A FAST SMART-CROPPING METHOD AND DATASET FOR VIDEO RETARGETING

Konstantinos Apostolidis, Vasileios Mezaris

Information Technologies Institute (ITI), CERTH
Thermi 57001, Thessaloniki, Greece
Email: {kapost,bmezaris}@iti.gr

ABSTRACT

In this paper a method that re-targets a video to a different aspect ratio using cropping is presented. We argue that cropping methods are more suitable for video aspect ratio transformation when the minimization of semantic distortions is a prerequisite. For our method, we utilize visual saliency to find the image regions of attention, and we employ a filtering-through-clustering technique to select the main region of focus. We additionally introduce the first publicly available benchmark dataset for video cropping, annotated by 6 human subjects. Experimental evaluation on the introduced dataset shows the competitiveness of our method.

Index Terms— video aspect ratio retargeting, cropping, saliency detection, clustering

1. INTRODUCTION

Videos created for traditional TV and desktop computer monitors are typically consumed in landscape aspect ratios (16:9 or 4:3). With the rise of mobile devices (mobile phones and tablets), these historical aspect ratios do not deliver the optimal user experience. Due to the widespread usage of such devices many video sharing platforms now dictate the use of specific video aspect ratios. In order to be published on these platforms, existing videos would have to be transformed to comply with their specifications. A straightforward approach for retargeting a video to a different aspect ratio would involve either static cropping of content or padding the frames with black borders to reach the target aspect ratio. However, static cropping can lead to a significant loss of visual content that might even be in the center of attention, while padding shrinks the original video by introducing large borders in the output video. Ultimately, the results of these simple approaches are often unsatisfactory. Furthermore, common video aspect ratio transformation methods of the literature often introduce distortions and may alter the semantics of the video.

In this paper we present a method for video aspect ratio transformation that minimizes the loss of semantically important visual content. The contributions of this work are two:

- We present a new, rather simple, yet fast and well-performing, video cropping method, which selects the main focus out of the multiple possible salient regions of the video by introducing a new filtering-through-clustering processing step.
- We introduce the first publicly available benchmark dataset for video retargeting, comprising ground-truth video cropping results for 200 videos, where each video was annotated by multiple human annotators considering two possible target aspect ratios.

2. RELATED WORK

The video aspect ratio transformation algorithms of the literature can be divided in three main categories: a) warping [1, 2], b) cropping [3, 4, 5, 6], and c) seam carving [7, 8]. Warping methods, instead of resizing the entire image uniformly, determine scaling factors in a content-adaptive way: the image is divided using a grid and important image regions are left untouched, while scale factors are applied to other less-important areas. Cropping techniques select a rectangular area in the image/frame and discard visual content outside of it. Seam carving algorithms remove seams of uninteresting pixels, i.e., connected paths of pixels inside the image are discarded. Finally, there are also multi-operator techniques that combine two or more operations (e.g. cropping and warping [9], or seam carving and cropping [10]).

It is easily understood that when applying warping or seam carving to the frames of a video, apart from undesirable artifacts introduced [11], the original video content is distorted significantly [12]. For example, let us consider a video frame depicting two persons at the edges of the image with the content between them being a uniform background. A warping method would shrink the uniform area while a seam carving method would probably entirely remove this area, making the two persons appear as if they were standing side-by-side. There are certain usage scenarios where such strong semantic distortions are unacceptable. Based on the above, we argue that cropping methods are more suitable for video aspect ratio transformation when the minimization of semantic distortions is a prerequisite, as they select a region of interest in the video frames but do so without introducing

This work was supported by the EU Horizon 2020 research and innovation programme under grant agreement H2020-780656 ReTV.

DOI:10.1109/ICIP42928.2021.9506390

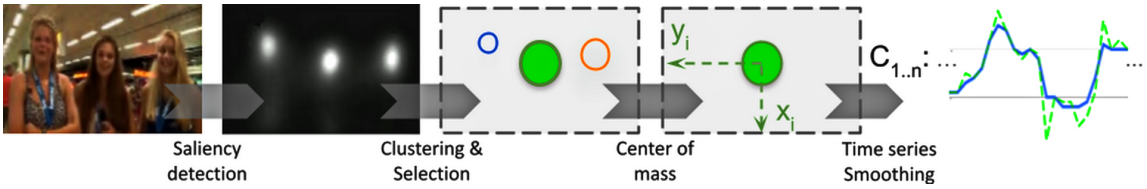


Fig. 1: A simplified overview of the proposed method's stages.

any distortion to the visual content.

Cropping methods most commonly extract some kind of feature to assess the importance of regions in the frames. Then a crop window is fitted in the frame so as to contain the most important regions. Additional effort is taken to ensure the smooth motion of this crop window throughout the video (e.g. in [3] camera operations are derived by optimizing the path of this window, seeking to adhere to the principles of cinematography). In [13] a Structural Similarity feature is proposed based on blur detection to identify whether an image contains a blurred background. In [5] low-level features are employed while in [4, 3] eye-gaze information is utilized. In Google's AutoFlip [14], a solution to smart video reframing, face and object detection results are employed - note that this is the only method, to our knowledge, for which its source code is publicly available. Face detection is also used in other methods (e.g. [15]), where in the latter saliency maps are also calculated and used. Note that, to our knowledge, there is no literature work that considers segmenting user attention to multiple regions and explicitly selecting a single one to focus on.

It is worth highlighting that each of the aforementioned works uses an arbitrary selection of videos to test its results, and most of the time these videos are not provided, while the evaluation procedure relies on visual inspection of selected frames. As opposed to the task of image retargeting, where a standard benchmark dataset exists, the RetargetMe dataset [16], this is not the case for video retargeting. For this reason, objective comparisons of video retargeting methods are difficult and scarce in the literature. Motivated by this, we construct and release RetargetVid, a new benchmark dataset for video retargeting.

3. PROPOSED METHOD

We developed a smart-cropping method that can transform a video to a target aspect ratio different than that of the original. The pseudo-code of the proposed method is listed in Algorithm 1 and an overview is illustrated in Fig. 1. We start by removing the potential black borders in the video frames by excluding the rows and columns of video frames for which the variance of all frame pixels throughout the whole video is below a predefined threshold t_b.

We continue to calculate the dimensions of the crop window that would satisfy the target aspect ratio, while retaining as much as possible of the original video content. For example, when performing a landscape-to-portrait conversion, e.g. transforming a 16:9 video to a 4:5 target aspect ratio, the final crop window height will be equal to the original video's height and the crop window will be able to move only in the X-axis (Fig. 2.a).

To infer the viewer's attention we compute the saliency map for each frame by employing the UNISAL [17] method, which is a well-preforming and fast saliency detection method according to the leaderboard of [18]. We continue to eliminate regions of small saliency by zeroing the pixel values of the saliency map that are below a pre-specified threshold t_s. Note that even after the thresholding procedure, the saliency may be concentrated in a small region of the whole frame or be in the form of multiple blobs. Aiming to select the main part of the viewer's focus, we employ a filtering-through-clustering procedure, where we cluster the location and value of the non-zeroed items of the saliency map. To do so, we employ the HDBSCAN [19] clustering algorithm, because a) it does not require specifying the number of clusters, and b) can identify data points as outliers - excluding these outliers from the rest of the procedure can help discard scattered salient pixels. We select the cluster with the highest weight, expressed as the sum of its items, and zero the rest of the items, producing a filtered saliency map. We go on to find the center of mass of the filtered saliency map and we consider this as a single point center of the viewer's attention. Since the crop window can move either in X-axis (when performing a landscape-to-portrait conversion - see Fig. 2a) or Y-axis (when performing a portrait-to-landscape conversion - see Fig. 2b), we only consider the dimension of interest.

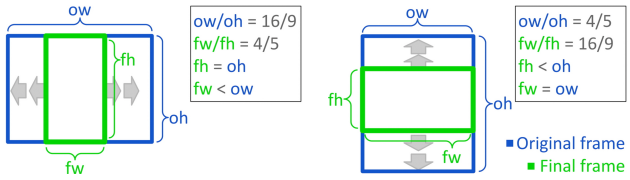


Fig. 2: Calculation of the crop window dimensions that respects the target aspect ratio when transforming a video from a) 16:9 to a 4:5, b) 4:5 to a 16:9 aspect ratio.

After the above described procedure has been conducted for all frames, we end up with C, a 1-dimensional time-series

DOI:10.1109/ICIP42928.2021.9506390

Algorithm 1 Smart-cropping pseudo-code

Input: Video *frames* and target aspect ratio dimensions
Output: Cropped video frames

```
Border detection using all frames
for every frame in frames, with n_skip + 1 step do
    Visual saliency detection
    Saliency map thresholding
    Clustering of saliency map
    Cluster selection
    C ← Center of mass calculation
end for
shots ← Shot detection
for every sub-series in C, corresponding to a single shot do
    Sub-series interpolation
    Sub-series low-pass filtering
    Sub-series LOESS smoothing
end for
Crop window inference for all shots
```

of the displacement of the crop window's center. We employ a variant of [20] to segment the input video to shots and we consider the respective sub-series of the displacement time-series for each detected shot in the video separately. We apply a low-pass filter at 2Hz. Consequently, we employ the LOESS [21] method, a well-known tool used in regression analysis, to fit a smooth curve to each sub-series of crop window movements. We infer the final crop window for each frame based on the smoothed time-series of displacements.

Finally, in an effort to further speed-up the whole process, and based on the fact that temporally-closeby frames of videos most often exhibit high visual similarity, we skip the n_{skip} next consecutive frames for each frame we process. We interpolate the sparse time-series of displacements, prior to the low-pass filtering process. The number of frames we skip was decided upon preliminary experiments and visual inspection of the results, noticing a negligible difference in the quality of the results when $n_{skip} = 4$. Based on the same preliminary experiments, we set $t_b = 30$ and $t_s = 200$, for all subsequent evaluations.

The proposed method was implemented using Python and PyTorch. In the publicly available source code we release[1], we have also implemented a technique to assess the quality of the cropped version and opt to resort to padding in the case where the results of cropping are unsatisfactory due to the nature of the video's content, i.e. only a small percentage of the salient visual information can be covered by any crop window. This is due to acknowledging that a cropped version of a video cannot always retain all regions of interest of the original video. However, this feature is not further discussed in this work, and we have disabled this functionality for all experiments reported in Section 5 in order to test the results of the cropping procedure per se.

[1]Source code and ground-truth annotations for the RetargetVid dataset are publicly available at `https://github.com/bmezaris/RetargetVid`

4. RETARGETVID DATASET CREATION

As discussed in Sec. 1, one of the contributions of this paper is the construction of a dataset for the task of video aspect ratio transformation. In addition to using this dataset to test our approach, we also provide it freely to the scientific community, which is missing a benchmark dataset for this task.

We selected a subset of 200 videos from the publicly available videos of the DHF1k dataset [22], specifically the first 100 videos of the training set (videos 001 to 100) and the 100 videos of the validation set (videos 601 to 700). All videos are in 16:9 aspect ratio and most of them consist of a single shot. The DHF1k dataset was originally constructed as a benchmark for evaluating visual saliency methods. Upon visual inspection, this dataset was considered ideal to provide a balanced set of both easy and challenging videos for the video cropping task.

We invited 6 human subjects and asked them to select the region of each frame that would be ideal to be included in a cropped version of the video. Specifically, we assigned them the task of generating two cropped versions for each video, one with target aspect ratio of 1:3 and another one with target aspect ratio of 3:1. We selected these extreme target aspect ratios (despite not being used in real-life applications) in order to identify human preferences under very demanding circumstances. Moreover, less extreme target aspect ratios can still be evaluated by assessing to what extent an e.g. 9:16 crop window includes the 1:3 manually specified window.

To assist the annotators in their task we implemented a graphical user interface tool (Fig. 3) which facilitates the navigation throughout the video, allows the user to set a crop window for each frame through simple drag-and-drop mouse operations, and overlays the crop window on the video frames to allow for the quick inspection of the user's decisions.

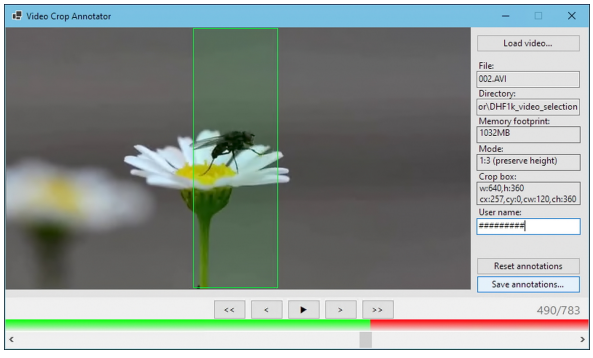
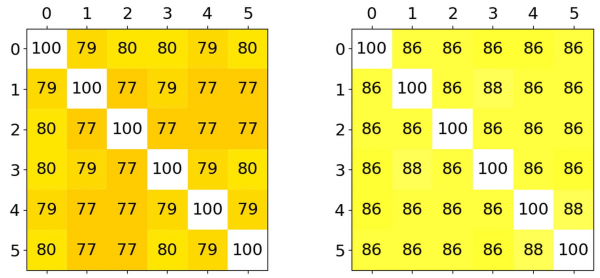


Fig. 3: Screenshot of the graphical interface of the tool implemented to assist the 6 annotators in their task.

We calculated the similarity of annotations between the 6 annotators in terms of the median of the Intersection over

DOI:10.1109/ICIP42928.2021.9506390



(a) (b)

Fig. 4: Similarity matrix of the 6 annotators in terms of the median of IoU score (%) between the crop windows, for (a) the 1:3, and (b) the 3:1 aspect ratio annotations.

Union (IoU) scores of all crop windows The resulting similarity matrices are illustrated in Fig. 4a for the 1:3 aspect ratio and in Fig. 4b for the 3:1 aspect ratio. The average IoU between subjects for the 1:3 aspect ratio is 82.57% while for the 3:1 aspect ratio is 89.18%. This showcases that our dataset is diverse and, while there is a good inter-annotator agreement, the annotations of the 6 subjects are not identical. The videos can be found in the website of the DHF1k dataset[2]. Our crop window ground-truth annotations for each video of the RetargetVid dataset are available online[1]. They are in the form of text files, where the i-th line contains the top-left and bottom-right coordinates of the crop window for the i-th frame.

5. EXPERIMENTS

We utilize the constructed RetargetVid dataset to compare our method to AutoFlip [14] using its default configuration. Regarding our method, we also evaluate an approach where the discussed filtering-through-clustering procedure is disabled and instead of calculating the center of mass of the selected cluster, we utilize the location of the max value in the saliency map. We employed all tested approaches to transform each video in the dataset to: a) 1:3 aspect ratio, and b) 3:1 aspect ratio. To evaluate an approach, we compute the IoU score of its resulting crop window for each frame and the respective ground-truth crop window of an annotator. We average the IoU scores for a video and proceed to take the mean of all videos' scores. We calculate this score for each annotator separately, and we record the best, worst as well as the mean for all annotators. All experiments were conducted on an Intel i5 9600K PC with 32GB of RAM, running Ubuntu 18, equipped with an Nvidia GeForce GPU (RTX 2080Ti); all tested approaches were configured to utilize the GPU.

In Table 1, we present the results of our comparisons. The second and third columns report the worst and the best scores considering the ground truth of a single annotator, respectively, while the fourth column is the mean IoU for all annotators. In the last column we report the ratio of the execution time to the video duration (the lower the better). We observe that AutoFlip provides the best results by a small margin, while the proposed method is over 100% faster than AutoFlip (the processing requires about one fifth of the duration of the original video). We also notice that the proposed filtering-through-clustering step of our method increases the mean IoU scores from 46.1% to 49.9% for the 1:3 target aspect ratio and from 68.1% to 71.4% for the 3:1 target aspect ratio.

Method	Worst	Best	Mean	t (%)
Results for 1:3 target aspect ratio				
AutoFlip	50.0	52.1	50.8	40
Ours (w/o clustering step)	45.3	47.4	46.1	17
Ours	48.6	50.9	49.9	19
Results for 3:1 target aspect ratio				
AutoFlip	70.9	74.7	72.2	41
Ours (w/o clustering step)	66.3	72.9	68.1	17
Ours	70.1	73.6	71.4	20

Table 1: Comparison of the proposed method and AutoFlip in terms on mean, worst and best IoU on the subjects annotations. All numbers are percentages.

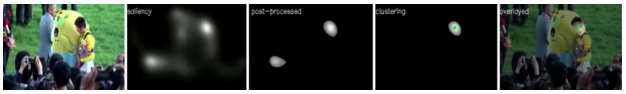


Fig. 5: Example frame from processing a video of the DHF1k dataset. Note how the clustering procedure filters out the salient region in the lower left area of the frame.

For a qualitative analysis regarding the importance of the clustering stage see Fig. 5, where the original frame, the inferred saliency map, the thresholded saliency map, the result of applying the clustering procedure and the filtered saliency map overlaid on top of the original frame, are depicted respectively. We observe that there are multiple salient blobs and the designed filtering-through-clustering procedure manages to select the blob of the main focus.

6. CONCLUSIONS

We presented a method that re-targets a video to a different aspect ratio by cropping unnecessary regions of the video frames based on visual saliency. We showed the merits of a filtering-through-clustering technique to select only a main object of focus. Additionally, we introduced a publicly available benchmark dataset for video cropping, RetargetVid. We strongly believe that the provision of such a dataset, as well as the availability of an easy-to-deploy source code for a fast baseline method, will help promote research in the domain of video aspect ratio transformation.

[2] https://github.com/wenguanwang/DHF1K

DOI:10.1109/ICIP42928.2021.9506390

7. REFERENCES

[1] H. Nam, D. Park, and K. Jeon, "Jitter-robust video retargeting with kalman filter and attention saliency fusion network," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 858–862.

[2] H.-S. Lee, G. Bae, S.-I. Cho, Y.-H. Kim, and S. Kang, "Smartgrid: Video retargeting with spatiotemporal grid optimization," *IEEE Access*, vol. 7, pp. 127564–127579, 2019.

[3] K.-K. Rachavarapu, M. Kumar, V. Gandhi, and R. Subramanian, "Watch to edit: Video retargeting using gaze," in *Computer Graphics Forum*. Wiley Online Library, 2018, vol. 37, pp. 205–215.

[4] E. Jain, Y. Sheikh, A. Shamir, and J. Hodgins, "Gaze-driven video re-editing," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, pp. 1–12, 2015.

[5] T. Deselaers, P. Dreuw, and H. Ney, "Pan, zoom, scan – time-coherent, trained automatic video cropping," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[6] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," in *Proceedings of the 14th ACM International Conference on Multimedia*, 2006, pp. 241–250.

[7] H. Kaur, S. Kour, and D. Sen, "Video retargeting through spatio-temporal seam carving using kalman filter," *IET Image Processing*, vol. 13, no. 11, pp. 1862–1871, 2019.

[8] S. Wang, Z. Tang, W. Dong, and J. Yao, "Multi-operator video retargeting method based on improved seam carving," in *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, 2020, pp. 1609–1614.

[9] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," in *ACM SIGGRAPH 2010 papers*, pp. 1–9. Association for Computing Machinery, 2010.

[10] S. Kopf, T. Haenselmann, J. Kiess, B. Guthier, and W. Effelsberg, "Algorithms for video retargeting," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 819–861, Jan. 2011.

[11] J. Kiess, B. Guthier, S. Kopf, and W. Effelsberg, "Seamcrop for image retargeting," in *Multimedia on Mobile Devices 2012; and Multimedia Content Access: Algorithms and Systems VI*. International Society for Optics and Photonics, 2012, vol. 8304, p. 83040K.

[12] S.-H. Nam, W. Ahn, I.-J. Yu, M.-J. Kwon, M. Son, and H.-K. Lee, "Deep convolutional neural network for identifying seam-carving forgery," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[13] Y.-C. Chou, C.-Y. Fang, P.-C. Su, and Y.-C. Chien, "Content-based cropping using visual saliency and blur detection," in *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)*. IEEE, 2017, pp. 1–6.

[14] "Google AutoFlip," `https://google.github.io/mediapipe/solutions/autoflip`, Accessed: 2021-01-25.

[15] X. Fan, X. Xie, H.-Q. Zhou, and W.-Y. Ma, "Looking into video frames on small displays," in *Proceedings of the eleventh ACM International Conference on Multimedia*, 2003, pp. 247–250.

[16] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," in *ACM SIGGRAPH Asia 2010 papers*, pp. 1–10. Association for Computing Machinery, 2010.

[17] R. Droste, J. Jiao, and J. A. Noble, "Unified Image and Video Saliency Modeling," in *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020.

[18] "DHF1K video saliency leaderboard," `https://mmcheng.net/videosal/`, Accessed: 2021-01-25.

[19] L. McInnes and J. Healy, "Accelerated hierarchical density based clustering," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 33–42.

[20] M. Gygli, "Ridiculously fast shot boundary detection with fully convolutional neural networks," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2018, pp. 1–4.

[21] W.-S. Cleveland and S.-J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American statistical association*, vol. 83, no. 403, pp. 596–610, 1988.

[22] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

DOI:10.1109/ICIP42928.2021.9506390