A New Model to Detect 2D Hand based on Multi-feature Skin Model

Abdullah Shawan Alotaibi[1]*
Computer Science Department
Shaqra University, Shaqra
Saudi Arabia

***Abstract*—Recognition of hand gesture is one of Human PCs most growing interfaces. In most vision-based signal recognition system, the initial phase is hand detection and separation. Because the hands are linked to a variety of day by day, local work experiences both extraordinary changes in the illumination and the innate unbroken appearance of the hand. In order to address these issues, we suggest another 2D hand position software that can be seen as a combination of multi-feature hand proposal generation and cascading neural system network characterization (CCNN). When considering various luminances we select color, Gabor, Hoard and Filter to separate the skin and produce a hand proposal. Therefore, we are selling a cascaded CNN that holds the deep setting information between the proposals. A mix of some datasets, including a few Oxford Hands Datasets, VIVA Hand Recognition, and Egohands Datasets, is tested as the positive example and image patch Net 2012, FDDEB dataset as a bad example; the proposed Multi-Feature Directed Cascaded CNN (MFS-CCNN) strategy. Aggressive results are achieved by the technique proposed. Our average sample dataset accuracy is considerably inferior to DPM. With an average of 43.55 and 51.78 percent accuracy, our CCNN and MFS-CCNN model perform DPM. Average accuracy of the CCNN model in a combined test set is 9.16% higher than the SSD model. Still, our model is faster than a DPM based on the statistical performance.**

Keywords—Hand detection; feature modeling; convolutional neural networks

I. INTRODUCTION

With the advances of astute processing lately, there has been a developing enthusiasm for growing increasingly instinctive and proficient methods for collaboration among human and PCs [1]. As hand signals structure a broad piece of characteristic human correspondence, vision-based motion acknowledgment is an engaging option in contrast to conventional presently utilized gadgets, for example, console and mouse [2]. As of late, countless examines [3, 4] about human hand exercises has increased exceptional consideration. 2D vison-based hand motion acknowledgment is conceivably a minimal effort data preparing instrument for human – PC interface.

Human hands are associated with a variety of tasks day after day with a lot of segmenting data. Hand identification is the biggest development in semantically awareness of hand operation. For instance [5-8] the color, region, Hoard [9], surface filter [10] or the district family CNN such as RCNN, Fast RCNN, Quicker R CNNN [11—13] and R-FCN [14] etc. are the key dependent features of usable recognition scans. The assignment of the recognition of human hands consistently experiences the multifaceted nature of the foundation which in addition causes rapid change in the condition of enlightenment [15].

Secondly, when (for example, gesture based communication or a command when associating with computers) or controlling objects constantly changes the form of human hands. Such changes in presentation impact the exploration and the understanding of hand exercises [6]. Thirdly, a specification for a perfect hand-location should be able to continue to work, similar to the steps for detection of hand motion. Extraordinary consideration is given to speed in business applications [16].

They recommend a MFS-CCNN approach to manual positioning and sorting in order to successfully supervise these problems. We could produce hand-area proposals to monitor a Cascaded CNN model by means of the viability of multi-feature registration. In the meantime, we use the cascade CNN system to keep information missing for classification beyond what others find feasible. We have built our CNN cascaded template with a few free datasets such as Oxford Hand Dataset[5], VIVA hands [18], and image patches Net 2012[19], FDDB[20].

The fundamental commitments of this paper can be condensed as pursues.

- To understand skin district cover and construct hand offering jumping boxes we use a multi feature skin model.
- We found that cascading CNN structure which aggregates attempts to improve the detection by setting the data for bouncing box regression.
- They demonstrated that mixing handcraft and CNN approaches can produce accurate results for hand recognition tasks.

This paper is organized as follows. In Section 2, we review state-to-art researches on hand detection and the average and limitations of feature based and region-based CNN work. Section 3 will address the methodology suggested for hand identification and classification. The findings of the test were summarized in Section 4. We will eventually reach a conclusion in Section 5 and prepare for the future.

*Corresponding Author

II. Related Work

In the individual PC community, improved fact and command driver behavior and so on, recognition and following of human hands are commonly used. In this article we concentrate on differentiating hands in multiple situations. Mittal et al. [5] exhibited a successful two-way approach to human hands. Three fundamental finders were used to introduce hand bouncing containers, which could be used to train a classifier for the final determination. Nevertheless, their skin-based features were profoundly respected in different conditions of enlightenment. The base pixels using the surface and limiting features were demonstrated by Fathi et al. [7]. We distinguish hands from various items using color histograms from the missing facial region pixels. Li and Kitani [6] have therefore prepared a pixel level indicator for objects with slowly realistic ego-centered characteristics, for example, lighting gestures and lighting switch. Their strategy combines super-pixels with color + surface features and invariance descriptors. While their methods have achieved a focused outcome, a few disappointments exist while hands are on bland premises. Also, as shown in the Fig. 1, it separates hands and other divisions of body, such as the lower arms and face.

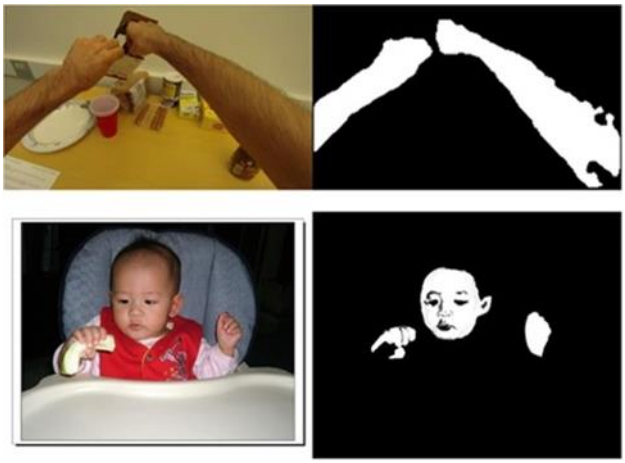

Fig. 1. The Hand Detector Pixel Stage not only Detects Hands but also other Field Skin Masks.

In [21] the two-arrange hand indicators of the color (RGB, HSV, LAB) and edge (HOG, GIST) features of Betancourt [24] were also proposed, using a classification (SVM, random forest, choice trees). As recently as this, Bambach et al.[23] introduced another egocentric dataset called EgoHands. This collection contains documents where a few people wear camera glasses each other while playing a pre-game. Their approach pipeline is like R-CNN, but they send the district a probabilistic plan and split it into a pixel point. Large estimated information requires the viability of profound learning approaches [31] to work consistently in difficulties. MS-FRCNN [24] was shown in THN le et al. to disable several features so that sincere hand seeking in vehicles can be done. They extend the Faster-RCNN system in the region Proposal Network (RPN) and organize discovery with a diverse scale of deep element extraction. The technologies referred to above are nevertheless centered on fixed vision hand discovery alongside their picked data set (egocentric or in a vehicle). Specific conditions cannot be checked. This problem can be dealt with by entering datasets which include more scenarios. The problem of datasets can be caused by the combination of datasets methodology. Yin, H. proposed [23, 32] included technical determinations to address this problem. In a number of applications, this principle could be used.

The problem of data malfunction in the RCNN family reduces the number of articles. For example, in RCNN's last item guide, a 32×32 Article stays 2×2. Hypernet [25] presented Hyper, which contains different levels of characteristic maps and has them packed into a single space. Although the Hyper Feature fuses overall information, the skip layer system also decreases data that is not associated with the Hyper Feature.

We are suggesting a new method that uses skin that demonstrates the generation of hand proposal and cascading CNN for the identification of hands in several situations in this paper. Little research apparently has sought to strengthen the display of the skin and the CNN approach for hand recognition. Given the pace of the planned technique implementation, they have altered the SSD [26] with a cascaded overall structure to retain in-depth settings. As positive example and photo patch Net 2012 [19], FDDB [20] data as bad case, we have built and are testing the proposed Oxford Hand Dataset [5], VIVA Hand Discovery [17] and EgoHands [18] mixed strategies. Our technique accomplishes aggressive hand identification results.

III. Proposed Method

For hand location, we present a multi-function controlled cascaded CNN structure. Fig. 2 shows the general flow. Next, by color region and histogram (to be multi-included specific), we focus the skin district A series of hand guidelines from the regional skin areas will be established at this stage. Eventually, the manual target is characterized by a cascaded CNN system. Skin Region Detection

For skin recognition task, Color is a generation included [27]. In the test [6], the low surface area can be isolated between hands and other similar hued items. We have used HSV and LAB color areas in conjunction with [6] to test the identification of the skin region. When color-based skin conditions are addressed, items giving equal color to skin are abused. In order to increase the bias of the skin color surface, they incorporate the response of 32 Gabor channels (4 directions and 4 scales). Therefore, the work for the skin display relies instead of a single pixel assessment on nearby pixel data. Since spatially modified neighbourhood-inclining histogram characteristics are capable of productively capturing the invariable characteristics of the area, we have chosen the Hoard [9] and Filter [10] descriptors to capture neighbouring hands [22] forms. Like [6], we train a gathering of regressors listed by a worldwide color histogram.

The posterior appropriation of a pixel i given a nearby appearance feature l and a global appearance include g, is registered by minimizing over various scenes s

$$p(i|\, l, g) = \sum_s p(i|\, l, s)\ p(s|g) \qquad (1)$$

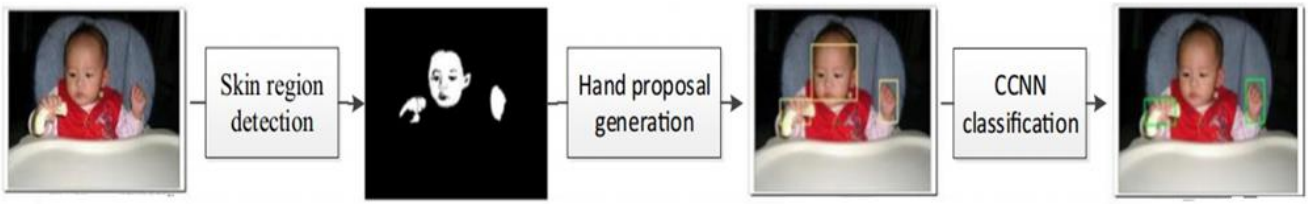

Fig. 2. Presents the Proposed Method Pipeline.

Where $p(i|\, l, s)$ is the yield of a discriminative worldwide appearance-explicit regressor and $p(s|\, g)$ is a contingent conveyance of a scene s given a worldwide appearance feature g. We use k-implies the bunch of different objects to construct global appearance models on the HSV histogram and each community is taught about various arbitrary tree regressors. Through scene unit codes both the look and the light of the scene by a histogram across each of the three HSV color streams. At the end, the skin saw will share a comparative transport appearance in the feature area. The restrictive $p(s|\, g)$ is approximated utilizing a uniform dispersion over the n closest models (in our examinations, $n = 5$) learned at preparing.

A. Hand Proposed Generation

In view of the skin mask, we determined the base encasing square shape (r_m) of each associated area (R). Likewise, equivalent to the explanation organization of article location task, The Hand proposal box (B) is defined as the most severe x and y directions provided by xmax, xmin, ymax and ymin to potentially useful objects. To generate a hand proposal from the hands that contribute to the arm area, we decide whether the length is over multiple times than the width of r_m, this associated district presumably speaks to an arm locale and the hand item is toward the finish of this area. If there should arise an occurrence of misdetection of multi-feature model (particularly limit pixel misfortune), we calculate a hand proposition box with t times the width of rm at each finish of this district. In our test, $t = 1.5$ created the best outcomes. The hand proposition generation is exhibited in algorithm 1. The hand proposal generation represented in Table I.

B. Hand Recognition with Cascaded Feature Aggregation CNN

We create a cascaded feature aggregation CNN to identify the hand area under the control of a hand proposition generated by multiple apps. As previously mentioned, the hand goal position has different hand photos in real situations. CNN's craft-based policies such as Quicker RCNN [13] and SSD [26] have unbelievable results in the discovery of the items on the PASCAL VOC dataset. However, the subsequent aims of the last layer feature guide, because it is much literal than the information picture (during the use of the pooling layer), generally do not contain the subtleties of a small article and setting of data. In the genius presented hand location CNN strategy, we pick SSD [26] as fundamental design in light of identifying speed and a cascaded feature conglomeration CNN structure is utilized to lessen the loss of setting data.

Assuming i and O represent the input and output of SSD network respectively. F_n is the feature maps of $n - th$ layer, $\theta_n(.)$represents the non-linear layers between the $(n - 1) - th$ layer and $n - th$ layer including convolutional layers, pooling layers and ReLU layers, and so on., $Y_n(.)$ is the capacity to change the $n - th$ layer feature maps to the recognition results for a specific scale range. f is the last operation to complete all the intermediate outcomes and produce the new recognition is therefore defined by multi-scale SSD feature maps:

$$F_n = \theta_n(F_{n-1}) = \theta_n\theta_{n-1}\,(\dots\theta_1(I))), \tag{2}$$

$$O = f\big(\gamma_n(F_n), \dots, \gamma_{n-i}(F_{n-i})\big), n > i > 0, \tag{3}$$

As per Eq. (3), it seems to rely strongly on a solid hypothesis of success. Since the feature charts in each layer are solely responsible for the performance of their size, each F alone is believed to be sufficiently advanced to support correct identification and restriction. The detailed map of component F must contain appropriate data that can heartfully capture the careful area of the low, shielding, obstructed or dark objects, including hands [13, 26]. We present a gradually sensitive element reproduction that takes account of the data setting in a cascaded system structure. Fig. 3 shows the nature of the cascaded structure. The function is characterized as follows:

$$O = \hat{f}\left(\gamma_n\left(\hat{F}_n(S)\right), \gamma_{n-1}\left(\hat{F}_{n-1}(S)\right), \dots, \gamma_{n-i}\left(\hat{F}_{n-i}(S)\right)\right), \tag{4}$$

$$S = \{F_n, F_{n-1}, \dots, F_{n-i}\}, n > i > 0,$$

$$size\,(F_{n-i}) = size\,\left(\hat{F}_{n-i}(S)\right), for\ all\ i \tag{5}$$

TABLE. I. SHOWS THE HAND PROPOSAL GENERATION

Hand proposal generation
Input: skin region mask set ***R***
Output: hand proposal box set ***B***.
For $i = 1: num(R)$
1. Calculating of minimum enclosing rectangle r_m along with the width and length of r_m.
2. For each r_m
If $\frac{1}{3}\ width \le length \le 3\ width$
Then ***B*** = xmax, xmin, ymax, ymin of R
3. Else
B_1 =xmin+1.5 width, xmin, ymax, ymax-1.5width of ***R***
B_2 =xmax, xmax-1.5 width, ymin+ 1.5 width, ymin of ***R***
End for
4. Return ***B***
End

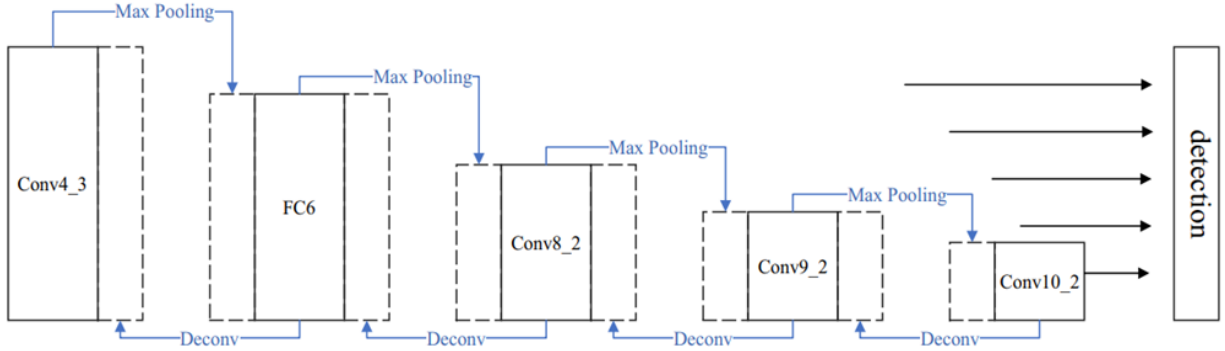

Fig. 3. Structure of the CCNN Network. The Function Aggregation in Reduced VGG-16 is shown in this figure. The Strong Boxes Represent SSD Characteristics Such as Conv4_3, FC6, Conv8_2, Conv9_2 and Conv10_2. The Dotted boxes Selected by Arrows were Added to the Next Feature Maps. Inspired by Hypernet [25], we are Integrating Max Pooling and Deconvolution Operations, as shown by Arrows, to Add up and Down Capabilities.

Where S is a lot of all the feature maps contribute to the detection function $f(\bullet)$. Dissimilar to in eq. (3), $\hat{F}_n(.)$ is a function that considers all the contributing feature maps and yields another include portrayal of a similar dimensionality to F_n.

An immediate mapping from S to $\hat{F}_n(S)$ need to be depend on an impressive size deep network with various layers of non-linearity. This will cost a lot of algorithm and it will be difficult to organize one generation. The alternative is to design an iterative approach where each move advances a little but substantially and consistently. Enlivened by [25], the map maps from conv1, conv3 and conv5 are used for Max pooling and Deconvolution. We use a similar activity in CCNN to add the neighbouring layer feature maps (as shown in Fig. 3). The following is a mathematical overview of this approach,

$$\widehat{F_p^{t+1}} = M\left(\widehat{F_p^t}, \widehat{F_{p-1}^t}, \widehat{F_{p+1}^t}; W\right), t > 0, \quad (6)$$

$$\widehat{F_n^t} = F_n, for\ all\ n \text{ where } t = 1, \quad (7)$$

Where M is a function maps only $\widehat{F_n^t}$ and its adjacent (higher and lower level) counterparts at step t to a new $\hat{F}_p$ at step $t + 1$. The variable M has a weight parameter W. The learning goal failure feature is similar to SSD with the lack of position and confidence. Assuming $x_{ij}^k = \{1,0\}$is an indicator for matching the $i - th$ default box to the $j - th$ ground truth box of category k:

$$L(x, c, h, g) = \frac{1}{N}(L_{conf}(x, y) + \alpha L_{loc}(x, h, g)) \quad (8)$$

Where, N is the organized standard box number. If N= 0 is omitted, the loss is set to 0. The loss of restrictions is a Smooth L_1 loss between the box (h) and the parameters for the bottom truth box (g). The Softmax loss for many classes of confidences is the loss of certainty (c). Through cross-approval, the weight variable α is set at 1.

C. Summary of Proposed Method

As indicated by the past description, the algorithm flow of MFS-CCNN can be summarized by Fig. 4. We extract the skin mask of input image patch by Multi-feature model. In light of skin mask, the hand proposition bouncing boxes could be produced by algorithm 1. At last, CCNN could characterize and relapse the hand protests alongside exact position. The assessment of MFS-CCNN will be introduced in Section 4.

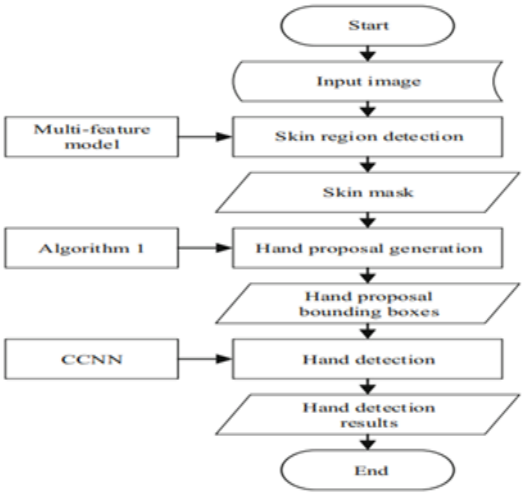

Fig. 4. MFS-CCNN Structure Graph.

IV. Experimental Results

In multi-scenes, this work aims to recognize hands. Our skin recognition model uses a sparse combination of HSV, LAB Gabor, Hoard and filter features. We created the template using the GTEA dataset and the skin district with a 9×9 image patch update, as suggested by [6]. The hand district plan is based on the skin template over-division that is appropriate for multi-situations with different foundations. The CCNN base system is a 2012 object patch network pre-trained VGG16 [28]. Our models were developed and tested using a mixture of Oxford Hand Dataset [5], VIVA hand location 17] and EgoHands [18] and neutral examples were also used to distinguish hands from various grades [19] and faces [20]. Remember that this research is just a set of commands, so EgoHands' "left own," "wrong say," "left side" and "other right" are listed as one category – hand. Table II, which contains 14,151 image patches for preparation and approval and another 7283 photographs for testing, shows how subtle the joined hand dataset is. For evaluation of the proposed technique we receive a 0.5 IoU Sift. The Stochastic Slope Plummet network has been optimized to maximize 60,000 simulations, with an energy of 0.9. We have reduced weight and the underlying study rate is 0.0005. The sample sequence of the analysis is shown in Fig. 5. The main section is the first selected image from the hand datasets of EgoHands, VIVA and

Oxford. We use the multi-function skin recognition model right outside the bat to focus skin cover as shown in the following section. We establish the proposition of the hand item as shown in the third column in light of the cover on body. Ultimately, the hand exploration findings are the last segment. The implementation of the DPM-based strategy [5], SSD [26], CCNN and MFS-CCNN is outlined in Table III. The details of the combined hand dataset are shown in Table II; it contains 14,151 images for training and validation, and another 7283 images for testing. A 64-bit Ubuntu 14.04 PC with the CPU Intel (R) Core(TM) i7-5960X, 320 G memory and TITAN Xp GPU will assess the method proposed, but the results showed that it can improve the recognition with skin model proposal and the conglomeration of information within the CNN. The proposed technology is also evaluated. Our normal precisions on the test datasets are altogether superior to [5]. Our normal data set accuracy is totally greater than [5]. With standard precision of 43.55 % and 51.78 %, our CCNN models alone and MFS-CCNN beat [5]. In consolidated test set, the CCNN model achieved normal accuracy of 9.16 % higher than the SSD model. In the combined sample collection Oxford Hand Dataset and Egohands, MFS-CCNN obtained the most notable performances. However, the VIVA test results show that the best possible performance for the CCNN model is created (0.86 per cent above CCNN-MFS). In addition, the consequences were assessed using the precision check bends outlined in the Fig. 6 for MFS-CCNN, CCNN and SSD. For red, green and blue respectively, the after-effects of MFS-CCNN, CCNN and SSD are spoken of. We use 11 accuracy points to create the plot and check information. As shown in line two Fig. 5, the right position is in a low luminances state, the whole locale of this hand can hardly be differentiated by our skin template, so that CCNN refuses to recognize it. Fig. 7 demonstrates the efficiency of the minor hand on the right side by CCNN. While the skin position model in low luminance is shaky, it has improved hand discovery (the MFS-CCNN model has averaged CCNN by 8.23%).

We also contrasted the execution time of DPM [5], SSD [26], CCNN and MFS CCNN versions. They also compared the time. It's much faster than our methodology [5]. The conglomerate structure with cascaded components and the skin template with many inclusive improved calculations in comparison to SSD. For instance, Squeezenet [29] and MobileNets [30], we accept that our approach can be time efficient by using a lightweight CNN as the basis system.

TABLE. II. DISTRIBUTION OF EXPERIMENT DATASET

Datasets	Training	validation	testing
Oxford Hand Dataset	3031	1780	823
VIVA hand detection	3465	2035	5500
EgoHands	2458	1382	960
Total	8954	5197	7283

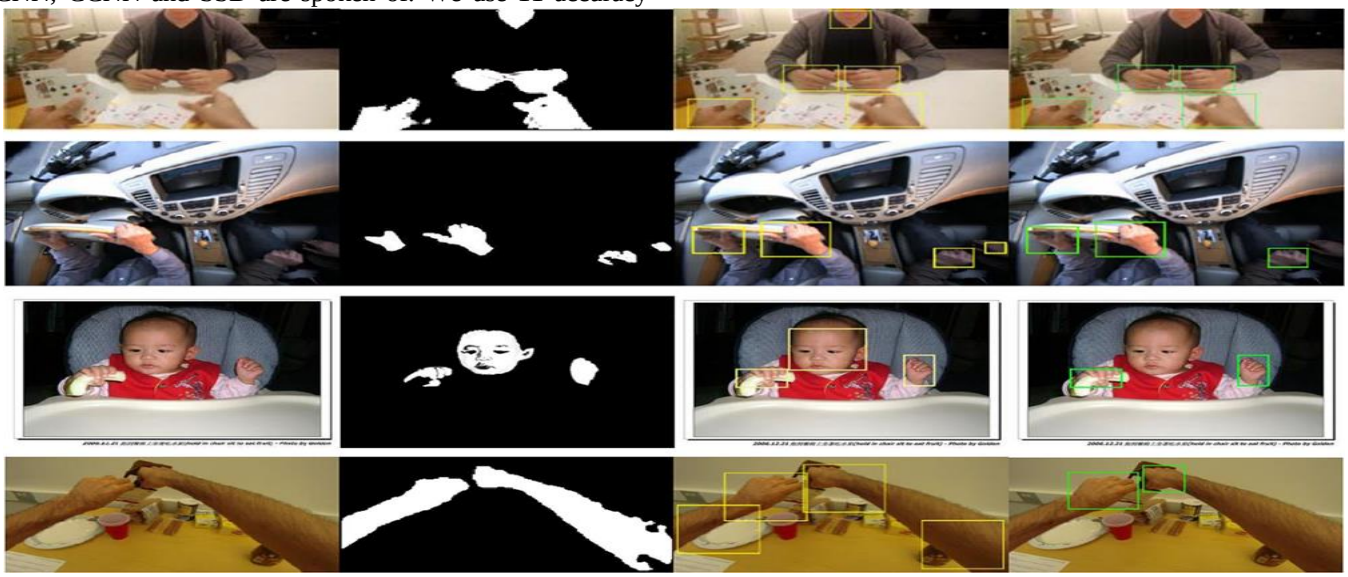

Fig. 5. The Original Patch of the Selected Datasets is Column 1. Column 2 is the Mask Produced by our Skin Model for the Skin Region. Column 3 is a Skin Mask Hand Proposal. The Detection Results are in Column 4.

TABLE. III. RESULTS ON THE TEST SET

Methods	Combined test set	Oxford Hand Dataset	VIVA hand detection	EgoHands	Running time
DPM based [5]	46.62%	48.20%	40.15%	47.13%	26 s
SSD [26]	79.43%	79.78%	82.49%	77.96%	0.018 s
CCNN model (ours)	89.67%	85.65%	92.17%	88.81%	0.046 s
MFS-CCNN model (ours)	91.22%	89.26%	91.31%	91.76%	0.057 s

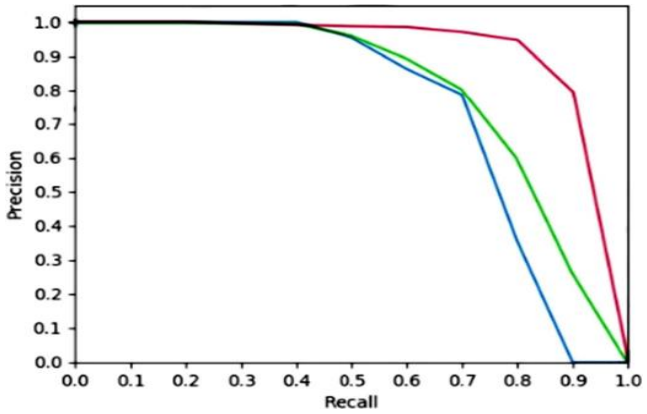

Fig. 6. Curve Notification Accuracy for Hand-Detection Correlation test. The Tests of MFSCCNN, CCNN and SSD are shown by a Black, Green and Blue Curve.

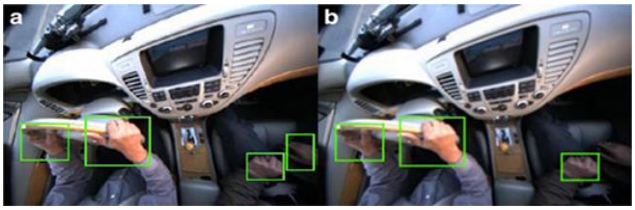

Fig. 7. Comparison between CCNN and MFS-CCNN for Weak Luminance Detection Tests. The CCNN and MFS-CCNN Results are Respectively (a) and (b).

V. Conclusion

In this paper we propose an approach for hand detection based on the multi-feature cascaded CNN model for the skin. We first use a HSV, LAB, Gabor, Hoard and Filter skin model for the identification of the skin area. Secondly, hand suggestion boxes depending on skin masks are produced. Finally, we propose to classify hand objects by cascading the CNN feature aggregation. The test shows that both the skin model and CCNN are able to enhance recognition and achieve state-of-the-art findings on mixed datasets. The technology proposed achieves aggressive results. Our average data collection accuracy is substantially better than DPM. With an overall accuracy of 43.55 and 51.78 % respectively, our MFS-CCNN system and CCNN CCNN alone outperform. The average accuracy of CCNN model in the combined test set was 9.16 per cent better than that of the SSD. Furthermore, the numerical results show our model is faster than DPM. Future work will focus on timely assessment. The current evaluation and recognition of operation will also be updated.

References

[1] Stergiopoulou, K. Sgouropoulos, N. Nikolaou, N. Papamarkos, and N. Mitianoudis, "Real time hand detection in a complex background," Engineering Applications of Artificial Intelligence, vol. 35, pp. 54-70, 2014.

[2] A. Ebert, N. D. Gershon, and G. C. van der Veer, " Human-computer interaction: introduction and overview," Künstliche Intelligenz, vol. 26, no. 2, pp. 121-126, 2012.

[3] J. Zariffa, and M. R. Popovic, "Hand contour detection in wearable camera video using an adaptive histogram region of interest," J NeuroEng Rehab), vol. 10, no. 1, pp. 114–114, 2013.

[4] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from RGB-D image patches," IEEE International Conference on Computer Vision, vol. 22, pp. 3889–3897, 2015.

[5] A. Mittal, A. Zisserman, and P. Torr, "Hand detection using multiple proposals," British Machine Vision Conference, vol. 40, pp. 1–11, 2011.

[6] C. Li, and K. M. Kitani, "Pixel-level hand detection in ego- centric videos. Computer Vision and Pattern Recognition," vol. 9, pp. 3570–3577, 2013.

[7] A. Fathi, and J. M. Rehg," Learning to recognize objects in egocentric activities," IEEE Conference on Computer Vision and Pattern Recognition, vol. 42, pp.3281-3288, 2011.

[8] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara," Hand segmentation for gesture recognition in EGO-vision," ACM International Workshop on Interactive Multimedia on Mobile & Portable Devices, vol. 24, pp. 31–36, 2013.

[9] N. Dalal, and Triggs,"Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol. 1, pp. 886–893.

[10] D. G. Lowe, "Distinctive image patch features from scale-invariant key points," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region- based convolutional networks for accurate object detection and segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp.142–158, 2015.

[12] R. Girshick, "Fast R-CNN," IEEE International Conference on Computer Vision, pp.1440-1448, 2015.

[13] S. Ren, K. He, R. Girshick, and J. Sun," Faster R-CNN: Towards real-time object detection with region proposal networks," International Conference on Neural Information Processing Systems, vol. 39, pp. 91–99, 2015.

[14] W. Dai, Y. Li, K. He, and J.Sun, " R-fcn: Object detection via region-based fully convolutional networks," 2016.

[15] A. Erol, G. Bebis, M. Nicolescu, R.D. Boyle, and X. Twombly,"Vision based hand pose estimation: A review," Computer Vision & image patch Understanding, vol. 108, no. 1, pp. 52–73. 2007.

[16] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan," Vision-based hand-gesture applications," Communications of the ACM, vol. 54, no. 2, pp. 60– 71. 2011.

[17] The Vision for Intelligent Vehicles and Applications (VIVA) Challenge, Laboratory for Intelligent and Safe Automobiles, UCSD. http://cvrr.ucsd.edu/vivachallenge/, accessed on (9 feb 2019).

[18] S. Bambach, S. Lee, D.J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," IEEE International Conference on Computer Vision, pp.1949-1957, 2016.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei," image patch net large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[20] V. Jain, and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," UMass Amherst Technical Report, 2010.

[21] A. Betancourt, "A sequential classifier for hand detection in the framework of egocentric vision,"Computer Vision and Pattern Recognition Workshops, pp.600-605, 2014.

[22] Q. Wang, and G. Zhang, "Ore image patch edge detection using hog-index dictionary learning approach," Journal of Engineering, vol. 1, no. 1, 2017.

[23] H. Yin, and K. Gai," An empirical study on preprocessing high-dimensional class-imbalanced data for classification," International Conference on High PERFORMANCE Computing and Communications, pp.1314-1319.

[24] T. H. N. Le, C. Zhu, Y. Zheng, K. Luu, and M. Savvides," Robust hand detection in vehicles," International Conference on Pattern Recognition, pp.573-578, 2017.

[25] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," Computer Vision and Pattern Recognition, pp.845-853, 2016.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C.Y. Fu,"SSD: Single shot MultiBox detector," European Conference on Computer Vision, pp.21-37, 2016.

[27] P. Kakumanu, S. Makrogiannis, N. Bourbakis,"A survey of skin-color modeling and detection methods," Pattern Recognition, vol. 40, no. 3, pp. 1106–1122, 2007.

[28] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image patch recognition," arXiv preprint arXiv:1409.1556, 2014.

[29] F. N. Iandola, S. Han, M. W. Mickiewicz, K. Ashraf, W.J. Dally, and K. Keutzer," Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," arXiv preprint arXiv:1602.07360, 2016.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, and T. Weyand," Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[31] K. Gai, M. Qiu, and X. Sun," A survey on fintech," Journal of Network & Computer Applications, vol. 103, pp. 262-273, 2018.

[32] H. Yin, K. Gai, and Z. Wang," A classification algorithm based on ensemble feature selections for imbalanced-class dataset, "International Conference on Big Data Security on Cloud (pp.245-249), 2016.