IMPROVED OPTIMAL SEAM SELECTION BLENDING FOR FAST VIDEO STITCHING OF VIDEOS CAPTURED FROM FREELY MOVING DEVICES

Motaz El-Saban, Mostafa Izz, Ayman Kaheel and Mahmoud Refaat

Cairo Microsoft Innovation Lab

ABSTRACT

We investigate the problem of stitching timely synchronized video streams captured by freely moving devices. Recently, it was shown that using frame-to-frame correlation can greatly enhance the efficiency and effectiveness of video stitching algorithms [19]. In this paper, we address some of the shortcomings in [19], namely the simple blending approach, causing almost a third of stitching errors and the fact that the stitching algorithm is only tested on a frame-by-frame basis which does not realistically mimic the user perception of the output system quality as a complete video. We propose the use of a modified blending technique based on optimal seam selection and experimentally validate its superiority using precision, recall and F1 measures on a frame-by-frame basis, while maintaining low computational complexity. Furthermore, we validate that the performance gains measured on a frame-by-frame basis are also evident when the stitched video output is evaluated as a single unit.

Index Terms— Image Stitching, Video Stitching, Blending, optimal seam selection.

1. INTRODUCTION

The goal of video stitching is to construct a single panoramic video output by stitching together *N* input video streams. This technology is important in several domains, such as security (through surveillance cameras) and entertainment. In the case of security applications, the cameras are mostly fixed in location, while in entertainment applications video feeds, simultaneously captured by different users in the same scene, are obtained by free-to-move capturing devices. The latter is particularly interesting given the wide availability of consumer-level capturing devices, such as mobile phones and digital camcorders.

The case of video stitching from fixed camera positions can be assumed solved to a large extent as the "image-based" stitching problem needs to be solved once during initialization and then occasional adjustments (mostly photometric) may be needed on a very low frame rate basis. In contrast to the fixed cameras case, when the capturing cameras are free to move, the problem becomes much more challenging since it needs to be solved for every frame. In scenarios that require producing the panoramic video in real-time, naively using an image-based stitching algorithm on every frame would be too computationally expensive and would not fully consider the specific nature of stitching videos, such as the perception of alignment/compositing errors. Utilizing time information in stitching videos was investigated in [19], and it was shown that the use of an optical flow-based tracker improves both the speed of video stitching algorithms and the quality of the output video. One of the main results of the work in [19] is that a decent percentage (approx. 30%) of the perceived video sititching errors was due to blending. In order to have a fast stitching pipeline, the authors in [19] used a rather simple feathering [2] approach for blending. This illustrates the need for a fast and effective blending technique for video streams in scenarios with freely moving cameras. In this paper, we make the following contributions:

- We propose to use of an improved blending technique and show experimentally that it increases precision/recall and F1-measures while maintaining low computational complexity.
- We propose a novel video-based method for judging video stitching results as opposed to the frame-by-frame evaluation used in [19]. Simply using a frame-by-frame does not fully mimic real user perception of the video stitching output.

The paper is organized as follows. The closest related work is summarized in Section 2. An overview of the proposed algorithm for video stitching is given in Section 3. Section 4 describes the used dataset, the experimental results and the error analysis. Conclusions and future research points are given in section 5.

2. RELATED WORK

While the literature on image stitching is very wealthy [20][2], very little work has been done on the front of producing a video composite, especially considering the case when the capturing devices are allowed to move freely, as opposed to the case where the cameras are fixed [4][5]. The work in [1] described fast techniques for stitching videos, however, the authors only showed objective alignment results for pure translation transformation videos which is not a realistic situation for freely moving cameras. In our case, we report objective results for the full affine case on a large video dataset collected under varying

conditions. The work in [9][12] solved the video stitching problem (both alignment and composition) on a frame-by-frame basis, which is highly inefficient and does not fully exploit the correlation between video frames. Content correlation between frames was exploited in [19], however this approach used a rather simple feathering technique responsible for as much as third of the stitching errors.

Blending techniques for stitching can be categorized into two main categories: 1) smooth-transition techniques and 2) optimal seam finding techniques. Smooth-transition techniques range from simple techniques such as feathering, which are very fast but don't usually produce acceptable results, to more involved techniques such as gradient domain techniques which can be considered as a smooth-transition but in the gradient domain. Gradient domain techniques were firstly introduced in 1983 [25], and regained popularity again in 2003 [24]. Although gradient domain techniques give very good results they require recovering the final composite from the gradient domain at a high computational cost. The second category of blending techniques are based on optimal-seam finding, in other words, they try to find the best pixels from each image that reduce the artifacts and intensity misalignments in the final composite [3][22]. These techniques yield good results (better than simple feathering and somewhat worse than gradient domain techniques), but their calculations are much faster than the gradient domain techniques. There are also hybrid techniques combining ideas from smooth transition and optimal seam finding methods [23]. Blending results for this technique are quite satisfactory however with a very high computational cost.

3. VIDEO STITCHING ALGORITHM OVERVIEW

The video stitching algorithm starts with two synchronized video streams. For the first pair of frames the algorithm applies the two main steps, alignment and compositing as in any regular image stitching algorithm. For the alignment step the algorithm computes interest points (IPs) then calculates a transformation matrix from one frame to the other. For later frames, the calculation of the geometric transformation is speeded up by utilizing temporal redundancy information extracted from the video streams.

The stitching algorithm proposed in [19] utilizes video information during the alignment phase for speed up purposes. It utilizes the optical flow-based tracking algorithm. As reported in [19], the most expensive part in the video stitching algorithm is the IP detection process, so the proposed speed up techniques focused on reducing the time used in computing IPs. The most successful approach is based on using Lucas-Kanade optical flow [11] to track IPs from frame pairs (n-1) to frame pairs (n) with the SURF descriptor [8], which is a speeded-up version of SIFT [7], performing favorably in many of the descriptors evaluations [21]. Once the new locations of IPs in frames (n) are found, their descriptors are obtained from frames (n-1) in order to avoid re-computation. Tracked IPs are filtered by discarding foreground moving objects, which leads to a more stable stitching using background IPs. In the compositing step; we propose a blending technique to address shortcomings related to using a simple feathering (linear weighting) in [19] while maintaining relatively low computational complexity. The proposed technique use an optimal seam finding method, particularly a region-of-difference-based (ROD) [3] method.

3.1. Enhanced Blending

The goal of the proposed blending technique is to enhance the final video composite quality, while maintaining low computational time complexity. The proposed blending approach draws ideas from the region-of-difference (ROD) approach [3] with some modifications, that are deemed experimentally beneficial, in the compensation and pixel selection phases, and hence we will refer to it as MROD (modified-ROD). Our approach is divided into two main steps. The objective of the first step is to perform exposure compensation by removing color differences between stitched pair of images based on mapping color channels in the overlapping region [18]. The second step aims at removing any ghosting artifacts in the overlapping area based on calculating the region of difference (ROD) between the stitched frame pair [3]. In this step, the algorithm is initialized by binarizing the images to only have the pixels with intensity difference more than a certain threshold. RODs from each frame are weighted using two measures: 1) the pixel difference from the frame pairs and 2) the distance between the pixel and the nearest edge in this video frame. To avoid pixel selection on edges, these pixels are penalized with a negative weight, which proved beneficial in our experiments and is different from the original ROD approach. Several post-processing steps are performed to avoid small fragments coming randomly from different frames through morphological erosion, dilation and a clustering step to form the final contiguous RODs. Finally, regions are selected from the source image where these regions are farther from edges.

4. EXPERIMENTAL RESULTS

4.1. Dataset and experiments baseline

The data required to test the proposed video stitching algorithm requires time synchronized videos. There is no such data available as of now and hence we resorted to collect our own dataset using commonly available capturing devices, namely mobile phones with video cameras. This dataset will be made freely available for download. Human data collectors were asked to capture time-stamped videos simultaneously at multiple locations, in different day times, with unrestricted camera motion conditions (both in-plane

and out-of-plane rotations allowed), and from different distances. Videos were captured using mobile phone cameras with a CIF video resolution (352x288) and with an average frame rate of 11 frames/sec (after video encoding using H.263+ encoder). Time synchronization between simultaneously captured videos was performed using a NTP (network time protocol) server [13]. Some individual frames are shown in Fig. 1 to illustrate the frames quality and some of the variability within the dataset.

Fig. 1 Sample frames from the field-trials video set

After data collection, a human judge was asked to manually label a sample of 1288 pairs of frames[1]. The judge was asked to record whether the frames are stitchable by human eye or not; then indicate if the frames are correctly stitched by the proposed algorithm. This evaluation aims at measuring precision/recall and speed for each blending method. The precision (Pr) and recall (Re) values were calculated as $Pr = N/P$ and $Re = N/M$, where N is the number of correctly stitched pairs by the proposed algorithm, M is the total number of possibly stitched pairs (i.e. ground truth which have some overlap) and P is the number of stitchable pairs by the algorithm (sum of the correctly stitched and incorrectly stitched frames). Based on precision and recall values, the $F1$-measures are computed.
In the evaluation dataset, the total number of stitchable frames is 900 frame pairs. Based on the work in [19], the baseline for our experiments is taken to be the optical flow - based video stitching method using feathering for blending and SURF as a descriptor (called "Baseline" thereafter).

4.2. Experiments on improved blending

Given the high percentage of blending errors reported in [19], we experimented with multiple methods for improved blending constrained by reasonably low computational complexity. We report here on two approaches, our proposed blending MROD and another blending approach based on graph cuts [22]. The graph-cut approach was selected as it showed in our experiments good accuracy while maintaining low computational cost. In [22], a graph is constructed from the overlapping region with weights set as the differences in the grayscale from overlapping pixels. Regions are selected form one frame or the other by solving a max flow problem over the graph.
The best parameters for the ROD-based technique were found to be: ROD threshold of 50, one morphological erosion iteration, and one blending iteration. With these parameters, we evaluated the precision/recall, F1-measures and time for the video stitching algorithm as shown in Table 1. The reported results clearly show the superiority of the proposed video stitching method in precision, recall and F1 measures. The cost being paid is a slight increase in computation time, however when compared to that imposed by other blending approaches such as graph-cuts and the interactive digital photomontage framework [23] which achieves very good results but adds an overhead of few seconds over the baseline.

Table 1 Video stitching frame-by-frame results by different blending and alignment methods

Method	Recall	Precision	F1	Time (msec)
Baseline (SURF + Optical Flow + Feathering)	0.55	0.8	0.65	102
SURF + OpticalFlow + MROD	**0.6**	**0.9**	**0.72**	**129**
SURF + Optical-Flow + Graph Cut	0.57	0.83	0.68	133

4.3. Evaluation using the whole video

Evaluating the video stitching results on a frame-by-frame basis is a good first step in optimizing system performance, as it is less forgiving for errors happening on any frame and can identify detailed issues worth to tackle. However, the consumer of such video stitched result will deal with the video as a single unit rather than a number of separate frames. Hence, we conducted another experiment to measure the quality of the whole stitched video to approximate as much as possible the perception of the output by a real user (and as a verification of the frame-by-frame results presented earlier). Here we use the optical flow-based compositing and the ROD-based blending as the best candidates from the frame-by-frame evaluation results.

We show the resulting video to three judges and ask them to rate if the output quality is acceptable or not and to note down any observations. The results show the level of user satisfaction reading three main dimensions: 1) composite video size stability (a video is considered stable in its frame size if the frame size is not constantly changing over time), 2) alignment quality and 3) blending quality. Three human judges were asked to give percentage of satisfaction for each

[1] In this paper, we restrict ourselves to stitching two time aligned video streams.

of the above dimensions on a scale from 0 to 100. The average of their satisfaction scores is presented in Table 2.

Table 2 Satisfaction levels on stitched videos (0 to 100)

Criteria/ Method	Alignment	Blending	Size stability
SURF + optical-flow + feathering	79	78	74
SURF + optical flow + MROD	84	83	76

Results in Table 2 show an increase in user satisfaction level with the improved blending technique. What is more interesting is that the perception of alignment quality and output video size stability has increased though no changes have been done for these parts specifically. This suggests that improved blending can accommodate for some mis-alignment.

5. CONCLUSIONS AND FUTURE WORK

In this work we have presented an algorithm for stitching two timely synchronized video streams using an improved seam selection-based blending method. The proposed algorithm addresses blending issues reported in [19]. Our evaluation of the proposed techniques was carried on both a frame-by-frame basis as well as for the resulting video as a whole to both mimic real-user judgment and pinpoint issues difficult to discern when judging the video as a whole. Based on the reported results the F1 measure is improved by more than 10% compared to [19]. Though this does not fully handle all the stitching error cases resulting from poor blending, it attacks around a third of the cases. It is worth noting that some of the blending errors reported in [19] are not perceived due to poor composition solely, but also due to mis-alignment as evidenced in the reported results in this paper. As far as future work is concerned, better methods are needed for compensating for large difference in 3D viewpoints, and depth/occlusion problems while not adversely affecting efficiency. Furthermore, it would be interesting to compare to efficient image mosaicking approaches from video frames such as [26].

6. REFERENCES

[1] T. Shimizu, A. Yoneyama and Y. Takishima, “A fast video stitching method for motion-compensated frames in compressed video streams”, International Conference on Consumer Electronics, 2006.

[2] R. Szeliski, “Image Alignment and Stitching: A Tutorial”, MSR Tech Report (last updated 2006)

[3] M. Uyttendaele, A. Eden and R. Szeliski, “Eliminating ghosting and exposure artifacts in image mosaics”, CVPR 2001

[4] Kolor autopano, http://www.autopano.net/blog-en/tag/video-stitching/, retrieved April 28, 2011.

[5] MindTree, http://www.slideshare.net/MindTreeLtd/mindtree-video-analytics-suite-real-time-image-stitching-1135870, retrieved April 28, 2011.

[6] P. Paalanen, J. K. Kamarainen and H. Kalviainen, “Image Based Quantitative Mosaic Evaluation with Artificial Video”, Lappeenranta University of Technology, Research Report 106, 2007.

[7] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, IJCV 04.

[8] H. Bay, T. Tuytelaars and L. Gool, “SURF: Speeded Up Robust Features”, ECCV, 2006.

[9] M. El-Saban, M. Refaat, A. Kaheel and A. Abdul Hamid, “ Stitching videos streamed by mobile phones in real-time”, ACM-MM 09.

[10] W. Zeng and H. Zhang, "Depth Adaptive Video Stitching,", Eigth IEEE/ACIS International Conference on Computer and Information Science, ICIS 2009.

[11] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision”, Proceedings of Imaging understanding workshop, 1981

[12] A. Kaheel, M. El-Saban, M. Refaat, and M. Izz, “Mobicast - A system for collaborative event casting using mobile phones”, in ACM Mobile and Ubiquitous Multimedia - MUM '09.

[13] D. Mills, Simple Network time protocol (SNTP), version 4 for IPv4 and IPv6 and OSI. RFC 2030 IETF.

[14] J. Kopf, M. Uyttendale, O. Deussen and M. Cohen, “Capturing and Viewing Gigapixel Images”, SIGGRAPH 2007.

[15] M. Brown and D. Lowe, “Automatic Panoramic Image Stitching using Invariant Features”, ICCV 2007.

[16] C. Doutre and P. Nasiopoulos, “Fast vigneting correction and color matching for panoramic image stitching”, ICIP 2009.

[17] A. Agarwala, “Efficient gradient-domain compositing using quadtrees”, ACM Transactions on Graphics, 2007.

[18] GY Tian, D Gledhill, D Taylor, “Colour correction for panoramic imaging”, Proceedings of the Sixth International Conference on Information Visualization, IV’02. pp. 483-488.

[19] M. El-Saban, M. Izz and A. Kaheel, “Fast stitching of videos captured from freely moving devices by exploiting temporal redundancy”, ICIP 2010.

[20] A. Mills and G. Dudek, "Image stitching with dynamic elements", Image and Vision Computing, Volume 27, Issue 10, 2,2009.

[21] K. Mikolajczyk and C. Schmid, “A Performance Evaluation of Local Descriptors”, TPAMI, October 2005.

[22] Y. Boykov, O. Veksler and R. Zabih, “Fast approximate energy minimization via graph cuts”, IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 11, 2001.

[23] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin and M. Cohen. “Interactive Digital Photomontage”. ACM Transactions on Graphics (Proceedings of SIGGRAPH 2004), 2004.

[24] P. Pérez , M. Gangnet and A. Blake, Poisson image editing, ACM Transactions on Graphics (TOG), v.22 n.3, 2003.

[25] P. Burt and E. Adelson,"A Multiresolution Spline with Application to Image Mosaics", ACM Transactions on Graphics, 2(4): (1983).

[26] W. Zhao, “Flexible image blending for image mosaicing with reduced artifacts”, in International Journal of Pattern Recognition and Artificial Intelligence 2006.