ABSTRACT

Video retargeting (resolution adaptation) is a challenging problem for its highly subjective nature. In this paper, a nonlinear saliency fusion approach, that considers human perceptual characteristics for automatic video retargeting, is being proposed. First, we incorporate features from phase spectrum of quaternion Fourier Transform (QPFT) in spatial domain and global motion residual based on matched feature points by the Kanade-Lucas-Tomasi (KLT) tracker in temporal domain. In addition, under a cropping-and-scaling retargeting framework, we propose content-aware information loss metrics and a hierarchical search to find optimal cropping window parameters. Results show the success of our approach on detecting saliency regions and retargeting on images and videos.

Index Terms— Video retargeting, saliency, nonlinear, QPFT, KLT

1. INTRODUCTION

Nowadays, the development of digital video applications has two opposite trends. On one hand, people appreciate the fantastic video contents in cinemas and on HDTVs with resolution higher than $1920 \times 1080$. On the other hand, people enjoy the flexibility of watching videos on their portable devices (iPhone, Blackberry, etc.) with resolution smaller than $480 \times 320$. These two trends evoke growing interests in automatic video adaptation that seeks to change the resolution (generally from the larger to the smaller) of video sources while faithfully convey the original information. This process, named “Video retargeting”, aims at preserving the viewers’ experience when the resolution changes. Video retargeting is naturally a hard problem, since it is a very subjective task to map human cognition into the automated process. Two major challenges are “saliency detection” to indicate the importance of pixels, and “retargeting optimization” to map original video contents into a target display.

Prior Arts In literature, saliency analysis methods could be classified as the top-down (object or event driven), or the bottom-up, like feature-driven [1, 2]. Features may come from both spatial and temporal domains. In the latter class, the spectral residue approach [2] suggests utilizing the spectral residue for an image to obtain the attention model. Later, Guo et al. showed that the phase spectral alone is good enough as a spatial feature, and they extend [2] for video by incorporating motion features under a QPFT framework [3] (we call it baseline-PQFT). This work provides a new insight for a spatial-temporal saliency detection. However, the naive motion feature obtained by frame differencing does not truly reflect human’s perceptual experience: it’s the local motion, not the global motion, that triggers the human interests mostly. Considering this, many people use optical flow based approaches [4] to factor out the global motion. While these methods are more reliable than naive differencing, they have to deal with the heavy computation burden and the aperture problem. Besides, those approaches usually adopt a linear combination scheme to fuse features in different domains, while the weighing factors need to be carefully selected.

Methods for “retargeting optimization” can be classified as warping [5], seam-carving [6] or cropping-scaling [4, 7]. The former two faithfully maintain the pixels with high visual saliency, while squeezing less salient pixels to make up for the source-target resolution difference. They are most successful for images with natural scenarios. Their applications, nevertheless, are limited when geometric distortion and computation load are the major concerns. Cropping-scaling, on the other hand, is an efficient and distortion-free method. It finds a cropping window with minimum information loss in original frame and scales the window as the retargeted frame.

Overview our approach We propose a novel nonlinear scheme for saliency detection: spatial saliency is detected by the phase spectrum of quaternion Fourier Transform (QPFT) on a color image, which utilizes the multiple channels as a vector field to exploit conspicuous spatial features (color, intensity, etc.); motion saliency is measured by local motion (global motion residue), while the global motion parameters are estimated by robust affine fitting with Least median of Squares (LMedS) [8], from a set of matched feature points by the KLT tracker [9]. Unlike the dense optical flow approaches, the KLT tracker works on sparse feature points, and thus is more efficient. The innovation for this nonlinear fusion is based on human perceptual properties: 1) When excitation is absent (texture uniformly distributed), people tend to focus on the center of the frame, instead of the borders. 2) The human perception process consists of a “stimulating” phase and a “tracking” phase, defined as “saccade and pursuit” in human vision theory. First, spatial-salient regions pop up as
“stimulus”. If a spatial-salient region has significant local motion activities, this motion stimulus will strengthen the spatial stimulus, and cause higher attention. Otherwise, “lazy” human eyes will continually focus on the spatial-salient regions. 3) Occasionally, spatial and motion saliency regions are not consistent. Our scheme treats this as a “prohibited case” since the motion stimulus will distract the spatial stimulus and make a rapid change of focus points, which will cause an eye fatigue. Thus, professional photographers and moviemakers will make their efforts to avoid the situation.

Our innovation points also include content-aware information loss metrics, and a hierarchical search to find optimal retargeting parameters on a single frame, under the efficient cropping-scaling framework. Compared to the content-independent scaling penalties [4], our metric can sharply adjust the scaling factor corresponding to different contents. Our scaling metric also outperforms the content-aware scaling metric in [7], as we take into account not only the anti-aliasing filter in [7] but also the true resizing process. The hierarchical search, can greatly save computation costs.

Fig. 2 illustrates the procedure of our approach: Taken original video and target size as inputs, the spatial-temporal saliency map are first computed, the optimal scaling-cropping window parameters are found thereafter, and the target frame is generated finally. Sec. 2 explains the technique details. Experimental results are shown in Sec. 3 and our website while Fig. 1 provides a snapshot. Conclusions are drawn in Sec. 4.

2. TECHNICAL DETAILS

2.1. Spatial-temporal saliency detection

Spatial saliency Denote the n-th frame in the video sequence \( F^n \). The frame can be represented as a quaternion image [10] which has four channels, \( q^n = Ch^n_1 + Ch^n_2µ_1 + Ch^n_3µ_2 + Ch^n_4µ_3 \), where \( µ_i, i = 1, 2, 3 \) satisfies \( µ_i = -1, µ_1 ⊥ µ_2, µ_2 ⊥ µ_3, µ_1 ⊥ µ_3, µ_1 ⊥ µ_2 \). \( Ch^n_1, Ch^n_2, Ch^n_3, Ch^n_4 \) are the channels of the quaternion image. If choosing \( µ_1 \) along the luminance axis, i.e., \( µ_1 = (i + j + k)/√3 \), the color image is thus decomposed into luminance and chrominance components \( Y^n, C^n_0 \) and \( C^n_1 \), and the quaternion image is pure \( (Ch^n_1 = 0) \) [10]. We can further represent \( q^n \) in symplectic form: \( q^n = q^n_0 + q^n_1µ_2, q^n_0 = Ch^n_1 + Ch^n_2µ_1, q^n_1 = Ch^n_3 + Ch^n_4µ_1 \). The Quaternion Fourier Transform (QFT) of the quaternion image \( q^n(x, y) \) can be calculated by two complex fourier transforms of the symplectic parts: \( Q^n_{Ch}[u, v] = Q^n_{Ch}[u, v] + Q^n_{Ch}[u, v]µ_2 \). The forward and inverse fourier transform of each part are:

\[
Q^n_{Ch}[u, v] = \frac{1}{\sqrt{WH}} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} e^{-j2\pi(uH+vh)/WH} q^n(x, y)
\]

\[
q^n(x, y) = \frac{1}{\sqrt{WH}} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} e^{j2\pi(uH+vh)/WH} Q^n_{Ch}[u, v]
\]

where \((x, y)\) is the spatial location of each pixel, \(W\) and \(H\) are image’s width and height, and \([u, v]\) is the frequency.

The phase spectrum of \( Q^n_{Ch}[u, v] \) (\( Q \) for abbreviation) can be calculated by \( \tilde{Q} = Q/||Q|| \). Taking the inverse Transform of the phase spectrum \( Q_{P} \), as in Eq. (2), the spatial saliency map is obtained by smoothing out the squared \( L_2 \) norm of \( q^n_r \) with a two-dimensional Gaussian smoothing filter \( g \).

\[
SM = g * ||q^n_r||^2
\]

The advantages of the PQFT approach over traditional multi-channel PFFT (phase spectrum of the 2D Fourier Transform) is shown in Fig. 3. PQFT not only achieves better saliency detection results by treating a color image as a vector field, but also consumes less computation time since only two complex 2D Fourier Transforms are conducted for the symplectic equations, while PFFT has three (for each channel).

Motion saliency There are two steps to obtain the motion saliency map: a) Kanade-Lucas-Tomasi (KLT) tracker to get a set of matched good feature points [9] and b) robust affine parameter estimation by Least Median Squares (LMedS) [11].

Denote the displacement of a point \((x, y)\) at previous frame \( F^{n-1} \) to current frame \( F^n \) as \( d = (dx, dy)^T \). A six-parameter affine model is adopted to estimate the global motion: \( d = Dx + t \), where \( t \) is the translation vector \((x_t, y_t)^T \) and \( D \) is a \( 2 \times 2 \) deformation matrix. The point \( x \) in \( F^{n-1} \) moves to point \( x' = Ax + t \) in \( F^n \), where \( A = I + D \) and \( I \) is a \( 2 \times 2 \) identity matrix. The model parameters are estimated by minimize the dissimilarity in each feature window. We adopt the Least Median Squares to estimate the affine parameters robustly [11]. The global compensated image is generated by warping with the estimated \( \hat{A} \) and \( \hat{t} \), and the absolute difference of the original frame with its global-compensated version is used to generate the motion saliency map.

\[
SM_m = g(x) * |F^{n-1}(x) - F^n((A^{-1}[x - \hat{t}])|
\]

Spatial-temporal saliency fusion When both the spatial and temporal saliency maps are available, the final saliency map is generated by a spatial-masked nonlinear manner which imitates the human vision features. First, a 2D Gaussian layer \( G \) centered at the frame center is fused to the spatial saliency map: \( SM_{spatial} = SM_{spatial} \times G \). A binary mask \( M_s \) of spatial saliency significance is generated by thresholding. The final saliency map \( SM \) is obtained by:

\[
SM = MAX(SM_s, SM_m \cap M_s)
\]

The reasons to use the MAX operator with a binary masked motion saliency map and a Gaussian layer are many. 1) Gaussian layer is used to adjust the descendiong importance from the center of a frame.
to the border. 2) Mask is used to exclude the spatial-temporal inconsistent cases (the “prohibited cases”). 3) The mask enhances the robustness of the spatial-temporal saliency map, when global-motion parameters are not estimated correctly. 4) The MAX operation avoids the depression of insignificant salient regions caused by renormalization if using a linear combination scheme. 5) The MAX operation avoids the selection of weighting factor between spatial and motion saliency maps. Fig. 4 shows the comparison of linear combination, naive nonlinear (unmasked MAX) fusion and our scheme, where in the first video the global motion parameters are correctly estimated but in second video are wrong. The comparison shows the robustness of our scheme.

Hierarchical brute-force search

For optimal parameter set \((x,y,s)\)

Retargeting squeezing

Retargeting squeezing

Hierarchical brute-force search for optimal parameter set \((x,y,s)\)

2.2. Retargeting optimization

Under a cropping-scaling framework, the retargeting optimization for a video frame is to find best parameters \((x, y, s)\) with minimum information loss, where \((x, y)\) is the location of the top-left point of the cropping window over original frame, and \(s\) is the scaling factor. Isotropic scaling is used to avoid geometric distortion.

Content-aware information loss metric In our scheme, we consider the saliency loss and scaling penalty as our major concern and propose a content-aware information loss metric \(L\), which consists of two terms: content-aware cropping loss \(L_c\) and content-aware scaling loss \(L_s\).

\[
L = (1 - \lambda)L_c + \lambda L_s
\]

where \(\lambda\) is a factor to balance the importance of cropping and scaling. The proposed \(L_s\) better measures the scaling loss for different contents than \([4, 7]\), by taking into account the real resizing process. Integral images are used for computing the \(L_c\) and \(L_s\) by a “look up” operation on integral images.

Hierarchical search for retargeting window parameters After the information loss metric is well formulated, a hierarchical brute-force search is used to find the best retargeting parameters \((\hat{x}, \hat{y}, \hat{s})\).

\[
P(\hat{x}, \hat{y}, \hat{s}) = \arg \min_{x, y, s} L(x, y, sW_t, sH_t)
\]

Note the search range of \((x, y)\) is constrained by \(s\), and \(s\) is constraint by \(1 \leq s \leq \min(W_t/W_i, H_t/H_i)\), where \(W_t, W_i, H_t, H_i\) are the width and height for source and target frames, respectively. The searching range is a group of surfaces in \((x, y, L)\) space. Fig. 5 shows the searching space. Each surface corresponds to a particular scaling factor \(s\). The point which yields the minimum \(L\) (the lowest point in space) is the best parameter \((x, y)\) and the surface it belongs to is the best \(s\).

For computation saving purposes, the search for \((x, y, s)\) is first on a coarse \((x, y)\) grid (a 10x10 grid search will save 99% computation over 1x1 grid search), a target parameter set \((x_1, y_1, s)\) is found after the coarse search. The second search is a fine search within a range around \((x_1, y_1)\) with the fixed scaling factor \(s\) found previously. After the hierarchically search, best parameter set \((x_2, y_2, s)\) is obtained.

3. EXPERIMENTAL RESULTS

In addition to the intermediate results shown in Fig. 3 and Fig. 4, we implement our scheme in C++ with OpenCV (http://opencv.willowgarage.com/wiki/), FFTW3 (http://www.fftw.org/) and KLT (http://www.ces.clemson.edu/~stb/klt/) libraries, and design three experiments to test the performance compared to some representative prior arts. We evaluate the comparison result using the popular subjective metric MOS (mean opinion score), which collect scores (ranges from 1 to 5 while 1, 2, 3, 4, 5 stands for bad, poor, fair, good, excellent, respectively) from 20 randomly selected people.

1) Saliency detection performance on images. We use a collection of images (multiple resolutions and aspect ratios) to present saliency detection results of our scheme, human labeled (as reference) and benchmark software Saliency Toolbox (STB http://www.saliencytoolbox.net/) in Fig. 6. In addition, “proto-regions”, shown in the red-circled regions in Fig. 6 are found by a thresholding method in [2]. MOS score shows that our saliency map better approximates humans’ experiences.

2) Cropping-scaling retargeting performance on images. We compare our retargeting performance with two representative image retargeting algorithms: Bidirectional Similarity (BS) [12] and Seam
Fig. 6. Comparison of saliency detection on images. Col.1: original image. Col.2: human labeled salient regions. Col.3: proto-regions detected by STB. Col.4: saliency map by STB. Col.5: proto-regions detected by our method. Col.6: saliency map of our method.

Carving (SC) [6]. We use test data from [12] and show the results in Fig. 7. It is clear that our approach does the best to keep the most important regions without distortion.

3) saliency detection and retargeting results on video sequences. Fig. 8 shows the spatial-temporal saliency map detected by our approach and the baseline-PQFT [3] and the corresponding video retargeting result using our retargeting framework. We can observe our saliency map better incorporates the motion features thus yields a better retargeting result.

We should mention that due to the page limitation, our work on video retargeting is not fully presented in this paper. We have another paper [13] in ICIP2010 simultaneously. This paper focuses on saliency detection and single frame video retargeting, and the other one focuses on the temporal smoothness of the single frame retargeting window path. We test our algorithm on a variety of videos including clips from popular movies “UP” and “Avatar”. Fig. 1 provides a snapshot of some demo results. More demos can be found at our website 1.

4. CONCLUSIONS

We propose a nonlinear approach to fuse the spatial and temporal saliency maps for video retargeting considering the human vision characteristics. We also present the new content-aware information loss metrics and a hierarchical search scheme we proposed under a cropping and scaling retargeting framework. Experimental results on images and videos show the benefits of our approach. Future works may include incorporating more features, like face and text detection into our saliency framework, and utilizing tracking techniques for a more stable saliency detection.

5. REFERENCES


1http://plaza.ufl.edu/lvtaoran/retargeting.htm