A DYNAMIC SALIENCY BASED METHOD FOR VIDEO RETARGETING

by

Hatice Çiğdem Koçberber

B.S, Computer Engineering, Boğaziçi University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2015

ACKNOWLEDGEMENTS

First of all, I would like to thank to my supervisor Assist. Prof. Albert Ali Salah for his endless support and patience during my study. He has guided me along the steps of this thesis while also teaching me how to be curious, and extract research ideas from any conversation. He not only helped me with this thesis, but also gave me an insight to observe, search and accomplish. While this thesis has enriched me as a student, his leading has also enhanced me as a human. He has a big part in who I am now, and who I will become later in life.

Secondly, I would like to thank all my family, my father Seyit Koçberber my mother Süreyya Suzan Koçberber, my little sister Çağla Koçberber, and my older sister Selin Dur who have always been there for me with their love.

I want to thank my friends, Sevgi Şen, Akif Cem Heren, Özgül Emine Vatan, Dila Hocaoğlu and Dilek Doğruer for their comments and advices, while listening my endless discussions about this thesis.

Finally, I want to thank to my husband Umur Kontacı, who held my hand and walked with me towards the completion of this thesis. His existence has always soothed me, his blessings, love and support has always encouraged me to achieve better.

ABSTRACT

A DYNAMIC SALIENCY BASED METHOD FOR VIDEO RETARGETING

With increased usage of smartphones, tablets and small displays to play multimedia content, video retargeting becomes an important tool for better user experience. In this thesis, we propose a novel content-based approach for video retargeting that relies on spatio-temporal saliency to estimate relevant information in videos. Our method preserves spatial saliency as well as temporal coherence. We also propose a spatio-temporal saliency algorithm designed for this application domain that combines spatial saliency with motion trajectories. We demonstrate the quality of the proposed approach through quantitative and qualitative evaluation, contrasting it with five different video retargeting methods. Quantitative evaluation is done using generic image/video quality metrics, so that they can be applied on any video retargeting solution. We have extracted the correlation between the quantitative and qualitative evaluation, to propose a new metric that is a combination of the existing quantitative metrics. The proposed metric is proven to be the best approximation to the qualitative results, thus can be used as a benchmark to evaluate video retargeting methods.

ÖZET

DİNAMİK BERCESTELİK TABANLI VİDEO UYARLAMA

Akıllı telefon, tablet ve küçük ekranların multimedya içerik için kullanımının artmasıyla birlikte video uyarlama, kullanıcı deneyimini zenginleştirmek için önemli bir araç haline geldi. Bu tezde, videolardaki önemli içeriği tespit etmek ve video uyarlama yapabilmek için konumsal ve zamansal olarak seyircinin dikkatini çeken noktalara dayanan, yeni bir içerik bazlı yaklaşım sunuyoruz. Önerdiğimiz metot görüntülerde dikkati çeken bölgelerin uyarlama sırasında korunmasını sağladığı gibi, videonun kareleri arasında da zamansal uyumu kaybetmemektedir. Ayrıca tezde bu uygulama için özel olarak tasarlanmış bir dinamik bercestelik metodu öneriyoruz. Bu metot görüntülerde dikkati çeken noktaları zaman içinde izleyerek tutarlılığı sağlıyor. Sunduğumuz yaklaşımın kalitesini beş farklı video hedeflendirme metoduyla niteliksel ve niceliksel olarak kıyaslayarak gösteriyoruz. Niceliksel değerlendirme diğer tüm video hedeflendirme çözümlerine uygulanabilmesi için genel görüntü/video kalite ölçütleriyle yapılmıştır. Nicel ve nitel değerlendirme arasında bulduğumuz korelasyonu kullanarak hâlihazırda bulunan niceliksel ölçütlerin birleşiminden oluşan yeni bir ölçüt sunuyoruz. Sunduğumuz bu ölçüt ile nicel sonuçlara olabildiğince yakın sonuçlar verilmiştir. Bu ölçütün video uyarlama metotlarını kıyaslamak için kıyaslamak için kullanılabileceğini düşünüyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$+L_R$	The difference between left right neighbour pixels
$+L_U$	The difference between left up neighbour pixels
$-L_U$	The reverse difference between left right neighbour pixels
b_i	Start frame of a trajectory
B_{ij}	Blue channel extracted from an image
$c(x, y)$	Contrast channel for SSIM
C^i	Center of saliency of frame i
C_i	Crop window center of frame i
C_M	Camera motion
C_o	Crop window flow
d_k	Distance between two consecutive polygon points z_k and z_{k+1}
D	Quad energy function defined over each quad of a mesh
D_l	Mesh line bending error
D_u	Quad deformation measure
$D_{i,j}$	Distance between trajectory i and j
E^f	Saliency dispersion
E_M	Energy map
F	Contrast sensitivity function
G_{ij}	Green channel extracted from an image
G_x	Gradient of Sobel Operator in x dimension
G_y	Gradient of Sobel Operator in y dimension
h_i	Homography matrix between frame i and $i-1$
h_k	Measure added for the distance between two consecutive polygon points z_k and z_{k+1}
H	Set of homography matrices
H_0	The null hypothesis
H^i	Homography matrix i in a set of five consecutive homography matrices
I	Image

I_i	i^{th} frame of a video
J_i	Method i in a pairwise comparison
K	Set of key frames extracted from a shot
K_P	Key points extracted from a seam
l_{ij}	Edge deformation between vertex i and vertex j
$l(p)$	Log-likelihood of p
$l(x, y)$	Luminance channel for SSIM
L	Length of a trajectory
M_A	Matching area that surrounds Key Points
M_i	Motion class of i^{th} shot
M_I	Matching index that is computed over Matching Area of Key Points
M_W	Window size to define the Matching Area
M_{Blur}	Blur metric
$M_{Brightness}$	Brightness metric
$M_{Compress}$	Compress metric
$M_{Div.Norm}$	Division Normalization metric
$M_{Jerkiness}$	Jerkiness metric
M_{MSE}	Mean Squared Error metric
M_{PSNR}	Peak Signal to Noise Ratio metric
$M_{Sharpness}$	Sharpness metric
M_{SSIM}	Structural Similarity Index metric
M_{UQI}	Universal Quality Index metric
M_{VIF}	Visual Information Fidelity metric
n_σ	Size of the trajectory window for improved trajectories
n_t	Length of the trajectory window in time dimension for improved trajectories
N_{J_i}	Number of times method J_i won over method J_j in a pairwise comparison
N^f	Total number of pixels that a saliency map has a non-zero value
$\widetilde{O}$	Smoothened optical flow map

$\mu_{\widetilde{O}}$	Mean of the thresholded optical flow map
π_i	BL worth parameter of object i
$\tilde{\pi_i}$	Approximated BL worth parameter of object i
p_i	Probability of event i
P	Pixel
$P(i > j)$	Probability that event i occurs more frequent than event j
$r_k(q)$	The weighting function for each landmark z_k
r_ij	Number of times object i is preferred over object j in pairwise comparisons
R_ij	Red channel extracted from an image
R_P	Reward Punish Map defined to select the best possible seam
s_i	Static saliency map
s_v	Dynamic saliency map
s_f	Scale factor of quad f
s_{ij}	Scale factor between quads i and j
$s(x, y)$	Structure channel for SSIM
S	Saliency map
$\tilde{S}$	Predicted saliency map
t	Time
T	Motion trajectory
T_S	Seed trajectories
U	Wavelet transforms
v_i	Vertex i of the mesh
$v_i^{'}$	Deformed vertex position of vertex i
w_k	k^{th} landmark between frames
W	Significance map
W_α	Gradient of the image used for computing the significance map
Y_ij	Brightness value of pixel i, j
z_i	Interest points in optical flow calculation
α	Coefficient of regularization in optical flow calculation

β	Mean distance factor between consecutive landmarks z_i and z_{i+1}
δ	Distance threshold of the trajectories
δ_P	The displacement of the trajectory P
ϵ	Error factor
η	Threshold for the optical flow map
μ	Optimization factor between spatial and temporal constraints
μ_{H^i}	Mean of H^i
ρ	Loss function of spatio-temporal saliency algorithm
σ_{H^i}	Variance of H^i
τ	Set of trajectories extracted from a shot
ϕ	Motion classifier SVM
ϕ_i	Neural network to for static saliency
ϕ_v	Neural network to for dynamic saliency
ψ	Threshold on optical flow map to define moving parts
χ	Chi-squared statistics
Λ	Polygon of interest points
Σ	SSIM score
Ω_V	Temporal constraint function
Ω_I	Spatial constraint function

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
AUC	Area Under Curve
BT	Bradley-Terry Model
CAMO	Camera Motion Dataset
DM	Deformation Minimization
HOG	Histogram of Oriented Gradients
HOF	Histogram of Optical Flow
HVS	Human Visual System
IM	Information Maximization
MBHx	Motion Boundary Histogram in x direction
MBHy	Motion Boundary Histogram in y direction
mRMS	Minimum Redundancy Maximum Relevance
MSE	Mean Squared Error
PSNR	Peak Signal to Noise Ratio
QMF	Quadrature Mirror Filters
RAM	Random Access Memory
RANSAC	Random Sample Consensus
ROC	Receiver Operating Characteristic
SAD	Sum of Absolute Differences
SSIM	Structural Similarity Index
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
UQI	Universal Quality Index
VIF	Visual Fidelity Index

1. INTRODUCTION

1.1. Motivation

The amount of available multimedia content grows rapidly, with sharp increase in hand-held device usage. The varying display sizes and aspect ratios of these devices make it harder to view these different structured content. The major problem occurs when fitting an image or a video to a screen having a different aspect-ratio then the original content, which is quite common with 16 : 9 televisions, most smartphones, unique sized in-flight screens etc. An illustration of this issue can be seen in Figure 1.1.

The popularity of viewing any kinds of multimedia content on different devices increases in recent years, and fitting movies, videos or images to different aspect ratios is becoming a daily concern. This matter has a much older history. With the introduction of $in-flight\ entertainment$, fitting movies to uncommon aspect ratios had started to become a problem. In-flight entertainment covers all kinds of offers made to the aircraft passengers, including food, drinks, objects of comfort, in-flight radio, noise canceling headphones, in-flight screens etc. Although the idea of an in-flight movie screen may be older, the earliest found record of such an event is in 1925. Taken from the *Flight International Magazine* [1]:

> **An aerial "Picture Theatre":** An interesting experiment was carried out on April 7, when a Handley Page aeroplane ascended from Croydon Aerodrome, with 12 passengers, and during half-an-hour's flight the film version of Sir Arthur Conan Doyle's "The Lost World," was "shown" on a screen fitted up in the cabin of the machine.

While back in 1925, an abnormal display size was probably a rare situation, today, each device has its own *unique* screen, which raises the question of "*what is the most suitable way to view these content on different sized displays*?" Fitting them directly, by changing their aspect ratios can cause distortions. There is a need to

Figure 1.1. The first row illustrates viewing a movie in various sized displays. The second row contains three aspect ratios; 21 : 9, 16 : 9 and 4 : 3 respectively. The original frame is taken from Star Wars: The Empire Strikes Back movie where the actual aspect ratio is 2.35 : 1, which is quite close to the first image. In the third image, the objects are stretched and squeezed, which will decrease the viewing quality.

perform this task in a smarter way, which can be done discarding unimportant parts of each frame, e.g. cropping. While editing the original movie/video is an unpleasant task, it must be preferred in cases when the viewing experience is decreased with distortions caused by aspect ratio changes. Video retargeting arises at this point, as a way to automatically fit videos to different sized displays.

Video retargeting focuses on understanding the important parts of videos, so that the unimportant parts can be discarded when fitting it into a different size/aspect ratio. An easier alternative to video retargeting is to use humans to perform this task manually. A human annotator can do re-editing on the video to make it more suitable for a target display. While a human annotator may perform better than any automated system, this option is not scalable when we consider the high number of videos/movies available. While video retargeting is a fully automated process, it must ensure that the automatically re-edited videos should not contain artifacts.

Video retargeting has started as an extension of image retargeting. Image retar-

geting is applied on images with the same aim: finding, and selecting the important parts of an image. While image retargeting should keep salient parts of an image, video retargeting is more complicated, as it also should ensure temporal smoothness and coherence, as well as keep distortion low. Any distortion in the motion flow can create unwanted waving and shaking effects and decrease the output video quality. While early works on video retargeting focus mostly on keeping important parts of the video [2], recent works focus on decreasing the distortion [3–5].

In this thesis, we propose a content based video retargeting method. Our solution involves cropping important parts of a video with motion flow considerations. Existing cropping methods [6, 7] sometimes produce virtual camera motions (Section 1.2) and artificial scene cuts, and subsequently, important objects might be discarded. These deficiencies can impair the presentation of the visual concept of the original video, e.g. the tone and the mood. The most important parts of a frame are always retained, while virtual scene cuts are barely perceivable.

1.2. Challenges of Video Retargeting

The major trade-off among various retargeting approaches is the trade-off between quality and information loss. If a video retargeting method focuses on keeping the important parts of each frame, it may cause temporal incoherence. On the other hand, if the method focuses on a higher output quality, with no distortions and preserved temporal smoothness, it may fail to include the important parts. A major challenge of any video retargeting method is to balance these constraints.

Another challenge of video retargeting is to understand the important parts of the video automatically. This is a completely different task that is named as $spatio-temporal\ saliency\ extraction$. While the term *saliency* refers to the important parts of the video, $spatio-temporal$ indicates that these important parts are extracted considering the motion in the video. The details of the leading video retargeting approaches, and their possible deficiencies are discussed in detail in Section 2.3.

While numerous studies focus on video retargeting, proposed video retargeting methods lack of common quantitative measures to estimate the output video quality. There are several studies that propose video quality measures, but these are generally method-specific and cannot be used in all video retargeting methods [8–10]. The details of the background of video quality evaluation measures can be found in Section 5.1.

1.3. Organization of the Thesis

This thesis is organized as follows; we describe recent video retargeting applications and spatio-temporal saliency methods in Chapter 2. Since video retargeting is closely related with image retargeting, we also provide some background information on image retargeting. Chapter 3 introduces the preliminary study that helped us to draw the boundaries of this thesis. Chapter 4 describes the proposed retargeting approach along with the enhanced spatio-temporal saliency algorithm. Chapter 5 details our experimental setup. In Chapter 5, we first provide a background in video retargeting evaluation to set the ground of the evaluation of the proposed method. We describe the dataset we have used and proceed with the results of qualitative and quantitative evaluation, followed by our conclusions in Chapter 6.

1.4. Contributions

Four major contributions of this thesis are as follows:

(i) *A Novel Content Based Video Retargeting Method:* Our video retargeting method ensures temporal and spatial coherence. It is designed to overcome the limitations of the existing video retargeting approaches described in Section 1.2.

(ii) *A Novel Spatio-temporal Saliency Algorithm:* We have designed a spatio-temporal saliency algorithm to overcome the information loss problem. The spatio-temporal saliency algorithm uses a state-of-the-art spatial saliency algorithm to capture the important parts of each frame. In addition, we have extracted motion trajectories to ensure the selected portion of the video follows important objects.

(iii) *A Twofold User Study:* In order to assess the quality of the retargeted videos, we

have conducted a user study. Since we make two proposals: a video retargeting method combined with a spatio-temporal saliency algorithm, we have designed the experiment to evaluate both proposals. Firstly, we compare the proposed retargeting method with other several video retargeting approaches. Secondly, we compare the proposed saliency algorithm with another recent spatio-temporal saliency algorithm. We apply the proposed retargeting method on the competing spatio-temporal saliency algorithm and use the resulting videos in the experiment.

(iv) *Correlations of Qualitative and Quantitative Results:* We have used 12 image and video quality metrics and applied them on the videos used in user study. The videos include results of several video retargeting methods, and a spatio-temporal saliency algorithm. The quantitative results of each video are then compared with the results of the user study to reveal the correlation between quantitative and qualitative results. The correlations show the most important metrics to be used in video retargeting evaluation. In addition to that, a new metric, which is a combination of the quantitative metrics is proposed. The newly proposed metric is determined by applying regression on quantitative results, where the target is the user study results. While conducting a user study is a better way for evaluating a new video retargeting approach, it is a lengthy process when iteratively improving the new model. The resulting metric can not replace a user study, but it may be used as an indicator of the results of a user study.

(v) *Publications:*

Koçberber, Cigdem, and Albert Ali Salah. "Video Retargeting: Video Saliency and Optical Flow Based Hybrid Approach" Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

Koçberber, Cigdem, and Albert Ali Salah. "Video Retargeting with Motion Trajectories" ACM Transaction on Graphics (Submitted for publication)

2. RELATED WORK

This chapter includes some background information related with the methods used in this thesis. Main topics covered are image saliency, spatio-temporal saliency, image retargeting and video retargeting.

2.1. Computational Models of Visual Attention

Among various definitions of attention, one of the most ancient and accurate one is made by Aristotle;

> *"..it is impossible to perceive two objects coinstantaneously in the same sensory act unless they have been mixed, [when, however, they are no longer two], for their amalgamation involves their becoming one, and the sensory act related to one object is itself one, and such act, when one, is, of course, coinstantaneous with itself.."* [11]

Here in this thesis, we will concentrate on the visual aspect of attention, which is performed by the Human Visual System (HVS). We give here a brief explanation of visual attention concepts:

Covert Attention: is an expression of attention involving eye movements [12].

Overt Attention: is an expression of attention without eye movements, typically thought of as a virtual "spotlight" [12].

Bottom-up Attention: is derived by an instinct that focuses the attention to the salient point. The salient point is determined with the low level features gathered from the scene such as color, orientation etc. A red dot on a white background is an example that stimulates the bottom-up attention.

Top-down Attention: is to focus on a salient point intentionally. The act of changing the focus is based on a prior knowledge of what to look for. When we are

trying to find someone, our attention is mainly derived by top-down attention [13].

Visual Attention Models: Models that describe how attention is deployed within a given visual scene.

Computational Models of Visual Attention: A type of a visual attention model that computationally describes the steps of simulating, representing, and testing visual attention [12].

Units of Attention: Computational attention models rely on a unit in order to compute attention. Based on theories on HVS, various units are proposed such as space-based attention [14], feature-based attention [15] or object-based attention [16].

Saliency Map: Feature-based attention models compute selected features from a given visual scene and create feature maps. A saliency map combines the information of the individual maps into one global measure of conspicuity [17]. The saliency map not only enables computational attention models to successfully represent attention, it also gives the ability to compare and evaluate the results.

While saliency maps are produced by fusing the features extracted from a visual scene, we can also generate saliency maps from human fixations, thus creating a ground truth for attention mapping. This is done by recording the eye movements of subjects while making them observe a scene (or a picture) and using their fixations. After both the ground truth and computed saliency maps are generated, the evaluation of the computed saliency map is made over well known metrics such as the area under ROC curve [18]. Details of this process are described in Section 5.4. Example ground truth saliency maps extracted from human fixations can be seen in Figure 2.3.

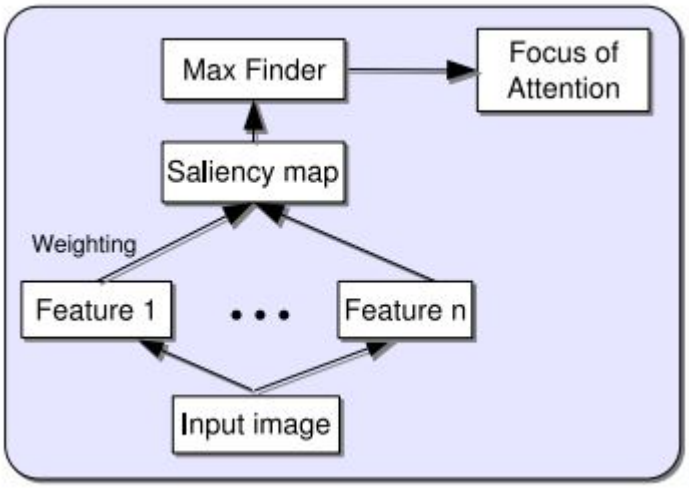


Figure 2.1. General structure of most bottom-up attention systems. The first step of a bottom-up attention models is to analyze the input image and extract features. Once the feature maps are gathered, they are combined into a overall saliency map with a suitable weighting methodology. Then, the most salient part of the saliency map is found, to be selected as the focus of attention. Figure is taken from [19].

2.1.1. Static Saliency

The process of computing static saliency starts with extracting feature maps from a given image. The features may represent bottom-up stimuli such as color and intensity as well as top-down stimuli like faces and horizon. After feature maps are generated, they are combined into a single saliency map. A general flow of static saliency map creation can be seen Figure 2.1.

One of the most recognized study on image saliency is proposed by Itti *et al.* [20]. They extract bottom-up features from three channels: color, intensity and orientation. 42 feature maps in overall are then fused into an single saliency map (Figure 2.2). The downside of this approach is that it fails to represent some top-down features such as faces.

An extension of [20] is proposed by Judd *et al.* [21] where in addition to the bottom-up features, top-down features are also integrated to the saliency map. The

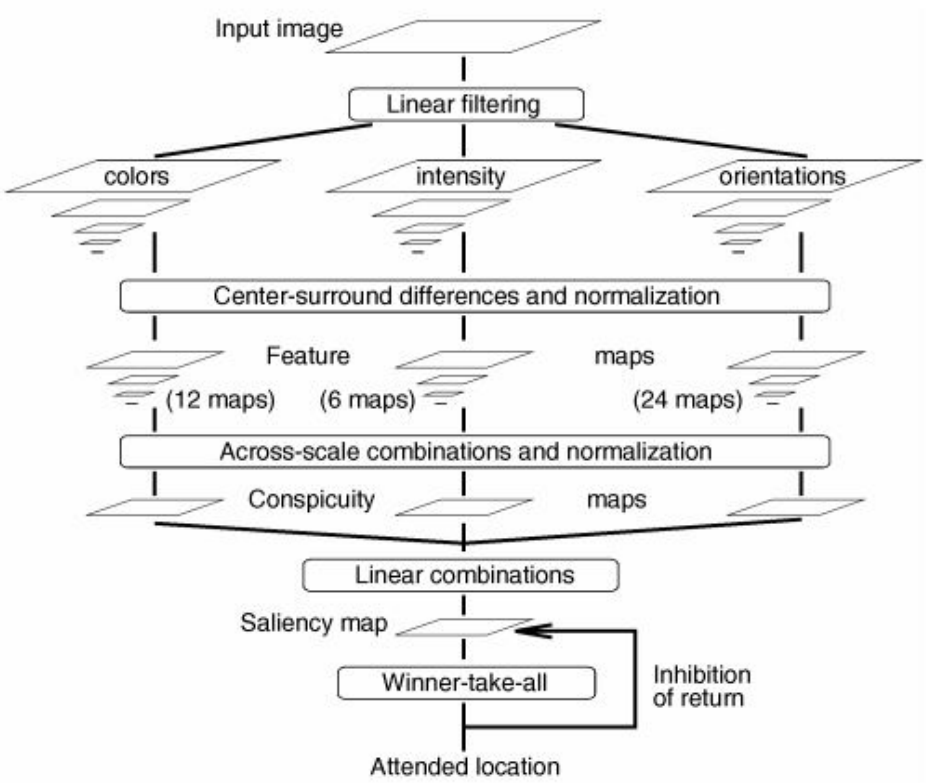


Figure 2.2. The flow of Static saliency algorithm of Itti *et al.* The general structure of the flow resembles the flow of Figure 2.1. Figure is taken from [20].

top-down features presented are faces, persons, cars, horizon and center. Output of face, person, car and horizon detectors are converted to feature maps and added to the overall saliency. In addition, a center bias is detected in human behavior, stating that humans tend to look more to the objects that are close to the center of an image. To include this behavior in their model, they add a feature map that represents the distance of each pixel to the center of the image.

Added top-down features includes faces, horizon and car, which can be applied to most images and produce good results. The downside of this method is that it is too generic, whereas important parts of an image may change according to the context. For example a basketball may be more important than a face in the audience in a basketball match photo.

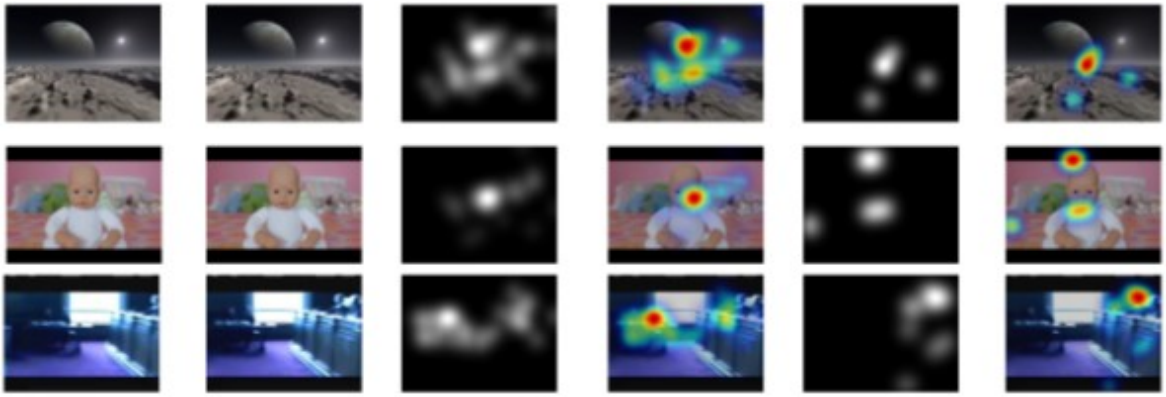


Figure 2.3. This figure shows the difference between the static and spatio-temporal saliency. First two columns are taken from three video sequences. The third and fourth column contain results of the images shown to the subjects. The fifth and sixth column are the results of video fixations of the same frames. The static fixation saliency and heat maps are more distributed. Figure is taken from [22].

2.1.2. Spatio-temporal Saliency

Current spatio-temporal saliency models have emerged as extensions of image saliency models. The most common way to create a spatio-temporal saliency map is to extract motion information from consecutive frames and to combine it with spatial saliency maps [23].

A saliency map is essentially a $2D$ representation. While videos have time as their third dimension, motion can be included in saliency computation, which results in $2D$ saliency maps for each frame. While this simplifies spatio-temporal saliency computation, saliency consistency across frames must be observed.

Humans' attention points in images and videos differ remarkably. We can observe the continuity of saliency in videos by comparing the fixation maps of frames with the fixation maps of images. [22] conducted an experiment to observe the difference of fixations in videos and images. Frame fixation data were collected while subjects watched videos and image fixation data were collected while subjects were shown random frames taken from the same videos.

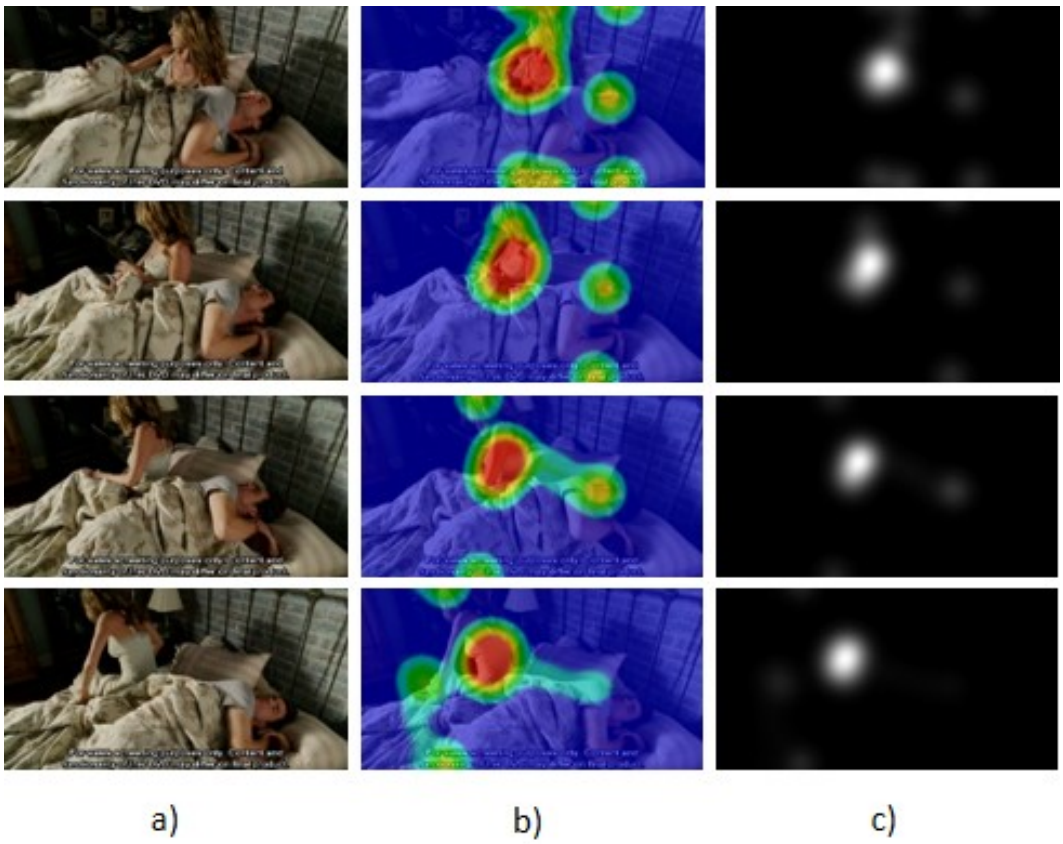


Figure 2.4. a) Original frames. b) Heat maps of the fixation data. c) Saliency maps of the fixation data. The consecutive fixation saliency maps are following the same path with the most important moving object, which is the woman. Frames are taken from Actions In The Eye dataset [24]. Saliency maps and heat maps are created with the Fixation Analysis Tool [25].

Figure 2.3 shows the difference between dynamic (video) and static (image) fixation maps.

We can see that the ground truth saliency maps of videos and images have a different structure. We can go one step further and check the eye fixations of consecutive frames. Figure 2.4 shows the fixation maps taken from Actions in the Eye Dataset [24]. It can be seen that the majority of eye-fixations are concentrated on a small area surrounding a center. In addition to this, the attention centers of consecutive frames are following a path.

Figure 2.5. The optical flow map shown in the third column is computed between the first and the second image. It is created by tracking the points between two consecutive frames in both x and y direction, and combining them into a single map.

This structure of continuous saliency is a result of the motion in the video. There are two main ways to utilize motion information: with optical flow fields or with homography matrices. A homography matrix contains mappings between two consecutive frames of a given video, but it does not include detailed information regarding the moving parts of the frame. As a result, homography matrices are generally used to represent relative camera motion [22, 26] between frames. On the other hand, optical flow maps give precise motion information of each pixel of a frame. Being rich in motion information, optical flow maps are being used more widely in video saliency models [27, 28].

Optical Flow Computation. An optical flow field can be computed by tracking the interest points across frames. Each optical flow algorithm has a different approach of doing this. An example optical flow algorithm, also used in this thesis is proposed in [29]. This method encapsulates the moving pixels in the frame, making it easier to identify the moving *objects*.

Let $\Lambda = \{z_k : z_k \in R^2\}_{k=1}^{N}$ be the polygon of interest points at frame I_1 and the flow vector w_k connects the k^{th} landmark between frames I_1 and I_2. $w_{N+1} = w_1$ constraint ensures that the polygons are closed.

$$E(w_k) = \sum_{k=1}^{N} \sum_{p \in N_k} r_k(p) ||I_2(z_k + w_k + p) - I_1(z_k + p)||^2 + \alpha \sum_{k=1}^{N} h_k ||w_k - w_{k+1}||^2 \quad (2.1)$$

is the objective function where h_k is added for the distance between two consecutive

interest points $h_k = \frac{\beta}{d_k+\beta}$, $d_k = ||z_k - z_{k+1}||$, $\beta = \bar{d_k}.N_k$. r_k is the weighting function for each landmark z_k and α coefficient of regularization. While Equation 2.1 is non-linear, the solution approach is unique to the optical flow algorithm. An example optical flow map can be seen in Figure 2.1.2.

Most spatio-temporal saliency algorithms use these optical flow maps extracted from each frame, and create feature maps. An example spatio-temporal saliency algorithm that combines motion with spatial features in a unique way is proposed by Nguyen *et al.* in [22]. As common in most spatio-temporal saliency algorithms, they compute static features and dynamic features, and combine them to achieve the overall spatio-temporal saliency map. They show that the importance of dynamic and static saliency maps changes with camera motion. An example may be a pedestal camera movement, where human fixations lie on the anticipated direction, not on the objects. In this kind of camera movement, dynamic saliency map has a higher weight in the overall spatio-temporal saliency calculation. They train two separate neural networks $\phi_i(C_M^j, x_j, y_j)$ and $\phi_v(C_M^j, x_j, y_j)$ such that

$$\tilde{S} = \phi_i(C_M^j, x_j, y_j)s_i + \phi_v(C_M^j, x_j, y_j)s_v, \tag{2.2}$$

where s_i is the static saliency map, s_v is the dynamic saliency map and $\tilde{S}$ is the predicted spatio-temporal saliency map.

In order to generate inputs of the neural networks, they divide each frame into 9×9 pixel patches and compute camera motion C_M of each patch j as stated in [30]. These patches are fed to the neural networks along with the position of the patches: x_j and y_j (3×3 inputs come from camera motion homography matrix and two inputs come from the position values sums up to 11 inputs for each neural network). The outputs of the neural networks are the weights of the static saliency map and dynamic saliency map of each patch ($9 \times 9 = 81$ pixels correspond to 81 output values).

The neural networks are trained iteratively to ensure minimizing the loss function ρ where,

$$\rho(\phi_i, \phi_v) = \sum_j ||\phi_i(C_M^j, x_j, y_j)s_i^j + \phi_v(C_M j, x_j, y_j)s_v^j - S^j||_2^2. \tag{2.3}$$

The training phase takes $40 - 50$ iterations.

Although training takes a long time because of the iterative neural network training, resulting saliency maps mostly match with human fixations, increasing the overall ROC score of the saliency algorithm. This method is a good representation of most spatio-temporal saliency methods; it is frame based, with additional motion constraints. When we consider applications of spatio-temporal saliency algorithms such as video retargeting, we can see that *salient* parts of frames are continuous, and a suitable spatio-temporal saliency algorithm for video retargeting should focus more on the *continuity* of the saliency. An improved spatio-temporal saliency algorithm should be able to focus on tracking important objects across frames as a human viewer would do.

2.2. Image Retargeting

Image retargeting is changing the aspect ratio of images by selecting the important parts. There are three main image retargeting approaches (Figure 2.6):

- **Seam Carving.** Removing seams that contain the least amount of information from the image. A seam is defined as an irregular line of connected pixels (vertical or horizontal) [31, 32].
- **Cropping.** Finding a rectangle box that encapsulates the most important portion of the image and discarding the rest [33].
- **Warping.** Distorting (squeezing) the least important part of the image and keeping the important parts undistorted [34, 35].

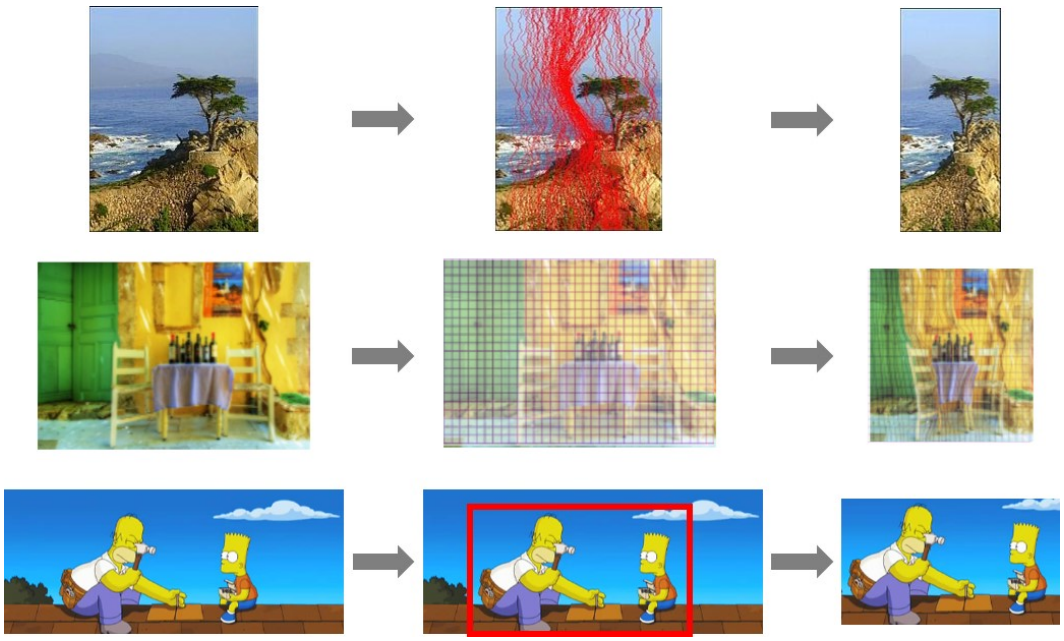


Figure 2.6. The first row is an example of seam carving. The red lines are the removed seams. The second row is an example of warping. The image is warped with a mesh where unimportant parts of the mesh are narrowed while important parts are widened. The third row is an example of cropping. The important part is selected in the red box and the rest of the image is discarded.

The major difficulty of image retargeting is to estimate spatial saliency effectively. There are numerous saliency methods that are based on HVS [20, 21, 36–39]. These approaches computationally model the selective attention process, and mainly depend on bottom-up (data-driven) features. With the help of saliency extraction, image retargeting methods are able to achieve satisfying results [31, 32, 35].

2.3. Video Retargeting

Video retargeting has gained importance with the introduction of smartphones and tablets as well as with movies being retargeted for small screens in airplanes. These have limited display sizes, and are frequently used in the display of visual content.

Figure 2.7. An example zoom-in effect. The rows are: original frames, crop window, retargeted frames successively. Although in the camera is still in the original frames, the retargeted frames show a zoom-in effect. The reason is the decreasing crop window size, which is represented with a red box in the second row.

Studies on video retargeting has started as an extension of image retargeting. Many studies of video retargeting use methods of image retargeting with adaptations to maintain the motion flow [3, 40]. These methods work well with videos that contain small amounts of motion. When the video contains fast motion, they fail to adapt to the flow and create unexpected cuts and waves.

Seam carving and warping can remove/shrink unimportant parts of frames while the removed parts are not necessarily connected. If applied on a single image, the integrity can be established with several basic constraints, but in case of a video where motion is also present as a 3^{rd} dimension, preserving the coherence across frames becomes a challenge. They can cause distortions and waving effects when the method cannot adapt to the motion of the video. If the motion is slow, distortions are generally not visible since the video is like a still image, whereas fast motion in the background or foreground can cause serious distortions and waving effects.

Cropping methods do not cause these problems. They rely on a crop window selected from each frame, which can change size or move in any direction. Since the selected parts of each frame is preserved as original, there is no quality loss as in seam carving and warping. The limitation of cropping methods occurs in a different way; when the viewers' attention in the frame is focused in a small area, the crop window

is able to capture the important parts but when viewers' attention is distributed, cropping may cause information loss. Cropping methods can also cause virtual camera movements (Figure 2.7), which is avoided in video editing since it changes the original setting and mood.

2.3.1. Recent Video Retargeting Methods

We describe here several video retargeting methods that are important in literature, and represent different approaches in video retargeting.

Rubinstein *et al.* [41] proposed one of the early works of video retargeting based on seam carving that uses forward energies to compute the most suitable graph cut. Each graph cut removes a single seam from all the frames of the video. While the original seam carving method [32] removes the seam that will cause the minimum *energy loss*, [41] proposes to select the seam that will cause minimum *energy insertion*. The approach is named as $forward\,energy$ and combined with the graph cut algorithm, it enables the method to adapt to the motion of the video in a dynamic manner. While calculating the energy gain of a possible seam removal, three possible scenarios are considered. These scenarios can be seen in Figure 2.8.

For the three cases, following adjustments are made on the pixel edges of the frames; the difference between **L**eft and **R**ight neighbors (Figure 2.8(b)), **L**eft and **U**p neighbors for upward arc (Figure 2.8(a)) and **L**eft and **U**p neighbors for downward arc (Figure 2.8(c)).

$$\begin{aligned} +L_R &= |I(i,j+1) - I(i,j-1)| \\ +L_U &= |I(i-1,j) - I(i,j-1)| \\ -L_U &= |I(i+1,j) - I(i,j-1)| \end{aligned} \tag{2.4}$$

where $I(i,j)$ is the pixel value corresponding the i^{th} row and j^{th} column of the frame. After pixel edge weights are adjusted with Equation 4.6, seams with minimum energy are removed as in [32].

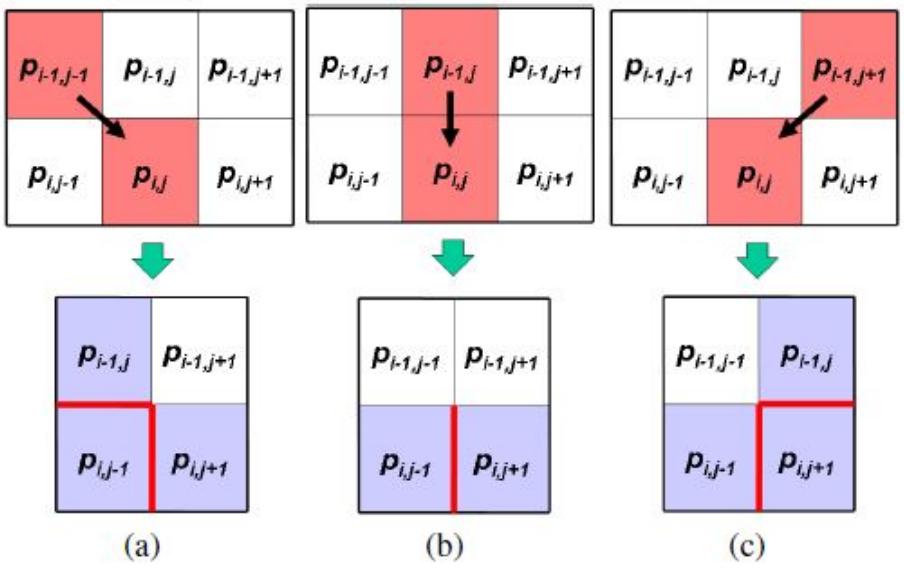


Figure 2.8. Three possible vertical seam step costs for pixel $p_{i,j}$ using forward energy. After removing the seam, new neighbors (in gray) and new pixel edges (in red) are created. In each case the cost is defined by the forward difference in the newly created pixel edges. Figure is taken from [41].

This method has become a baseline of comparison for recent video retargeting methods. It is proposed both for videos and images. In order to adapt the method to videos, the three dimensional frames in $x \times y \times t$ are rearranged into $x \times t \times y$ format. By this way, the two dimensional seam carving method is applied in the $x \times t$ dimension. While this attempt aims to adapt seam carving to video motion, it fails to produce consistent results temporally.

Yan *et al.* [5] proposes another seam carving method that improves [32] by adjusting the selected seams according to the motion of the video. They start by computing the Energy Map E_M of each frame and adjust the E_M by comparing it to the previous frame. At last, the seams that are connected, and having the minimum energy is selected and removed.

While computing the E_M at the first step, they use Sobel Operator where the gradient of each frame I_i is computed with gradient components

$$G_x = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -2 \end{bmatrix} I_i,\ G_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} I_i.$$

After the gradients of all frames are computed, the seam that will be removed from the first frame is computed directly, as in [32]. This single seam is then divided into equal vertical parts, and the points having the maximum energy within each part is selected as Key Point K_P. 10 K_P's are selected in [5].

In order to continue selecting seams from rest of the frames, the K_P's of the previous frames are used as reference points. A Matching Area M_A of 3×3 pixels is defined surrounding the K_P's of the previous frames. A Matching Index M_I defined over all pixels P in M_A of each K_P is computed as follows:

$$\begin{aligned} M_I^*(P, K_P) &= \frac{SAD(M_A^i(P), M_A^{i-1}(K_P))}{255 \times (2 \times M_W + 1)^2}, \\ M_I(P) &= min\{M_I^*(P, K_P)\}, \end{aligned} \tag{2.5}$$

where i is the current frame, SAD is Sum of Absolute Differences, M_I^* is possible M_I defined for each K_P. The minimum M_I^* is chosen to be M_I. M_W is the window size, and it is selected as three.

Starting from the second frame, a Reward-Punish map R_P is created where $M_I(P)$ is divided by 0.2. The resulting E_M for each frame starting from the second, is calculated as

$$E_M^{'}(P_{x,y}) = E_M(Px, y) \times R_P(P_{x,y}) \tag{2.6}$$

After the final E_M map is calculated, the seams having the minimum energy is removed.

This method is designed as an extension of [41], which is optimized to adapt to the motion. It can successfully remove seams from the background that are barely

noticeable. While the major downside of this behavior occurs when the motion is in the background, and the foreground objects remain unchanged. This limitation can be observed in Figure 3.2.

Another video retargeting method that uses warping based on meshes is proposed by Wang *et al.* [4]. General flow of the method is as follows:

(i) Scale frames according to the method proposed in [42]
(ii) Optimize motion pathlines of the resized frames
(iii) Resize again with computed pathline constraints
(iv) Apply cropping based on the determined *natural-width*

Resizing the image as proposed in [42] starts by generating the significance map of each frame. The significance map is denoted by $W = W_\alpha \times S$ where $W_\alpha = (\frac{\partial}{\partial x} I^2 + \frac{\partial}{\partial y} I^2)^{1/2}$ and S is the saliency map. A mesh that is applied over the significance map is used to calculate the deformation of the frame. The aim is to warp the least important quads of the mesh while applying near-linear scaling to the important quads.

An energy function $D = D_u + D_l$ is defined over quads is optimized iteratively, where D_u is quad deformation and D_l represents mesh line bending error. These functions are defined as follows:

$$\begin{aligned} D_u(f) &= \sum_{\{i,j\} \in E(f)} ||(v_i' - v_j') - s_f(v_i - v_j)||^2, \\ D_u &= \sum_{f \in F} W_f D_u(f), \\ D_l &= \sum_{\{i,j\} \in E} ||(v_i' - v_j') - l_{ij}(v_i - v_j)||^2, \\ l_{ij} &= ||v_i' - v_j'||/||v_i - v_j||. \end{aligned} \tag{2.7}$$

Here, v are the original and v' are the deformed vertex positions of the mesh. s_f is the scale factor of quad f, where the vertices v of f undergoes the transition $v' = s_f v + \epsilon$.

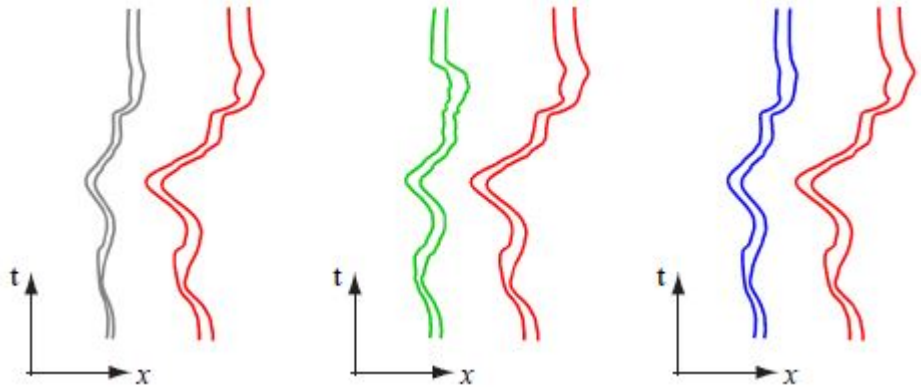


Figure 2.9. The original, linearly scaled, per-frame resized and the optimal motion pathlines are shown in red, gray, green and blue, respectively, projected onto the (x,t) plane. Note that the horizontal offsets between the pathlines are consistently reduced in the linearly scaled and the optimized trajectories. Figure is taken from [4].

After all the frames are resized by optimizing D, pathlines are extracted from the video according to the definition in [43]. The aim of this step is to ensure neighboring pathlines go under a similar transition in the resized version of the video. An explanatory plot can be seen in Figure 2.9.

The optimization is done by balancing the temporal and spacial constraints defined as:

$$\begin{aligned} \Omega_V &= \sum_{\{i,j\}\in\varepsilon} \sum_{t=m}^{n} ||((s_i p_i^t + \epsilon_i) - (s_j p_j^t + \epsilon_j)) - s_{i,j}(p_i^t - p_j^t)||^2, \\ \Omega_I &= \sum_{P_i} \sum_{t=m}^{n} ||((s_i p_i^t + \epsilon_i) - q_i^t||^2, \end{aligned} \tag{2.8}$$

.

where Ω_V represents the temporal coherence and Ω_I the spatial shape preservation. Optimization is done by minimizing $\Omega = \Omega_V + \mu\Omega_I$, and μ is selected as 0.5.

Let ε defines the set of neighboring pathlines, p_i is the position of the ϵ^{th} pixel of the pathline P, and q_i is the deformed version of the same pathline. By minimizing Ω, s_i and ϵ_i are obtained for each pathline P. The resulting optimized pathlines are then computed as $\tilde{S}_i = s_i P_i + \epsilon_i$. Optimized pathlines are used as positional constraint while resizing the video for the second time at step (iii).

The $natural - width$ is determined by applying warping with a soft constraint on the corners of the video. The soft constraint enables frames to get resized to a target size. This ensures to preserve the critical regions of each frame. The frames are then cropped so that the critical region of each frame remains.

This method is based on an optimization function that converges to the minimum error. When D is converged, the method can produce good results. This happens when the video is not highly dynamic; the possible distortions like waving effects are also not visible. When the motion is fast, the pathline constraints limit the optimization function, and it fails to converge. In such cases, the method steps back in optimization process and repeats the optimization with loosened constraints. This causes the method to perform slowly, and these cases mostly end with poorly optimized results with shaking effects between consecutive frames.

The common step in all video retargeting methods is extracting the important parts of the frames, which is typically done with saliency algorithms that automatically assign saliency to image parts. Some recent video retargeting methods [8] prefer to use image based saliency algorithms for the ease of computation whereas others [44, 45] prefer spatio-temporal saliency algorithms to increase the overall quality. Spatio-temporal saliency methods are not as successful as image saliency methods and this decreases the quality of the retargeted videos.

2.4. Discussions

In Chapter 2, we have introduced several concepts that are related with video retargeting; spatial saliency, spatio-temporal saliency and image retargeting. We have provided brief explanation of important concepts, and provided several state-of-the-art studies that are important for the following Chapters. We have discussed that the spatio-temporal saliency calculation is an important step for video retargeting applications, thus it effects the video retargeting results.

We have observed the limitations of several video retargeting methods which helps us shape the boundaries of our proposed method. We have seen that these limitations differ according to the retargeting approach used in the method, and they should be observed and be understood in detail in order to propose a new video retargeting method. The major limitations are related with the motion in the video, and different video retargeting approaches may handle the motion in different ways.

In addition to this, most video retargeting methods use an existing spatial or spatio-temporal saliency algorithm. We have seen that the methods that proposes their own saliency algorithm may achieve better results [4]. This also supports the importance of the motion in the video, on the results of the video retargeting method.

3. PRELIMINARY STUDY

3.1. Motivation

In order to propose a valid and meaningful video retargeting method that will produce good results, we first observed the current state-of-art video retargeting methods. We have implemented several video retargeting methods that uses different techniques, and detected the cases where they perform well, and the cases where they produced distortions.

While most video retargeting methods perform well on images, when applied to videos, they may cause waving or jumping effects. Thus, we have argued that the main limitation of all video retargeting methods is related with the ability to include motion information in the retargeting procedure. This observation leads us to test different video retargeting methods on dynamic and static scenes separately, and compare the results. A dynamic scene is defined to have a high amount of dynamic content, like a running person, or a fast camera motion, while static scenes mostly lack of any kind of motion.

Not surprisingly, the resulting quality of the static and dynamic videos differ remarkably. We have concluded that there is no video retargeting solution that works well with all types of videos. Depending on the distribution of the content, motion of the camera and amount of texture, the existing retargeting approaches will fail in some videos, and succeed in others. Thus, for the preliminary study, we propose a hybrid approach to remedy some of the shortcomings of current video retargeting methods.

Warping or seam carving based methods remove parts of each frame, which are not necessarily near the edges. When applied to videos, these method must ensure that the parts they remove from consecutive frames are consistent. This is a challenging task, and if this condition is not satisfied, retargeted videos can end up with artifacts. Thus, seam carving and warping methods cause distortions and waving effects when

the method cannot adapt to the motion of the video. If the motion is slow, distortions are generally not visible, whereas fast motion in the background or foreground can cause serious distortions and waving effects.

Cropping methods do not cause distortions or waving effects. They rely on a crop window selected from each frame, which can change size or move in any direction. The size and the movement of the crop window should be perfectly consistent with the original camera motion of that shot. Any change in crop window that differs from the original camera motion causes virtual camera movements. An expanding crop window causes additional zoom-out effect or a sudden jump is perceived as an artificial scene cut. Thus, the output quality of a crop based method can be measured with the preservation of motion flow, and its success in avoiding virtual camera motion. The output quality is not affected by the fast motion in the video. It can be argued that videos containing fast motion are more suitable for crop based methods, since the center of attention in a frame is focused. In videos that contain slow motion, the attention is distributed across the frame, making it harder to crop.

3.2. Preliminary Approach

In the preliminary study, we propose a hybrid video retargeting method that will analyze the input video, and apply the most suitable video retargeting method per shot. The analysis is aimed to determine the amount of dynamic content of a shot of a given video. Applying the most suitable retargeting method helps remedy the limitations related with the motion, or distributed content.

While a video clip may contain multiple shots, we analyze each shot separately. Let n be the number of shots in a given video. Each shot has a dominant motion class M_i, $i = 1, 2, ..., n$ where $M_i \in \{fast, slow\}$. Our aim is to detect M_i, $\forall i$ to apply the most suitable video retargeting algorithm per shot.

We have trained a Support Vector Machine (SVM) ϕ in order to identify the motion class of a given frame. The output of ϕ gives the motion class of each frame.

The class that occurs most frequently in a given shot becomes the dominant class of that shot. The ground truth class labels are assigned by annotation. For training the SVM, we have used a radial basis function kernel with sequential minimal optimization method [46].

We have applied two different video retargeting algorithms according to the motion class we get from ϕ. We propose a novel cropping approach for shots belonging to the $fast$ motion class. For shots belonging to the *slow* motion class, we have applied an improved seam carving approach proposed in [5]. Note that this hybrid approach can also be used with different video retargeting methods.

3.2.1. Identifying the Motion Class

Let $H = h_1, h_2, ..., h_{i-1}$ be the set of homography matrices between consecutive frames of a shot, where i is the number of frames of a shot. The motion class of a frame i can be found by taking into account a window of frames, expressed by the homography matrices.

$$H^i = \{h_{i-2}, h_{i-1}, h_i, h_{i+1}, h_{i+2}\} \tag{3.1}$$

$$M_i = \phi(\sigma_{H^i}, \mu_{H^i}), \tag{3.2}$$

The *slow* motion class represent the frames that contain minimum action and a rather slow camera motion, whereas the $fast$ motion class contains frames having a rapid camera movement or an active action in the frame. The classes do not give any information about the action occurring in a frame or about the type of the camera motion. They rather represent the overall amount of motion that the frame contains.

Two motion classes $\{fast, slow\}$ are sufficient for our purpose since they are able to represent the limitations of current video retargeting algorithms. Seam carving and warping based approaches have a good quality performance when the motion is

classified as *slow*. Distortions in the background and waving effects are minimal. When the motion class is $fast$, the quality of the output video decreases (Figure 3.2).

We have chosen to use homography matrices as a representative of the overall motion. A homography matrix of an affine transformation provides a mapping between two consecutive frames indicating a general information about the change between frames. This change is used as a measure of the dynamic content. The reason that we have decided to use the affine homography matrices to represent the dynamic content is that they give general representation of camera motion [30]. We did not prefer using the optical flow maps since they provide a local representation of motion, whereas the dynamic content is defined over the whole frame.

3.2.2. Choosing Video Retargeting Method

For the shots belonging to the *slow* class, we have applied [5], for $fast$ videos, we propose a cropping based video retargeting method. We first run a recently proposed spatio-temporal saliency algorithm [22] on each frame of a given shot and threshold the saliency maps to get the important points of each frame that the cropping method must cover. The selected spatio-temporal saliency algorithm is proposed specifically for videos and also takes into account of the camera motion. The details of this method can be found in Section 2.3.

After extracting the important points to cover in each frame, we proceed with defining a valid crop window location, represented with the crop window center. The last step is to define the crop window size for each shot, and to apply cropping. The steps following the spatio-temporal saliency estimation can be found in Section 4.3, Figure 4.1.

3.3. Results

3.3.1. Dataset

We have evaluated our approach on the Camera Motion (CAMO) dataset [22]. CAMO dataset contains 120 short video clips of six different camera motions. Each video contains a single camera motion in a given shot. The camera motion labels provided by CAMO dataset [22] are as follows:

- Tilting: the camera is stationary and rotates in a vertical plane.
- Panning: the camera is stationary and rotates in a horizontal plane.
- Dolly: the camera is mounted to the dolly and the camera operator and focus puller or camera assistant, usually ride on the dolly to operate the camera.
- Trucking: roughly synonymous with the dolly shot, but often defined more specifically as movement, which stays a constant distance from the action, especially side-to-side movement.
- Pedestal: moving the camera position vertically with respect to the subject.
- Zooming: Technically this is not a camera move, but a change in the lens focal length with gives the illusion of moving the camera closer or further away.

3.3.2. Output Quality Evaluation

We verify the performance of the proposed approach visually, on a set of videos selected for their diversity of motion and other conditions like scene clutter, and content. While visual evaluation is only a preliminary step for an extensive evaluation, as stated in Section 3.1, the aim of the preliminary study is to observe the possible deficiencies of different video retargeting methods. In order to achieve our aim, we have selected both *fast* and *slow* videos, and apply two different retargeting approaches. Figure 3.1 shows several examples taken from the visual evaluation First two rows are taken from the *slow* class. Columns show the original frames, results of linear scaling, seam carving and proposed cropping approach. Since the video belongs to *slow* motion class, salient points are not focused on a specific object, but are rather distributed

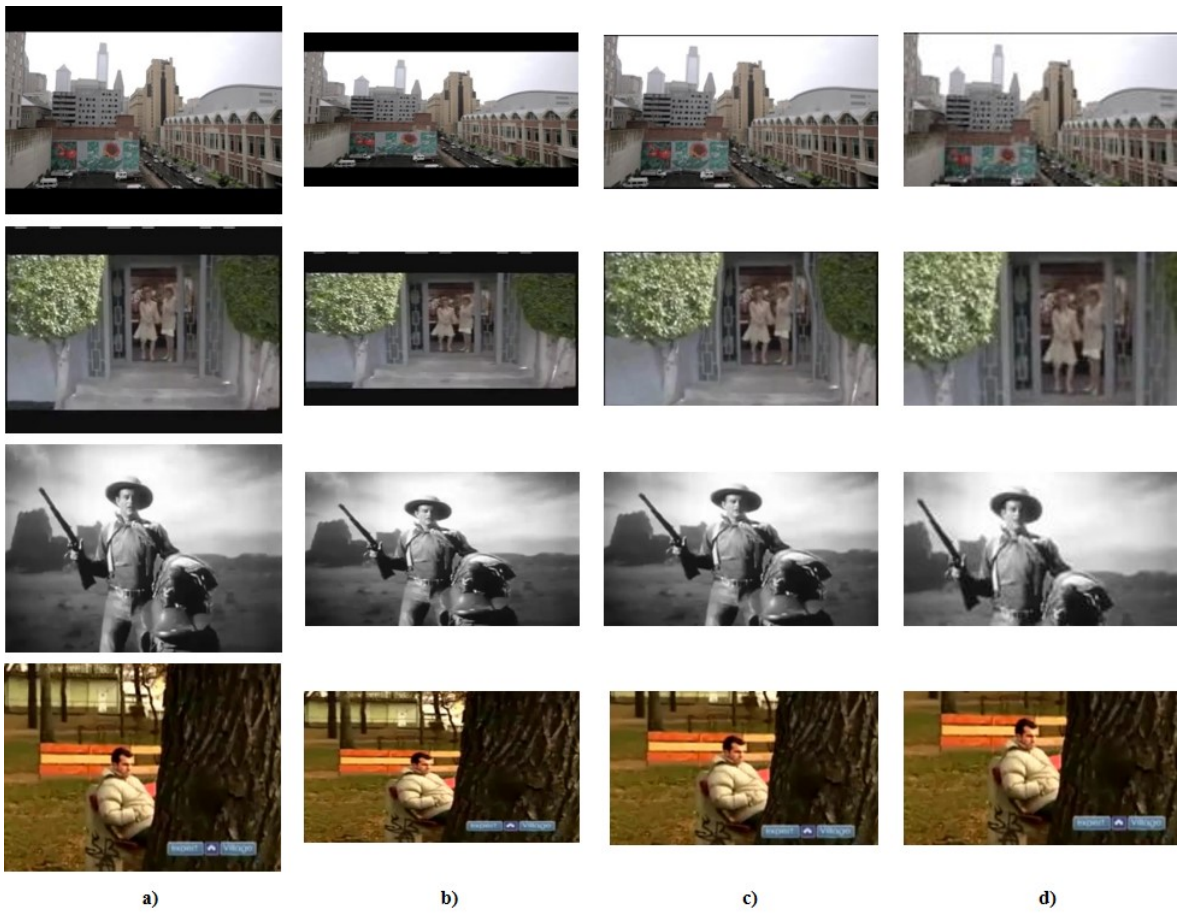


Figure 3.1. a) The original frames. b) Results of linear scaling. c) Results of seam carving by Yan *et al.* d) Results of proposed crop based retargeting. Figure is best viewed in color, where problems of both approaches become obvious.

across the frame. These two cases illustrate the limitations of the crop based method, which tries to capture all the salient points and producing inefficient results. Since the motion in these videos is *slow*, there are no waving effects on seam carving results.

The last two rows in Figure 3.1 are frames taken from videos of the $fast$ motion class and illustrate the limitations of seam carving. In both frames, saliency is focused around a center, more specifically around faces in this case. This makes the crop based method more effective since the location to crop is easy to determine. Because of the fast zooming motion in the videos, seam carving is not able to adapt to the camera motion and produces waving effects in the background.

Figure 3.2. a) The original frames. b) Results of linear scaling. c) Results of seam carving by Yan *et al.* d) Results of proposed crop based retargeting. Figure is best viewed in color, where problems of both approaches become obvious.

One of the main artifacts seam carving causes is the waving effect that occurs on the background. These effects are visible in the video, but hard to demonstrate on frame representation as in Figure 3.1. In order to show these waving effects seen in the video, we have also provided a set of consecutive frames taken from a $fast$ and a $slow$ video on Figure 3.2. Figure 3.2 contains frames from two video sequences. In the first case, the video belongs to the $fast$ motion class. We can observe the distortions due to the high dynamic structure of the video. On the second frame sequence, the camera motion is slow and the salient content is distributed. This is the main characteristic of a $slow$ video. In such case, cropping misses some important parts of the frame.

Visual evaluation supports our observation that video retargeting methods have different limitations, occurring on different types of videos. Visual evaluation on two main motion classes of videos, with two main types of video retargeting methods shows that seam carving approaches are more suitable for $slow$ videos, whereas cropping approaches suits better to the $fast$ videos.

3.3.3. Evaluation of the Motion Classifier

Originally, CAMO dataset was annotated according to the camera motion. In addition to this, we have annotated 36 of the videos according to their motion class; $fast$ motion or $slow$ motion. While the camera motion information helps the annotation process (e.g. most zoom-in videos belong to the $fast$ motion class), the camera motion does not necessarily indicate the motion class of a video. For example, a zoom-in camera motion can occur both in a fast, or a slow way, and this affects the motion class of the overall video.

The motion classifier is being tested on annotated movies. We have divided movies into sets of five consecutive frames such that each frame is included in only one set. The frames belonging to a shot inherit the shot's motion class. We have used 200 frames from each class for training, and 100 frames from each class for testing the SVM. The confusion matrix, precision, recall and F-Score results of the test can be seen in Tables 3.1 and 3.2.

Table 3.1. Classification results. Columns represent the ground truth and the rows represent test results.

	Fast	Slow
fast	60	14
slow	40	86

Table 3.2. Performance measures of SVM.

	Precision	Recall	F-Score
fast	.81	.60	.68
slow	.68	.86	.75

Precision, *Recall* and $F-Scores$ are generic metrics used to evaluate classifiers [47–49]. For our case, *Precision* answers the question of how many $fast/slow$ motion class labels that SVM predicts are actually belongs to that motion class. While *recall* gives information about the amount of the $fast/slow$ frames that the classifier predicts correctly. The formulas of the metrics are as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives},$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}, \tag{3.3}$$

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

Table 3.2 indicates the that the *slow* motion class has slightly better results. The movies belonging to *slow* motion class tend to keep a low amount of motion throughout the video. As opposed to that, the movies belong to the $fast$ motion class do not necessarily contain a fast motion all the time. There can be times where the camera motion decreases or the action slows down. Subsequently, the slow class has a higher accuracy. The samples that correspond to these times can be classified as *slow*, even though the overall movie is in the $fast$ class. While SVM classifies the motion classes of separate frames, the video retargeting method is preferred according the the motion class of the shot.

3.4. Discussion

In the preliminary study, we proposed to use homography to identify a given video according to the rate of change in its dynamic content, and applied seam carving or cropping based video retargeting approach depending on the result. We used a recent video saliency approach to keep track of relevant content, and proposed a novel cropping method to eliminate virtual camera motion. The resultant hybrid algorithm produced good qualitative results on the CAMO benchmark. We have published the preliminary study in [50].

After completing the preliminary study, we have identified a major limitation of the cropping based video retargeting method, which is preventing information loss, as well as keeping the parts of the frame that are not necessarily salient, but may be important for human viewers (such as the hair on top of a head may not not appear as salient in spatio-temporal saliency maps, but it is important to include it in the retargeted video for a better viewing experience). While keeping all salient points in videos having low dynamic content, a successful cropping method should be able to do so. Equipped with these insights from the preliminary study, we continue to improve the cropping-based video retargeting method to perform with a more suitable spatio-temporal saliency algorithm that will prevent the limitations of the current method.

4. PROPOSED METHOD

4.1. Motivation

After completing the preliminary study, we have seen that the major limitation of the proposed cropping method is its inability to capture all important parts of a frame when the video contains low amount of dynamic content. These videos can be considered as lightly moving images, where viewers take time to observe the scene, thus, the attention is distributed in a frame. The root of this limitation lies in the spatio-temporal saliency algorithm. Even thought the spatio-temporal saliency algorithm used in the preliminary study produces high quantitative results (ROC score), we were not able to achieve satisfying results in videos having a low dynamic content. As a result, we continue focusing on improving the spatio-temporal saliency part of the cropping method.

When a video is being retargeted, important objects must be identified correctly, and then captured as a whole. Even though some parts of the followed object do not appear as important in the traditional spatio-temporal saliency maps, a suitable spatio-temporal saliency algorithm should amplify the whole object, so that the resulting cropped video does not contain any disturbing cuts.

Since our aim in this thesis is a full stack video retargeting method, including a suitable spatio-temporal saliency algorithm, the saliency algorithm is designed specifically to improve results of video retargeting. We already have a cropping method at hand after the preliminary study; the improvements to preliminary study mainly focus on the spatio-temporal saliency algorithm. In order to represent the overall proposed method as a whole, we have included the description of both spatio-temporal saliency and the cropping method in this chapter. The detailed flow of the proposed method can be seen in Figure 4.1.

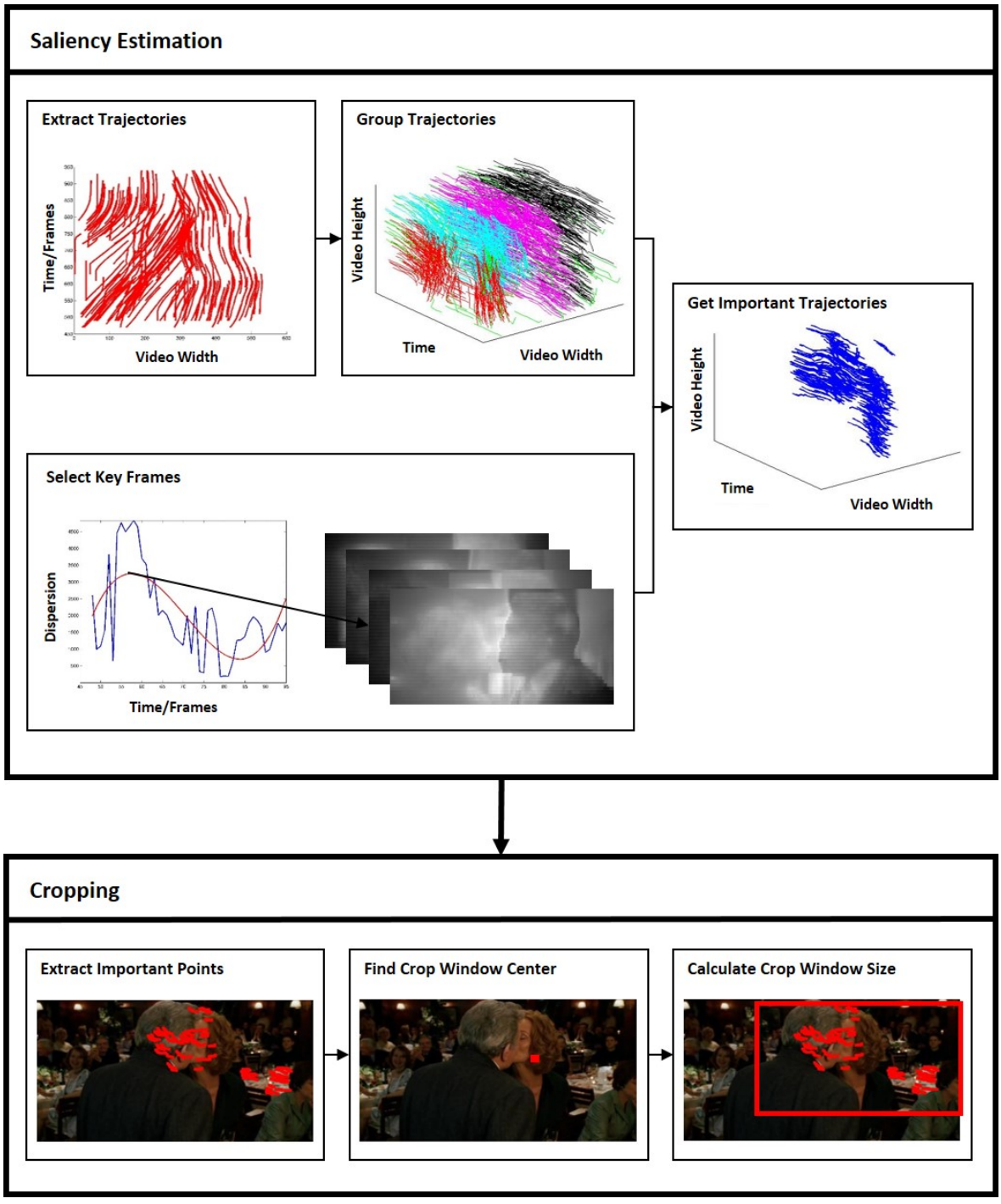


Figure 4.1. Overall flow of the proposed method.

4.2. Extracting Spatio-Temporal Saliency

For video retargeting, motion saliency of each frame provides a limited representation of motion continuity. The reason that motion continuity is more important than discrete motion representation such as optical flow maps or homography matrices can best be seen in eye fixation data of the videos. Figure 2.4 shows the eye fixations of consecutive frames. As discussed in Section 2.1.2, fixations of videos have a continuous structure, so viewers do not directly concentrate on the *moving parts* of the frame, but *follow* a specific moving object.

Motion continuity can best be seen in video cubes rather than in discrete frames, as shown in Figure 4.2. A suitable way to represent motion continuity is using motion trajectories. Motion trajectories are a representation of the optical flow map, and they also provide connectivity between consecutive frames, thus, we have used motion trajectories to represent the connection in the video cube.

When we observe the Figure 4.2, we can see that the trajectories form groups, while a group can be defined as trajectories that are *following a similar path* and also *close to each other*. Each group of trajectories correspond to an object in the scene; for the example in Figure 4.2, these objects are a man, a women and the gap between them. Some trajectory groups are overlapping with the eye fixation data. This points that trajectories can be used to represent the eye fixations and provide a spatio-temporal saliency map, but it is necessary to select the important trajectories.

Gathered from the observations, the steps of the proposed spatio-temporal saliency algorithm can be listed as; (i) Extracting the trajectories, (ii) Grouping the trajectories, (iii) Selecting the important groups to preserve in the retargeted video. The steps ensure that the important objects will remain in the retargeted video as a whole, since they are followed through the shot, even though the the video has low dynamic content, and attention is distributed. Details of these steps can be found below.

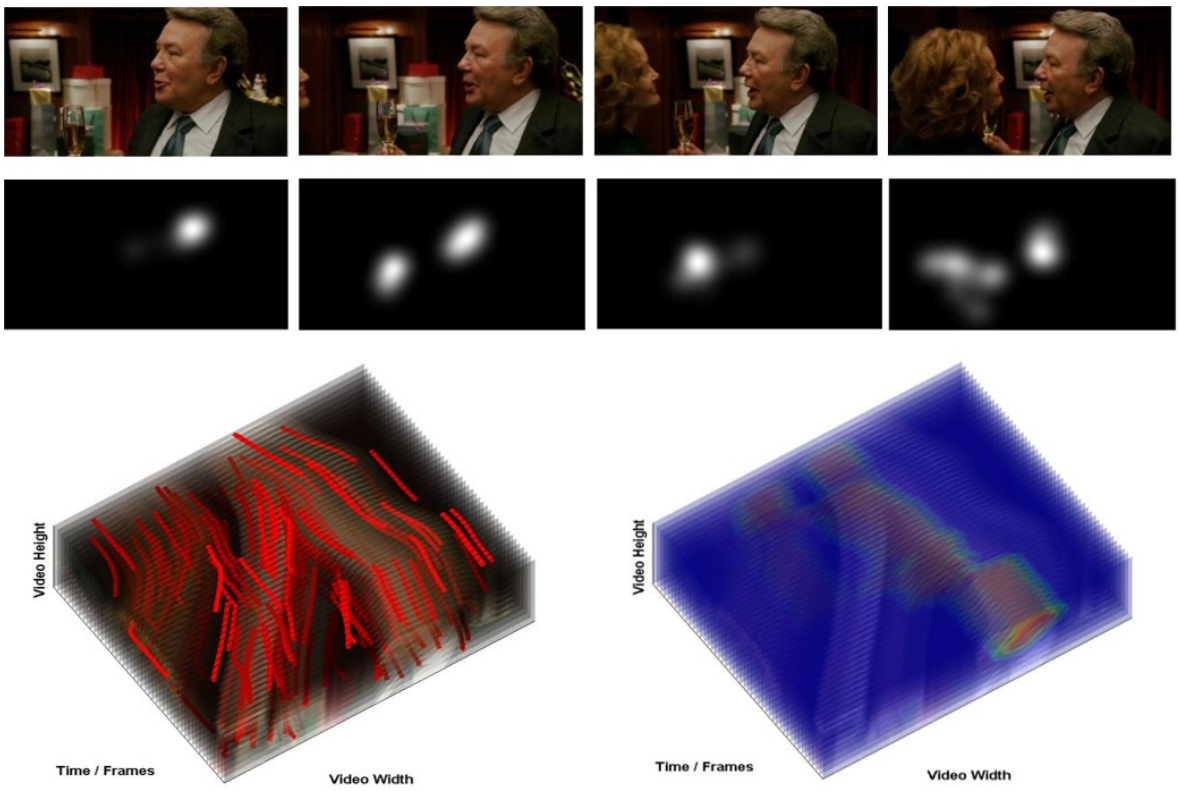


Figure 4.2. Illustration of spatio-temporal saliency. Frames are taken from one shot of a clip of the Hollywood2 dataset [51]. The first row shows a sequence of frames and the second row is the corresponding ground truth saliency maps. On the third row, we can see the trajectories mapped into a spatio-temporal video cube and heat maps extracted from fixation data.

4.2.1. Extracting Trajectories

The first step of the spatio-temporal saliency algorithm is to compute trajectories from a given video. We have used Improved Trajectories proposed in [53], which are the improved versions of Dense Trajectories [52]. While computing Dense Trajectories, they sample feature points from a dense grid, and follow the feature points with the help of a dense optical flow algorithm [54]. At each frame, for each tracked feature point, they use local descriptors designed to characterize shape, appearance and motion of the neighboring pixels.

The shape descriptors of the trajectories are defined over the displacement of the

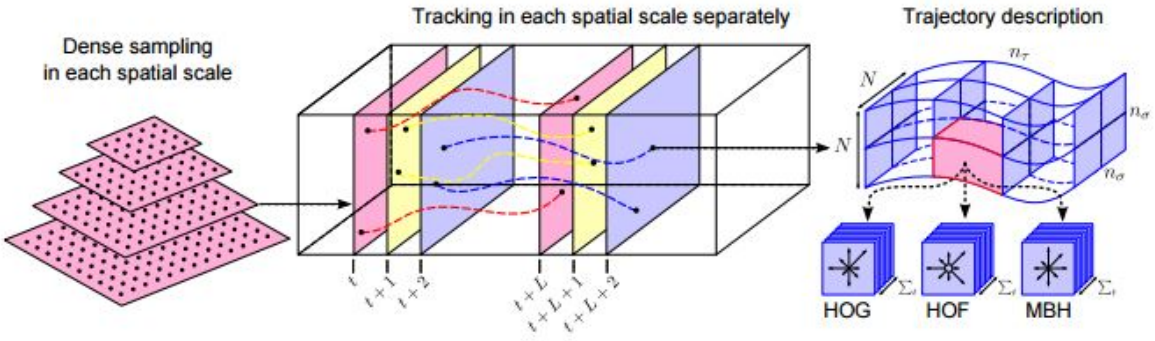


Figure 4.3. Illustration of the approach to extract and characterize dense trajectories. Left: Feature points are sampled. Middle: Tracking is carried out. Right: The trajectory shape is represented by relative point coordinates, and the descriptors (HOG, HOF, MBH) are computed along the trajectory in a $N \times N$ pixels neighborhood, which is divided into $n_\sigma \times n_\sigma \times n_t$ cells. Figure is taken from [52].

trajectory. Let

$$\Delta P_t = (P_{t-1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t) \tag{4.1}$$

be the shape of the trajectory at time t, the trajectory itself is defined as

$$T = \frac{(\Delta P_t, \ldots, \Delta P_{t+l-1})}{\sum_{j=t}^{t+l-1} ||\Delta P_j||}, \tag{4.2}$$

where l is the length of the trajectory. In addition to the above trajectory definition, at each point, Histogram of Oriented Gradients (HOG) [55], Histogram of Optical Flow (HOF) [56] are used as appearance and motion descriptors, respectively. Motion Boundary Histograms (MBHx and MBHy) [57] are computed over the derivatives of the optical flow.

Improved Trajectories is implemented over Dense Trajectories by taking into account of the camera motion. They have extracted SURF features [58] and estimated the homography matrix between frames with RANSAC [59]. The homography matrix represents the camera motion, which is removed from the optical flow field. Removing camera motion enables them to capture the moving parts in the video.

While there are several descriptors accompanying the raw motion trajectories, in action recognition field -which is the actual field of both Dense and Improved Trajectories studies-, HOG, HOF and MBH descriptors improve action recognition results remarkably. While these descriptors are important for action recognition, we have not used them in our study, but we have extracted the original trajectories instead.

The reason we have used Improved Trajectories is to ensure that the trajectory groups represent the whole objects that they follow. Dense trajectories extracted from a dense optical flow field yield a successful to a successful grouping. On the other hand, the major limitation of the improved trajectories method is its time complexity. Extracting improved trajectories with all features enabled can take up to 200 times of the length of the original video on a two core machine with 16GB RAM (e.g. a 10 second video can take up to 2000 seconds).

4.2.2. Grouping Trajectories

Trajectories that are following the same object tend to move together. We propose to define *trajectory groups* that will follow a similar path. When a trajectory group is defined as important, all trajectories belonging to than group are labeled as important. A trajectory can only belong to a single group since trajectory groups represent objects.

Let

$$\tau = \{T^1, T^2, \ldots, T^n\}, n \in \mathbb{N} \tag{4.3}$$

be the set of trajectories extracted from a shot of a given video where n is the number of trajectories. Each T^i defines a line of pixels

$$T^i = \{P_1, P_2, \ldots, P_l\}, \tag{4.4}$$

where l is the length of the trajectory, spanning the time slot between $t^{th} - (t+l)^{th}$ frames. Each point P of a trajectory is a pixel location x, y. The trajectories are

computed according to [53]. The length of the trajectories are an input to the system, which is chosen to be $l = 15$ since it was the default trajectory length defined in [53]. When a trajectory T^i ends at pixel P_l^i, a new trajectory is seeded. We define the distance between trajectories as follows:

$$D_{i,j} = \begin{cases} \frac{1}{l}\sum_{t=1}^{l} ||P_t^i - P_t^j||, & \text{if } b_i = b_j. \\ \frac{1}{l-1}\sum_{t=2}^{l} ||P_{t-1}^i - P_t^j||, & \text{if } b_i = b_j - 1. \\ \infty, & \text{otherwise,} \end{cases} \tag{4.5}$$

where b_i is the start frame of the trajectory T^i. T^i and T^j belong to the same group, if $D_{i,j} < \delta$.

It is important to adjust the δ value to be able to distinguish between objects. When it is too high, then most of the trajectories will belong to a single group and when it is too low, a trajectory group will fail to represent the whole object. The appropriate value for the δ depends on how dense the trajectories are seeded. During our studies, we find that δ should be smaller than 10 pixels in order to distinguish between objects successfully. In Figure 4.1, Grouping Trajectories part shows an example with a threshold of 10.

4.2.3. Selecting important groups

The key point is to understand which objects are important enough to be tracked. We use saliency maps as a guide to detect these objects.

Detecting important objects have two main cases, (i) following an object that is present from the beginning of the shot, and (ii) detecting an important object that is not present at the beginning, but enters the scene during the shot. First case is rather straightforward since we know the frames that are important to be investigated. For the second case, we should first detect the frame that the new important object is entering the scene.

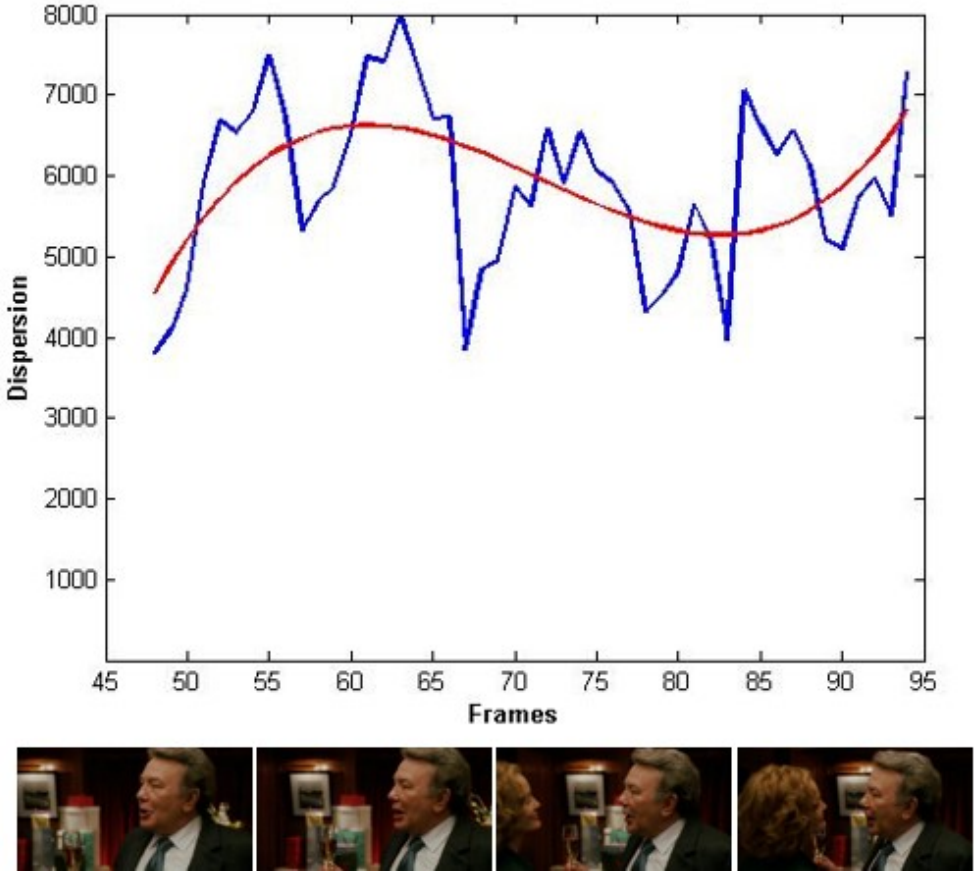


Figure 4.4. An example dispersion graph. Images correspond to the 50^{th}, 60^{th}, 70^{th} and 80^{th} frames respectively. As the woman enters the scene between $60^{th} - 70^{th}$ frames, the dispersion starts to increase.

As we have observed the ground truth saliency maps of videos, we have seen that when a new important object enters the scene, the eye fixations start moving to the newly entering object and fixations becomes more *distributed* across the frame. This shows that the frames having more distributed eye fixations than the rest of the shot may contain newly introduced important objects.

In order to estimate the attention distribution among the frames, we compute static saliency maps of the frames by using [21], apply a threshold ζ and calculate the saliency dispersion E of each frame. We have experimented with different ζ values, since the static saliency maps are normalized between $0 - 1$ values, we have chosen

$\zeta = 0.7$.

$$E^f = \frac{1}{(N^f)^2} \sum_{i,j<i} ||P_i^f - P_j^f||^2 \tag{4.6}$$

where P^f are the pixel locations of frame f for which the saliency map has a non-zero value and N is the total number of such pixels. The reason that we first apply a threshold is to eliminate the unimportant and *noisy* parts of the saliency map. While the saliency map produces a lot of salient points, we are looking for the distribution of the *most important parts*. Dispersion quantifies the distribution of the important content across the frame [28]. An increase in the dispersion rate may be a sign of distributed attention, such as an important object entering the scene and competing with existing objects. An example can be seen in Figure 4.4.

4.2.4. Selecting key frames and getting seed trajectories

After assessing the dispersion of attention for each frame, we proceed to choose the frames that are important enough to search for newly entering objects. We name these important frames as $key\ frames$. K is the set of $key\ frames$ where $\frac{dE}{df} = 0$, meaning that we are searching for the frames where dispersion has a peak.

Key frames are thresholded and used to find the set of seed trajectories T_S.

$$T^i \in T_S \iff K^k(P_1^i) > 0 \ \wedge \ b^i = k \tag{4.7}$$

Seed trajectories show us which objects are important, and worth following. Trajectories that belong to the same group with seed trajectories should be preserved. They will be used at the next step for video retargeting.

4.3. Cropping Based Retargeting

After computing the spatio-temporal saliency map, we proceed to explain the details of the cropping method. Main challenge of a cropping based video retargeting method is to avoid virtual camera movements and to preserve salient objects. We propose an optical flow based cropping algorithm that is able to adapt to the camera motion to minimize virtual camera movements while preserving salient objects.

The flow of the cropping method can be found in Figure 4.1. We start by defining the important points to be kept in the retargeted video, and proceed with defining center of crop window for each frame. The centers are aligned to be in correspondence with the camera motion. The last part is to define the crop window size while ensure preserving the aspect ratio of the original video.

4.3.1. Extracting important points

Important points are extracted from the selected trajectories at the spatio-temporal saliency computation step. All points covered by the selected trajectories are defined as important, and preserved in the retargeted video. First, seed trajectories are used to create spatio-temporal saliency maps S.

$$S_{i,j}^{n} = \begin{cases} 1, & \text{if } P_t^k = \{i,j\} \ \& \ b_k + t = n. \\ 0, & otherwise. \end{cases} \tag{4.8}$$

where n is the number of frames. The paths that the seed trajectories pass through become salient. After computing S for each frame, we calculate the center of saliency C^i by taking the mean of the salient pixels in S^i. The center of saliency is used as a guide the select a crop-window at the retargeting step.

4.3.2. Finding the center of crop window

The center of crop window should follow the camera motion to avoid artifacts. To estimate the camera motion, we first calculate the optical flow map O as proposed in [29], we then define a threshold ψ that is used to define the moving parts of a frame. Once we remove the moving parts of each frame, we obtain the camera motion.

$$\eta = \overline{O} * \psi, \tag{4.9}$$

where $\overline{O}$ is the mean of the optical flow map and ψ is chosen to be 0.9. Any region of the O that exceeds η in both positive and negative directions is defined as moving parts.

$$\widetilde{O}_{x,y} = \begin{cases} O_{x,y} & (-\eta < O_{x,y} < \eta) \\ 0 & otherwise \end{cases} \tag{4.10}$$

$\widetilde{O}$ is the updated optical flow map, where pixels having a value smaller than threshold $-\eta$ (or greater than η) are removed. Removed pixels correspond to moving objects in the frame, since their value diverges from the average. The mean of the updated optical flow map $\mu_{\widetilde{O}}$ corresponds to the camera motion. We have normalized $\mu_{\widetilde{O}}$.

For each frame, we calculate the center of the crop window. These centers create a flow C^o for the crop window. This flow should be smooth, and should follow the camera motion in order to avoid virtual camera movements in x and y directions.

$$C_1^o = (C_1^i + C_2^i + C_3^i)/3, \tag{4.11}$$

where C^i is the center of the saliency map. The centers of saliency maps of the first three frames are used to define the beginning of the center of crop window flow C_1^o.

$$C_i^o = C_{i-1}^o + \mu_{\widetilde{O}_i} \quad i = 2, 3, ..., n. \tag{4.12}$$

For the rest of the frames, centers are shifted by the camera motion $\mu_{\widetilde{S}^o}$ of the current frame using Equation (4.12).

Crop window size estimation. After determining the center of crop window for each frame, we estimate the size of the crop window, which is fixed for each shot in order to avoid virtual camera movements in the z direction.

$$\Gamma = (x, y) \quad \textit{such that} \quad S_{x,y} > 0 \tag{4.13}$$

The crop window size Γ is determined as the minimum size possible that will include all points in Γ. The points in Γ that cause the crop window to exceed the frame boundaries are discarded. Since the size of the crop window is fixed per shot, virtual camera movement in the z direction is eliminated. The video saliency computation ensures that the relevant objects are included in the cropped frame.

4.4. Conclusion

As a summary, the proposed approach contains two main steps; a spatio-temporal saliency calculation using motion trajectories, and a cropping video retargeting method. The spatio-temporal saliency algorithm designed to improve video retargeting results, while the cropping method relies on the spatio-temporal saliency algorithm to define the important parts of the video to capture. As a result, the most important objects are always kept in the cropped video by avoiding disturbing cuts.

5. EVALUATION

This chapter includes quantitative and qualitative evaluation of the proposed method as well as the correlation between these results. The organization of the chapter is as follows; we first conduct a research on recent video retargeting evaluation methods. We then proceed with the quantitative analysis. For the spatio-temporal saliency algorithm, we have used a common quantitative metric to compare it with two other saliency methods. For the results of cropping, we have used 12 general video quality measures applied to the results of six different retargeting methods. For qualitative evaluation, both visual evaluation and a user study are provided. At the end of the chapter, we provide the correlation between the user study and the video quality metrics in order to provide a common ground for video retargeting evaluation, and propose a new metric by fusing the results of 12 measures.

5.1. Related Work In Video Retargeting Evaluation

Video retargeting results can be evaluated both qualitatively and quantitatively. While quantitative evaluation is not as common as qualitative evaluation, conducting a user study is the most common way to do the quantitative analysis. We have surveyed the recent video retargeting studies to find out about the common ways of evaluating these methods. We summarize our findings in Table 5.1.

As can be seen in Table 5.1, performing a user study is common, while only few studies do a quantitative evaluation. The dominance of user study over quantitative metrics points to a lack of baseline for quantitative evaluation. Several state-of-the-art methods that include quantitative studies propose new metrics for output video quality assessment.

Lin *et al.* [10] uses correlations between motion trajectories. They extract motion trajectories from both original and retargeted videos, match the trajectories and compute their correlation. The correlation should be high for a successful video re-

targeting. This metric is proposed as an intermediary step of their video retargeting approach. The correlations are calculated, and they are optimized during the retargeting process, so the resulting video has the best possible results. This metric is not suitable to apply to all video retargeting methods because it is based on pathline correspondence between original and retargeted video. If the compared retargeting method is not based on pathline optimization, their results will decrease directly regardless of their visual quality.

One of the main limitations of most video retargeting methods is the temporal incoherence. Bo *et al.* [8] define a jittery metric that measures the amount of warping occurred on the retargeted video. They utilize a grid structure in their video retargeting method and warp frames according to the informativeness of each grid cell. The change of each grid cell in the retargeted video is calculated and used to compute the Jittery Metric. Although the idea of a jittery metric that is designed for video retargeting is brilliant, this metric shares the same limitation with [10]. In order to compute it for an arbitrary video retargeting method, the retargeted videos should be reverse engineered and the grid structure should be extracted. This may be problematic for example in crop based methods since the retargeted video and the original video will not have comparable grids.

Wang *et al.* [60] measures deformation by comparing salient curves extracted from original and retargeted videos. In addition, they propose a temporal consistency metric that compares the optical flow maps of retargeted frames by taking the original optical flow maps as ground truth. While extracting salient curves can not be applied on all retargeting approaches, extracting optical flow maps are straightforward, so the temporal consistency metric can be extended for all video retargeting methods. The downside of this approach is its runtime to perform the comparison of two retargeting methods. Since the first step is to compute optical flow maps of both original and retargeted videos, the overall procedure is lengthy.

Two major limitation of various video retargeting approaches are information loss and distortion. Because of this fact, Wang *et al.* [9] defines the quality of a video

retargeting method over two metrics; Information Maximization (IM) and Deformation Minimization (DM). Cropping based methods do not cause distortion or any related artifacts, but they suffer from information loss, which can be measured with IM. On the other hand, the opposite applies for seam carving and warping methods. DM metric aims to measure the distortion of the latter group. They have made an extensive study to create ground truth labels of informative areas in videos to measure IM. DM is defined over aspect-ratio changes of grid cells. While this measure for video retargeting is appropriate for all cropping, warping and seam carving approaches, it requires an extensive annotation process.

Most of the metrics in video retargeting are defined specifically for their own retargeting methods. They either require a detailed annotation, or for different reasons cannot be applied uniformly to different retargeting methods.

We propose to use a selection of video quality metrics that will be applicable to all videos. We use these metrics to reveal the correlation between video quality and the conducted user study. These metrics can be used as a baseline to evaluate future video retargeting methods. We have used 12 generic image and video quality metrics and compare our results with four other video retargeting methods. In addition, we have used a state of the art spatio-temporal saliency algorithm and applied cropping video retargeting method.

Table 5.1. Recent Approaches in Video Retargeting Evaluation. * correspond to the comparison methods used in our evaluation.

	Number Of Subjects in User Study	Number of Videos in User Study	Total Number of Comparisons	Quantitative Evaluation	Visual Evaluation
[Rubinstein *et al.* 2008]*	*no user study*	*no user study*	*no user study*	*no quantitative metric*	3 videos & 7 images
[Wang *et al.* 2011]*	*no user study*	*no user study*	*no user study*	*no quantitative metric*	4 videos
[Nguyen and Won 2013]	*no user study*	*no user study*	*no user study*	*no quantitative metric*	3 videos & 4 images
[Nie *et al.* 2013]	*no user study*	*no user study*	*no user study*	*no quantitative metric*	4 videos
[Liu *et al.* 2014]	3	3	**9**	*no quantitative metric*	2 videos
[Wang *et al.* 2014a]	20	10	**117**	Temporal Consistency & Salient Curve Deformation Metrics	7 images
[Wang *et al.* 2014]	9	10	**270**	Information Maximization & Deformation Minimization Metrics	25 videos & 4 images
[Yan *et al.* 2013]*	30	5	**450**	*no quantitative metric*	4 videos
[Yan *et al.* 2014]	20	5	**600**	Jittery Metric	19 videos
[Lin *et al.* 2013]	90	8	**2520**	Temporal Correlation Metric	
Proposed Method	**34**	**5**	**2550**	**12 Video & Image Quality Metrics**	**4 videos**
[Qu *et al.* 2013]	46	18	**4968**	*no quantitative metric*	9 videos

5.2. Dataset

We have selected Hollywood2 dataset [51] for both quantitative and qualitative evaluation. Hollywood2 dataset contains 3669 short clips taken from 69 different Hollywood movies. The dataset is annotated with 12 classes of human actions occurring in 10 different settings (i.e. scenes). Various human actions contain different amounts of dynamic content, combined with different settings provide a rich resource for video retargeting evaluation.

The Hollywood2 database has a wide variety of actions and scene settings that makes it suitable for this thesis. Dynamic content affects quality results of video retargeting remarkably (Chapter 3); a shot having a fast motion may result with a shaky retargeted video. Since Hollywood2 videos both have static and dynamic scenes, we are able to investigate the quality results of different methods on different levels of dynamic content.

Another major benefit of Hollywood2 dataset is it's readily available eye fixation dataset *Actions in the Eye* [24]. The dataset contains eye fixation of 16 subjects recorded while watching Hollywood2 videos, summing up to 669.187 fixations in total. Existence of an available eye fixation dataset enables us to perform quantitative analysis of the proposed spatio-temporal saliency without performing an eye tracking experiment.

We have used Hollywood2 dataset across all steps of evaluation. This way, we are able to compare the results of separate evaluations. We have selected five videos to use at all steps of evaluation. The videos are selected for their diversity of motion and other conditions like scene clutter, and content in order to address the strengths and limitations of video retargeting methods. Selected videos are used through all steps of evaluation.

5.3. Video Quality Measures

For the first step of quantitative evaluation, we have used a set of popular metrics for video quality assessment. Typically, video quality is assessed frame-by-frame using image quality measures. We also use several video-based measures, which typically look at change across consecutive frames. Image quality measures are calculated separately for each frame, normalized with the corresponding original video for ease of comparison with other measures, and a single average value is calculated for each video.

The selected videos are retargeted with the comparison methods, and 12 metrics are applied on the retargeted videos. The explanations of the metrics in more detail can be found in the following subsections. The results of the metrics on the selected videos can be found in Appendix (Tables 5.2 - 5.6).

5.3.1. MSE

Mean squared error is one of the most commonly used metrics for image quality. In order to apply this to video quality, we have computed the MSE between consecutive frames.

5.3.2. UQI

The Universal Image Quality Index [61] defines image distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion. As can be understood from its name, UQI increases with image quality.

5.3.3. Blur

Blur is one of the most commonly used metrics for image quality measurement. We have used a state-of-art no-reference blur measure [62].

Table 5.2. Quantitative results of test video 1. ↑ means the higher results indicate better performance while ↓ means lower results are better. ↔ means the better results are closer to the baseline method.

	MSE↓	UQI↑	Blur↓	Focus↔	Sharpness↔	Brightness↔
Original	126.80	0.76	0.54	5.67E+17	1212.09	0.13
Linear Scaling (Baseline)	123.93	0.78	0.55	7.06E+17	748.93	0.13
[Rubinstein *et al.* 2008]	282.98	**0.76**	0.58	5.81E+17	**859.65**	**0.11**
[Yan *et al.* 2013]	222.24	0.64	0.55	9.90E+17	405.57	0.19
[Wang *et al.* 2011]	**172.39**	0.69	0.54	7.83E+17	554.58	0.16
Proposed Method + [Nguyen *et al.* 2013]	208.91	0.64	**0.52**	**7.41E+17**	454.71	0.18
Proposed Method	224.78	0.62	**0.52**	7.84E+17	348.98	0.19

	Compress↔	PSNR↑	Jerkiness↓	VIF↑	Divisive Normalization↓	SSIM↑
Original	58.86	29.74	0.34	-	-	-
Linear Scaling (Baseline)	54.74	29.93	0.23	1.00	0	1.00
[Rubinstein *et al.* 2008]	52.16	**28.17**	**0.15**	0.16	1.44	**0.50**
[Yan *et al.* 2013]	**53.65**	26.56	0.72	**0.42**	1.78	0.46
[Wang *et al.* 2011]	57.62	27.81	0.43	0.24	1.43	**0.50**
Proposed Method + [Nguyen *et al.* 2013]	60.35	27.29	0.71	0.24	1.20	0.43
Proposed Method	58.43	27.02	0.68	0.22	**0.90**	0.38

Table 5.3. Quantitative results of test video 2. ↑ means the higher results indicate better performance while ↓ means lower results are better. ↔ means the better results are closer to the baseline method.

	MSE↓	UQI↑	Blur↓	Focus↔	Sharpness↔	Brightness↔
Original	370.42	0.52	0.52	6.82E+17	745.11	0.17
Linear Scaling (Baseline)	364.95	0.55	0.53	8.56E+17	437.22	0.17
[Rubinstein *et al.* 2008]	768.95	**0.49**	0.55	9.15E+17	332.53	0.19
[Yan *et al.* 2013]	596.19	0.41	0.51	1.16E+18	230.99	0.23
[Wang *et al.* 2011]	**435.89**	0.48	0.52	7.72E+17	**435.70**	**0.18**
Proposed Method + [Nguyen *et al.* 2013]	500.00	0.43	**0.48**	7.87E+17	234.61	0.24
Proposed Method	484.55	0.47	0.51	**8.93E+17**	297.56	0.20

	Compress↔	PSNR↑	Jerkiness↓	VIF↑	Divisive Normalization↓	SSIM↑
Original	68.47	24.59	0.87	-	-	-
Linear Scaling (Baseline)	54.98	24.68	0.50	1.00	0	1.00
[Rubinstein *et al.* 2008]	56.47	23.56	**0.73**	0.29	1.54	0.39
[Yan *et al.* 2013]	50.23	21.94	1.05	**0.57**	1.78	0.53
[Wang *et al.* 2011]	**55.57**	23.56	0.96	0.27	1.51	**0.69**
Proposed Method + [Nguyen *et al.* 2013]	60.94	23.51	1.05	0.18	1.36	0.36
Proposed Method	56.19	**23.82**	1.09	0.29	**0.81**	0.45

Table 5.4. Quantitative results of test video 3. ↑ means the higher results indicate better performance while ↓ means lower results are better. ↔ means the better results are closer to the baseline method.

	MSE↓	UQI↑	Blur↓	Focus↔	Sharpness↔	Brightness↔
Original	416.77	0.63	0.71	1.27E+18	227.92	0.42
Linear Scaling (Baseline)	326.91	0.67	0.72	1.52E+18	128.04	0.42
[Rubinstein *et al.* 2008]	804.16	**0.60**	0.70	**1.56E+18**	**140.53**	**0.41**
[Yan *et al.* 2013]	**498.53**	0.57	0.71	1.66E+18	106.51	0.45
[Wang *et al.* 2011]	607.78	0.54	0.71	1.59E+18	141.25	0.47
Proposed Method + [Nguyen *et al.* 2013]	540.69	0.57	**0.69**	1.34E+18	101.83	0.46
Proposed Method	542.44	0.57	**0.69**	1.34E+18	101.38	0.46

	Compress↔	PSNR↑	Jerkiness↓	VIF↑	Divisive Normalization↓	SSIM↑
Original	36.12	23.09	1.06	-	-	-
Linear Scaling (Baseline)	31.56	24.42	0.63	1.00	0	1.00
[Rubinstein *et al.* 2008]	**31.70**	21.61	1.40	0.31	1.96	0.59
[Yan *et al.* 2013]	30.38	21.83	**0.77**	**0.50**	2.05	**0.66**
[Wang *et al.* 2011]	29.67	21.04	1.07	0.16	1.99	0.47
Proposed Method + [Nguyen *et al.* 2013]	29.82	21.81	0.89	0.12	1.60	0.39
Proposed Method	29.64	**21.89**	0.88	0.12	**1.43**	0.39

Table 5.5. Quantitative results of test video 4. ↑ means the higher results indicate better performance while ↓ means lower results are better. ↔ means the better results are closer to the baseline method.

	MSE↓	UQI↑	Blur↓	Focus↔	Sharpness↔	Brightness↔
Original	58.98	0.88	0.72	1.18E+18	722.57	0.27
Linear Scaling (Baseline)	53.58	0.90	0.73	1.26E+18	342.69	0.27
[Rubinstein *et al.* 2008]	322.94	0.81	0.73	1.32E+18	452.45	0.29
[Yan *et al.* 2013]	267.51	0.76	0.74	1.55E+18	386.27	0.30
[Wang *et al.* 2011]	311.35	0.78	0.73	**1.26E+18**	**322.27**	**0.26**
Proposed Method + [Nguyen *et al.* 2013]	**136.58**	**0.82**	**0.71**	1.28E+18	472.19	0.24
Proposed Method	175.02	0.80	**0.71**	1.37E+18	466.01	0.22

	Compress↔	PSNR↑	Jerkiness↓	VIF↑	Divisive Normalization↓	SSIM↑
Original	31.92	31.13	0.15	-	-	-
Linear Scaling (Baseline)	28.55	31.62	0.08	1.00	0	1.00
[Rubinstein *et al.* 2008]	31.09	29.72	0.32	0.21	2.05	0.24
[Yan *et al.* 2013]	24.85	24.27	0.27	0.34	2.29	0.40
[Wang *et al.* 2011]	**29.71**	**23.85**	0.40	**0.54**	1.90	**0.56**
Proposed Method + [Nguyen *et al.* 2013]	32.19	28.24	0.12	0.15	**0.86**	0.18
Proposed Method	29.78	27.00	**0.11**	0.15	0.88	0.13

Table 5.6. Quantitative results of test video 5. ↑ means the higher results indicate better performance while ↓ means lower results are better. ↔ means the better results are closer to the baseline method.

	MSE↓	UQI↑	Blur↓	Focus↔	Sharpness↔	Brightness↔
Original	782.33	0.67	0.67	1.47E+18	581.68	0.34
Linear Scaling (Baseline)	757.40	0.68	0.68	1.81E+18	253.37	0.34
[Rubinstein *et al.* 2008]	2044.16	**0.56**	0.67	1.77E+18	243.04	**0.34**
[Yan *et al.* 2013]	**1129.97**	0.55	0.67	2.00E+18	206.01	0.36
[Wang *et al.* 2011]	2243.79	0.38	0.65	**1.82E+18**	**253.35**	0.44
Proposed Method + [Nguyen *et al.* 2013]	1423.52	0.46	**0.62**	1.65E+18	254.91	0.46
Proposed Method	1281.87	0.51	0.64	1.73E+18	300.81	0.45

	Compress↔	PSNR↑	Jerkiness↓	VIF↑	Divisive Normalization↓	SSIM↑
Original	39.61	22.05	0.94	-	-	-
Linear Scaling (Baseline)	33.37	22.41	1.72	1.00	0	1.00
[Rubinstein *et al.* 2008]	34.66	**19.35**	17.90	0.30	1.90	0.46
[Yan *et al.* 2013]	29.99	18.50	0.60	**0.45**	1.92	**0.56**
[Wang *et al.* 2011]	28.48	15.57	**0.42**	0.21	1.95	0.23
Proposed Method + [Nguyen *et al.* 2013]	**33.16**	18.85	1.31	0.13	1.57	0.19
Proposed Method	30.27	19.29	1.37	0.14	**0.97**	0.18

5.3.4. Focus

Focus determines the depth of every point on the frame from the camera lens. Depth map calculation is an indicator of the amount of information at each depth and a change in depth map signals a change in the quality of the frame.

We have used the focus computation proposed in [63], which uses steerable filters [64]. Frames are convolved with steerable Gaussian filters spaced at 45° orientations {0°, 45°, 90° ...}. The maximum value generated by different filters are used to define the depth value of each pixel in the depth map.

5.3.5. Sharpness

Sharpness is an indicator of the amount of detail in an image. We have used the method proposed in [65], which is a fast variant of the Global Phase Coherence [66] used for image quality assessment and automatic image restoration.

5.3.6. Brightness/Contrast

Brightness is one of the first metrics that is used in image quality assessment. While computing the brightness map of each RGB frame, we have used the following equations:

$$\begin{aligned} Y_{i,j} &= 0.299 * R_{i,j} + 0.587 * G_{i,j} + 0.114 * B_{i,j}, \\ D_{i,j} &= \bar{Y}_{i,j}, \end{aligned} \tag{5.1}$$

where $R_{i,j}$, $G_{i,j}$, $B_{i,j}$ values correspond to the Red, $Green$ and $Blue$ values and $D_{i,j}$ is the brightness value at pixel i, j.

5.3.7. Compressiveness

We propose compressiveness as a possible measure of information loss. Compressiveness is a proposed metric that measures the information loss. We have applied H.264 compression to both retargeted and original videos [67]. The decrease in the size of the video gives a clue about the information in the video.

5.3.8. PSNR

Peak Signal to Noise Ratio is another popular image quality metric. PSNR measures the ratio of the maximum MSE, and the actual MSE of an image. As MSE, PSNR is computed over errors between consecutive frames.

5.3.9. Jerkiness

Jerkiness measures the amount of motion change across frames. This is computed by following the location of various points through frames. The mean change between frames correspond to the mean change of the tracked points. As jerkiness increases, comprehensibility of the video decreases.

5.3.10. VIF

Visual Information Fidelity measure is an information theoretic measure that utilizes natural scene statistics [68]. VIF does not define distortion, but defines the fidelity of the image. VIF can take values between $[0,\infty]$, where zero means no fidelity and infinity means perfect fidelity. We have normalized the VIF value of a retargeted frame with the VIF value of the corresponding original frame in order to compare retargeting methods. A VIF value closer to 1 indicates a closer match to the original frame and a higher quality.

5.3.11. Divisive Normalization

Divisive Normalization image quality metric proposed in [69] that is inspired by the human visual system. An input image x is first analyzed in terms of scale and orientation by a set of orthogonal 4-scales QMF wavelet transforms, $U_{i,j}$.

5.3.12. SSIM

Structural similarity index (SSIM) [70] is an image quality metric that is based on the human visual system (HVS). They state that HVS is highly capable of extracting structural information such as luminance and contrast, thus, the SSIM provides information about local luminance, contrast and structure patterns in an image. An overall SSIM score is computed using

$$\Sigma(x, y) = f(l(x, y), c(x, y), s(x, y)), \tag{5.2}$$

where l, c and s represents channels for luminance, contrast and structure channels . Details of the calculation of l,c,s and f can be found in [70].

5.4. Spatio-Temporal Saliency Evaluation

The second step of the quantitative evaluation is assessing the success of the proposed spatio-temporal saliency algorithm. One of the most popular metrics used in saliency evaluation is the area under the Receiver Operating Characteristic Curve [18].

ROC represents the ratio of false positive decisions versus true positive decisions. In image saliency, an ROC curve is the correctness of the saliency map compared to the human fixations of the same image. For a selected region of the image, the percentage of the false positives (points that saliency map shows as salient, but does not contain any human fixation) on the saliency map is plotted against the true positives (points that saliency map shows as salient, and contain human fixations).

The steps to draw an ROC curve can be seen in Figure 5.1. While drawing the ROC curve, the saliency map is thresholded, and the ROC value of the thresholded portion is added to the plot, resulting a curve starting from zero and ending at one as we continue to cover all parts of the image. At the point where 100% of the saliency map is covered, 100% of the human fixations must also be covered, so the finishing point of the ROC curve is always one. An ideal saliency map that perfectly represents the human fixations does not contain any false positives at any threshold level. On the other hand, the worst case scenario includes a random saliency map where half of the image is salient. This case is called *Chance Condition* where the saliency map covers half true and half false positives for all threshold levels. The ROC curve for the best and worst case scenarios can be seen in Figure 5.1. For the best case scenario, the area under ROC curve is 1 while for the chance condition, the ROC score is 0.5.

The proposed spatio-temporal saliency algorithm has been run on each frame of the selected five videos. The ROC of each frame is calculated separately (by utilizing the fixation maps of [24]) and combined into one ROC curve map. The results can be found in Table 5.7 along with the results of two comparison methods: Nguyen *et al.* and Itti *et al.* baselines. The details of the compared spatio-temporal saliency algorithms can be found in Section 2.1.

As can be seen in Table 5.7 the quantitative results of the proposed spatio-temporal saliency method lie between Nguyen *et al.* and Itti *et al.* The proposed spatio-temporal saliency algorithm is designed to improve video retargeting quality, so improving ROC results was not our primary concern. Although the AUC results of the proposed method is lower than Nguyen *et al.*, what is more important is the results of the user study since it is much more reliable for evaluating the video quality.

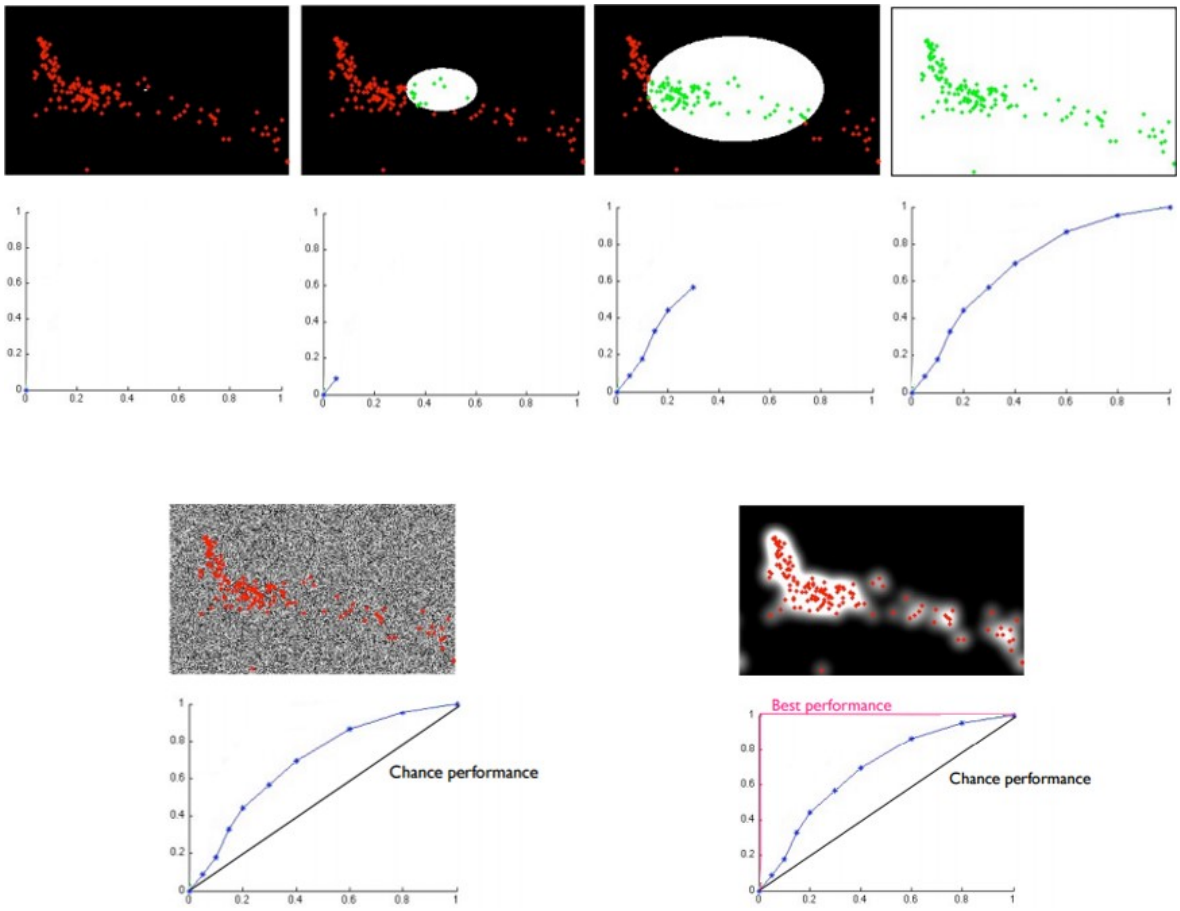
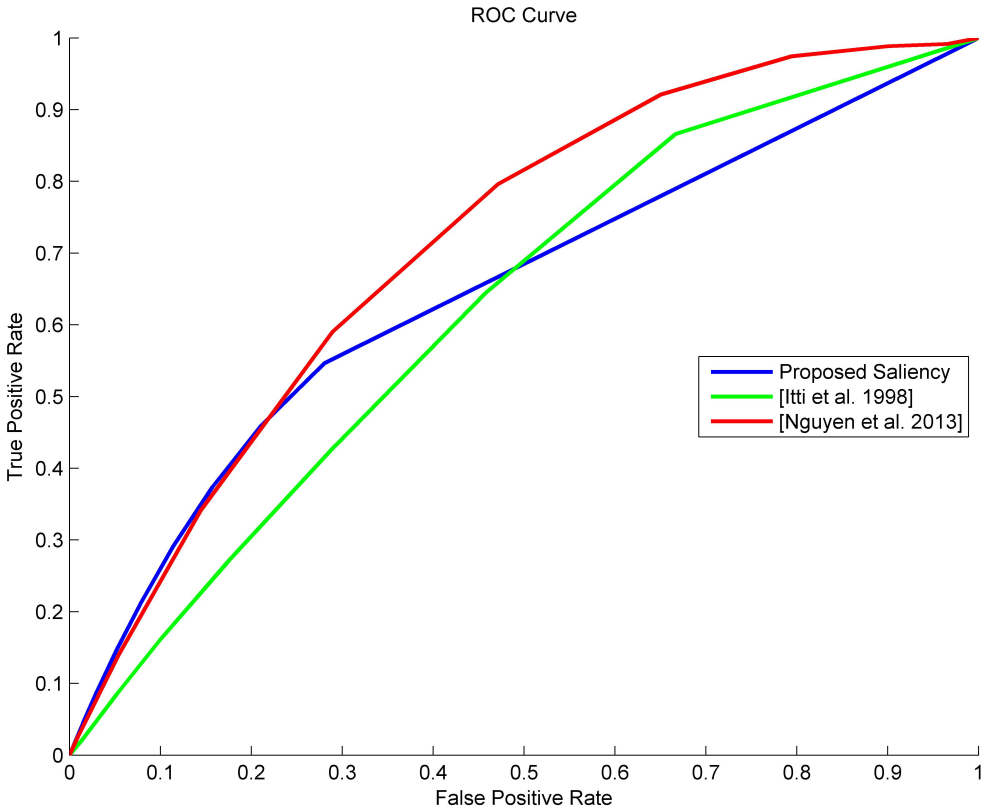


Figure 5.1. The first two row shows the steps while drawing the ROC curve of a selected image. The x axis of the ROC curve is the proportion of covered salient points (calculated) and the y axis is the proportion of the human fixations (green marks). The best possible saliency map and the random saliency map can be seen on the second row, along with their ROC curves.

Table 5.7. ROC curve and AUC results. Methods are run on the videos selected for user study and the graph shows the mean values over five videos.



	[Itti *et al.* 1998]	[Nguyen *et al.* 2013]	Proposed Method
AUC	0.6234	0.7125	0.6459

5.5. Visual Evaluation

We have compared the results of the proposed method with the following approaches; linear scaling, Rubinstein *et al.* [41], Wang *et al.* [4] and Yan *et al.* [5] on the selected five videos. These methods are selected to cover different types of video retargeting techniques such as seam carving and warping. In addition to this, we have used a recent spatio-temporal saliency algorithm proposed by Nguyen *et al.* [22], to apply to the proposed cropping method. This is done to visually evaluate the success of the proposed spatio-temporal saliency algorithm. Details of the retargeting methods and the spatio-temporal saliency algorithm can be found in Section 2.3.

Figure 5.2 shows some example frames taken from the selected videos. It is important to note that while some of the distortions are visible in the figure, it is challenging to capture the shaking and waving effects in the still images.

The first column includes a close up shot of a man's face. This video has similar results for all retargeting methods, since it is mostly static, the seam carving and warping solutions produces good results, and the content is in the middle of the scene, which makes it easier to crop.

The second column includes one of the challenging cases for all retargeting methods. Rubinstein *et al.* produces distortions in the background while Yan *et al.* produces distortions on the foreground objects, such as woman's legs. While Wang *et al.* produces rather good results by keeping the foreground objects (people) undistorted, minor distortions in the background are visible such as bended curtains and walls. While in the image case, these sorts of distortions are not perceivable, in the videos, they may cause waving effect, which catches attention. This video is also challenging for both cases of the proposed methods, as the people are moving throughout the scene. This causes the woman to be cut, while only her face and the baby she is holding are visible.

Figure 5.2. Results of comparison methods on selected videos from Hollywood2 dataset. Same videos are used in user study. Figure is best viewed in color.

The third column belongs to a scene where the most important object is the man on the foreground, while there are also other people in the background. While both Yan *et al.* and Wang *et al.* produces distorted results, Proposed Method + Nguyen *et al.* and Rubinstein *et al.* produces information loss. While the Proposed Method cuts the top of the man's hat, all important information including the stick he is holding in included in the cropped video.

The last column includes a scene where the woman enters from the right and approaches to the window, then turns back to make a call. Yan *et al.* produces some distortions on the areas around the woman, and Wang *et al.* causes distortions around the object in front of the window, while the artifacts of Wang *et al.* are barely perceivable. While both cases of the proposed method can capture the woman, Proposed Method + Nguyen *et al.* provides a wider representation of the scene. On the other hand, Rubinstein *et al.* fails to adapt to the motion of the woman returning back to the desk from the window, thus cuts the woman out. After a short time, the woman appears back in the scene as Rubinstein *et al.* adapts to the motion, causing an effect of the woman disappearing from the scene and appearing back, from nowhere.

5.6. User Study

We have conducted a two-fold psycho-visual evaluation that measures the visual quality of the proposed method. The two-fold evaluation is designed for assessing the success of both the proposed spatio-temporal saliency algorithm and the overall proposed cropping video retargeting method. Since there is no widely accepted quantitative metric for video retargeting evaluation, psycho-visual evaluation is highly important for estimating the success of the proposed method.

The first component of the psycho-visual evaluation includes five video retargeting methods; Linear Scaling, Rubinstein *et al.* [41], Wang *et al.* [4], Yan *et al.* [5] and the overall proposed video retargeting method. The second component aims to measure the success of the proposed spatio-temporal saliency algorithm alone. Since it is designed to improve the results of the video retargeting, the best way to evaluate its success

is to replace it with an alternative spatio-temporal saliency method. Thus, second component includes the results of a recent spatio-temporal saliency algorithm proposed by Nguyen *et al.* [22], combined with the proposed cropping method.

Psycho-visual evaluation includes a user study where subjects are asked about their preference of the retargeted videos to determine the methods that are preferable. While presenting the retargeted videos to subjects, we have used the paired comparison approach presented in [71].

5.6.1. Experiment Setting

We have designed a web-page for the user study, which ensures all subjects view the videos with the same setting (same size, same distance from each other etc.). The evaluation web-site can be reached from the author's personal web-site.

Five videos are selected from Hollywood2 dataset. Selected videos are downsized to 225×400 pixels. The size of the retargeted videos is chosen to be an approximation of a smartphone screen, since video retargeting targets limited sized displays. Example frames from the selected videos along with the retargeted versions can be seen in Figure 5.2.

During the user study, at each video comparison, two retargeted versions of videos are shown along with the original version. The retargeted versions are shown at their exact sizes (225×400) and the original versions are downsized to a width of 400 pixels, preserving the aspect ratio. Both original and the retargeted versions cannot be enlarged to be viewed in full-screen. The subjects are given two choices for each retargeted video as *Better* and *Much Better*. Figure 5.3 shows our video comparison setup.

Subjects are not given a chance to select the original video, but can only prefer between the retargeted versions. While linear scaling is also added as a comparison method, and it represents the downsized version of the original video, the subjects

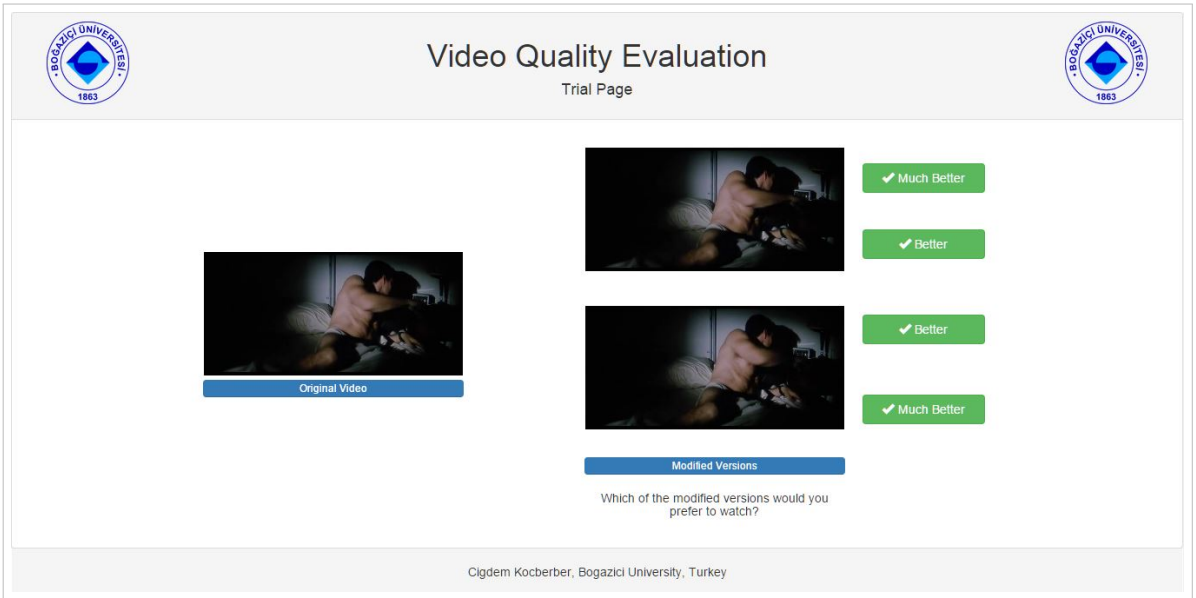


Figure 5.3. An example video comparison page. The original video that is positioned at equal distance to the retargeted versions is on the left side. Two different retargeted versions are on the right with two options per video.

are unaware of the situation, making them present an equal interest in all retargeted versions, rather than selecting the original version directly.

An equity option between the videos is also not presented to the subjects, which can end with a draw for the comparison. Equity option is omitted to enforce subjects examine the videos in detail. There are some video pairs that slightly differ especially viewed in a small sized display. The comparison results of these videos are assumed to be uniformly distributed between two choices, since all comparisons are shown in a shuffled manner.

Subjects are asked to select the video that is most appealing to them, and not been provided with any other prior information regarding the aim of the experiment. They are informed about the setup of the experiment: how the videos are going to be shown, and how they should make a selection. They are also provided with a trial page that has the same setup with the video comparison pages. The answers on the trial page are not recorded and not added to the results of the experiment.

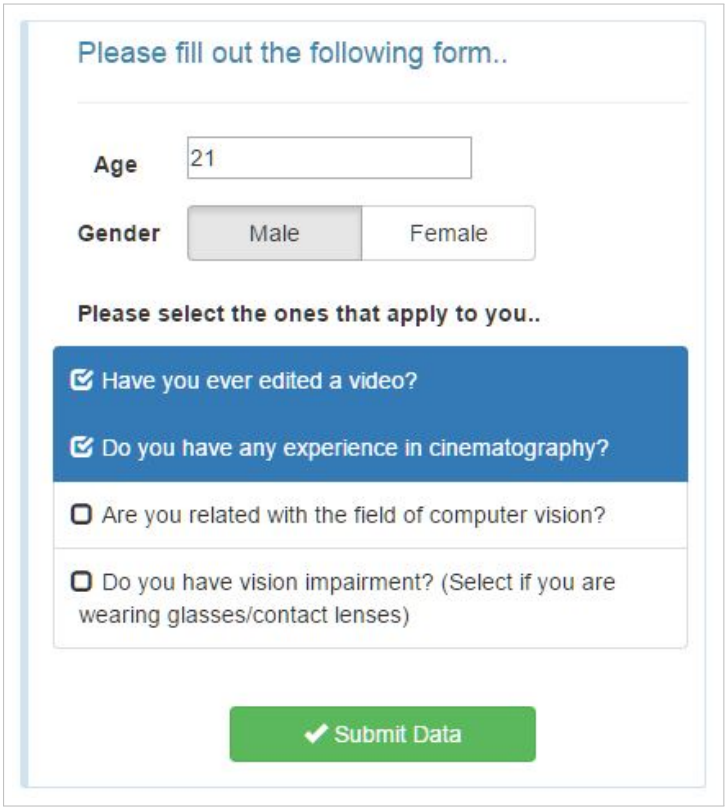


Figure 5.4. The form provided at the beginning of the user study.

Before video comparisons, subjects are given a brief form (Figure 5.4) that includes *Age* and *Gender* input values along with four questions that may affect the answers of the subjects. The questions that are asked to subjects are:

- Have you ever edited a video?
- Do you have any experience in cinematography?
- Are you related with the field of computer vision?
- Do you have vision impairment? (This applies if you are myopic/astigmatic, even if you are wearing glasses/contact lenses)

After subjects complete the experiment, they are also given a survey (Figure 5.5) to understand the most disturbing artifact in the retargeted videos. The list of possible impairments contains shakiness, cuts and jumps, information loss, distortion and blur.

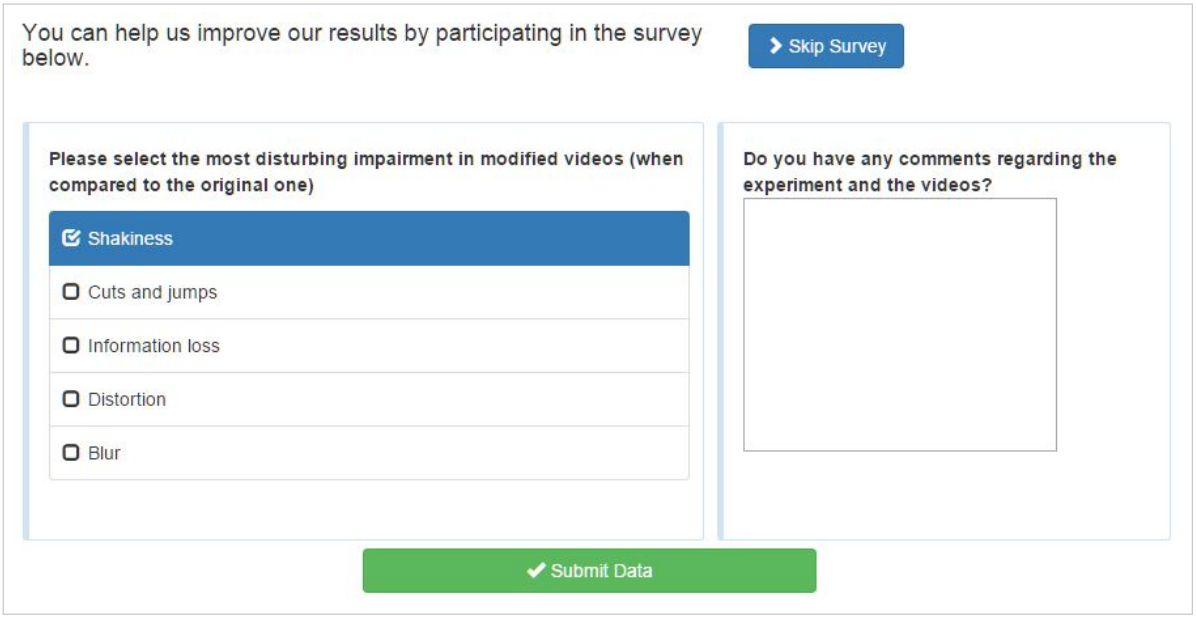


Figure 5.5. The survey provided at the end of the user study.

5.6.2. Results

Table 5.9 shows the overall pairwise comparison results of the methods (summed over videos and subjects). 34 subjects have participated in the experiment, each voting in all 75 comparisons. There were in total of 2550 pairwise comparisons while each method is subjected to 850 comparisons, and each method pair is compared 170 times. There were 17 female and 17 male subjects participated in the survey and their ages are between $18 - 54$.

Table 5.8. Overall preference ratios of the comparison methods.

	Overall Ratios (%)
Linear Scaling	88.59
[Rubinstein *et al.* 2008]	11.06
[Wang *et al.* 2011]	31.41
[Yan *et al.* 2013]	30.82
Proposed Method +[Nguyen *et al.* 2013]	65.29
Proposed Method	72.82

Table 5.9. Pairwise comparison results of the user study.** indicates that the compared methods are significantly different with a .01 significance level and * indicates .05 significance level.

	[Rubinstein *et al.* 2008]	[Wang *et al.* 2011]	[Yan *et al.* 2013]	Proposed Method + [Nguyen *et al.* 2013]	Proposed Method
Linear Scaling	168/2/**	160/10/**	165/5/**	134/36/**	126/44/**
[Rubinstein *et al.* 2008]		43/127/**	32/138/**	11/159/**	6/164/**
[Wang *et al.* 2011]			83/87/*	23/147/ **	24/146/**
[Yan *et al.* 2013]				18/152/**	14/156/**
Proposed Method +[Nguyen *et al.* 2013]					61/109/**

As can be seen from Tables 5.9, linear scaling is the most preferred method, which is expected since it represents the original version of the videos. Among other video retargeting methods, the proposed video retargeting method outperforms the state-of-the-art video retargeting methods.

While both proposed method combined with Nguyen *et al.* and the overall proposed method is highly preferred over other methods, there is also a smaller, but a significant difference between Nguyen *et al.* and proposed spatial-saliency algorithm.

Significance Test: We have conducted McNemar's test [72], which is suitable for dependent binomial data. For each method pair J_i, J_j the hypotheses are as follows,

$$\begin{aligned} H_0 &: p_{J_j} = p_{J_i} \\ H_0 &: p_{J_j} \neq p_{J_i}, \end{aligned} \tag{5.3}$$

where p_{J_i} denotes the probability of the occurrence of event J_i. The test statistic χ^2 has a chi-squared distribution with one degree of freedom is defined as

$$\chi^2 = \frac{(N_{J_i} - N_{J_j})^2}{N_{J_i} + N_{J_j}}. \tag{5.4}$$

Here, C_{J_i} is the number of the times that method J_i won over method J_j.

After applying McNemar's Test on each pair of methods, we have reported the significance levels to reject the null hypothesis H_0. Results can be seen in Table 5.9. The significance level to reject H_0 for "Proposed Method + Nguyen *et al.*" and "Proposed Method" is remarkably low. This states that the proposed spatio-temporal saliency algorithm outperforms Nguyen *et al.*'s spatio-temporal saliency algorithm despite having lower results in quantitative comparison (Table 5.7).

Another remarkable fact is the small difference between Wang *et al.* and Yan *et al.* These methods are preferred nearly equally, and the difference between them can be interpreted as chance.

Rank Analysis: We have applied the Bradley-Terry Model [73], which is designed for ranking the objects in pairwise comparisons. The model assigns the probability P that object i obtains top ranking over object j as

$$P(i > j) = \frac{\pi_i}{\pi_i + pi_j}, \quad i \neq j, \tag{5.5}$$

where π_i represents the worth parameter of the object i. Since worth parameters of objects depend on each other, an object is selected to have zero worth. The worth of the other objects are relative differences (delta) with the baseline object.

Worth parameters of objects are estimated by iteratively solving the negative log-likelihood l:

$$\min_{p} \qquad l(p) = -\sum_{i<j} \left(r_{i,j} log \frac{\tilde{\pi}_i}{\tilde{\pi}_i + \tilde{\pi}_j} + r_{ji} log \frac{\tilde{\pi}_j}{\tilde{\pi}_i + \tilde{\pi}_j} \right), \tag{5.6}$$

with respect to $\sum_i \tilde{\pi}_i = 1, \;\; \tilde{\pi}_i > 0$, where $\tilde{\pi}_i$ is the approximated worth of object i and r_{ij} is the number of times object i is preferred over object j.

Table 5.10 shows the BT worth parameters. Since Rubinstein *et al.* has the lowest overall preference score, we have selected it having a zero worth, and calculate others according to Rubinstein *et al.*

As can be seen in the Table 5.10, BT worth results are in full correspondence with the preference ratios (Table 5.8). Each subject has two options while selecting a video, *Better* and *Much Better* (Figure 5.3). Up to this point in evaluation, we have used the binary version of the selection map, since it is more appropriate for significance level tests, and preference ratio calculations. But while assessing the worth scores, this information will reveal not only if one method is preferred over another, but also *how* it was preferred.

We have calculated the BL worth score for two cases, the first case is the straight-

Table 5.10. Worth scores of the compared methods according to the BT Model [73].

	Overall Score	Binary BT Worth	Rewarded BT Worth	Difference
Linear Scaling	753	2.60	2.68	0.082
[Rubinstein *et al.* 2008]	94	0	0	0
[Wang *et al.* 2011]	267	0.08	0.06	-0.017
[Yan *et al.* 2013]	262	0.07	0.05	-0.021
Proposed Method + [Nguyen *et al.* 2013]	555	0.63	0.61	-0.018
Proposed Method	619	0.96	0.97	0.008

forward BL worth calculation, whenever a method is *selected* over another, it gets a single point, thus, creating a binary selection map for each subject. The second case is BL worth calculation of the rewarded score, where a selection with *Better* option is rewarded with one point and a selection of *Much Better* option is rewarded with two points. Thus, each time a method is selected with a *Much Better* option, it gets double points against to the compared methods.

The results of both BL scores can be found in Table 5.10 among with the change between them. When a method is selected with a *Much Better* option, its worth score for the *rewarded* case increases over the *binary* case. The difference column in the table is calculated to show the change between binary and rewarded versions. As can be seen, only the worth values of linear scaling and the proposed method has a positive change while all other methods' scores decrease. This shows that the cases where the proposed method and linear scaling is selected with high certainty dominates other methods.

5.6.3. Further Inspections of the Results

We have analyzed the results of user study by means of survey, and here are some interesting facts and findings: 17 of the subjects said they are mostly disturbed

by shakiness, 13 of them selected cuts and jumps, 12 selected distortions and only three of them selected information loss. While all subjects have watched the same videos and see the same artifacts, their preference of the most disturbing artifact changes. This confirms the fact that preferences are unique to human, and they are hard to be estimated.

Another observation is that, among the subjects who have selected shakiness and cuts and jumps as the most disturbing artifact, liner score is much higher while proposed methods are lower. This may point out that the cases where subjects are sensitive for these artifacts, our method is not preferred. Shakiness and cuts and jumps should be observed in the proposed method, and this should be marked as an improvement for our method. Another similar observation is that among the subjects who have selected information loss as the most disturbing artifact, original and the proposed results are higher. This shows that our method achieves to preserve information integrity.

Another important fact came out of the survey is; subjects that are related with cinematography has a higher ranking in the proposed method and Rubinstein *et al.* while their ranking for the Linear Scaling is remarkably low. This observation can be interpreted as the experience in this field brings the insight that the original videos should not be fitted into the target screen directly, and some post-editing is required. While it is important that experienced people are open to the idea of post-editing, it is also important that they have preferred the proposed method for a suitable automated post-editing option.

5.7. Correlation of the Qualitative and Quantitative Results

The last part of the evaluation includes the correlation of the user study and the quantitative video quality metrics. Our aim is to find a general image or video quality metric that can be applied to all videos, independent of the retargeting methodology that is used to create them. Since there is no common ground in video retargeting evaluation, such general metric may serve as an indication for user study results, and fasten the video retargeting method evaluation.

We have extracted the preference results of separate videos that are used in user study, and run a correlation analysis with the results of quantitative metrics presented in Tables 5.3 - 5.6. The results of the correlation analysis can be found in Table 5.11. The rows represent the correlation of separate videos and *mean* represent the total mean correlation among all videos. While most traditional image quality metrics such as Blur, Contrast and Jerkiness does not produce consistent correlation results, VIF, Divisive Normalization, MSE and SSIM has consistent results among videos.

The last row of Table 5.11 contains the mRMR [74] results that show the importance of each metric on the user ratings. mRMR is a feature selection technique that maximizes the relevancy while minimizing redundancy of features against a given target. While only several features may be selected as the result of mRMR, we have not limited the number of output features, so our mRMR results show the order of importance of the quantitative metrics.

We have also applied linear regression on aggregated and standardized results of the quantitative metrics, and the user study. The resulting function V (Equation 5.7) has $R^2 = 0.8866$ and it creates a combined metric that converges most to the results of user study.

$$\begin{aligned} V = & M_{MSE} * (-11.9404) \ + \ M_{UQI} * (0.0016) \ + \ M_{Blur} * (-0.0040) + \\ & M_{Sharpness} * (-0.4500) \ + \ M_{Brightness} * (0.0024) \ + \ M_{Compress} * (-0.3855) + \\ & M_{PSNR} * (-0.0286) \ + \ M_{VIF} * (0.0045) \ + \ M_{Jerkiness} * (0.0328) + \\ & M_{Div.Norm.} * (-0.0211) \ + \ M_{SSIM} * (-0.0057), \end{aligned} \tag{5.7}$$

where M_x stands for the result of the x metric, and $*$ is the multiplication operator. The higher results of V indicate better user study results. V can be used as a benchmark for future video retargeting methods.

Table 5.11. Correlation between quantitative metrics and user study results. First five rows represent results of separate videos, and the last line is the mean of all videos.

	MSE↓	UQI↑	Blur↓	Focus↔	Sharpness↔	Brightness↔	Compress↔	PSNR↑	Jerkiness↓	VIF↑	Divisive Normalization↓	SSIM↑
(1)	-0.70	-0.21	-0.69	0.27	-0.41	0.42	0.54	0.20	0.40	0.58	-0.66	0.41
(2)	-0.64	0.26	-0.32	0.04	-0.02	-0.02	-0.03	0.39	-0.15	0.27	-0.80	0.42
(3)	-0.78	0.38	0.01	-0.63	-0.40	0.23	-0.24	0.62	-0.74	0.39	-0.78	0.18
(4)	-0.97	0.67	-0.48	-0.29	-0.01	-0.54	0.08	0.43	-0.89	0.40	-0.92	0.36
(5)	-0.48	0.36	-0.16	-0.60	0.78	0.20	0.26	0.61	-0.23	0.43	-0.88	0.19
Mean	**-0.71**	**0.29**	**-0.33**	**-0.24**	**-0.01**	**0.06**	**0.12**	**0.45**	**-0.32**	**0.43**	**-0.81**	**0.31**
mRMR	1	11	2	5	7	10	3	9	8	12	4	6

6. CONCLUSION

Video retargeting is an application that is gaining importance with increasing usage of different sized displays. The main aspect of video retargeting is to avoid possible artifacts such as artificial cuts, shaking effects, virtual camera movements as well as information loss. In order to ensure these constraints, the first step is to define important parts of the video, which is done with spatio-temporal saliency algorithms. Spatio-temporal saliency maps are defined over each frame, which aim to mimic human fixations, as well as capturing important objects. A suitable spatio-temporal saliency algorithm should be able to integrate the motion continuity, as well as keeping the important parts of frames.

6.1. Contributions

We propose a crop based video retargeting method relying on a crop window for each frame that ensures the following constraints:

- Covers the important objects in the frame
- Follows a continuous path during a shot to avoid artificial cuts
- Moves in sync with the camera motion, and preserves its size during a shot to prevent artificial camera movements

While last two items are satisfied with the help of optical flow maps, in order to meet the first constraint, a spatio-temporal saliency algorithm is required.

We propose a spatio-temporal saliency algorithm based on motion trajectories to ensure motion continuity. In order to capture all important objects, we detect key frames by observing the change in spatial saliency maps, then use key frames to detect the important trajectory groups. This method guarantees to follow important objects, and keep them as a whole in the retargeted video.

While the most common qualitative evaluation method is performing a user study, it is costly in terms of time and work. In order to overcome this challenge of video retargeting, a common quantitative metric that can be applied to all retargeting methods, and is able to represent the results of user study is needed. Thus, we use 12 image/video quality metrics to perform the quantitative evaluations, and find the correlation between these metrics and the results of user study.

The correlation reveals that some metrics used in quantitative analysis such as Divisive Normalization, Visual Fidelity Index (VIF) and Mean Square Error (MSE) have a higher importance on the results of user study. A linear regression that is run on quantitative metrics targeting the results of the qualitative evaluation outputs a new quantitative metric that is a combination of the available metrics. The proposed metric can be used in further studies on video retargeting, as an indication of the user study results.

6.2. Lessons Learned

While most criticized artifact of crop based video retargeting methods is the artificial camera movements, several straightforward constraints can ensure preventing these artifacts. Information loss stands out as the most challenging task, so we proceed to design a spatio-temporal saliency algorithm for video retargeting.

The most popular way to integrate the motion information in videos is to utilize optical flow maps. Optical flow maps are a good way to discretize the motion into separate frames, which is not suitable for application of auto-editing videos. The continuity of the motion is the most crucial part in defining the important parts, which can be accomplished better with motion trajectories.

The most lengthy process while producing a video retargeting method is to evaluate its efficiency. During the user study, we have observed that the most disturbing artifacts vary vastly according to the subject. Some subjects report that cutting the top of the heads of people was the most disturbing, while other subjects were annoyed

by waving effects. This aspect makes the qualitative evaluation a crucial step in new video retargeting methods, while it is also the most costly one.

6.3. Future Work

A major extension of the proposed cropping method is to make it faster. The bottleneck of the method is the trajectory extraction step, so in order to decrease the time complexity of this method, one must optimize the trajectory extraction step, or use a different representation of motion that can be computed faster.

Another improvement may be on extending the quantitative metrics. Since two outstanding metrics in correlation study are Divisive Normalization and MSE, we can say that metrics that try to minimize an error function better represent the results of the user study. More metrics can be found based on this finding, and be included in the quantitative evaluation.

REFERENCES

1. Spooner, S., "The Blackburn Flying School", *Flight Magazine*, Vol. 17, No. 16, p. 224, 1925.

2. Wolf, L., M. Guttmann and D. Cohen-Or, "Non-homogeneous Content-driven Video-retargeting", *International Conference on Computer Vision*, pp. 1–6, IEEE, 2007.

3. Grundmann, M., V. Kwatra, M. Han and I. Essa, "Discontinuous Seam-carving for Video Retargeting", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 569–576, IEEE, 2010.

4. Wang, Y.-S., J.-H. Hsiao, O. Sorkine and T.-Y. Lee, "Scalable and Coherent Video Resizing with Per-frame Optimization", *ACM Transactions on Graphics*, Vol. 30, p. 88, ACM, 2011.

5. Yan, B., K. Sun and L. Liu, "Matching-area-based Seam Carving for Video Retargeting", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 23, No. 2, pp. 302–310, 2013.

6. Liu, F. and M. Gleicher, "Video Retargeting: Automating Pan and Scan", *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 241–250, ACM, 2006.

7. Deselaers, T., P. Dreuw and H. Ney, "Pan, Zoom, Scan - Time-coherent, Trained Automatic Video Cropping", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

8. Yan, B., B. Yuan and B. Yang, "Effective Video Retargeting with Jittery Assessment", *IEEE Transactions on Multimedia*, Vol. 16, No. 1, pp. 272–277, 2014.

9. Wang, B., H. Xiong, Z. Ren and C. W. Chen, "Deformable Shape Preserving Video Retargeting with Salient Curve Matching", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, Vol. 4, No. 1, pp. 82–94, 2014.

10. Lin, S.-S., C.-H. Lin, I.-C. Yeh, S.-H. Chang, C.-K. Yeh and T.-Y. Lee, "Content-aware Video Retargeting Using Object-preserving Warping", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 10, pp. 1677–1686, 2013.

11. Aristotle, "On Sense and the Sensible", `http://classics.mit.edu/Aristotle/sense.mb.txt`, [Accessed August 2015].

12. Itti, L. and C. Koch, "Computational Modelling of Visual Attention", *Nature Reviews Neuroscience*, Vol. 2, No. 3, pp. 194–203, 2001.

13. Desimone, R. and J. Duncan, "Neural Mechanisms of Selective Visual Attention", *Annual Review of Neuroscience*, Vol. 18, No. 1, pp. 193–222, 1995.

14. Posner, M. I., "Orienting of Attention", *Quarterly Journal of Experimental Psychology*, Vol. 32, No. 1, pp. 3–25, 1980.

15. Treisman, A. M. and G. Gelade, "A Feature-integration Theory of Attention", *Cognitive Psychology*, Vol. 12, No. 1, pp. 97–136, 1980.

16. Duncan, J., "Selective Attention and the Organization of Visual Information", *Journal of Experimental Psychology: General*, Vol. 113, No. 4, p. 501, 1984.

17. Koch, C. and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry", *Matters of Intelligence*, pp. 115–141, Springer, 1987.

18. Nevin, J. A., "Signal Detection Theory and Operant Behavior: A Review of David M. Green and John A. Swets' Signal Detection Theory and Psychophysics", *Journal of the Experimental Analysis of Behavior*, Vol. 12, No. 3, pp. 475–480, 1969.

19. Frintrop, S., E. Rome and H. I. Christensen, "Computational Visual Attention Sys-

tems and Their Cognitive Foundations: A Survey", *ACM Transactions on Applied Perception*, Vol. 7, No. 1, p. 6, 2010.

20. Itti, L., 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, No. 11, pp. 1254–1259, 1998.

21. Judd, T., K. Ehinger, F. Durand and A. Torralba, "Learning to Predict Where Humans Look", *IEEE 12th International Conference on Computer Vision*, pp. 2106–2113, IEEE, 2009.

22. Nguyen, T. V., M. Xu, G. Gao, M. Kankanhalli, Q. Tian and S. Yan, "Static Saliency vs. Dynamic Saliency: A Comparative Study", *Proceedings of the 21st Annual ACM International Conference on Multimedia*, pp. 987–996, ACM, 2013.

23. Fang, Y., Z. Wang and W. Lin, "Video Saliency Incorporating Spatiotemporal Cues and Uncertainty Weighting", *IEEE International Conference on Multimedia and Expo*, pp. 1–6, IEEE, 2013.

24. Stefan Mathe, C. S., "Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition", *European Conference on Computer Vision*, 2012.

25. Le Meur, O. and T. Baccino, "Methods for Comparing Scanpaths and Saliency Maps: Strengths and Weaknesses", *Behavior Research Methods*, Vol. 45, No. 1, pp. 251–266, 2013.

26. Zhai, Y. and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues", *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 815–824, ACM, 2006.

27. Fang, Y., W. Lin, Z. Chen, C.-M. Tsai and C.-W. Lin, "Video Saliency Detection in the Compressed Domain", *Proceedings of the 20th ACM International Conference*

on Multimedia, pp. 697–700, ACM, 2012.

28. Marat, S., T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin and A. Guérin-Dugué, "Modelling Spatio-temporal Saliency to Predict Gaze Direction for Short Videos", *International Journal of Computer Vision*, Vol. 82, No. 3, pp. 231–243, 2009.

29. Liu, C., *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*, Ph.D. Thesis, Massachusetts Institute of Technology, 2009.

30. Ghanem, B., T. Zhang and N. Ahuja, "Robust Video Registration Applied to Field-sports Video Analysis", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

31. Hwang, D.-S. and S.-Y. Chien, "Content-aware Image Resizing Using Perceptual Seam Carving with Human Attention Model", *IEEE International Conference on Multimedia and Expo*, pp. 1029–1032, IEEE, 2008.

32. Avidan, S. and A. Shamir, "Seam Carving for Content-aware Image Resizing", *ACM Transactions on Graphics*, Vol. 26, p. 10, ACM, 2007.

33. Suh, B., H. Ling, B. B. Bederson and D. W. Jacobs, "Automatic Thumbnail Cropping and Its Effectiveness", *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*, pp. 95–104, ACM, 2003.

34. Glasbey, C. A. and K. V. Mardia, "A Review of Image-warping Methods", *Journal of Applied Statistics*, Vol. 25, pp. 155–171, 1998.

35. Liu, F. and M. Gleicher, "Automatic Image Retargeting with Fisheye-view Warping", *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, pp. 153–162, ACM, 2005.

36. Wolfe, J. M., "Guided Search 2.0 a Revised Model of Visual Search", *Psychonomic Bulletin & Review*, Vol. 1, No. 2, pp. 202–238, 1994.

37. Bruce, N. and J. Tsotsos, "Saliency Based on Information Maximization", *Advances in Neural Information Processing Systems*, pp. 155–162, 2005.

38. Liu, T., Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang and H.-Y. Shum, "Learning to Detect a Salient Object", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, pp. 353–367, 2011.

39. Goferman, S., L. Zelnik-Manor and A. Tal, "Context-aware Saliency Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 10, pp. 1915–1926, 2012.

40. Kopf, S., J. Kiess, H. Lemelson and W. Effelsberg, "FSCAV: Fast Seam Carving for Size Adaptation of Videos", *Proceedings of the 17th Annual ACM International Conference on Multimedia*, pp. 321–330, ACM, 2009.

41. Rubinstein, M., A. Shamir and S. Avidan, "Improved Seam Carving for Video Retargeting", *ACM Transactions on Graphics*, Vol. 27, p. 16, ACM, 2008.

42. Wang, Y.-S., C.-L. Tai, O. Sorkine and T.-Y. Lee, "Optimized Scale-and-stretch for Image Resizing", *ACM Transactions on Graphics*, Vol. 27, p. 118, ACM, 2008.

43. Werlberger, M., W. Trobin, T. Pock, A. Wedel, D. Cremers and H. Bischof, "Anisotropic Huber-L1 Optical Flow", *The British Machine Vision Conference*, Vol. 1, p. 3, 2009.

44. Qu, Z., J. Wang, M. Xu and H. Lu, "Context-aware Video Retargeting via Graph Model", *IEEE Transactions on Multimedia*, Vol. 15, No. 7, pp. 1677–1687, 2013.

45. Nie, Y., Q. Zhang, R. Wang and C. Xiao, "Video Retargeting Combining Warping and Summarizing Optimization", *The Visual Computer*, Vol. 29, No. 6-8, pp. 785–794, 2013.

46. Platt, J. C., "Fast Training of Support Vector Machines Using Sequential Minimal

Optimization", *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1998.

47. Joachims, T., "A Support Vector Method for Multivariate Performance Measures", *Proceedings of the 22nd International Conference on Machine Learning*, pp. 377–384, ACM, 2005.

48. Zhuang, L., F. Jing and X.-Y. Zhu, "Movie Review Mining and Summarization", *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 43–50, ACM, 2006.

49. Fawcett, T., "An Introduction to ROC Analysis", *Pattern Recognition Letters*, Vol. 27, No. 8, pp. 861–874, 2006.

50. Koçberber, C. and A. A. Salah, "Video Retargeting: Video Saliency and Optical Flow Based Hybrid Approach", *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

51. Marszalek, M., I. Laptev and C. Schmid, "Actions in Context", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2929–2936, IEEE, 2009.

52. Wang, H., A. Kläser, C. Schmid and C.-L. Liu, "Action Recognition by Dense Trajectories", *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3169–3176, Colorado Springs, United States, 2011.

53. Wang, H. and C. Schmid, "Action Recognition with Improved Trajectories", *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

54. Farnebäck, G., "Two-Frame Motion Estimation Based on Polynomial Expansion", *Image Analysis*, pp. 363–370, Springer Berlin Heidelberg, 2003.

55. Dalal, N. and B. Triggs, "Histograms of Oriented Gradients for Human Detection", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*,

Vol. 1, pp. 886–893, IEEE, 2005.

56. Laptev, I., M. Marszałek, C. Schmid and B. Rozenfeld, "Learning Realistic Human Actions from Movies", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

57. Dalal, N., B. Triggs and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance", *European Conference on Computer Vision*, pp. 428–441, Springer, 2006.

58. Bay, H., T. Tuytelaars and L. Van Gool, "Surf: Speeded Up Robust Features", *European Conference on Computer Vision*, pp. 404–417, Springer, 2006.

59. Fischler, M. A. and R. C. Bolles, "Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Communications of the ACM*, Vol. 24, No. 6, pp. 381–395, 1981.

60. Wang, J., M. Xu, X. He, H. Lu and D. Hoang, "A Hybrid Domain Enhanced Framework for Video Retargeting with Spatial-temporal Importance and 3D Grid Optimization", *Signal Processing*, Vol. 94, pp. 33–47, 2014a.

61. Wang, Z. and A. C. Bovik, "A Universal Image Quality Index", *IEEE Signal Processing Letters*, Vol. 9, No. 3, pp. 81–84, 2002.

62. Crete, F., T. Dolmiere, P. Ladret and M. Nicolas, "The Blur Effect: Perception and Estimation with a New No-reference Perceptual Blur Metric", *Electronic Imaging*, pp. 64920I–64920I, International Society for Optics and Photonics, 2007.

63. Minhas, R., A. A. Mohammed, Q. J. Wu and M. A. Sid-Ahmed, "3D Shape from Focus and Depth Map Computation Using Steerable Filters", *Image Analysis and Recognition*, pp. 573–583, Springer, 2009.

64. Freeman, W. T. and E. H. Adelson, "The Design and Use of Steerable Filters",

IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 9, pp. 891–906, 1991.

65. Blanchet, G. and L. Moisan, "An Explicit Sharpness Index Related to Global Phase Coherence", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1065–1068, IEEE, 2012.

66. Blanchet, G., L. Moisan and B. Rougé, "Measuring the Global Phase Coherence of an Image", *IEEE International Conference on Image Processing*, pp. 1176–1179, IEEE, 2008.

67. Wiegand, T., G. J. Sullivan, G. Bjøntegaard and A. Luthra, "Overview of the H. 264/AVC Video Coding Standard", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 13, No. 7, pp. 560–576, 2003.

68. Sheikh, H. R. and A. C. Bovik, "Image Information and Visual Quality", *IEEE Transactions on Image Processing*, Vol. 15, No. 2, pp. 430–444, 2006.

69. Laparra, V., J. Muñoz-Marí and J. Malo, "Divisive Normalization Image Quality Metric Revisited", *JOSA A*, Vol. 27, No. 4, pp. 852–864, 2010.

70. Wang, Z., A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, 2004.

71. David, H. A., *The Method of Paired Comparisons*, Vol. 12, DTIC, Fort Belvoir, VA, 1963.

72. McNemar, Q., "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages", *Psychometrika*, Vol. 12, No. 2, pp. 153–157, 1947.

73. Bradley, R. A. and M. E. Terry, "Rank Analysis of Incomplete Block Design the Method of Paired Comparisons", *Biometrika*, Vol. 39, No. 3-4, pp. 324–345, 1952.

74. Peng, H., F. Long and C. Ding, "Feature Selection Based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226–1238, 2005.