Video Retargeting: Video Saliency and Optical Flow Based Hybrid Approach

Çiğdem Koçberber and **Albert Ali Salah**
Computer Engineering Department
Boğaziçi University
Istanbul/Turkey
{hatice.kocberber, salah}@boun.edu.tr

Abstract

As smart phones, tablets and similar computing devices become an integral part of our lives, we increasingly watch various types of streaming visual data on the display of those devices, especially as cloud video services become ubiquitous. One challenge is to present videos on diverse devices with an acceptable quality. Video retargeting is the key technology in video adaptation of cloud based video streaming. The most important challenge of video retargeting is to retain the shape of important objects, while ensuring temporal smoothness and coherence. We propose in this paper a new approach that adopts to the content of the video. We describe a cropping video retargeting method that ensures temporal coherence while enforcing spatial constraints by a saliency method. The average motion dynamics is calculated for each frame with optical flow and merged with the information of the user attention model for a given video. The resulting information is used to estimate a cropping window size. The output is a video that preserves important actions and the important parts of the scene. The results are promising in respect to overcoming the temporal and spatial challenges of video retargeting.

1 Introduction

Video retargeting is changing the aspect ratio of a video in order to fit it in a target display. While small sized hand devices as tablets and smartphones become popular, increasing number of people are using these devices to view videos. Considering the various sizes of the displays, video retargeting is gaining importance since it ensures a better viewing experience for diverse aspect ratios.

A good retargeting solution should keep salient parts of each frame, ensure temporal smoothness and coherence, and keep distortion low. While early works on video retargeting focus mostly on the first constraint (Wolf et al. 2007), recent works also focuses on the third aspect (Wang et al., 2011; Grundmann et al., 2010; Yan et al., 2013). Especially while retargeting movies, there are some additional constraints, such as preserving the atmosphere and the mood of a shot created by the director.

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. In this work, we propose a hybrid approach to remedy some of their shortcomings for video retargeting. In addition, we propose a cropping video retargeting algorithm. Our solution involves cropping of the salient parts with action flow considerations. Existing cropping methods (Deselaers, Dreuw & Ney, 2008; Liu & Gleicher, 2006) sometimes produce virtual camera motions and artificial scene cuts, and subsequently, important objects might be discarded. These deficiencies can cause the inability to convey the visual concept of the original video, e.g. the tone and the mood. It is important to preserve the visual concept while balancing between keeping salient parts. Our method adapts to the video saliency while satisfying the action flow in the scene, which is a result of camera motion plus the object motion. The most important parts of a frame are always retained while virtual scene cuts are barely perceivable.

This work is organized as follows; we describe video retargeting applications in Section 2. Section 3 describes the proposed retargeting approach. Section 4 details our experimental setup, including the data and annotations we have used, followed by our conclusions in Section 5.

2 Related Work

2.1 Image Retargeting

We can divide image retargeting approaches into three main categories: Seam carving (Hwang & Chien, 2008; Avidan & Shamir, 2007), image warping (Glasbey & Mardia, 1998; Liu & Gleicher, 2005) and cropping (Suh, Ling, Bederson, & Jacobs, 2003). Early works of image retargeting mainly concentrate on keeping the salient parts of the image, and removing unnecessary parts.

Cropping is the most basic way to resize an image into a smaller display. The important part is selecting where to crop. This method has been criticized for creating virtual camera movements (Yan et al., 2013; Wang et al., 2011). Seam carving is an example method, where consecutive pixel lines that cut each frame vertically or horizontally (called seams) are removed from the image. The lines are not necessarily straight lines, which enables the preservation

This is an unedited author copy. Please cite as: Kocberber, C., A.A. Salah, "Video Retargeting: Video Saliency and Optical Flow Based Hybrid Approach," 3rd Workshop on Intelligent Cinematography and Editing (at AAAI), Quebec City, July 27, 2014.

of salient parts even if they reside at the corners of the image. Image warping, on the other hand, is done by distorting the image in order to transform it to a different scale.

2.2 Video Retargeting

Video retargeting has gained importance with the introduction of smartphones and tablets. These have limited display sizes, and are frequently used in the display of visual context.

The work 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 (Grundmann et al., 2010; Kopf et al., 2009). 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.

Recent studies focus on motion flow of the video as well as spatial saliency. Wang et al. (2011) propose a method to maintain the motion flow. They first resize the video, taking only salient parts, and then reconstruct the motion flow by examining defected motion flow instances. This method produces good results but it is computationally costly.

The work of (Yan et al., 2013) presents a motion aware seam carving technique, which can conserve temporal coherence. The first step is to calculate seams for each frame. For each frame, the seams of the previous frame creates a matching area that needs to be preserved. Seams are recalculated taking into account the matching area. The resulting seams keep the moving objects undistorted, thus preserving the motion flow of the video. This method is a good adaptation of seam carving, but produces distortion and waving effects in the background for some camera movements, especially with zooming.

2.3 Challenges of Retargeting

The quality of retargeted videos can be measured in terms of distortion, waving effects and motion flow preservation. (Wang et al. 2011; Yan et al., 2013) Major limitations of the main video retargeting approaches can be listed as follows:

- *Seam carving* can fail to adapt to dynamic content.
- *Warping* can cause distortions in background.
- *Cropping* can cause virtual camera movements.

Seam carving and warping 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 enlarging 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 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 Method

We propose a hybrid video retargeting method that will analyze the input video, and apply the most suitable video retargeting method per shot.

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.

For training the SVM, we have used a radial basis function kernel with sequential minimal optimization method (Platt, 1998).

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 class $fast$. For shots belonging to $slow$ class, we have applied an improved seam carving approach (Yan et al. 2013). Note that this hybrid approach can also be used with different video retargeting methods.

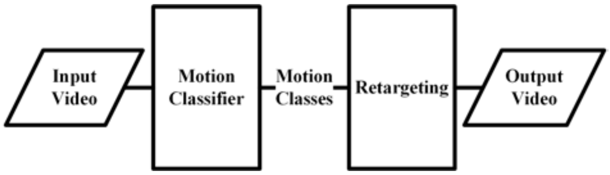

Figure 1: Overall process flow of the proposed method.

3.1 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.

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

$$M_i = \phi(\sigma_{H^i}, \mu_{H^i}), \tag{2}$$

The $slow$ class represent the frames that contain minimum action and a rather slow camera motion, whereas the $fast$ 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

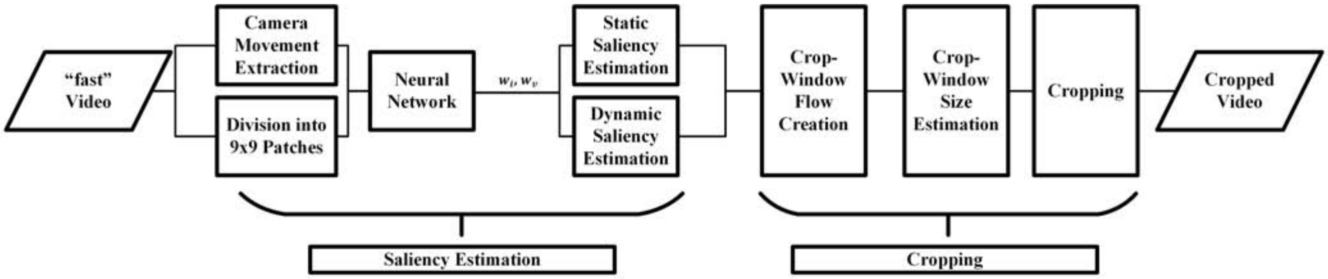
"fast" Video
Camera Movement Extraction
Division into 9x9 Patches
Neural Network
w_i, w_v
Static Saliency Estimation
Dynamic Saliency Estimation
Crop-Window Flow Creation
Crop-Window Size Estimation
Cropping
Cropped Video
Saliency Estimation
Cropping

Figure 2: The process flow of the proposed cropping method.

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).

We have chosen to use homography matrices as a representative of the overall motion, since they contain information about both the action in the movie and the camera 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.

3.2 Video Retargeting

According to the result of the motion classifier, we apply two different video retargeting algorithms. For shots belonging to $fast$ class, we apply a novel cropping approach, and for shots belonging to $slow$ class, we apply a recent seam carving approach (Yan et al., 2013).

Cropping with Optical Flow and Video Saliency : The main problem of crop based methods 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 camera motion to minimize virtual camera movements while preserving salient objects.

The first step of retargeting process is extracting the saliency map. We assume the use of a computational saliency algorithm to produce from each video frame a saliency map S to represent the most interesting and informative parts of the frame. A recent study on computational saliency can be found in (Nguyen et al., 2013).

Let $S = S^1, S^2, ..., S^n$ denote the set of video saliency maps, where n is the number of frames. $S^i = S^i_1, S^i_2, ...S^i_n$ is the set of image saliency maps and $S^o = S^o_1, S^o_2, ...S^o_n$ is the set of optical flow maps. Video saliency S can be obtained as follows:

$$S = w_i * S^i + w_v * S^o, \tag{3}$$

where w_i and w_v are the weights of the corresponding saliency map and optical flow map. For finding S^i and S^o we have utilized saliency detection algorithm of Judd et al. (2009), and the optical flow extraction algorithm of Liu (2009). Details for finding weights w_i and w_v can be found in the study of Nguyen et al. (2013).

After computing S for each frame, we use non-maxima suppression to reduce the processing load. We calculate the center of video saliency C^i by taking the mean of the resulting pixels.

We update the optical flow map S^o for each frame as follows:

$$T = \overline{S^o} * t \tag{4}$$

$$\widetilde{S}^o_{x,y} = \begin{cases} S^o_{x,y} & (-T < S^o_{x,y} < T) \\ 0 & otherwise \end{cases} \tag{5}$$

$\widetilde{S}^o$ is the updated optical flow map, where pixels having a value greater/smaller than threshold T is removed. Removed pixels correspond to moving objects in the frame, since their value diverge from the average. We have set t as 0.9. Mean of the updated optical flow map $\mu_{\widetilde{S}^o}$ corresponds to the camera motion.

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 able to follow the camera motion in order to avoid virtual camera movements in x and y directions.

$$C^o_1 = C^i_1, \tag{6}$$

$$C^o_i = C^o_{i-1} + \mu_{\widetilde{S}^o_i} \quad i = 2, 3, ..., n \tag{7}$$

For the first frame of each shot, we use the center of video saliency map C^i_1 as the center of the crop window (6). For the rest of the frames, centers are shifted by the camera motion $\mu_{\widetilde{S}^o}$ of the current frame (7).

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

$$W = (x, y) \quad s.t. \quad S_{x,y} > 0 \tag{8}$$

Crop window size is determined as the minimum size possible that will include all points in W. The points in W that cause the crop window to exceed the frame boundaries are

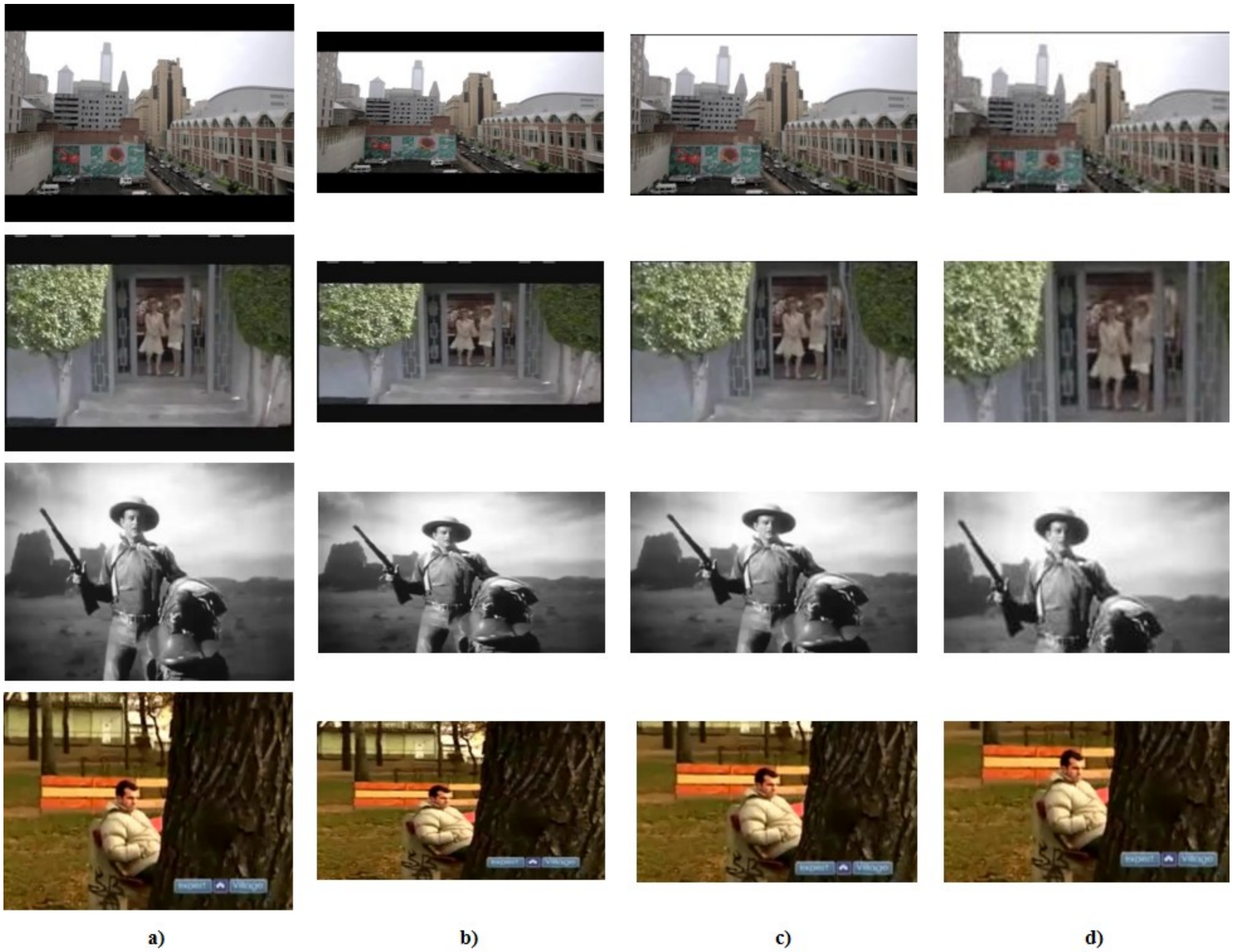

Figure 3: a) The original frames. b) Results of linear scaling. c) Results of seam carving (Yan et al., 2013) d) Results of proposed crop based retargeting. Figure best viewed in color, where problems of both approaches become obvious.

discarded. Since the size of the crop window is fixed per shot, virtual camera movement in z direction is eliminated. The video saliency computation ensures that the relevant objects are included in the cropped frame.

The flow of cropping process can be seen in Figure 2.

Seam Carving : The major limitation of a classical seam carving method is to preserve motion flow. We have implemented a recent study on an improved seam carving algorithm that minimizes this problem (Yan et al., 2013).

Let $E_i = \{(1, p_1), (2, p_2), ..., (H, p_H)\}$ be a seam selected from ith frame. Elements of E_i are the pixel coordinates of the seam where H is the height of the frame. We perform a key point selection where N_{KP} is the number of key points. For each seam S_i, we divide each frame horizontally into N_{KP} equal regions, denoted with R_i.

$$R_i = \left\{ x | (i-1) \times \frac{H}{N_{KP}} < x \leq i \times \frac{H}{N_{KP}}, x \in N \right\}, \tag{9}$$

$$\begin{aligned} K_i = \{(x_j, y_j) | 1 \leq j \leq N_{KP}, \\ (x_j, y_j) \in E_i, x_j \in R_j, \forall (x_k, y_k) \in E_i, \\ x_k \in R_j : EM(x_j, y_j) \geq EM(x_k, y_k)\} \end{aligned} \tag{10}$$

EM is the energy function that contains the gradient and saliency information of the image. K_i is the set of key points of ith frame. These key points are used to update E_{i+1}.

The area surrounding K_i on the ith frame is compared pixel by pixel that surrounds E_{i+1} on the $(i+1)$th frame. Pixels of $(i+1)$th frame that exceed a threshold t in their match scores are rewarded. EM is updated with the rewarded values, and E_i is recalculated.

Since the seams are calculated by taking into account of the consecutive frames, this method is able to adapt to the motion flow of the video. This decreases the motion artifacts, by shifting the distorted areas to the background of the frame. The distortions are not visible in rather stationary shots. Just as in warping based methods, this method can cause waving effects and distortions in the background when it fails to adapt to the fast motion.

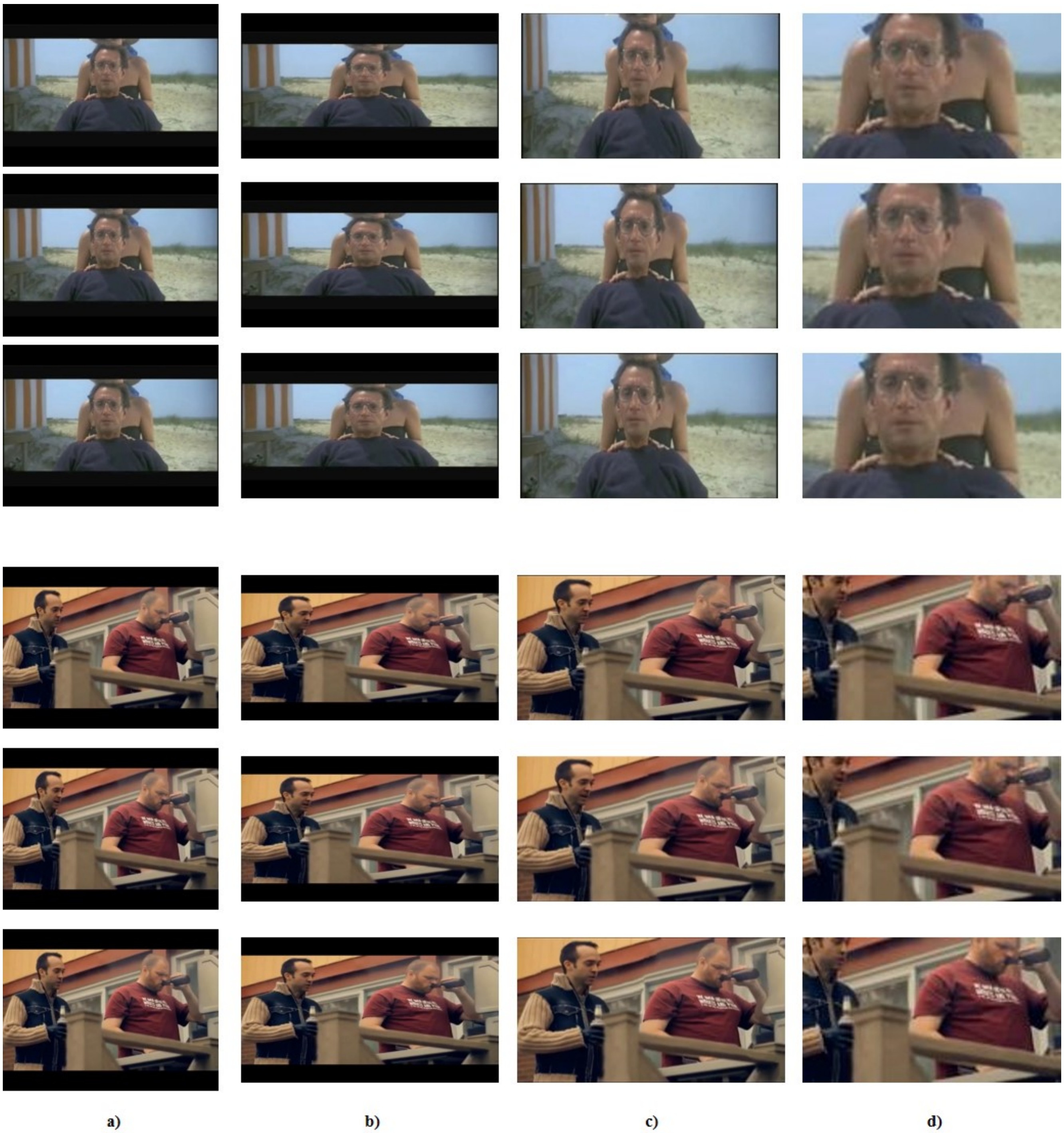

Figure 4: a) The original frames. b) Results of linear scaling. c) Results of seam carving (Yan et al., 2013) d) Results of proposed crop based retargeting. Figure best viewed in color, where problems of both approaches become obvious.

4 Evaluation & Results

4.1 Dataset

We have evaluated our approach on the CAMO dataset (Nguyen et al., 2013). CAMO dataset contains 120 videos of 6 different camera motions: dolly, zoom, trucking, tilt, pan and pedestal. Each video contains a single camera motion in a given shot.

4.2 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. Figure 3 shows several examples.

The first two rows are taken from the $slow$ class. Columns show the original frame, results of linear scaling, seam carving and proposed cropping approach. Salient points are not focused on a specific object, but are rather distributed across the frame. These two cases illustrate the limitations of the crop based method, which tries to capture all the salient points, resulting in an inefficient result. Since the motion in these videos is slow, there is no waving effects on seam carving results. The last two rows in Figure 3 are taken from the $fast$ class and illustrate the limitations of seam carving. In both frames, saliency is focused around a center, making crop based method more effective. Seam carving is not able to adapt the camera motion and produces waving effects in the background.

We have illustrated some limitations of both algorithms on Figure 4. The results supports each method should be applied to separate cases. Frames are taken from two video sequences. In the first case, the class of the video is $fast$. 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. In such case, cropping misses some important parts of the frame.

4.3 Evaluation of the Motion Classifier

CAMO dataset was annotated according to the camera motion. Additionally, we have annotated 36 of the videos as $fast$ or $slow$.

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. We have used 200 samples from each class for training, and 100 samples from each class for testing SVM. The confusion matrix and the results of the test can be seen in Table 1 and Table 2.

	Fast	Slow
$fast$	60	14
$slow$	40	86

Table 1: Classification results. Columns represent the gold standard and the rows represent test results.

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

Table 2: Performance measures of SVM.

The movies annotated as $slow$ tend to keep a low amount of motion throughout the video. As opposed to that, the movies annotated as $fast$ 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.

5 Conclusions & Future Work

In this study, we propose to use homography to identify a given video according to the rate of change in its content, and apply seam carving or cropping based video retargeting approach depending on the result. We use a recent video saliency approach to keep track of relevant content, and propose a novel cropping method to eliminate virtual camera motion. The resultant hybrid algorithm produces good qualitative results on the CAMO benchmark.

References

Avidan, S., & Shamir, A. 2007. Seam carving for content-aware image resizing. In *ACM TOG*, Vol. 26. No. 3.

Shamir, A., & Avidan, S. 2009. Seam carving for media retargeting *Communications of the ACM*, 52(1), 77-85.

Grundmann, M., Kwatra, V., Han, M., & Essa, I. 2010. Discontinuous seam-carving for video retargeting. In *CVPR*, 569-576.

Kopf, S., Kiess, J., Lemelson, H., & Effelsberg, W. 2009. FSCAV: fast seam carving for size adaptation of videos. In *ACM MM*, 321-330.

Hwang, D. S., & Chien, S. Y. 2008. Content-aware image resizing using perceptual seam carving with human attention model. In *IEEE ICME*, 1029-1032.

Liu, F., & Gleicher, M. 2005. Automatic image retargeting with fisheye-view warping. In *ACM UIST*, 153-162.

Krähenbühl, P., Lang, M., Hornung, A., & Gross, M. 2009. A system for retargeting of streaming video. In *ACM TOG* Vol. 28, No. 5, p. 126.

Glasbey, C. A., & Mardia, K. V. 1998. A review of image-warping methods. *Journal of applied statistics*, 25(2), 155-171.

Wang, Y. S., Hsiao, J. H., Sorkine, O., & Lee, T. Y. 2011. Scalable and coherent video resizing with per-frame optimization. In *ACM TOG* Vol. 30, No. 4, p. 88.

Wang, Y. S., Lin, H. C., Sorkine, O., & Lee, T. Y. 2010. Motion-based video retargeting with optimized crop-and-warp. *ACM TOG*, 29(4), 90.

Liu, C. 2009. Beyond pixels: exploring new representations and applications for motion analysis. Doctoral dissertation, Massachusetts Institute of Technology).

Rubinstein, M., Gutierrez, D., Sorkine, O., & Shamir, A. 2010. A comparative study of image retargeting. In *ACM TOG* Vol. 29, No. 6, p. 160.

Deselaers, T., Dreuw, P., & Ney, H. 2008. Pan, zoom, scantime-coherent, trained automatic video cropping. In *CVPR*,. 1-8.

Liu, F., & Gleicher, M. 2006. Video retargeting: automating pan and scan. In *ACM MM*, 241-250.

Yan, B., Sun, K., & Liu, L. 2013. Matching-area-based seam carving for video retargeting. *Circuits and Systems for Video Technology, IEEE Transactions* on 23(2), 302-310.

Wolf, L., Guttmann, M., & Cohen-Or, D. 2007. Non-homogeneous content-driven video-retargeting. In *IEEE ICCV*, 1-6.

Zhai, Y., & Shah, M. 2006. Visual attention detection in video sequences using spatiotemporal cues. In *ACM MM*, 815-824.

Li, J., Tian, Y., Huang, T., & Gao, W. 2010. Probabilistic multi-task learning for visual saliency estimation in video. *International journal of computer vision* 90(2), 150-165.

Nguyen, T. V., Xu, M., Gao, G., Kankanhalli, M., Tian, Q., & Yan, S. 2013. Static saliency vs. dynamic saliency: a comparative study. In *ACM MM*, 987-996.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. 2009. Learning to predict where humans look. In *IEEE ICCV*, 2106-2113.

Suh, B., Ling, H., Bederson, B. B., & Jacobs, D. W. 2003. Automatic thumbnail cropping and its effectiveness. In *ACM UIST*, 95-104.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *ISO 690*