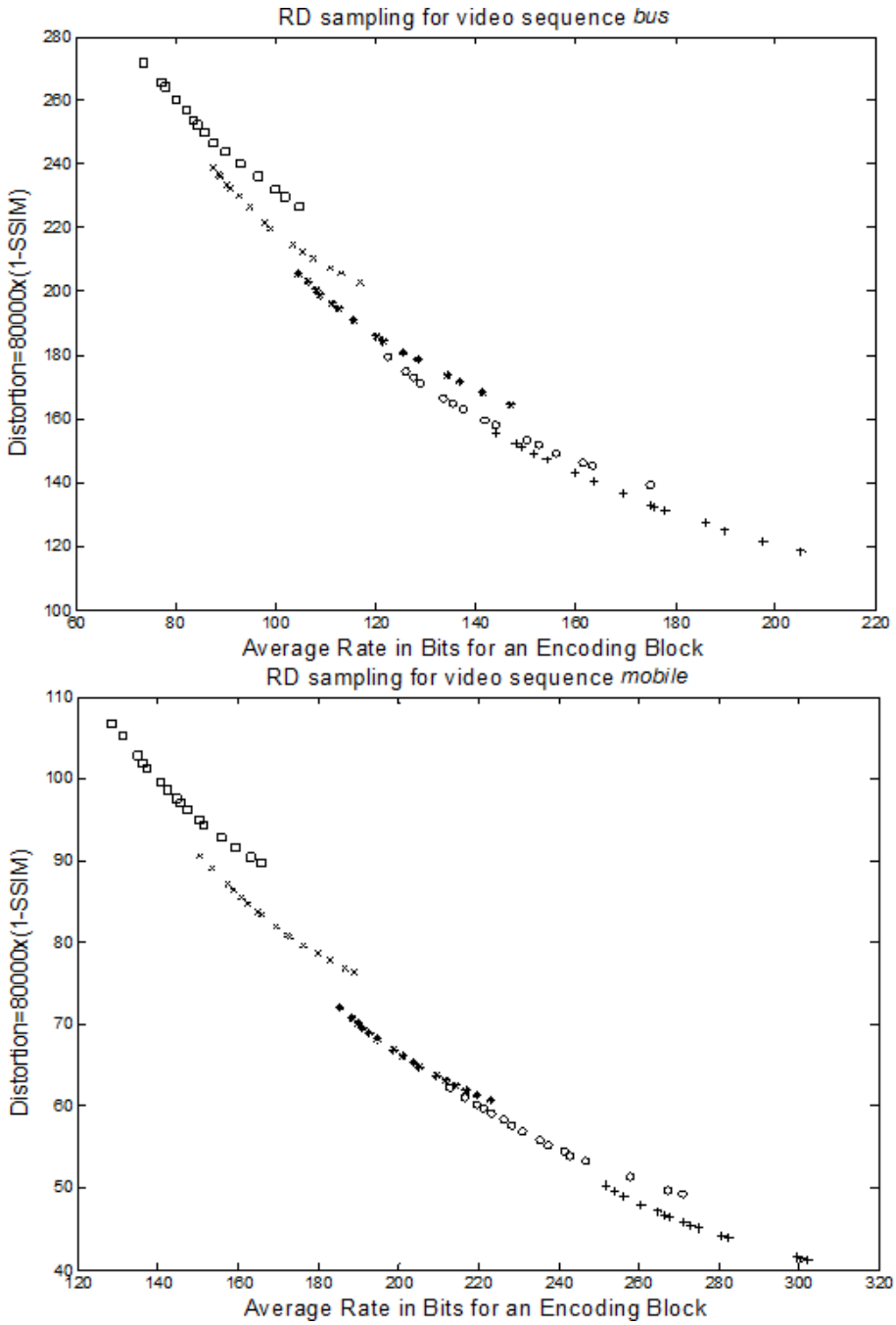

Figure 5-5. RD samples over different QP and varying Lagrange multipliers, video Bus and Mobile.

segments (blue line in Fig. 5-7), each of which is on the common tangent line of two neighboring local RD curves.

Since the local RD curve belongs to a circle, we use the following procedure to find the common tangent line of two circles. Suppose the two circles have the following form,

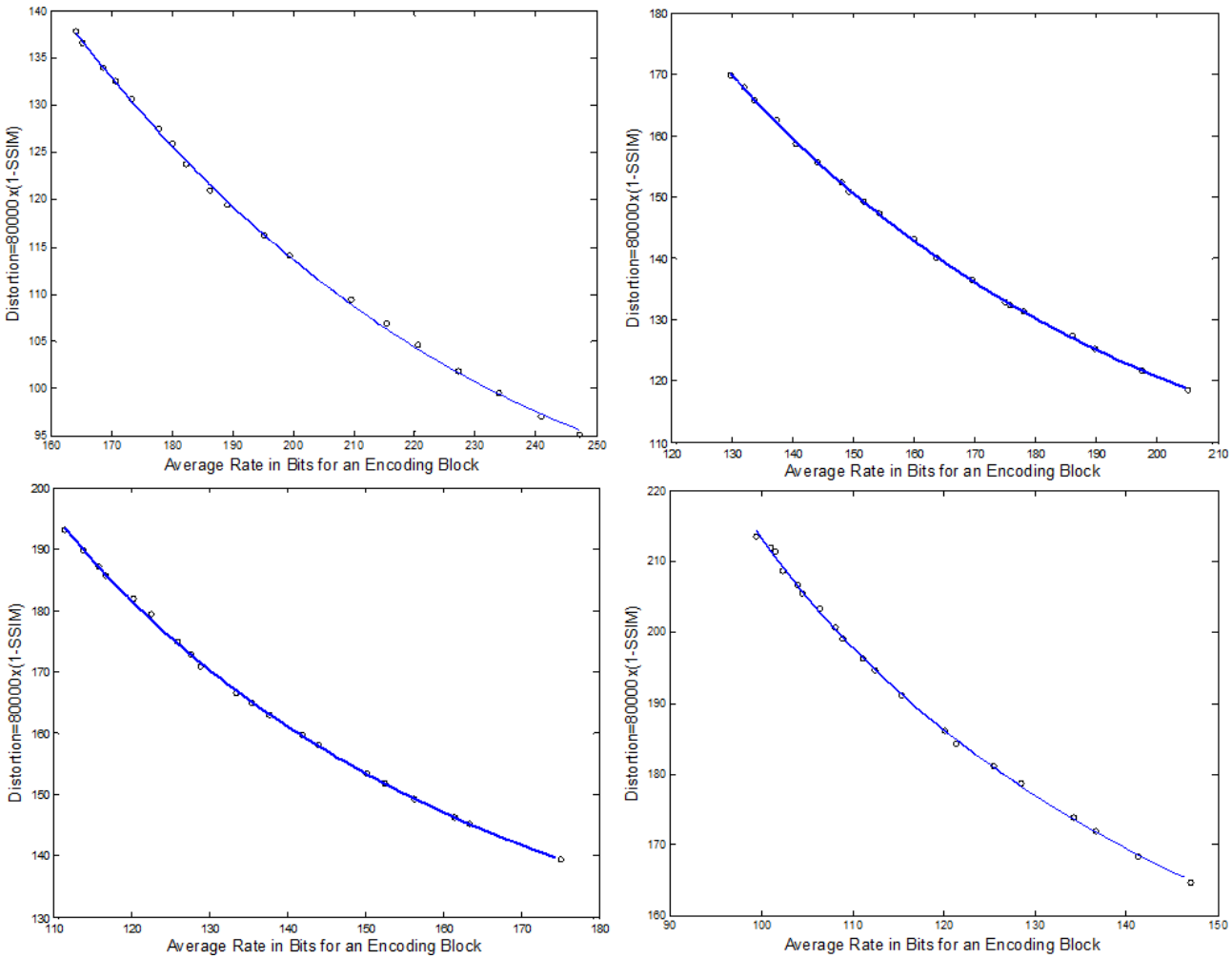

Figure 5-6. Samples from video Bus with the same QP, varying Lagrange multipliers. Circle is used as the fitting model. Clockwise: QP = 23, 24, 25, 26, respectively.

$$
\begin{aligned}
(x - x_1)^2 + (y - y_1)^2 &= r_1^2 \\
(x - x_2)^2 + (y - y_2)^2 &= r_2^2
\end{aligned}
\tag{5–8}
$$

where (x_1, y_1) and (x_2, y_2) are the centers of the two circles, r_1 and r_2 are the radius of the two circles. Suppose its common tangent line is $a \times x + b \times y + c = 0$, then $a = RX - kY\sqrt{1-R^2}$, $b = RY + kX\sqrt{1-R^2}$, $c = r_1(ax_1 + by_1)$, where $X = \frac{x_2 - x_1}{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}}$, $Y = \frac{y_2 - y_1}{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}}$, $R = \frac{r_2 - r_1}{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}}$.

Based on the fitted local curves in Section 5.4 and the procedure above, we obtain the common tangent line for every two neighboring local RD curves. Fig. 5-8 shows the

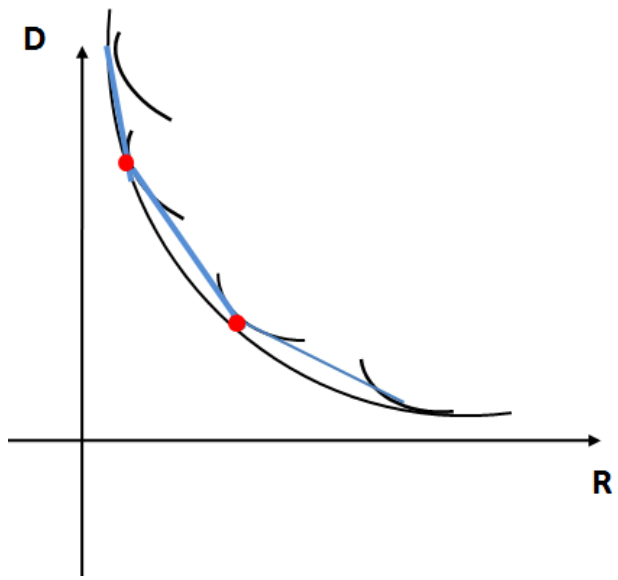

Figure 5-7. The piece wise approximation of RD envelope by tangent line segments.

tangent line over two local RD curves. The tangent line demonstrates the gradient of global RD envelope at QP of the two neighboring local RD curves. Therefore, the family of tangent lines forms an approximation of the global RD envelope. The intersections of every two neighboring tangent lines are on the piecewise envelope. As shown in Fig. 5-10, the red markers are such intersection points and they are well positioned on the envelope that encloses all RD samples (black markers).

5.6 Experiment Results

We compare the RD performance of our proposed perceptual RDO with [86] and baseline JM 16.0 [78]. In the experiment, we use H.264 baseline profile to encode each video. The encoding configurations are GOP structure IPPP with 12 frames as a group, RDO enabled, maximum 3 reference frames for inter coding and QP from 20 to 36. Our proposed framework is implemented in JM 16.0 and we test five video sequences in cif size (352x288). As shown in Fig. 5-10, the RD curve of our method locates in the most lower left among the three RD curves, which means that our method outperforms both [86] and the baseline JM. In the low bitrate range (QP 28-36), our method has a large saving margin, with 5% and 15% bitrate reduction. In the high bitrate range

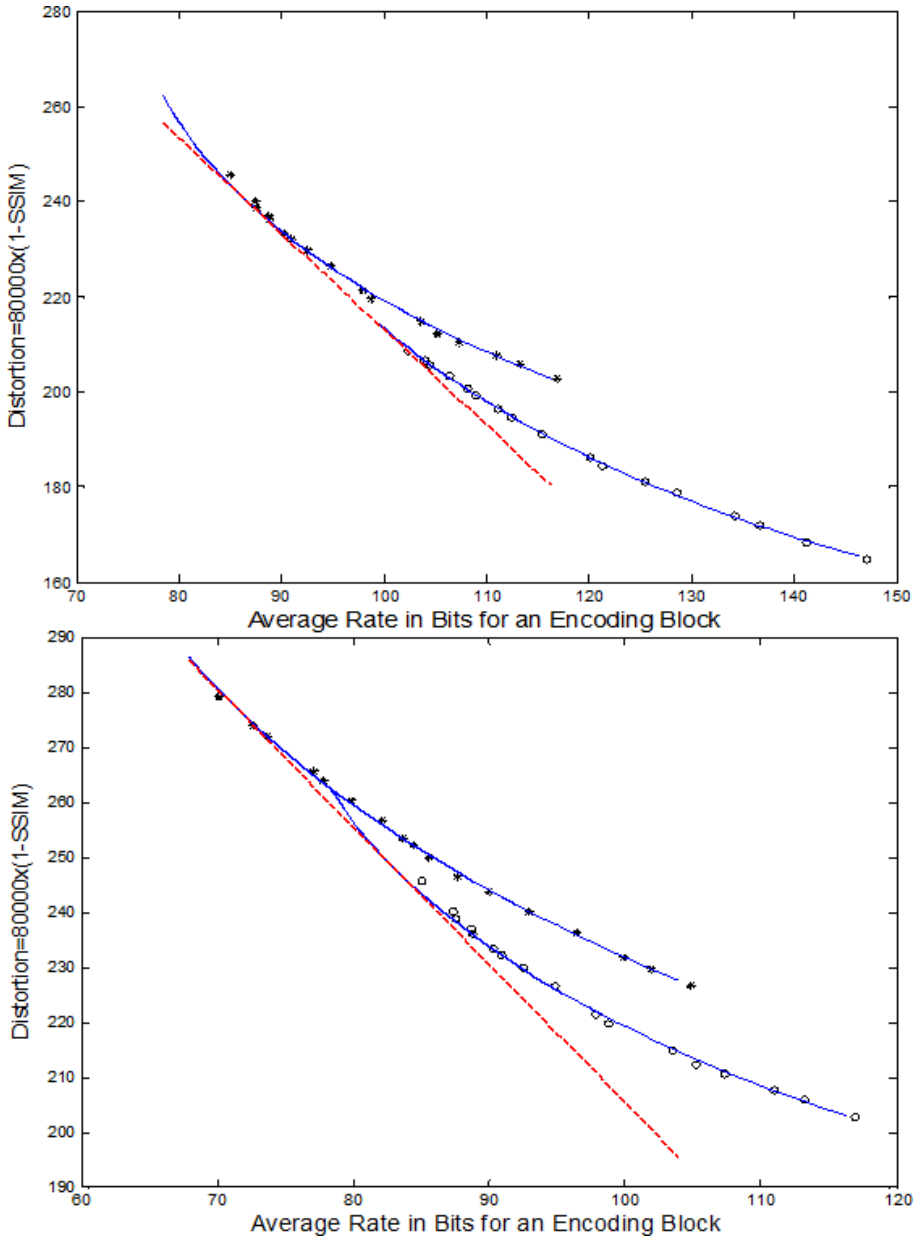

Figure 5-8. The common tangent line of two local RD curves (circle segment). Left: QP = 26. Right: QP = 28.

(QP 20-27), since the bitrate is less prioritized than distortion, all three methods may choose the most complex modes to encode a block, which results in similar distortion. However, our method still can save 2% and 12% bitrate from the two reference methods. Table. 5-1 shows that the bitrate reduction rate of our proposed method comparing with the two reference methods.

Table 5-1. Bitrate reduction (%) of our proposed RDO for inter-frame coding under Baseline profile. Courtesy of the video trace library [3].

Videos	QP 28-36		QP 20-27	
	Ref-[86]	Ref-JM	Ref-[86]	Ref-JM
Foreman	-5.42	-15.29	-3.84	-12.2
Coastguard	-6.97	-12.56	-2.27	-9.10
Bus	-9.78	-16.6	-3.86	-7.59
Mobile	-7.63	-11.21	-3.58	-8.42
Akiyo	-6.25	-10.63	-2.61	-9.86

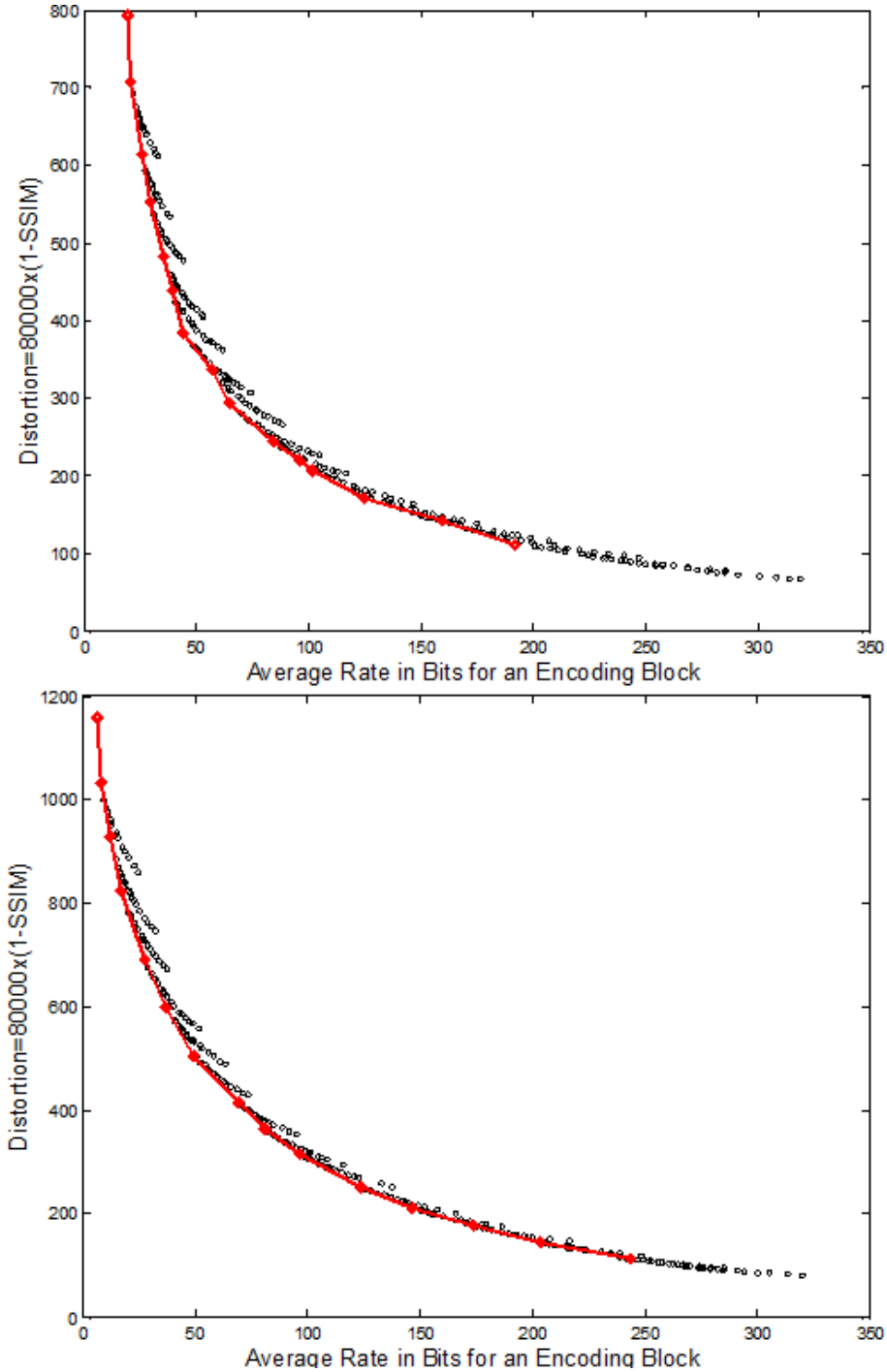

Figure 5-9. Piecewise linear approximation of the RD envelope. Left: Bus. Right: Coastguard.

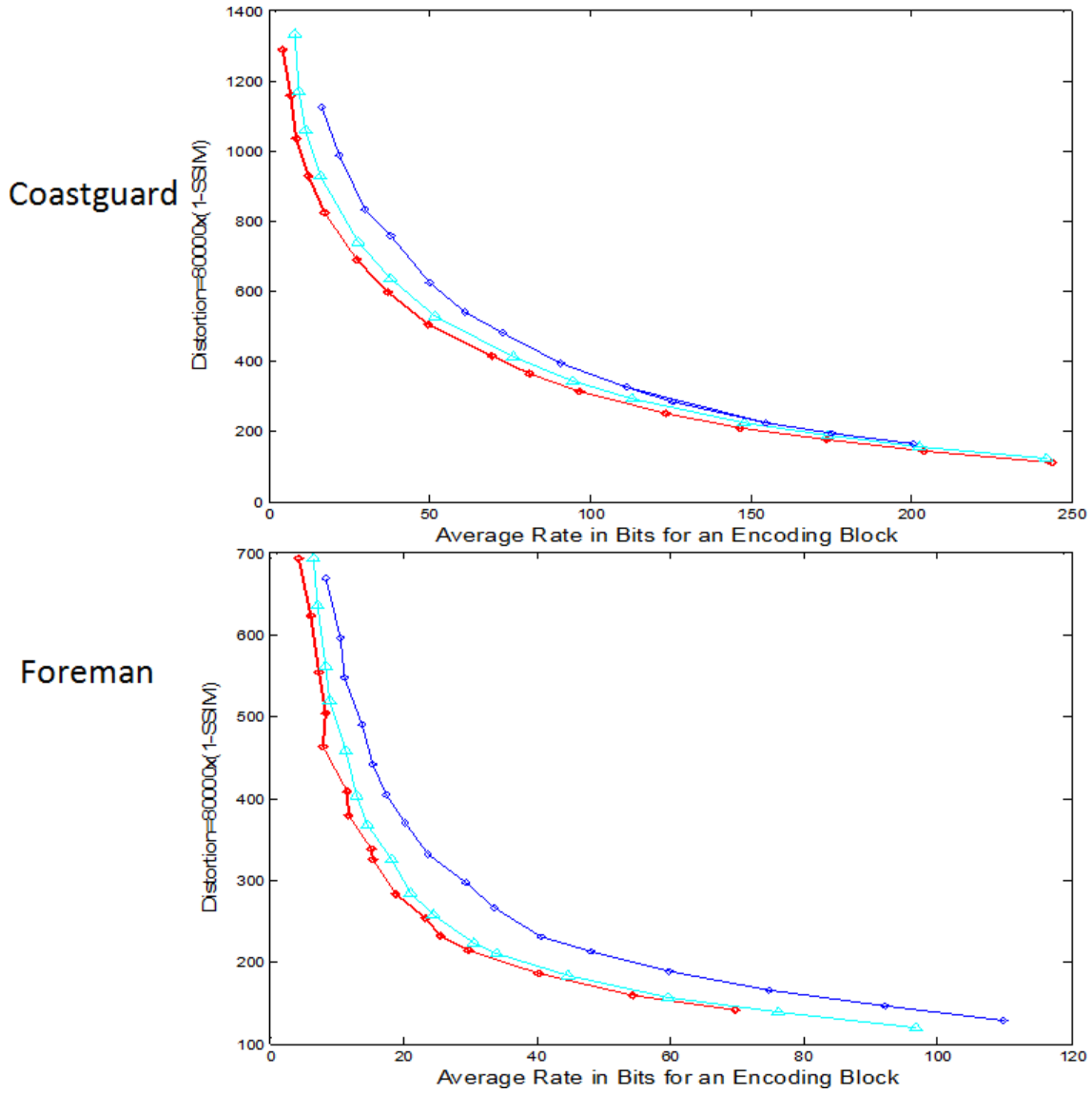

Figure 5-10. Perceptual RDO performance comparison. Red: the proposed RDO. Cyan: the framework in [86]. Blue: JM-MSE based RDO.

CHAPTER 6
CONCLUSION

This dissertation studies four advanced video processing techniques in video transmission system. The first two techniques, video retargeting and video summarization, enhance the adaptivity when transmitting video content across different platforms. The other two techniques, perceptual quality assessment and perceptual based encoder optimization, exploit human visual characteristics to reconfigure video transmission system components.

In Chapter 2, the author presents a novel video retargeting system that is suitable for long videos with generic contents. Different from many existing approaches that focus on visual interestingness preservation, the author identifies that the temporal retargeting consistency and non-deformation play a dominant role. Also, the statistical study on human response to the retargeting scale shows that seeking a global retargeting view as in heterogeneous approaches is not necessary. The refined homogeneous approach for the first time comprehensively addresses the prioritized temporal retargeting consistency and achieves the most possible preservation of visual interestingness. In particular, the author proposes a volume retargeting cost metric to jointly consider the two objectives and formulated the retargeting as an optimization problem in graph representation. A dynamic programming solution is given. To measure the visual interestingness distribution, the author also introduces a non-linear fusion based attention modeling. Encouraging results have been obtained through the image rendering of the proposed attention modeling and video retargeting system on various images and videos. The subjective test statistically demonstrates that the proposed attention model can more effectively and efficiently extract salient regions than conventional methods and also the video retargeting system outperforms other homogeneous methods.

In Chapter 3, the author aims at a summary that guarantees viewers' understanding of the video plot. The author proposes a novel method to represent the semantic structure as concepts and instances and argued that a good summary maintains all concepts complete, balanced and saliently illustrated. It is shown in the experiments the effectiveness of the proposed concept detection method. Also, the human subjective test justifies the strength of the summarization method.

In Chapter 4, a visual psychological factor based perceptual quality model is proposed to measure the video quality degraded by the packet loss events during transmission. The model focuses on exploring two dominant psychological factors that directly correlate with human response in a linear fashion. For the first factor the glittering block effects, the author proposes a structural edge detection method to describe its strength. For the second factor the human face distortion, the author proposes a motion vector entropy based face deformation estimation method to describe its value. Then the author uses linear regression with cross validation to train the perceptual model. The experiments show that the proposed method correlates with the actual human score more than the conventional bitstream/QoS factor based approach and also can increase prediction accuracy by 4% in terms of RRMSE.

In Chapter 5, a perceptual RDO framework based on piecewise linear approximation is proposed. The author starts from the MSE-RDO and rescale and offset its Lagrange multiplier to suit for the dynamic range of perceptual distortion in the proposed perceptual RDO. Based on the collected RD samples, the author finds the envelope curve correspond to the best achieve RD model. The author then approximates the RD envelope with piecewise line segments, each segment is from a common tangent line of two circles fitted from RD samples. Experiments illustrate that the proposed RDO featuring RD envelope approximation outperforms the conventional methods by 2% to 5%.

REFERENCES

[1] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," *ACM Transactions on Graphics*, vol. 29, pp. 160, 2010.

[2] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *Proc. of IEEE Computer Vision and Pattern Recognization*, vol. 20, pp. 1–8, 2007.

[3] Video trace library, "Yuv video sequences," *http://trace.eas.asu.edu/yuv/*, 2010.

[4] CBS Broadcasting Inc., "The big bang theory," *http://the-big-bang-theory.com/*, 2009.

[5] the Blender institute, "Big buck bunny," *http://www.bigbuckbunny.org/*, 2008.

[6] Cable News Network, "Student news," *http://www.cnn.com/studentnews/*, 2009.

[7] S. Kanumuri, P. Cosman, A. Reibman, and V. Vaishampayan, "Modeling packet-loss visibility in mpeg-2 video," *IEEE Transactions on Multimedia*, vol. 8, pp. 341–355, 2006.

[8] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," *Proc. of IEEE Computer Vision and Pattern Recognization*, pp. 1–8, 2008.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, 1998.

[10] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," *Advances in Neural Information Processing Systems*, vol. 21, pp. 681–688, 2008.

[11] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395–1407, 2006.

[12] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, pp. 907 – 919, 2005.

[13] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," *Proc. of IEEE International Conference on Computer Vision*, vol. 7, pp. 1 – 6, 2007.

[14] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Transactions on Graphics*, vol. 26, pp. 10, 2007.

[15] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," *Proc. of ACM Multimedia*, pp. 241–250, 2006.

[16] G. Hua, C. Zhang, Z. Liu, Z. Zhang, and Y. Shan, "Efficient scale-space spatiotemporal saliency tracking for distortion-free video retargeting," *Computer Vision-ACCV 2009*, pp. 182–192, 2010.

[17] T. Deselaers, P. Dreuw, and H. Ney, "Pan, zoom, scan – time-coherent, trained automatic video cropping," *Proc. of IEEE Computer Vision and Pattern Recognization*, pp. 1–8, 2008.

[18] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Transactions on Graphics*, vol. 27, pp. 1–9, 2008.

[19] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Discontinuous seam-carving for video retargeting," *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 569–576, 2010.

[20] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Transactions on Graphics*, vol. 28, pp. 23:1–23:11, 2009.

[21] Y. S. Wang, C. L. Tai, O. Sorkine, and T. Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Transactions on Graphics*, vol. 27, pp. 1–8, 2008.

[22] Y. S. Wang, H. Fu, O. Sorkine, T. Y. Lee, and H. P. Seidel, "Motion-aware temporal coherence for video resizing," *ACM Transactions on Graphics*, vol. 28, pp. 127, 2009.

[23] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Transactions on Graphics*, vol. 28, pp. 126:1–126:10, 2009.

[24] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *Journal of Electronic Imaging*, vol. 5, pp. 122–128, 1996.

[25] N. V. Patel and I. K. Sethi, "Video shot detection and characterization for video databases," *Pattern Recognition*, vol. 30, pp. 583–592, 1997.

[26] T. A. Ell and S. J. Sangwine, "Hypercomplex fourier transforms of color images," *IEEE Transactions on Image Processing*, vol. 16, pp. 22–35, 2007.

[27] J. Shi and C. Tomasi, "Good features to track," *Proc. of IEEE Computer Vision and Pattern Recognization*, vol. 29, pp. 593–600, 1994.

[28] P. J. Rousseeuw and A. M. Leroy, "Robust regression and outlier detection," *Wiley-IEEE*, vol. 589, 2005.

[29] DreamWorks Animation, "Madagascar," *http://www.madagascar-themovie.com/main.php*, 2005.

[30] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1915–1926, 2012.

[31] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–416, 2011.

[32] R. Rensink, "Seeing, sensing and scrutinizing," *Vision Research*, vol. 40, pp. 1469–1487, 2000.

[33] Michael Kors, "runway model," *http://www.youtube.com/watch?v=QvKHuZfJzN0*, 2010.

[34] Y. Li, S. Lee, and et al., "Techniques for movie content analysis and skimming," *IEEE Signal Processing Magazine*, vol. 23, pp. 79, 2006.

[35] Y. Li, T. Zhang, and et al., "An overview of video abstraction techniques," *HP Laboratories Palo Alto, Tech. Report No. HPL-2001-191s*, 2001.

[36] D. Dementhon and D. Doermann, "Video summarization by curve simplification," *ACM Interntional Conference on Multimedia*, pp. 211–218, 1998.

[37] Y. Gong and X. Liu, "Video summarization using singular value decomposition," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 174–180, 2000.

[38] M. Padmavathi, R. Yong, and et. al, "Keyframe-based video summarization using delaunay clustering," *International Journal on Digital Libraries*, vol. 6, pp. 219–232, 2006.

[39] Z. Cernekova and Ioannis Pitas, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circular Systems for Video Technology*, vol. 16, pp. 82–91, 2006.

[40] Yuxin Peng and Chong-Wah Ngo, "Clip-based similarity measure for query-dependent clip retrieval and video summarization," *IEEE Transactions on Circular Systems for Video Technology*, vol. 16, pp. 612–627, 2006.

[41] Zhu Li, Guido M. Schuster, and et. al, "Minmax optimal video summarization," *IEEE Transactions on Circuits Systems for Video Technology*, vol. 15, pp. 1245–1256, 2005.

[42] Y. F. Ma, X. S. Hua, and et al., "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, pp. 907–919, 2005.

[43] C. W. Ngo, Y. F. Ma, and et al., "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits Systems for Video Technology*, vol. 15, pp. 296–305, 2005.

[44] G. Evangelopoulos, A. Zlatintsi, and et al., "Video event detection and summarization using audio, visual and text saliency," *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 3553–3556, 2009.

[45] Y. F. Ma, L. Lu, and et al., "A user attention model for video summarization," *ACM Conference on Multimedia*, pp. 533–542, 2002.

[46] L. Itti, C. Koch, and et al., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, 1998.

[47] T. Lu, Z. Yuan, and et al., "Video retargeting with nonlinear spatial-temporal saliency fusion," *IEEE International Conference on Image Processing*, pp. 1801–1804, 2010.

[48] T. Wang, Y. Gao, and et al., "Video summarization by redundancy removing and content ranking," *ACM International Conference on Multimedia*, pp. 577–580, 2007.

[49] D. Lewis, "Naive bayes at forty: The independence assumption in information retrieval," *Machine Learning, Springer*, pp. 4–15, 1998.

[50] D. G. Lowe, "Object recognition from local scale-invariant features," *IEEE International Conference on Computer Vision*, pp. 1150–1157, 1999.

[51] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing, Springer*, vol. 17, pp. 395–416, 2007.

[52] L. Herranz and J. M. Martinez, "A framework for scalable summarization of video," *IEEE Transactions on Circuit Systems for Video Technology*, vol. 20, pp. 1265–1270, 2010.

[53] J. Asghar, F. Le Faucheur, and I. Hood, "Preserving video quality in iptv networks," *IEEE Transactions on Broadcasting*, vol. 55, pp. 386–395, 2009.

[54] International Telecommunication Union (ITU), "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," *ITU-T Recommendation J.246*, 2008.

[55] M. Claypool and J. Tanner, "The effects of jitter on the perceptual quality of video," *ACM International Conference on Multimedia-Part 2*, pp. 115–118, 1999.

[56] "Objective perceptual multimedia video quality measurement in the presence of a full reference," *International Telecommunication*, vol. J.341, 2011.

[57] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 71–83, 2008.

[58] S. Argyropoulos, A. Raake, M. Garcia, and P. List, "No-reference video quality assessment for sd and hd h.264/avc sequences based on continuous estimates of packet loss visibility," *Proc. 3rd International Workshop Quality of Multimedia Experience (QoMEX)*, pp. 31–36, 2011.

[59] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, pp. 660–668, 2008.

[60] T. L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. Cosman, and A. Reibman, "A versatile model for packet loss visibility and its application to packet prioritization," *IEEE Transactions on Image Processing*, vol. 19, pp. 722–735, 2010.

[61] F. Pan, X. Lin, S. Rahardja, W. Lin, E. Ong, S. Yao, Z. Lu, and X. Yang, "A locally-adaptive algorithm for measuring blocking artifacts in images and videos," *Proceedings of the International Symposium on Circuits and Systems*, vol. 3, pp. 23–26, 2004.

[62] Tao Liu, Yao Wang, J.M. Boyce, Hua Yang, and Zhenyu Wu, "A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, pp. 280–293, 2009.

[63] Y.J. Liang, J.G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Effect of burst losses and correlation between error frames," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 861–874, 2008.

[64] N. Staelens, G. Van Wallendael, and et al., "No-reference bitstream-based visual quality impairment detection for high definition h.264/avc encoded video sequences," *IEEE Transactions on Broadcasting*, vol. 58, pp. 187 – 199, 2012.

[65] Video Quality Experts Group (VQEG), "Report on the validation of video quality models for high definition video content," *IEEE Signal Processing Magazine [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/*, pp. 577–580, 2010.

[66] "Live mobile quality database," *http://live.ece.utexas.edu/research/Quality/index.htm*, 2012.

[67] "Xiph.org video test media," *http://media.xiph.org/video/derf/*, 2012.

[68] "Vqeg hybrid test plan," *http://www.its.bldrdoc.gov/vqeg/projects/hybrid-perceptual-bitstream/hybrid-perceptual-bitstream.aspx*, 2012.

[69] Paul Viola and Michael J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.

[70] "Weka," *http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html*, 2012.

[71] Gary J. Sullivan and Thomas Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, pp. 74–90, 1998.

[72] Tihao Chiang and Ya-Qin Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, pp. 246–250, 1997.

[73] Hung-Ju Lee, Tihao Chiang, and Ya-Qin Zhang, "Scalable rate control for mpeg-4 video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 878–894, 2000.

[74] Liang-Jin Lin and Antonio Ortega, "Bit-rate control using piecewise approximated rate-distortion characteristics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 446–459, 1998.

[75] P. H. Westerink, R. Rajagopalan, and C. A. Gonzales, "Two-pass mpeg-2 variable-bit-rate encoding," *IBM Journal of Research and Development*, vol. 43, pp. 471–488, 1999.

[76] Chengsheng Que, Guobin Chen, and Jilin Liu, "An efficient two-pass vbr encoding algorithm for h. 264," *IEEE International Conference on Communications, Circuits and Systems*, vol. 1, pp. 118–122, 2006.

[77] Dimitri P. Bertsekas, "Constrained optimization and lagrange multiplier methods," *Computer Science and Applied Mathematics, Boston*, 1982.

[78] "H.264/avc software jm 16.0," *http://iphome.hhi.de/suehring/tml/*, 2012.

[79] Thomas Wiegand and et al., "Overview of the h. 264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, 2003.

[80] Kalva Hari and Lakis Christodoulou, "Using machine learning for fast intra mb coding in h. 264," *Electronic Imaging 2007*, 2007.

[81] Siwei Ma, Wen Gao, and Yan Lu, "Rate-distortion analysis for h. 264/avc video coding and its application to rate control," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1533–1544, 2005.

[82] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, pp. 98–117, 2009.

[83] Winkler Stefan, Animesh Sharma, and David McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," *Proc. International Symposium on Wireless Personal Multimedia Communications*, pp. 547–552, 2001.

[84] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of iptv

and video on demand services in real-life environments," *IEEE Transactions on Broadcasting*, vol. 56, pp. 458–466, 2010.

[85] Z. Wang and et al., "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, 2004.

[86] Yi-Hsin Huang and et al., "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 1614–1624, 2010.

[87] C. Wang, X. Mou, and L. Zhang, "A perceptual quality based rate distortion model," *2012 Fourth International Workshop on Quality of Multimedia Experience*, pp. 74–79, 2012.

[88] S. Wang and et al, "Ssim-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 516–529, 2012.

[89] X. Wang, L. Su, Q. Huang, and C. Liu, "Visual perception based lagrangian rate distortion optimization for video coding," *IEEE International Conference on Image Processing*, pp. 1653–1656, 2011.

[90] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 605–618, 2012.

[91] Ben James Winer, Donald R. Brown, and M. Michels Kenneth, *Statistical principles in experimental design*, McGraw-Hill, 1991.

BIOGRAPHICAL SKETCH

Zheng Yuan was born in 1984 in Zhangjiakou, Hebei Province, China. He received his B.S. degree in electronic and information engineering at Jilin University, Changchun, China in 2006 and the M.S. degree in electronic information and electrical engineering at Shanghai Jiao Tong University, China in 2009. He received his Ph.D. degree in electrical and computer engineering at the University of Florida, Gainesville, USA in the summer of 2013. His research interests include video and image analysis/processing, video compression, multimedia, computer vision and machine learning.