Hand segmentation under different viewpoints by combination of Mask R-CNN with tracking

Dinh-Ha Nguyen*, Trung-Hieu Le†, Thanh-Hai Tran* Hai Vu*, Thi-Lan Le* and Huong-Giang Doan‡
*Computer Vision Department
International Research Institute MICA, Hanoi University of Science and Technology
†Faculty of Information Technology, Dainam University
‡Faculty of Control and Automation, Electrical Power University

***Abstract*—This paper presents a new method for hand segmentation from images and video. The method based mainly on an advanced technique for instance segmentation (Mask R-CNN) which has been shown very efficient in segmentation task on COCO dataset. However, Mask R-CNN has some limitations. It works on still images, so cannot explore temporal information of the object of interest such as dynamic hand gestures. Second Mask R-CNN usually fails to detect object suffered from motion blur at low resolution as hand. Our proposed method improves Mask R-CNN by integrating a Mean Shift tracker that tracks hands in consecutive frames and removes false alarms. We have also trained another model of Mask R-CNN on cropped regions extended from hand centers to obtain a better accuracy of segmentation. We have evaluated both methods on a self-constructed multi-view dataset of hand gestures and show how robust these methods are to view point changes. Experimental results showed that our method achieved better performance than the original Mask R-CNN under different viewpoints.**

***Index Terms*—hand segmentation, neural network, deep learning, tracking**

I. Introduction

Hand segmentation and gesture recognition have originated several decades ago. However it remains an active research topic in recent years due to its wide range of applications in human robot interaction or entertainment. Despite many impressive results have been obtained, there exists some issues that should be carefully taken into account if ones would like deploy a hand segmentation method to a practical application. The main challenges come from low resolution of hand, huge variation of posture as deformable hand with high dofs, camera view point changes, complex background, occlusion, etc.

In the literature, a number of methods for hand segmentation has been proposed. The simplest ones based on skin detection. However, as analyzed in [1], skin based hand detection is usually sensitive to lighting condition. Hand crafted features based methods aiming to represent a hand by some features (e.g. shape context, Haarlike, HoG) have difficulties to deal with complex lighting or background [2]. Recently, deep learning technique such as Mask R-CNN proved its excellent performance on segmentation problem [3]. However, Mask R-CNN was not tested on hand object which often suffers from challenges as mentioned previously.

Thank to the impressive achievement by Mask R-CNN for object segmentation, in this work, we will adopt Mask R-CNN for hand segmentation. Firstly, we will investigate the performance of Mask R-CNN at different viewpoints of camera. This finds that Mask R-CNN has some limitations. It works on still images, so cannot explore temporal information of the object of interest such as dynamic hand gestures. Beside, Mask R-CNN usually fails to detect objects suffered from motion blur at low resolution as hand

We then improve Mask R-CNN by proposing a new framework which integrates a Mean Shift tracker that tracks hands in consecutive frames and removes false alarms. We have also trained another model of Mask R-CNN on cropped regions extended from hand centers to obtain a better accuracy of segmentation. We have evaluated both methods on a self-constructed multi-view dataset of hand gestures and show how robust these methods are to view point changes. Experimental results showed that our method achieved better performance than the original Mask R-CNN. Our contributions are two-fold: i) investigate the performance of Mask R-CNN on a new type of challenging object which is hand; ii) propose a new method for hand segmentation that combines semantic segmentation by Mask R-CNN with tracking, the performance of the proposed method is improved compared to the original Mask R-CNN.

The paper is organized as follows. In section II, we present related works on hand segmentation. In section III, we describe Mask R-CNN method and our proposed method that combines Mask R-CNN with tracking using Mean Shift. We will evaluate both methods on our self-constructed multi-view dataset of hand gestures. Experimental results will be shown in Section IV. Section V concludes and gives ideas for future works.

II. Related works

Hand detection is a process that aims at determining the hand region in frames/images. This is an inevitable and important step in hand postures/gestures recognition since the quality of this step will affect the performance of the whole system. However, accurate detection of hands in still images or video remains a challenging problem, due to the variability

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-17-1-4056

978-1-5386-7678-3/18/$31.00 ©2018 IEEE

of hand appearances and environments. In this section, we review work closely related to our method: hand detection from RGB images. An intensive survey of hand detection and gesture recognition is available at [4].

Prior works on hand detection have attempted to exploit hand characteristic in order to distinguish hand from others objects in the scene. In [5], the authors used a saliency map to detect hands in the images. Mittal et al. [6] proposed a hand detector using a two-stage hypothesize and classify framework. In the first stage, hand hypotheses were proposed from three independent methods including a sliding window hand-shape detector, a context-based detector, and a skin-based detector. In the second stage, the proposals are scored by all three methods and a discriminatingly trained model is used to verify them. This method obtains improvements in precision and recall (average precision: 48.20%; average recall: 85.30%on PASCAL VOC 2010 dataset). However, the computation time is too expensive. The time taken for the whole detection process is about 2 minutes for an image of size 360 × 640 pixels on a standard quad-core 2.50 GHz machine. Nguyen et al. [7] tried to accelerate the detection speed by introducing a new feature named Internal Haar-like in Adaboost and cascade architecture.

Recently, the impressive results of deep learning for object detection have led to emergence of works that learn discriminative features using CNN (Convolutional Neural Network) for hand detection. In [8], [9], the authors proposed a multiple scale deep feature extraction approach for robust hand detection. Deep feature is extracted through a Multiple Scale Faster Region-based Convolutional Neural Network (MS-FRCNN). The authors have proved that the proposed method is able to handle the the challenging factors of hand detection. Taking into account the fact that FRCNN produces many false positives while working with small objects such as hand, Roy et al. [10] incorporated skin information in two-stage hand detection. In the first stage, RCNN (Region-based convolutional neural network) and FRCNN are used to estimate the spatial location of hands in the input image. Then, a patch-based convolutional neural network skin detector is applied in the second stage to reduce the false positives of the first one. In [11], the authors designed a convolutional neural network(CNN) which handles object rotation explicitly to jointly solve the object detection and rotation estimation tasks. The proposed network allows to detect hands and calibrate in-plane rotation under supervision at the same time. To guarantee the recall, they introduced a context aware proposal generation algorithm which significantly outperforms the selective search.

Recent works on hand detection based on CNN have improved significantly the accuracy of hand detection especially when environment is cluttered. However, in these works, temporal information is not exploited. In this work, we investigate the impact of viewpoint change in CNN-based hand detection method and propose to combine CNN-based detection with tracking method in order to secure a good detection result.

III. Proposed method for hand detection

The general framework of our proposed method is illustrated as in Fig. 1. It consists of two main blocks:

- Hand segmentation: In this block, hand is segmented from still images using a fine tuned neural network (Mask R-CNN). However, as hand is of low resolution, sometime confused with other objects in the scene and/or overlapped on face or skin like region. In those cases, the segmentation result could be not correct.
- Hand tracking: Mask R-CNN is applied only on still images which does not capture the temporal constraint of hand movement in a video. In this phase, we utilize a hand tracking technique to reduce false alarms and recover some missed detection from the first step. The result of tracking will be used to update the segmentation result. Details about these steps will be presented in the following sections.

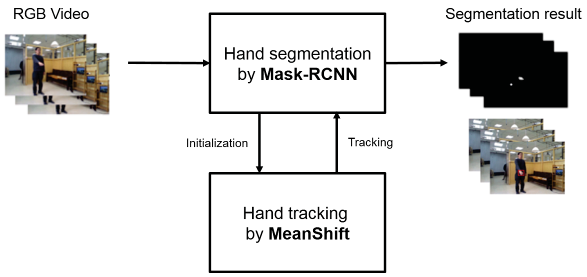


Fig. 1. Framework of proposed method for hand segmentation

A. Hand detection using Mask R-CNN

1) Brief review of Mask R-CNN: Mask R-CNN was first introduced by K. He et al. in 2017 [3]. The network efficiently detects object and simultaneously generating segmentation mask for each instance. This network extends Faster R-CNN by adding a brand for bounding box recognition. Faster R-CNN consists of two stages [12]. The first stage is Region Proposal Network (RPN) that proposes candidate object bounding box. The second stage is Fast R-CNN that extracts features using RoIPool from each candidate box and performs classification and bounding box regression. In Mask R-CNN architecture, in the second stage, instead of generating the class label and box offset Mask R-CNN outputs an additional binary mask for each bounding box. To this end, the authors have defined a multi-task loss on each sampled RoI as $L = L_{cls} + L_{box} + L_{mask}$ where L_{cls}, L_{box} are classification loss and bounding-box loss respectively. L_{mask} allows the network to generate masks for every class. So given an input image, the network will generate a class label for each region. In case of hand segmentation, we consider only two class: hand and non-hand. Fig.2 illustrates the principle of Mask R-CNN. For more technical details, the readers are recommended to refer to the original paper [3].

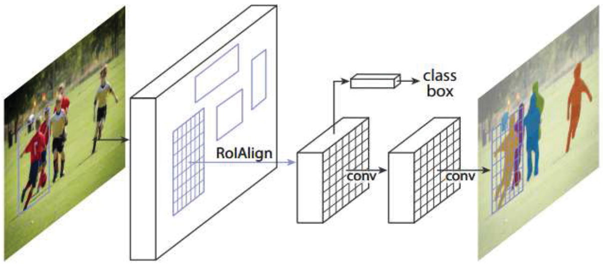


Fig. 2. Principle of Mask R-CNN [3].

2) Fine tuning Mask R-CNN for hand segmentation: The original Mask R-CNN [3] was trained on COCO dataset[1]. It consists 80 object classes but none of these contains human hands. To be utilized in our framework, we have to fine tune Mask R-CNN with our dataset. We use the opensource of Mask R-CNN[2]. To fine tune Mask-CNN, we have to prepare pairs of images. Each pair includes an original image containing hand and a binary image that contains only the hand region (without background) (see an example in Fig. 3). This phase was carried out manually with the support of Interative Segmentation Tool[3].

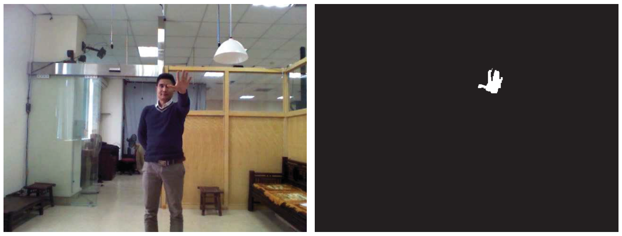


Fig. 3. A pair of original image (left) and binary image (right) used to fine tune Mask R-CNN. These images are presented at full resolution. These images are used to fine tune Mask R-CNN_640x480.

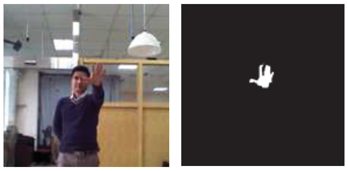


Fig. 4. A pair of original image (left) and binary image (right) used to fine tune Mask R-CNN. These images are cropped from original images to 256x256 from the center of groundtruth hand region. These images are used to fine tune Mask R-CNN_256x256.

Our annotated dataset contains about 2400 pairs of images containing hands at different viewpoints (see more detail in the experiment section). The resolution of our original images is 640x480. We fine tune two types of model of Mask R-CNN with two different sizes. The first model was fine tuned with original resolution while the second we cropped 256x256 sub-images centered at ground truth hand regions (see Fig.4). Henceforth, these are refered as Mask R-CNN_640x480 and Mask R-CNN_256x256.

The reason for which we have fine tuned an additional model of Mask R-CNN on centered 256x256 images sizes is that if in case we know wheather there is a hand candidate (by a very simple detector such as skin based one), we can extend the region of hand candidate to 256x256 and apply Mask R-CNN to obtain very high accuracy of segmentation. In our experiment, the accuracy of fine tuning and applying Mask R-CNN_256x256 on regions extended from hand center is 99.3 % at IoU = 0.5. Therefore, we will use this model to refine our detection result after tracking. We fine tune on GPU GTX 1080Ti (Vram 12Gb) with minibatch size is 16 for 40k iterations to fine tune the first layers of Mask R-CNN with a learning rate of 0.001, the next 80k iterations to fine tune the whole network with the same learning rate, which is decreased by 10 at the 160k iterations. We use a weight decay of 0.0001 and momentum of 0.9. The train loss and validation loss after 160k iterations are (0.09, 0.11) and (0.01, 0.08) for (256x256) and (640x480) respectively.

Mask R-CNN works with still images. It does not explore temporal characteristic of the object movement. Otherwise, when using only still images, Mask R-CNN has some limitations. First, sometime hand moves very fast and image of hand suffers from motion blur. In addition, hand could be occluded partially or totally (out of camera view) that cause Mask R-CNN fail to detect hands. Second limitation of Mask R-CNN is that Mask R-CNN usually fails to detect hands overlapped on face region. To overcome these drawbacks, we will utilize a tracker to constraint movement of hand then recover missed detection as well avoid false alarms.

B. Hand tracking using Mean Shift

Mean Shift was originally introduced in 1975 by Fukunaga and Hostetler [13]. It is now a well known nonparametric and iterative procedure that shifts each data to local maximum of density function. In 1995, it was revisited by Cheng and his colleagues in [14] that develops a more general formulation and demonstrates its potential uses in clustering and global optimization. After that, Mean Shift has been widely used in object tracking [15], [16], [17], [18].

The principle of Mean Shift for object tracking as follows. Let us consider a set of pixels in the ROI at current frame. This set of points is represented by a color distribution. The location of the target in the new frame is predicted based on the past trajectory. A search will be performed in its neighborhood for image regions whose distribution (color histogram) is similar to that of the model. Object tracking for an image frame is done by a combination of histogram extraction, weight computation and derivation of new location.

In our case of hand tracking, suppose that hand is the unique object in the scene to be tracked and we consider only the hand which performs the dynamic hand gestures. Each hand region is characterize by a HSV color histogram. The size of hand region is not changed during the tracking. Given the detection

[1] http://cocodataset.org/
[2] https://github.com/matterport/Mask_RCNN
[3] http://segmentit.sourceforge.net/

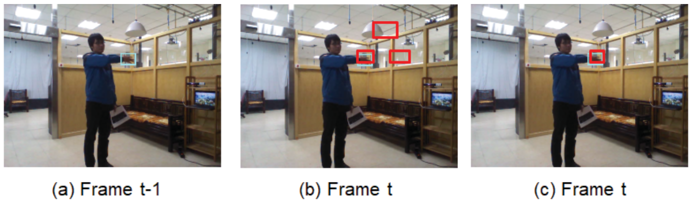
results provided by Mask R-CNN at the current frame which could contain multiple windows $ROI_{1t}, ROI_{2t}, ..., ROI_{Nt}$. Suppose M_{t-1} is the result of track from the previous frame. We select the best candidate that satisfies the minimum value:

$$ROI_{it} = \underset{i \in [1,N]}{\operatorname{argmin}}(\mathrm{f_{it}}) \tag{1}$$

where

$$f_{it} = \alpha_1 * d(C_{it}, C_{t-1}) + \alpha_2 * \delta H + \alpha_3 * \delta S \tag{2}$$

is a weighted function to measure the difference between two ROIs. $d(C_{it}, C_{t-1})$ is the euclidean distance between two centroids of the ROI_{it} and the predicted track M_{t-1}. δH and δS are differences in term of average Hue and Saturation of ROI_{it} and M_{t-1} respectively. $\alpha_1, \alpha_2, \alpha_3$ are experimentally selected. To emphasize the constraint on distance, in our experiment, we set $\alpha_1 = 10, \alpha_2 = 1, \alpha_3 = 1$. It noticed that before computing f_{it} all values of $d, \delta H, \delta S$ are normalized to [0, 1] to take the scale of the parameters into account.


(a) Frame t-1

(b) Frame t

(c) Frame t

Fig. 5. Example of Mean Shift tracking: a) Track result at time $t-1$ displayed in cyan rectangle; b) Multiple candidates produced by Mask R-CNN shown in red rectangles at time t; c) The best candidate is selected.

Fig. 5 illustrates an example of selecting the most suitable candidate. In Fig. 5a, the cyan rectangle is track result at time $t-1$. At time t, Mask R-CNN produces three candidates ROI_1, ROI_2, ROI_3 displayed in red rectangles in Fig.5b. We compute three scores: f_{1t}, f_{2t}, f_{3t}, and the one having smallest value will be selected. The best ROI is kept as the final detection result (red rectangle) as shown in Fig.5c.

C. Integration of detection and tracking

This section presents how tracking and segmentation are integrated in an unified framework for a more robust hand segmentation (see Fig.6).

For every coming frame at time t, we apply Mask R-CNN_640x480 on this frame. Mask R-CNN_640x480 could output a number of candidates N. There are two cases to be considered:

- If it is the first frame ($t = 0$), we determine the best candidate for initialization of Mean Shift tracker. If there is no candidate at that time, the frame is given up and we continue with the next frame. In case there are more than one candidate, we choose one candidate having the highest score returned from Mask R-CNN which is bigger than a pre-determined threshold θ. If only one candidate is detected, but its score returned by Mask R-CNN is not bigger than the threshold θ then we conclude that this candidate is unreliable and we move to the next frame as the case no candidate is detected.
- Otherwise, among the detected candidates by Mask R-CNN_640x480, we select the best candidate based on the function value f (eq. (1)) calculated previously and continue the tracking of this candidate for the next frame. If no candidate was detected, we extend the region predicted by the tracker to 256x256 and apply Mask R-CNN_256x256 on this region to try one more time for a higher accuracy of detection. If some candidates are detected, the best will be selected using the criterion in eq. (1). If no candidate is detected, we keep the predicted region as a candidate for the next frame.

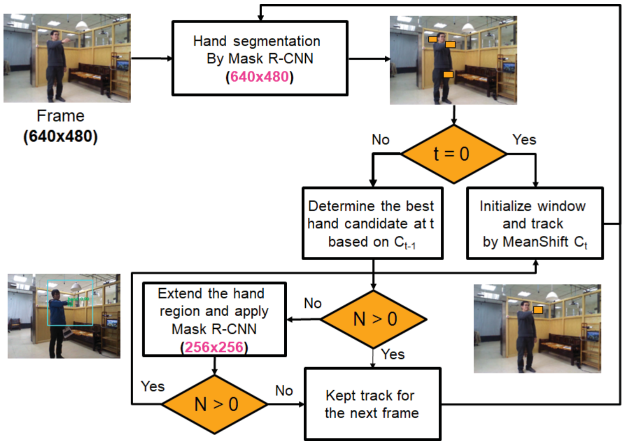

Frame
(640x480)
Hand segmentation
By Mask R-CNN
(640x480)
No
t = 0
Yes
Determine the best hand candidate at t based on C_{t-1}
Initialize window and track by MeanShift C_t
Extend the hand region and apply Mask R-CNN (256x256)
No
N > 0
Yes
Yes
N > 0
No
Kept track for the next frame

Fig. 6. Workflow of hand segmentation and tracking.

IV. Experiments

A. Multi-view dataset of hand gestures

This section presents the evaluation result of two methods: the original Mask R-CNN and our proposed method. To prepare data for finetuning and testing, we designed a new dynamic hand gestures dataset. This dataset consists of five dynamic hand gestures for controlling home appliances in a human machine interaction. These gestures correspond to controlling commands: ON/OFF, UP, DOWN, LEFT, RIGHT with movement of the hand in the corresponding direction. For each gesture, hand starts from one position with close posture, it opens gradually at half cycle of movement then closes gradually to end at the same position and posture as describe in [19]. Fig.7 illustrates the movement of hand and changes of postures during gesture implementation.

Five camera $\{C_1, C_2, C_3, C_4, C_5\}$ are setup at five various positions in a simulation room of 4mx4m with a complex background (Fig. 8). This work aims to capture hand gestures under multiple viewpoints at the same time. Subjects are invited to stand at a nearly fixed position in front of five cameras at an approximate distance of 2 meters. Five participants (3 males and 2 females) are voluntary to perform gestures. Each subject implements one gesture three times. Totally, the dataset contains 375 (5 views x 5 gestures x 5

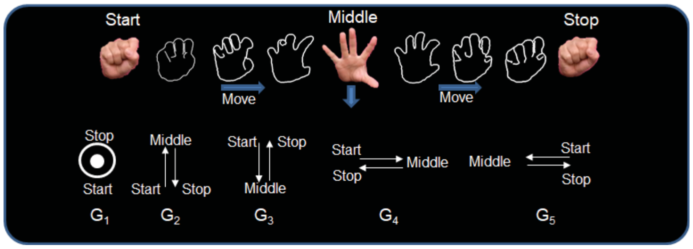


Fig. 7. Set of five dynamic hand gestures.

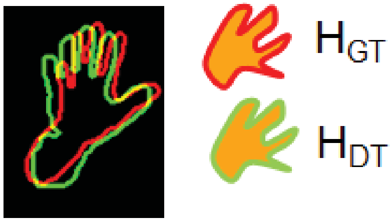


Fig. 9. Computation of IoU for segmentation evaluation.

subjects x 3 times) dynamic hand gestures. The frame rate is 20fps and frame resolution is set to 640x480. Each gesture's length varies from 50 to 120 frames (depending on the speed of gesture implementation). This leads to a huge number of frames to be processed.

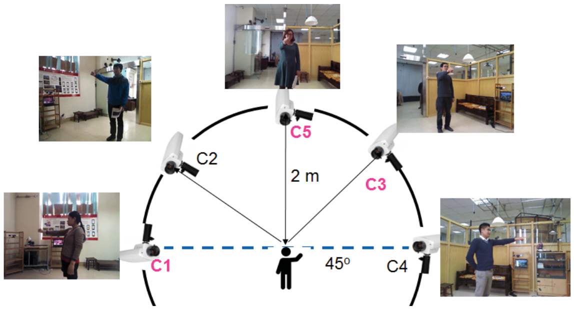


Fig. 8. Experimental setup.

To fine tune Mask R-CNN and evaluate our algorithm, ground truth of hand segmentation must be prepared. As the entire dataset contains a huge number of frames which requires huge annotation time, in this work, only three camera views C_1, C_3, C_5 will be processed. In addition, for each view, only 2400 frames taken from 5 subjects will be annotated. We take 2000 frames from four subjects for fine tuning and 400 frames of the remaining subject for evaluation.

B. Evaluation protocols

The proposed method aims to segment hand regions in images, we then evaluate the accuracy at a certain IoU. In general a IoU is computed as:

$$IoU = \frac{H_{DT} \cap H_{GT}}{H_{DT} \cup H_{GT}} \tag{3}$$

where H_{DT} is the hand region detected automatically by the proposed method and H_{GT} is the ground truth hand region predetermined manually. Fig.9 illustrates these regions. A detected region is considered positive if it has IoU with a ground-truth box of at least 0.5 and negative otherwise.

In this work, we perform single view evaluation. We use a part of data from one camera view for fine-tuning the network and the remaining part from the same camera view for testing. In addition, we apply one-leave-out subjects to prepare training and testing data. Specifically, if we have M subjects participating in the experiment, data from $M-1$ subjects will be used for training and the remaining subject is used for testing. This is subject independent test. In our dataset, $M = 5$.

C. Experimental results

In this section, we present the experimental results obtained with the original Mask R-CNN method and our proposed method.

1) Evaluation of Mask R-CNN_640x480: Fig.12 shows the accuracy of hand segmentation using original Mask R-CNN fine tuned at full resolution (640x480) with IoU = 0.5 (in blue). It is noticed that the best accuracy (98%) is obtained with C_3 having frontal view. This is an impressive result even gestures are performed with huge variation of postures. This means than Mask R-CNN is very good for segmenting different structures at small resolution.

When the view point deviates in the range of 45^o, the accuracy could be achieved to 92% (C_3 view). When the view point changes alot, the accuracy of hand segmentation is reduced to 70% (C_1 view). The reasons come from the motion blur of the hand, hand is out of camera view or totally occluded by body part. The two last are the main cause of missed or wrong detection of the hand. Fig.10 shows some of failed detections in the camera view C_1. In Fig.10.a, hand is detected nearly at the hand position but it is biased due to motion blur effect. In this case, the detected region has IoU < 0.5 so it is concluded as a negative detection. In Fig.10.b, the hand of interest is out of camera view, a hand is wrongly detected at arm region. In Fig.10.c, the hand of interest is totally occluded and the left hand is wrongly detected.

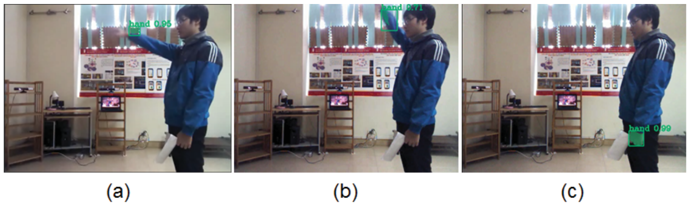


Fig. 10. Issues of the side view C_1. a) Missed detection due to the IoU $<$ 0.5; b) Missed detection due to out of camera view; c) Missed detection of right hand (interest hand) due to its occlusion.

2) Comparison of the proposed method with Original Mask R-CNN_640x480 : In this subsection, we compare the results obtained by our proposed method with the original Mask R-CNN on single view. Fig.12 shows the improvement of our proposed method compared to the original Mask R-CNN on

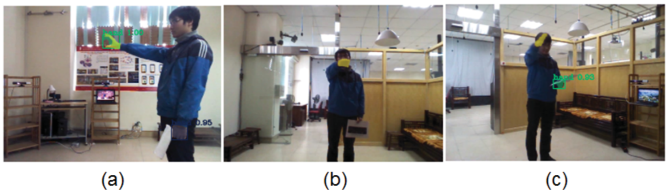


Fig. 11. Illustration of Mask R-CNN limitations. In this figure, we ovelap yellow region of ground-truth hand and green region of detected hand on the original image. a) Missed detection due to the IoU < 0.5; b) Missed detection due to overlapping of hand on face region; c) Missed detection of right hand (interest hand)

two views C_1 and C_3. The accuracy has increased by 9% in the camera view C_1 and 5% in the camera view C_3. That means that the most improvement is been achieved with the most difficult view point. However, the accuracy of 79% should be still improved. In the camera view C_5, the accuracy does not change. However, if we look at Fig.13, we find that the hand segmentation obtained a better precision (the average IoU increased by 4% in camera C_5).

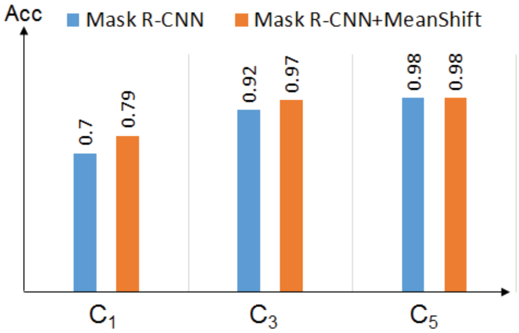


Fig. 12. Comparison in term of accuracy of the proposed method with Original Mask R-CNN_640x480 at IoU = 0.5 on separated view.

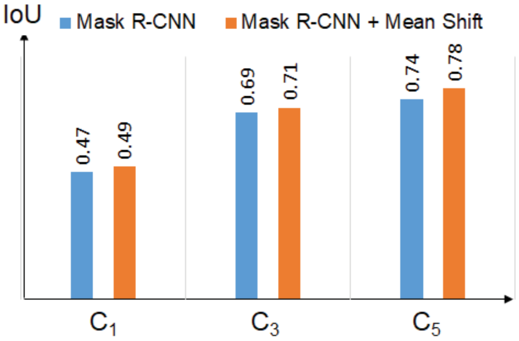


Fig. 13. Comparison in term of average IoU of the proposed method with Original Mask R-CNN_640x480 on separated view.

One of the reason that the accuracy does not improve much as expected is that to evaluate the hand segmentation, we need to prepare manually the ground-truth which are images containing only hand without background. The annotation is very time consuming and boring. To reduce the total images to be annotated, we have taken one frame for every three frames and for all video sequences. Therefore, for these frames, the results of tracking could be not really significant due to the sub-sampling. However, if we look at a full frame rate sequence (Fig. 14), we can see that all frames in the red block are missed detected by Mask R-CNN. However, using our method, all hands are correctly detected.

Fig. 15 illustrates clearly a case where Mask R-CNN_640x480 failed to detect the hand, but with our tracking technique, Mask R-CNN_256x256 successfully find out the hand on the extended region. Fig. 16 shows the case where Mask R-CNN gives many false alarms on face and paper regions with very high confidence score. Using our proposed method, these false alarms have been removed.

V. Discussions and Conclusions

In this paper, we have studied an advanced method for instance segmentation which is Mask R-CNN and evaluated it on a new type of object (hand region) in still images. We found that Mask R-CNN could perform well in frontal viewpoint or a deviated view in the range of 45^o on both single view or cross-view. In case of side view, Mask R-CNN can not achieved such performance. Our proposed method based mainly Mask R-CNN but explored temporal information of hand movement and keep the track of hand for refine detection of hand on the extended regions. This allowed to improve by from 5% to 9% compared to the original Mask R-CNN in term of accuracy at IoU = 0.5. The quality of segmentation is also improved on all investigated views. We concluded that the proposed method works very well when the human stands frontally or at a deviation angle of about 45^o to the camera view. In the future, we will evaluate this method on all camera views and apply the automatic detection result for hand gesture recognition in an application of controlling home appliance using hand gestures.

References

[1] Thanh-Hai Tran and Thanh-Mai Nguyen. Invariant lighting hand posture classification. In *Proc. of 2010 IEEE International Conference on Progress in Informatics and Computing*, pages 827–831, 2010.

[2] Van-Toi Nguyen, Thi-Lan Le, Thanh-Hai Tran, Rémy Mullot, and Vincent Courboulay. A method for hand detection based on internal haar-like features and cascaded adaboost classifier. In *The International Conference on Communications and Electronics (ICCE)*, pages 608–613, 2012.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017.

[4] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, May 2007.

[5] Pramod Kumar Pisharady, Prahlad Vadakkepat, and Ai Poh Loh. Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 101(3):403–419, Feb 2013.

[6] Andrew Zisserman Arpit Mittal and Philip Torr. Hand detection using multiple proposals. In *Proceedings of The British Machine Vision Conference (BMVC)*, pages 75.1–75.11, 2011.

[7] Van-Toi Nguyen, Thuy Thi Nguyen, Remy Mullot, Thi-Thanh-Hai Tran, and Hung Le. A method for hand detection using internal features and active boosting-based learning. In *Proceedings of the Fourth Symposium on Information and Communication Technology*, SoICT '13, pages 213–221, New York, NY, USA, 2013. ACM.

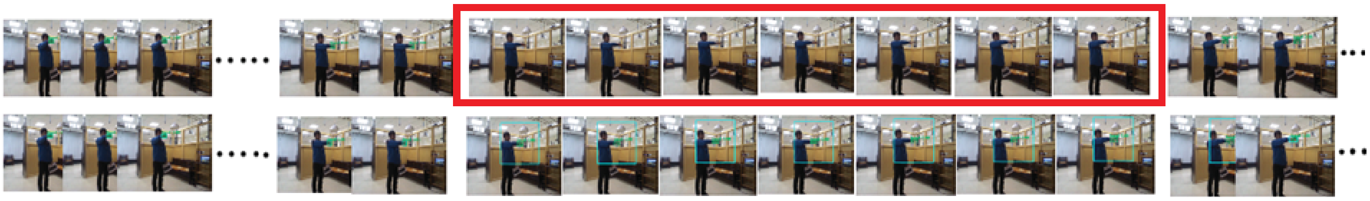


Fig. 14. The first panel shows the results detected by the original Mask R-CNN. In the red block, there are many missed detections in consecutive frames. The secon panel shows the results detected by our proposed method. Hands have been detected in all frames.

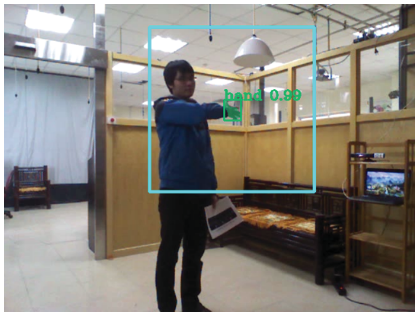
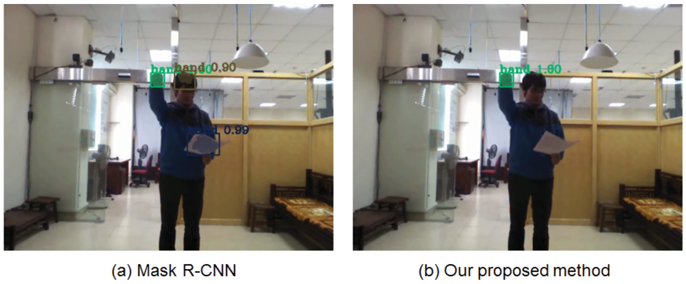


Fig. 15. An example of hand detection refinement using Mask R-CNN_256x256 on the extended region from the center of the region predicted by MeanShift.



(a) Mask R-CNN

(b) Our proposed method

Fig. 16. a) Mask R-CNN gives more false alarms. b) Our proposed method could removed the false alarms and keep only the right detection.

[8] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides. Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 46–53, June 2016.

[9] T. H. N. Le, Chenchen Zhu, Yutong Zheng, Khoa Luu, and M. Savvides. Robust hand detection in vehicles. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 573–578, Dec 2016.

[10] K. Roy, A. Mohanty, and R. R. Sahay. Deep learning based hand detection in cluttered environment using skin segmentation. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 640–649, Oct 2017.

[11] X. Deng, Y. Zhang, S. Yang, P. Tan, L. Chang, Y. Yuan, and H. Wang. Joint hand detection and rotation estimation using cnn. *IEEE Transactions on Image Processing*, 27(4):1888–1900, April 2018.

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[13] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

[14] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

[15] Dorin Comaniciu and Visvanathan Ramesh. Mean shift and optimal prediction for efficient object tracking. In *ICIP*, pages 70–73, 2000.

[16] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

[17] Robert T Collins. Mean-shift blob tracking through scale space. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–234. IEEE, 2003.

[18] Tomas Vojir, Jana Noskova, and Jiri Matas. Robust scale-adaptive mean-shift for tracking. *Pattern Recognition Letters*, 49:250–258, 2014.

[19] Huong-Giang Doan, Van-Toi Nguyen, Hai Vu, and Thanh-Hai Tran. A combination of user-guide scheme and kernel descriptor on rgb-d data for robust and realtime hand posture recognition. *Engineering Applications of Artificial Intelligence*, 49:103–113, 2016.