THE METHOD FOR DEFOCUSING SELFIE TAKEN BY MOBILE FRONTAL CAMERA USING BURST SHOT

Sun-Jung Kim, Beom Su Kim, Hong Il Kim, Tae-Hwa Hong and Joo-Young Son

Healthcare & Sensing Lab., DMC R & D Center, Samsung Electronics Co., Korea

ABSTRACT

In this paper, we present a novel human segmentation technique that automatically detects upper-body region in the selfie. To detect and segment upper-body without user interactions, we develop an initial tri-map by combining face detection results and upper-body shape prior in the selfie. Moreover, we employ motion vectors between two images captured in a short time interval to deal with various human poses and cluttered backgrounds. From motion vectors, we estimate the implicit depth layers without auxiliary hardware or time-consuming algorithms. By integrating information from the face detector, shape prior, and motion vectors, we detect and segment the human upper-body accurately. We also implement the proposed algorithm on the mobile phone. In the extensive experiments on selfie dataset, the proposed method shows competitive results in terms of accuracy / recall and outperforms the previous methods.

Index Terms— human segmentation, defocus, graphcut

1. INTRODUCTION

A selfie, a term for a self-portrait photograph, has gained astounding popularity over time by being shared on social networking services such as Facebook, Instagram and Twitter [1, 2]. Especially, defocus effect on background of selfie enables mobile phone images, the entire scene is in-focus due to the small lens, comparable to DSLR cameras which are capable of producing defocus images, and therefore provides a new selfie experience to users (Fig. 1). However, to detect and segment human for defocus without user interactions is a challenging problem due to the cluttered backgrounds, various poses of human, and a wide range of clothes. There are many approaches for defocusing images, some of them estimate the depth map and defocus an image by selecting a desired focal plane [3, 4, 5, 6, 7, 8] and others employ segmentation techniques to detect target objects [9, 10, 11, 12, 13, 14]. The approaches using depth map can be classified into three ways : the methods using 1) special hardware such as depth sensor or pre-calibrated stereo rig [3, 4], 2) segmentation technique with Phase Auto Focus (PAF) [5], and 3) Structure-from-Motion (SfM) algorithm with several images [7, 8]. Although these approaches allow us to handle a range of case using depth map, the needs of special hardware or a complex and time-consuming algorithm are their major drawbacks. Moreover, depth sensors based on infrared are limited to the indoor scene, and most of depth-measuring equipment produce low-quality and low-resolution depth map which may degrade the defocus quality. The methods using PAF require the function of modulating focal length, which is not available in low-cost mobile frontal cameras. The approaches using SfM are inconvenient to use since they require long sequence of images with large displacement and fail on moving objects. Furthermore, all of these methods require user interaction, such as a click or tap on display to select portions of in-focus.

In addition, there are many algorithms which segment certain objects automatically in general scenes [9, 10, 11]. The segmented results can be used to make defocus image by giving blur effect on background excluding target object. However, their pixel accuracies are only about 50%. Moreover, these approaches are time consuming and difficult to be applied in mobile devices since they are based on rich features to describe the region properties. To alleviate the difficulty of segmentation in the general environment, many segmentation methods employ learning scheme [12, 13] or utilize additional information such as shape prior [15]. Though the methods are more robust than general object segmentation, there are limits in applying to the objects with large variation. Particularly, human has various body appearance caused by changing viewpoints, clothing, and limb articulation.

In this paper, we propose a robust human segmentation scheme in the selfie taken by mobile frontal camera. For this, we exploit face detector results and the upper-body shape prior from the noticing that the viewpoints are limited in the selfie. This results in a robust and fully automatic initialization of tri-map, as opposed to interactive techniques [3, 4, 5, 6, 7, 8, 15]. Furthermore, we utilize motion vectors calculated from the two images captured using burst shot to deal with various human poses and patterns of clothes. Since two images are captured in a very short time (90 ms) with a single click, the users do not discern the difference to the general selfie shot. Therefore, the users inconvenience is minimized while maintaining quality of segmentation. By integrating an initial tri-map, RGB and motion information, we obtain the initial segmentation result. The initial segmentation mask

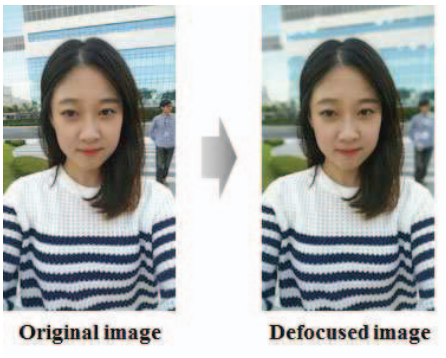

Fig. 1. Defocusing effect on background of selfie.

is refined by iteratively applying the graph cuts optimization [16]. In experiment, the proposed method shows the competitive performance on challenging dataset and outperforms the previous methods.

2. PROPOSED METHOD

Fig. 2 shows the overall flow of the proposed algorithm. Firstly, we calculate motion vectors from the consecutive two images. Then, the upper-body tri-map is estimated based on the facial information and the upper-body prior knowledge in the selfie. Finally, the segmentation is iteratively performed by combining tri-map, RGB information, and motion vectors.

The proposed algorithm uses two consecutive images obtained from the burst shot of mobile frontal camera. Since two images are captured in a very short interval, no additional effort is required after users press the shutter button. Though the time interval of two images is very short, there is inevitable motion due to hand shaking, which results in a motion vectors between two images. At a far distance to the camera center, the magnitude of motion vectors is small, while the magnitude of motion vectors at a near scene (i.e. human) is large. Thus motion vectors from foreground / background are separated as shown in Fig. 3, and it can be calculated using optical flow algorithm [17].

Then, the upper-body tri-map is estimated based on the facial information and the prior knowledge about selfie. The facial information including position of face, 36 facial landmarks, and pose of face (yaw / pitch / roll) are calculated from the first image using the method in [18]. We estimate the initial positions of head and shoulder based on the prior knowledge (ex, the relationship between facial width and shoulder width etc.) obtained from the learned result using selfie database. The tri-map consists of 4 regions (foreground (FG) / probably foreground (PRFG) / probably background (PRBG) / background (BG)). The rule to estimate each point of FG region in the tri-map is shown in Fig.4. The face width (a), shoulder center (c), and shoulder start (d) points are decided by the facial landmarks and the face height (b), half shoulder width (e), and shoulder end (f) are decided by points (a, b, d). The estimated shoulder points are refined by learning the RGB information from initial FG / BG region using Gaussian Mixture Model (GMM) [15]. The points are moved to nearby points in which foreground probability is bigger than that of background to extend or shrink shoulder region as shown in Fig. 5. The PRFG and PRBG region are located with a certain margin from the FG region. The head tri-map of PRFG and PRBG region is different from FG region in order to deal with various hair style. The process of generating tri-map is a fully automatic initialization scheme, as opposed to interactive techniques [3, 4, 5, 6, 7, 8, 15].

From the first image, we obtain the initial segmentation result by applying graph cuts segmentation [16], with RGB information (from the first image), motion vectors (calculated from two images) and an initial tri-map calculated in the previous step. To combine the color and motion vectors in graph cuts optimization [16], we modify the conventional cost function for segmentation as shown in (1).

$$E(A) = \lambda \cdot R_c(A) + \mu \cdot R_m(A) + B(A) \qquad (1)$$

The $R_c(A)$ is a regional term which estimates the property of the region A in terms of color, $R_m(A)$ is also regional term related to motion vectors, and $B(A)$ is a term which measures boundary property of the region A. The coefficients λ and μ specify the relative importance of the regional terms $R_c(A)$ and $R_m(A)$ against the boundary term. The cost function (1) is optimized using the graph cuts optimization technique similarly to [16].

After that, the union of segmented region in the previous step and initial tri-map is used to create the new tri-map. The eroded mask, original mask and dilated mask are used for the FG, PRFG and PRBG region in the new tri-map. With this manner, we iteratively apply graph cuts segmentation, and update each region. This iterative approach is very popular in the segmentation community. However, we expand or shrink the hard constrained region (FG or BG) unlike the conventional method in which hard constrained regions are fixed [15]. By doing so, foreground is updated so as to be closer to the upper-body region as the segmentation is repeated. Furthermore, we incorporate the motion vectors into segmentation by using RGB color and motion vectors alternatively in the iteration loop. In the segmenting step using motion vectors, the magnitude of motion vectors (2 dimension of x and y) is used to compute the regional term. In this step, foreground region is robustly extended or shrunk regardless of various colors or textures. On the other hand, the graph cuts segmentation using colors enables exact segmentation near edge. After the iteration, the final segmentation is performed with RGB information and the tri-map estimated only with the previous segmentation result, since the segmented result is reliable at the end of iteration loop. Finally, lens blur effect is applied to image using the integral image as in [19].

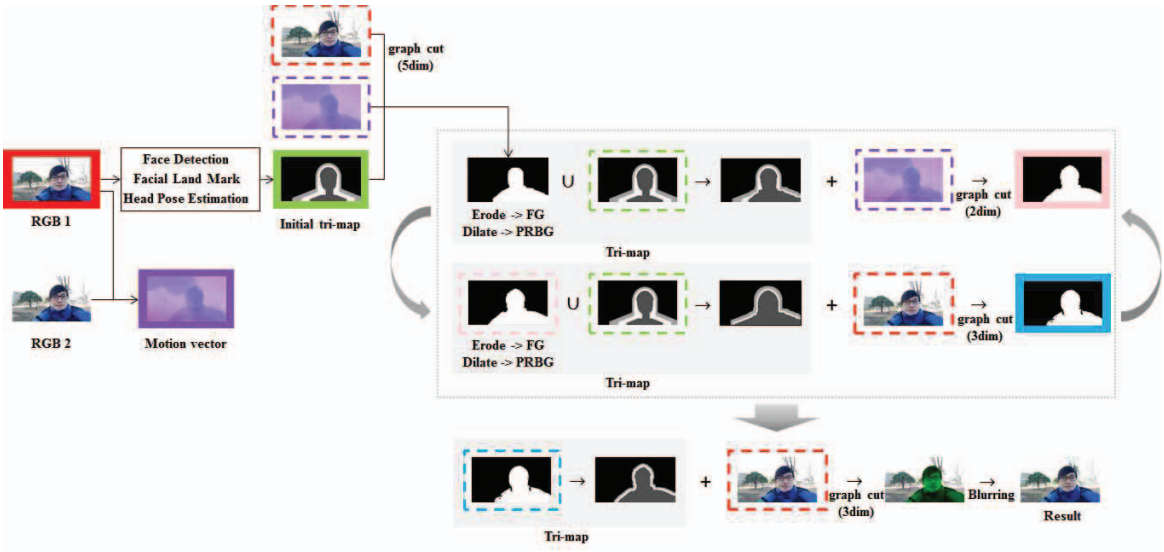

Fig. 2. The overall flow of the proposed algorithm.

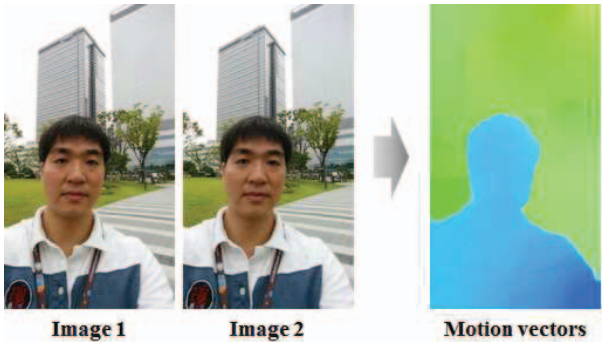

Fig. 3. The motion vectors between two consecutive images.

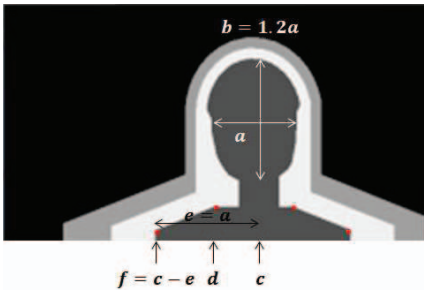

Fig. 4. The tri-map points (red dots) (a : face width, b : face height, c : shoulder center, d : shoulder start, e : half shoulder width, f : shoulder end).

3. EXPERIMENTAL RESULTS

3.1. Experiments on the selfie dataset

To demonstrate the performance of the proposed algorithm, we collected 286 pairs of selfie dataset containing a variety of human poses and cluttered backgrounds (Fig. 7). The time interval between two images is about 90 ms. We implemented our algorithm on mobile phone (Samsung Galaxy Note 4) and all the experiments are performed on the mobile. Our implementation takes about 10 sec to process an image (1920×1080).

Fig. 6. shows the tri-maps and consequential segmentation results. We can see that the foreground region is updated so as to be closer to the upper-body region as segmentation is repeated. In Fig. 6-(a), the head part of tri-map is adjusted to the long hair style, otherwise in Fig. 6-(b) the tri-map remains the same to fit short hair style. By segmenting iteratively with proper PRFG and PRBG region, we can deal with various hair style without additional hair model. Furthermore, the tri-map is properly adjusted to the thick winter clothes in Fig. 6-(b).

Fig. 7 shows the segmentation results and defocused images using proposed algorithm. In table 1, we compare the proposed method with the conventional method [15] using same initial tri-map in terms of the accuracy (precision / recall). In the result, the proposed method shows very competitive results due to the motion vectors.

3.2. Comparison with the previous methods

For comparison, we tested out-focus mode built in Samsung Galaxy Note 4 [5], out-focus mode in Google Camera [6], and the proposed method under similar scenes. Fig. 8 shows the comparison results. The Samsung out-focus mode built

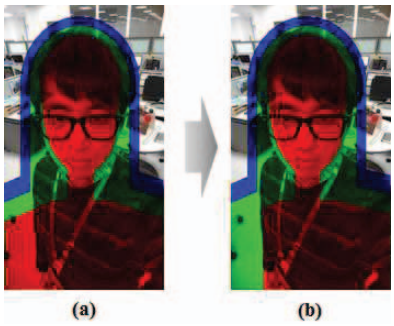

Fig. 5. The refined tri-map using GMM, (a) initial tri-map, (b) refined tri-map.

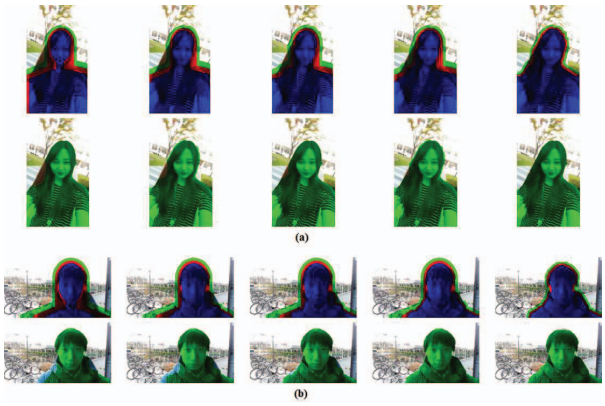

Fig. 6. The progress of tri-map and consequential segmentation result in the iteration loop (from left to right).

in Galaxy Note 4 has a weakness at texture-less wall regions since it estimates depth of the scene based on image sharpness. Moreover, it requires the auto-focus function to estimate the depth of the scene. Therefore, it doesnt work at the low-cost frontal camera for selfie. In the Google Camera result, a part of hair is defocused since the SfM algorithm is also vulnerable to texture-less regions. Moreover, the Google Camera requires users to move camera intentionally to obtain several images with different views for estimating depth map. This degrades the usability of the Google Camera. The proposed algorithm shows the best result, and it is robust to texture-less region in that it employs color and motion vectors

Table 1. Accuracy of the proposed algorithm in terms of precision / recall.

		Precision	Recall	f-measure
w/o motion vector	Mean	97.54%	97.04%	97.29
	Std.	4.64	2.97	N/A
w/ motion vector	Mean	97.34%	98.03%	97.68
	Std.	4.43	1.99	N/A

Fig. 7. The segmentation results (a) detected human region (painted in green) and (b) defocused images using proposed algorithm.

Fig. 8. Comparison result of three methods.

as prior cues for the segmentation. Moreover, unlike the Samsung built-in out-focus mode and the Google Camera which regard the central or nearest part as the target object, proposed method automatically detects human upper body and defocuses the background except detected upper body region.

4. CONCLUSION

We have presented a novel technique for defocusing selfie taken by mobile frontal camera. We automatically initialize the tri-map using prior knowledge in the selfie. Moreover, we utilize the burst shot of mobile frontal camera to obtain depth information without users additional effort. To enhance the performance, we iteratively segment upper-body region using both of RGB color and motion vectors. The proposed method works automatically without special hardware or users inconvenience and shows competitive results on challenging dataset. Moreover, the proposed algorithm outperforms the previous method using high-quality auto-focus camera.

5. REFERENCES

[1] B. Adewunmi, "The rise and rise of the 'selfie'," Apr. 2013, [Online; posted 2-April-2013].

[2] J. Hempel, "Contagionhow the "selfie" became a social epidemic," Aug. 2014, [Online; posted 22-August-2014].

[3] Google, "Project tango," Accessed: 2015-09-22.

[4] J. T. Barron, A. Adams, Y. Shih, and C. Hernandez, "Fast bilateral-space stereo for synthetic defocus," in *Proc. IEEE CS Conf. Comput. Vision and Pattern Recognition*, Jun. 2015.

[5] J.H. Na, K.H. Lee, W.Y. Lee, M.C. Kim, Y.K. Yoon, and H.J. Kang, "Method for synthesizing images and electronic device thereof," Mar. 4 2015, EP Patent App. EP20,140,177,057.

[6] C.H. Esteban, S.M. Seitz, S. Agarwal, and S. Fuhrmann, "Capturing and refocusing imagery," Sept. 25 2014, WO Patent App. PCT/US2014/021,862.

[7] F. Yu and D. Gallup, "3d reconstruction from accidental motion," in *Proc. IEEE CS Conf. Comput. Vision and Pattern Recognition*, June 2014, pp. 3986–3993.

[8] N. Joshi and C. L. Zitnick, "Micro-baseline stereo," Tech. Rep. MSR-TR-2014-73, May 2014.

[9] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Proc. the 12th European Conf. Comput. Vision*, Berlin, Heidelberg, 2012, ECCV'12, pp. 430–443, Springer-Verlag.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE CS Conf. Comput. Vision and Pattern Recognition*, June 2014, pp. 580–587.

[11] F. Wang, Q. Huang, M. Ovsjanikov, and L.J. Guibas, "Unsupervised multi-class joint image segmentation," in *Proc. IEEE CS Conf. Comput. Vision and Pattern Recognition*, June 2014, pp. 3142–3149.

[12] C. Scheffler and J.-M. Odobez, "Joint adaptive colour modelling and skin, hair and clothes segmentation using coherent probabilistic index maps," in *Proc. the British Machine Vision Conference*. 2011, pp. 53.1–53.11, BMVA Press.

[13] P. Julian, C. Dehais, F. Lauze, V. Charvillat, A. Bartoli, and A. Choukroun, "Automatic hair detection in the wild," in *Proc. Int'l Conf. Pattern Recognition*, Aug 2010, pp. 4617–4620.

[14] P. Kohli, J. Rihan, M. Bray, and P. Torr, "Simultaneous Segmentation and Pose Estimation of Humans Using-Dynamic Graph Cuts," *Int'l J. Comput. Vision*, vol. 79, no. 3, pp. 285–298, 2008.

[15] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut -interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, Aug. 2004.

[16] Y.Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary amp; region segmentation of objects in n-d images," in *Proc. 8th IEEE Int'l Conf. Comput. Vision*, 2001, vol. 1, pp. 105–112 vol.1.

[17] Andrs Bruhn, Joachim Weickert, and Christoph Schnrr, "Lucas/kanade meets horn/schunck: Combining local and global optic flow methods," *Int'l J. Comput. Vision*, vol. 61, no. 3, pp. 211–231, 2005.

[18] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE CS Conf. Comput. Vision and Pattern Recognition*, June 2013, pp. 532–539.

[19] O. Niemitalo, "Circularly symmetric convolution and lens blur," Accessed: 2015-09-22.