Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation

Kankana Roy[1], Aparna Mohanty[2], and Rajiv R. Sahay[2]
Computational Vision Laboratory,
[1]Department of Computer Science and Engineering, [2]Department of Electrical Engineering,
Indian Institute of Technology Kharagpur
{kankana.kankana.roy,aparnamhnty,rajivsahay}@gmail.com

Abstract

Robust detection of hand gestures has remained a challenge due to background clutter encountered in real-world environments. In this work, a two-stage deep learning based approach is presented to detect hands robustly in unconstrained scenarios. We evaluate two recently proposed object detection techniques to initially locate hands in the input images. To further enhance the output of the hand detector we propose a convolutional neural network (CNN) based skin detection technique which reduces occurrences of false positives significantly. We show qualitative and quantitative results of the proposed hand detection algorithm on several public datasets including Oxford, 5-signer and EgoHands dataset. As a case study, we also report hand detection results robust to clutter on a proposed dataset of Indian classical dance (ICD) images.

1. Introduction

In this work, we propose a robust hand detection technique for still images and extracted video frames using deep learning. Detecting hands in the images can be useful for many applications such as automatic sign language analysis [4], fine grained action recognition [5], movie interpretation and even for understanding dance gestures [6]. Note that we restrict ourselves to detection of hands in images and the scope of the work presented here does not involve recognizing and understanding of hand gestures. On the other hand automatic gesture recognition in an unconstrained image is a very challenging problem because it requires robust detection of hands despite background clutter, noise, poor illumination etc. Several challenging images are shown in Fig. 1 depicting the immense problem of hand detection in the wild. Recently, several researchers have attempted to address the interesting yet challenging problem of automatic semantic interpretation of small videos of Indian classical dance [7, 8, 9, 10, 6]. However, localization of hands in videos of ICD is a challenge because of the complexities in the costume, make-up, cluttered environment, people in background etc. The works of [9, 10] resorted to manual segmentation of hand regions in order to identify corresponding hand gestures of ICD. Hence, as part of our study we also show hand detection results over images of ICD. We have captured small videos of Indian classical dancers performing ICD[1]. We took frames containing vary-

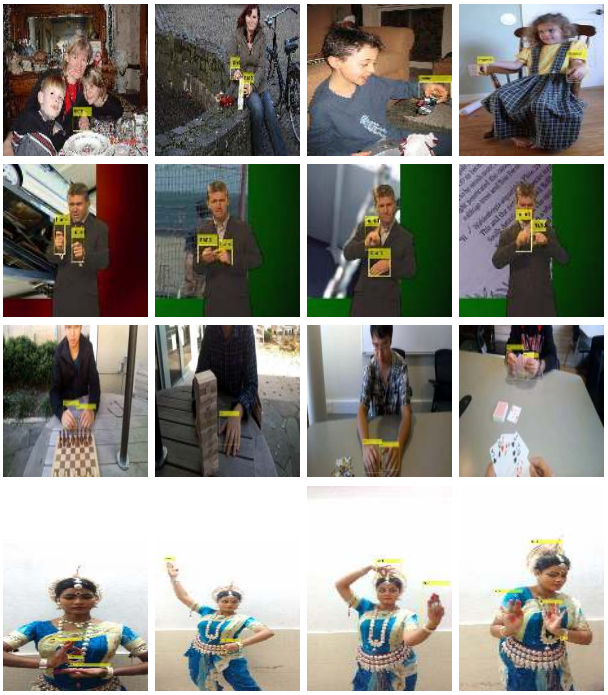


Figure 1. Some example outputs of our proposed hand detection algorithm on Oxford hand test dataset [1] (first row), 5-signer dataset [2] (second row), EgoHands dataset [3] (third row), our proposed ICD dataset (last row).

[1]https://github.com/kankanar/ICCVW-2017

ing poses, from these videos to build our ICD dataset. A total of 657 frames are collected from 7 videos. To address the problem of robust hand detection in unconstrained scenarios here we propose a two-stage image based approach using deep learning. In this paper we attempt to evaluate the performance of two recent CNN-based approaches for hand detection. Initially, we used the object detection algorithm, namely, region-based convolutional neural networks (RCNN) [11] and adapt it for the specialized task of hand detection. Chen *et al*. [12] showed that with proper context and dataset RCNN could detect small objects. But RCNN has some major drawbacks such as detection speed along with high disk space requirement for storing region proposals. Also, region proposals need to be sequentially passed into the classification network. Subsequently, we consider a more recent and advanced algorithm, namely, Faster-RCNN (FRCNN) [13] which supposedly overcomes the drawbacks of the RCNN. FRCNN uses the same architecture for both region proposal network and the classifier. There is no requirement for storing external region proposals as the network generates them simultaneously. But, fixed number of convolution layers and filter receptive fields hinder FRCNN from generating robust region proposals which is very important in the multi-scale problem of hand detection [14]. So, we compared both object detectors to decide which is most suitable for hand detection. As already mentioned in the work of Cai *et al*. [14] RCNN and FRCNN are not very good for detecting small objects on their own. We observe that in cluttered backgrounds, both RCNN and FRCNN algorithm produces many false positives. Given an input image, initially we use either RCNN [11] or FRCNN [13] to detect the hand regions. Thereafter, to improve the results obtained by hand detector, we propose a deep learning based skin segmentation method. Row one to four of Fig. 1 represent some example outputs of proposed hand detection algorithm on Oxford hand test dataset [1], 5-signer dataset [2], EgoHands dataset [3] and our proposed ICD dataset, respectively. Note that our algorithm can accurately localize hands in spite of different kinds of clutter in the background. The contributions of our work can be summarized as:

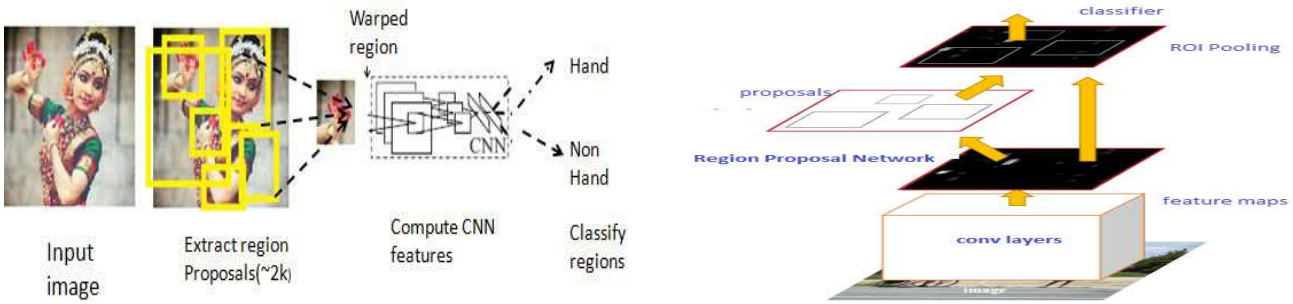
- We individually train two deep learning based object detectors, namely RCNN [11] and FRCNN [13], for obtaining the initial estimate of the spatial location of hands in the input image.
- To reduce false positives in the above first stage of our hand detector, we propose a patch-based convolutional neural network skin detector which is robust to background clutter and detects skin pixels accurately.
- To improve the computational efficiency of the skin classifier we also propose a regression based end-to-end full image segmentation CNN wherein the total variation regularization is incorporated.

The remainder of the paper is organized in the following manner. Section 2 describes related work. Section 3 outlines proposed methodology. In section 4 we present experimental results and finally conclusions are given in Section 5.

2. Related Works

Although several researchers have addressed the problem of detection of hands in an image, a robust algorithm is still elusive. This is primarily due to the fact that hands are deformable and articulated in nature. For high level computer vision tasks such as parsing hand poses, gesture understanding for robotics and human computer interaction, human layout detection [15], action recognition [5], sign language recognition [2] and human activity analysis [5], hand detection is inevitable. Many hand detection algorithms have been proposed in the past [16, 17]. Prior works on hand detection used external hardware such as depth sensors [18]. But the use of depth sensors might not be feasible in all environments such as during a live dance concert. Pisharady *et al*. [19] used a saliency map to detect hands in the image. Mittal *et al*. [1] proposed a state-of-the-art hand detection algorithm by fusing three techniques, namely, hand shape detector, context detector and skin-based detector. Ong and Bowden [20] proposed a tree classifier to detect and recognize hand pose. Kolsch and Turk [21] proposed cascade hand detector using Haar features. Buehler *et al*. [2] proposed a hand detection method using multiple cues. Do and Yanai [22] proposed a hand detection and tracking technique for fine grained action recognition in videos. The algorithm of [22] uses multiple cues for hand detection such as the work of Mittal *et al*. [1] and includes upper body detection and flow information. Hoang *et al*. [23] proposed to use a multiple scale FRCNN along with other cues such as face, steering wheel and cell-phone to detect hands inside a car for studying driver cell-phone usage. Deng *et al*. [24], have proposed a CNN based approach to detect hands and estimate rotation. The work in [24] proposes a context aware region proposal algorithm that uses a multi-component SVM.

To improve the results obtained by hand detector, we propose a deep learning based skin segmentation method. Skin detection has various applications in areas including face detection, hand gesture tracking, human computer interaction and objectionable image detection/blocking etc. Challenges for skin detection include skin tone variation, ambiguity in foreground background separation, occlusion and illumination [25]. In the literature, there exist broadly two approaches for skin detection, namely, pixel based and region based classification. In this paper we attempt to label



(a) (b)

Figure 2. (a) RCNN architecture. (b) Faster-RCNN architecture. (Original figure source [13]).

each pixel in the image as skin or non-skin. Skin detection methods that are reported in the literature include Gaussian models (single and mixture of Gaussians) [26], Bayesian classifiers [27], skin probability map [28], and neural networks [29]. A survey of the various skin classification algorithms is presented in [30]. Recently, Lei *et al*. [31] have used stacked autoencoder for skin segmentation. Zuo *et al*. [32] used CNN and recurrent neural networks (RNN) for human skin detection.

3. Methodology

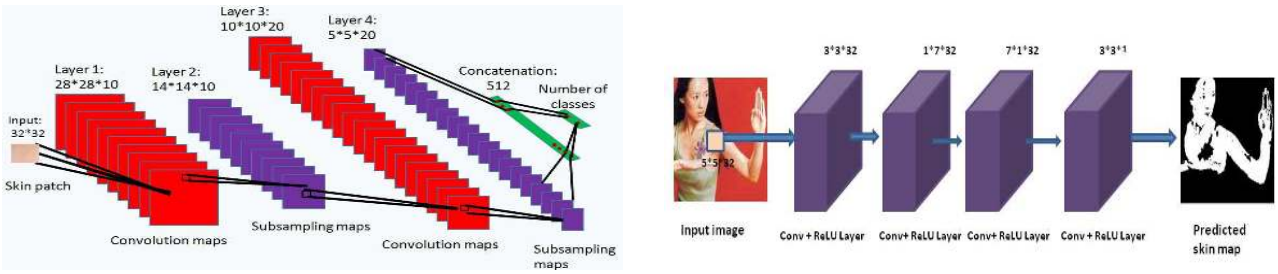
In this paper we attempt to address the problem of hand detection in two steps. The first stage of our algorithm attempts to obtain probable hand locations in an image by generating bounding boxes. The second stage of our approach attempts to reduce the false positives from the estimated bounding boxes. For the first part we trained two individual state-of-the-art object detectors, namely RCNN [11] and FRCNN [13] in similar fashion to detect hands. For the second stage we propose a skin detection algorithm to reduce false positive rates. Firstly, we propose a patch-based skin detector which takes as input images which are divided into overlapping patches and are classified as skin or non-skin. Although this method is quite accurate for detecting skin pixels, the computational cost is high because of the sliding window based detection protocol. Inspired by recent advancement in low level image processing tasks such as image deconvolution using CNN [33], we also resorted to a different CNN architecture which processes the whole image at once instead of using patches. In the next sub-section we will discuss both the hand detection and skin segmentation algorithms.

3.1. Hand detection using RCNN [11]

RCNN [11] is a state-of-the-art visual object detection system that combines bottom-up region proposals with rich features computed by a convolutional neural network. The utility of RCNN has been demonstrated previously in [11] wherein a substantial improvement in the mean average precision (MAP) from a little more than 30% to 53.3% is achieved on Pascal VOC 2012 challenge [34]. Previous research showed that CNNs have incredible ability to distinguish between different classes of objects irrespective of their shape, size and color [35]. Inspired by performance of RCNN for object detection, we attempt to use it for localization of hands. RCNN algorithm has three distinct parts. The first step generates region proposals using selective search [36] to create candidate regions for hand localization. The second stage of RCNN uses a pre-trained convolutional neural network which extracts features over the region proposals. Here we used a network with three convolutional layers and two fully connected layers pre-trained with CIFAR-10 dataset [37]. The motivation of taking a small architecture is that higher convolutional layers respond very weakly to small objects [14]. The third stage is the regression network where mean square error between predicted bounding box and given bounding box is minimized. This method has been proved to be very efficient when used for object detection in natural images. Fig. 2 (a) shows the schematic of the proposed RCNN based hand detection algorithm. We trained the baseline RCNN on a combined dataset containing $13,269$ images from Signer dataset [2], Oxford hand data [1], ImageNet [38], INRIA pedestrian dataset [39], EgoHands dataset [3] with annotated hand regions.

3.2. Faster-RCNN [13] based Hand Detector

A drawback of RCNN [11] is that it generates region proposals independent of the classification stage. These region proposals are given as input in a sequential manner to the CNN to obtain the classification scores. This procedure is slow and processing large input images is tedious. In contrast, as shown in Fig. 2 (b), the work in [13] proposes a fully covolutional network called region proposal network (RPN) which shares convolutional layers with the Fast-RCNN [40] object detection classifier. This algorithmic novelty in [13] overcomes the bottleneck of generating region proposals sequentially as done in previous ob-



(a) (b)

Figure 3. Architecture of the two networks used in our work: (a) Patch-based CNN. (b) Regression based end-to-end CNN for skin segmentation.

ject detection works. RPN slides over the shared convolutional feature maps to determine whether the region is an object or not. FRCNN consists of 4 stages. Firstly RPN is initialized with a pre-trained model and fine-tuned end-to-end for the task of generating region proposals. In the second step, the Fast-RCNN [40] object detection network is trained using the proposals generated by the RPN. In the third step RPN is again fine-tuned using detector network. Each region proposal of RPN is passed into a region of interest pooling layer to fix shared convolutional layers. In the fourth step fully connected layers of Fast-RCNN are fined-tuned keeping convolution layers fixed. It is to be noted that region proposals generated by RPN are less sensitive to scale due to a fixed set of filters over a fixed set of convolutional feature maps [14]. For the proposed FRCNN based hand detector we have used the same training data set containing $13,269$ images which has also been used for RCNN to provide an unbiased comparison.

3.3. Skin detection using convolution neural network

Detection of hands in unconstrained images is a challenging task because they are deformable in nature, have many degrees of freedom with substantial wrist movement, self occlusions etc. Due to these challenges, CNN based hand detection method as described in sections 3.1 and 3.2 tend to give some false positives. One way to reduce false positive is to check whether bounding boxes generated by the hand detection algorithm, contain skin pixels or not. We have proposed a deep learning based skin segmentation method to eliminate those bounding boxes, which do not contain skin or wherein skin regions are less. In our proposed approach, we have considered the problem of skin detection as a per pixel classification challenge. In this work, we have used two different CNN architectures to segment skin regions from images. The first skin classification approach which we propose is a patch-based CNN whose detailed architecture is shown in Fig. 3 (a). For every point in the input image we have considered a surrounding region of size 5×5 which provides contextual information to detect skin pixel. Following Ciresan *et al*. [41], who have used a sliding window based technique to segment brain cells, for robust skin segmentation we used a stride of 1 pixel. Note that we re-sized the input image patches from 5×5 to 32×32 pixels before feeding them to the proposed CNN. The network in Fig. 3 (a) consists of two convolution layers with kernels of size 5×5 followed by cross channel normalization, ReLU and pooling layer. Two fully connected layers are followed by a softmax classification layer. The network has two output classes skin and non-skin. For training this architecture we collected a large dataset consisting of approximately 1 million $(1,01,880$ to be exact) skin and non-skin patches which are obtained from ImageNet Large Scale Visual Recognition Challenge dataset [38] and annotated manually. Positive samples corresponds to skin image patches, and negative samples contain non-skin image patches. We made sure that our skin dataset is large and diverse enough to detect skin pixels despite varied race and color. All the skin and non-skin patches are collected from ImageNet [38] ensuring variation and diversity. Non-skin patches contain skin-like image regions such as fire, fur, wood and sand to increase the robustness of the skin detection algorithm.

During testing we divided the input image into overlapping patches with stride of one column and replace the center pixel by the output of classification layer. The input patches of 5×5 pixels are re-sized to 32×32 and are then convolved with 10 filter maps of size 5×5 resulting in 10

Dataset	Accuracy (%)	F-score	TPR	TNR
Pratheepan [42]	84	67	83	84
Uchile [43]	87	65	70	90
IBTD [44]	77	58	86	76
HGR [45]	93	85	80	97

Table 1. Performance measure of proposed patch-based CNN for skin segmentation on different datasets.

Data-set	Mittal *et al*. [1]	Deng *et al*. [24]	RCNN [11]	RCNN +Skin	FRCNN [13]	FRCNN +Skin
Oxford hand dataset [1]	42.3	48.3	31.23	**49.51**	14.22	31.12
EgoHands dataset [3]	—	75.7	57.27	92.96	50.00	**96.00**
5-signer dataset [2]	76.67	—	95.56	**97.27**	29.03	69.00
ICD dataset	—	—	25.69	35.33	24.39	31.88

Table 2. Performance (average precision) improvement obtained using the proposed skin segmentation algorithm along with RCNN [11] and FRCNN [13] based hand detection method on different public datasets.

output maps of size 28×28 in layer 1. The obtained feature maps are then down sampled with max pooling of 2×2 regions to yield 10 output maps in layer 2. These feature maps are again convolved with each of 20 kernels of size 5×5 pixels to obtain 20 output maps of size 10×10 pixels. These outputs are further down sampled by 2 via max-pooling to produce 20 output maps in layer 4 which are concatenated to obtain a single vector that is fed to the next layer. The number of neurons in the final layer depends on the number of classes. There is full connection between the neurons in the final layer and the previous layer. The outputs of neurons in the proposed architecture are modulated by the non-linear activation function (rectified linear unit i.e ReLU) to produce the resultant score for each class. Despite the small size of this network it yields state-of-the-art results on various standard datasets for skin segmentation as shown in Table 1.

However, the main drawback of this network is that it is slow because of the use of sliding windows. To overcome this problem we propose another deep CNN regression network which processes the whole image at once within seconds. This model is inspired by the CNN models used for low level image processing tasks such as image super-resolution [46]. The proposed CNN predicts the class of all pixels in the input image simultaneously at a time. We show full image skin segmentation CNN architecture in Fig. 3 (b). This network takes 227×227 pixels color image as input. It has total five convolution + ReLU layers. Note that there is no pooling layer as we want pixel-wise prediction. The first layer consists of 32 convolution kernels of size 5×5 followed by a ReLU layer. Second layer consists of 32 convolution kernels of size 3×3 followed by a ReLU layer. Third layer consists of 32 convolution kernels of size 7×1 followed by another ReLU layer. Fourth layer consists of 32 convolution kernels of size 7×1 followed by ReLU layer. The last layer consists of 1 convolution kernels of size 3×3. We call this a prediction layer as it predicts a skin map. We have performed several experiments changing the number of convolution layers. Specifically, we used $2, 3, 4, 5, 8$ convolution layers and find out that 4 convolution layers yields he highest accuracy. We propose to add a customized loss layer in order to minimize the error between predictions $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$ and train the network weights using stochastic gradient descent.

Specifically, for hand detection we propose to add the total variation prior [47] to obtain robust results. Suppose, we have to solve the task of skin segmentation using supervised learning with a training set of input-target pairs $x^{(i)}, y^{(i)}$. Our objective is to learn the parameters θ of a representation function G_θ which optimally approximates the input-target dependency according to a loss function $L(G_\theta(x), y)$. A typical choice is mean squared error (MSE) loss

$$L(G_\theta(\mathbf{x}), \mathbf{y}) = \| G_\theta(\mathbf{x}) - \mathbf{y} \|_2^2 \quad (1)$$

However, these lead to blurry and noisy skin predictions. In our work, we combine mean squared error (MSE) with total variation of the prediction as given by

$$L(G_\theta(\mathbf{x}), \mathbf{y}) = \| G_\theta(\mathbf{x}) - \mathbf{y} \|_2^2 + \lambda_{TV} TV(\hat{\mathbf{y}}) \quad (2)$$

where λ_{TV} is the regularization parameter which is set as 0.06 in our experiments and, $\hat{\mathbf{y}}$ is the prediction. For training the proposed regression CNN we used $1,141$ images from datasets in [42, 43, 44, 45] with ground truth labeling of skin pixels. The testing set consisted of 495 images from the same datasets also with ground truth skin pixels annotation. We change the loss layer by first implementing Eq. 1 and then Eq. 2 and concluded that TV prior improves average accuracy by 1%. We finally obtained an average accuracy of 78% over the test dataset of 495 images.

4. Experimental section

As shown in Fig. 1, hand detection is a formidably challenging problem due to cluttered backdrops. The robustness of the proposed hand detection algorithm is demonstrated on the publicly available images of Oxford hand dataset [1], EgoHand dataset [3] and 5-signer dataset [2]. Note that the results are reported on various standard datasets with varying degree of complexities in terms of clutter in the background. As a case study we also show the performance of the proposed hand detection on a dataset of ICD, containing 657 frames where the hand gestures are semantically meaningful.

Figure 4. Comparison with [31]: The first and sixth columns show original images. Second and seventh columns show the output of [31]. Third and eighth columns show skin segmentation obtained by our patch-based CNN (Fig. 3 (a)). Fourth and ninth columns show skin segmentation obtained by our regression based CNN (Fig. 3 (b)). Fifth and tenth columns show ground truth images.

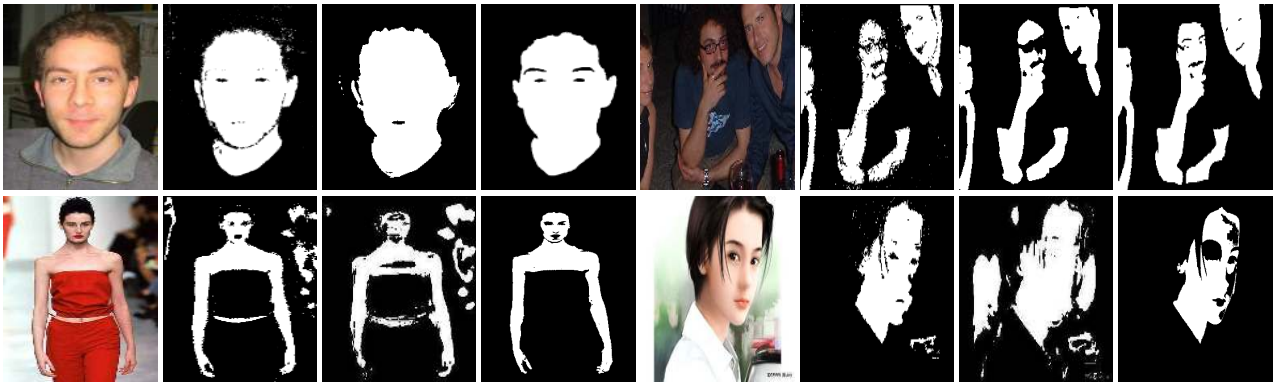


Figure 5. First row shows the skin detection result on Uchile dataset [43] and second row shows the result on IBTD dataset [44]. First and fifth columns show original images. Second and sixth columns show segmentation results of our patch-based skin detection algorithm (Fig. 3 (a)). Third and seventh columns show segmentation results of our regression based skin detection algorithm (Fig. 3 (b)). Fourth and eighth columns show ground truth images.

To address the challenging problem of detecting hands in images with cluttered background we adopt an approach consisting of two stages. The baseline RCNN or FRCNN trained on hand dataset gives the initial hand bounding box proposals, with some false positives. In natural images human face and hands contain maximum amount of skin pixels. Mathias *et al*. [48] showed that simple deformable part based model and detector based on rigid templates such as Viola and Jones [49] can detect faces robustly if properly trained. In this work, we have adopted the approach of [48] to detect and eliminate mis-classified bounding boxes containing faces. Non-maximal suppression, based on confidence score (> 0.9) of the obtained bounding boxes, is used to remove duplicate boxes.

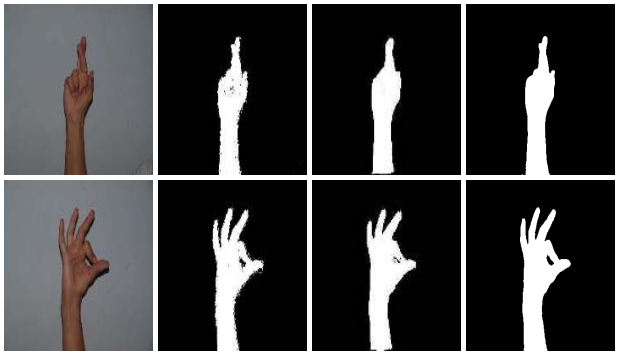
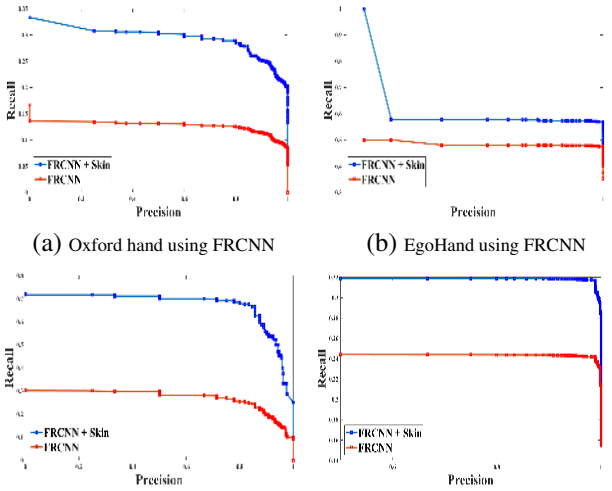


Figure 6. Results for the HGR dataset [45]: The first column show original images. Second column show segmentation results using proposed patch-based CNN. Third column show segmentation results using proposed regression based CNN. Fourth column show ground truth images.

For our hand detection experiments we collected $13,116$ images containing hands from Signer dataset [2], Oxford hand data [1], ImageNet [38], INRIA pedestrian dataset [39], EgoHands dataset [3]. We also collected 153 images from Flickr, Pinterest, DevianArt and annotated them manually. Our total training dataset consisted of $13,269$ images with hands. We have taken only those images which contain hands because RCNN and FRCNN

both generate negative examples using hard data mining. The entire proposed framework has been trained using a Nvidia GTX 1080 GPU with 8 GB memory. Initial learning rate was 10^{-6} and reduced by factor of 10 after every 100 epochs. Total number of epochs were fixed at 500. In the second stage of the proposed algorithm we refine the results of hand detection algorithm by using skin segmentation. Sometimes false positives occur at non-skin regions also. Hence, we reject these detected bounding boxes wherein skin pixels constitute less than 30% of the total pixels. Our first approach uses a patch-based CNN for skin detection while the second proposed method trains an regression-based CNN for full image segmentation which is achieved with the aid of total variation prior [47]. Skin detection experiments are conducted on four public skin classification databases: Pratheepan dataset [42], Uchile dataset [43], IBTD dataset [44] and the HGR dataset [45] in order to report various quantitative performance metrics in Table 1. Details regarding the training and testing of the patch-based and regression based skin classifier are provided in section 3.3. Fig. 4 shows the comparative result of skin detector with Lei *et al*. [31]. The first and sixth columns show the images to be segmented. Second and seventh columns show the output of the method by Lei *et al*. [31]. Third and eighth columns show the output of the proposed patch-based skin detection algorithm. Fourth and ninth columns show the output of the proposed regression based skin detection algorithm. Fifth and last columns show the ground truth images. Note that our patch-based skin segmentation algorithm produces superior segmentation results. The performance of the proposed patch based approach of skin segmentation on the Uchile dataset and IBTD dataset is reported in Fig. 5. The first row of Fig. 5 corresponds to the results obtained on Uchile datset [43], while the second row corresponds to the results obtained on IBTD data [44].

Fig. 6 shows the results of skin segmentation using HGR dataset [45], with both the proposed skin segmentation algorithms. First column show original images. Second column show results generated by our patch-based skin detection algorithm (Fig. 3 (a)). Third column represent results generated by our regression based skin segmentation algorithm (Fig. 3 (b)). The ground truth masks are shown in fourth column. Note that the full image CNN classifier obtains results comparable in quality to the patch-based CNN. However, the regression based CNN is substantially fast as compared to the patch-based CNN. A detailed quantitative assessment of the true positive rate (TPR), true negative rate (TNR), F-score and accuracy obtained on the individual datasets of Pratheepan [42], Uchile [43], IBTD [44] and the HGR [45] datasets is reported in Table 1. Lei *et al*. [31] also reported their performance on HGR dataset [45], achieving an accuracy of 93% which is same as ours. The work in [31] has also reported results



(a) Oxford hand using FRCNN (b) EgoHand using FRCNN

(c) 5-igner using FRCNN (d) ICD using FRCNN

Figure 8. Precison-Recall curve depicting the advantage of using the proposed skin segmentation algorithm to aid the FRCNN-based hand detection method over Oxford hand [1], EgoHand [3], 5-signer [2] and our proposed ICD dataset.

on Pratheepan dataset [42], achieving an accuracy of 89% on RGB images whereas we have achieved an accuracy of 84%.

Fig. 7 shows the hand detection result obtained using the proposed algorithm (fourth row represent results of FRCNN). First two rows of Fig. 7 denote the result on the Oxford hand test dataset [1]. Third row denote the result on the 5-Signer dataset [2]. Fourth row denote the result on the Ego hand test dataset [3]. Last row shows the result on proposed ICD dataset. Note that semantic interpretation of hand gestures is beyond the scope of the work presented here and we focus only on detection/localization of hands in images. Table 2 shows the average precision on four different datasets using standard intersection over union (IoU) procedure. As demonstrated in Table 2 we observe substantial improvement over the baseline of only RCNN or FRCNN methods by using our skin segmentation algorithm. It can be observed that with proper training using a diverse and large dataset RCNN based hand detector with skin segmentation can give superior results. Fig. 8 shows the precision-recall curve obtained using FRCNN with and without proposed skin segmentation algorithm over Oxford hand [1], EgoHand [3], 5-Signer [2] and our proposed ICD datasets, respectively. Prior work [14] has shown that FRCNN's performance is hampered when objects to be detected are small in size. However, in the work of Chen *et al*. [12], showed that RCNN with proper context and region proposals it is possible to detect small objects in an image. In this work, we are able to conclude that RCNN with a small baseline architecture is capable of detecting hands in images despite their low spatial resolution with proper train-

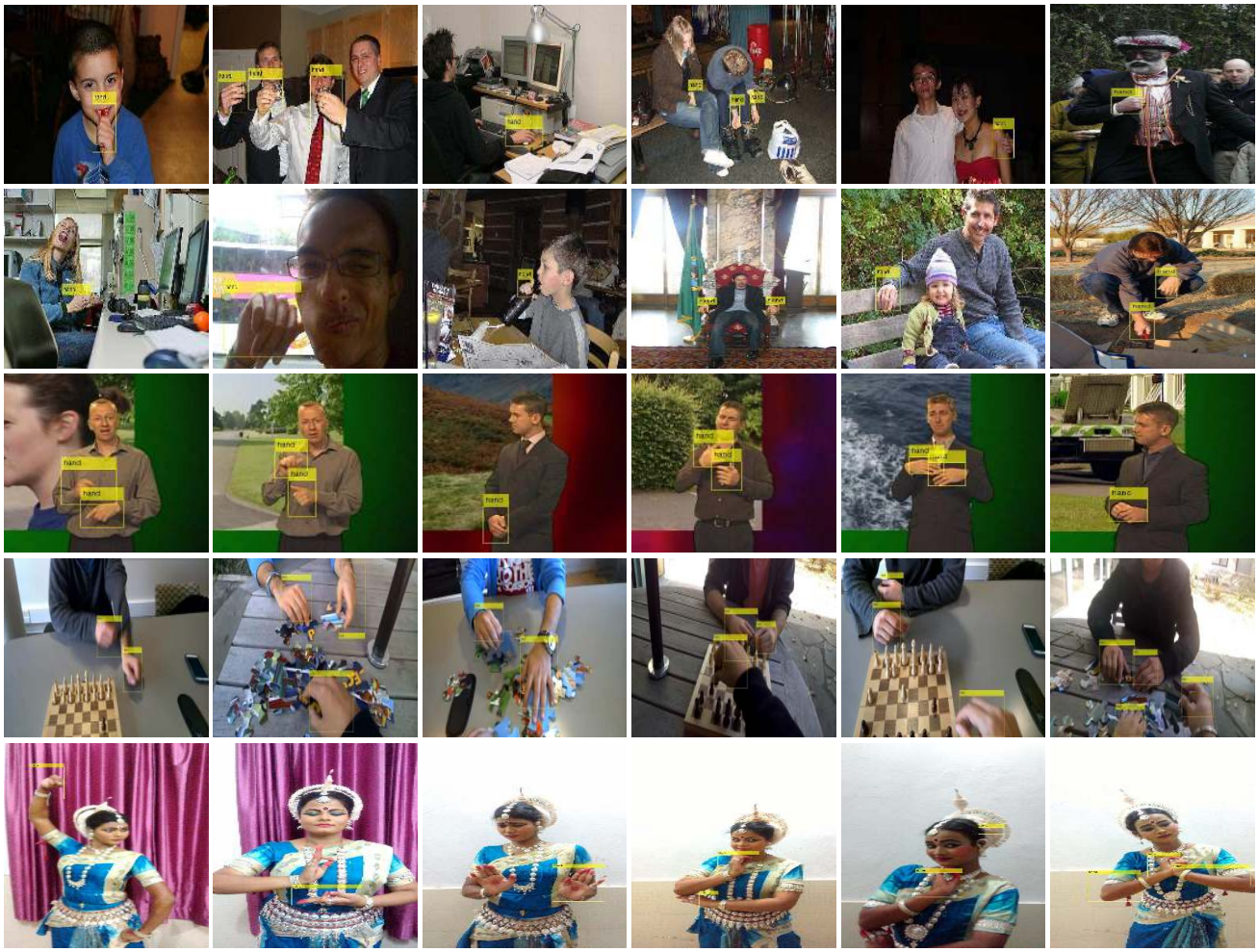


Figure 7. Performance of proposed hand detection algorithm on various datasets. Images in the first to second row show our results on the Oxford hand dataset [1], while in the third row we show the output of the proposed algorithm on 5-signer dataset [2]. The fourth row represents output of proposed approach on EgoHand dataset [3]. The last row shows output of proposed approach on some sample images from the proposed ICD dataset.

ing using a large dataset. This fact is established in the case of 5-signer dataset [2] where detection accuracy increases significantly compared with other methods. 5-signer dataset contains small images with tiny hands, where the spatial resolution of humans in the images are less. Here FRCNN performs poorly yielding high false positives. On the other hand EgoHands dataset [3] contains egocentric videos captured using Google glass, where images are large and hands are in foreground. Here FRCNN performs quite well. The proposed algorithm failed to detect hands particularly in the presence of heavy blurring, shadows and severe occlusions. We seek to investigate methods to overcome these issues.

5. Conclusion

Detection of hand in cluttered environment has remained a challenge due to complexities associated with it. Here we propose an algorithm for simultaneous hand region localization and skin detection. We show the performance of the proposed methodology on various public dataset such as, Oxford hand test dataset [1], 5-signer dataset [2] and EgoHand dataset [3], demonstrating its versatility. As a case study, we show the application of our algorithm on images from ICD affected by background clutter. Reduction in false positive at the output of hand localizer is further achieved using a deep learning based skin detection algorithm. As part of future work, we will continue to improve the robustness of our algorithm to poor illumination, shadows, blur etc.

References

[1] A. Mittal, A. Zisserman, and P. H. Torr, "Hand detection using multiple proposals.," in *British Machine Vision Conference*, pp. 1–11, Citeseer, 2011.

[2] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zis-

serman, "Long term arm and hand tracking for continuous sign language TV broadcasts," in *Proceedings of the British Machine Vision Conference*, pp. 1105–1114, BMVA Press, 2008.

[3] S. Bambach, S. Lee, D. Crandall, and C. Yu, "Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1949–1957, 2015.

[4] A. Farhadi and D. Forsyth, "Aligning ASL for statistical translation using a discriminative word model," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1471–1476, 2006.

[5] Y. Yang, C. Fermuller, Y. Li, and Y. Aloimonos, "Grasp type revisited: A modern perspective on a classical feature for vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 400–408, 2015.

[6] A. Mohanty, A. Ahmed, T. Goswami, A. Das, P. Vaishnavi, and R. R. Sahay, "Robust Pose Recognition Using Deep Learning," in *Proceedings of International Conference on Computer Vision and Image Processing*, pp. 93–105, Springer, 2017.

[7] A. Mallik, S. Chaudhury, and H. Ghosh, "Nrityakosha: Preserving the intangible heritage of Indian Classical Dance," *Journal on Computing and Cultural Heritage*, vol. 4, p. 11, 2011.

[8] S. Samanta, P. Purkait, and B. Chanda, "Indian classical dance classification by learning dance pose bases," in *IEEE Workshop on Applications of Computer Vision*, pp. 265–270, 2012.

[9] A. Mohanty, P. Vaishnavi, P. Jana, A. Majumdar, A. Ahmed, T. Goswami, and R. R. Sahay, "Nrityabodha: Towards understanding Indian Classical Dance using a deep learning approach," *Signal Processing: Image Communication*, pp. 529–548, 2016.

[10] A. Mohanty, S. S. Rambhatla, and R. R. Sahay, "Deep Gesture: Static Hand Gesture Recognition Using CNN," in *Proceedings of International Conference on Computer Vision and Image Processing*, pp. 449–461, Springer, 2017.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[12] C. Chen, M. Liu, O. Tuzel, and J. Xiao, "R-CNN for small object detection," in *Asian Conference on Computer Vision*, pp. 214–230, Springer, 2016.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[14] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*, pp. 354–370, Springer, 2016.

[15] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1014–1021, IEEE, 2009.

[16] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics*, vol. 33, p. 169, 2014.

[17] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt, "Fast and robust hand tracking using detection-guided optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3221, 2015.

[18] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using kinect.," in *British Machine Vision Conference*, p. 3, 2011.

[19] P. K. Pisharady, P. Vadakkepat, and A. P. Loh, "Attention based detection and recognition of hand postures against complex backgrounds," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 403–419, 2013.

[20] E. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 889–894, IEEE, 2004.

[21] M. Kölsch and M. Turk, "Robust Hand Detection," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 614–619, 2004.

[22] N. H. Do and K. Yanai, "Hand detection and tracking in videos for fine-grained action recognition," in *Asian Conference on Computer Vision*, pp. 19–34, Springer, 2014.

[23] T. Hoang Ngan 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 46–53, 2016.

[24] X. Deng, Y. Yuan, Y. Zhang, P. Tan, L. Chang, S. Yang, and H. Wang, "Joint Hand Detection and Rotation Estimation by Using CNN," *arXiv preprint arXiv:1612.02742*, 2016.

[25] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Proceedings of the Graphicon*, vol. 3, pp. 85–92, 2003.

[26] Z. Yu *et al.*, "Fast Gaussian mixture clustering for skin detection," in *IEEE International Conference on Image Processing*, pp. 2997–3000, 2006.

[27] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.

[28] M. Kawulok, "Fast propagation-based skin regions segmentation in color images," in *Proceedings of International Conference and Workshops Automatic Face and Gesture Recognition*, pp. 1–7, 2013.

[29] H. K. Al-Mohair, J. M. Saleh, and S. A. Suandi, "Hybrid human skin detection using neural network and K-means clustering technique," *Applied Soft Computing*, vol. 33, pp. 337 – 347, 2015.

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

[31] Y. Lei, W. Yuan, H. Wang, Y. Wenhu, and W. Bo, "A skin segmentation algorithm based on stacked autoencoders," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 740–749, 2017.

[32] H. Zuo, H. Fan, E. Blasch, and H. Ling, "Combining convolutional and recurrent neural networks for human skin detection," *IEEE Signal Processing Letters*, vol. 24, pp. 289–293, 2017.

[33] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advanced Neural Information Processing Systems*, pp. 1790–1798, 2014.

[34] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results." http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[36] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[37] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[38] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[40] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[41] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*, pp. 2843–2851, 2012.

[42] W. R. Tan, C. S. Chan, P. Yogarajah, and J. Condell, "A fusion approach for efficient human skin detection," *IEEE Transactions on Industrial Informatics*, vol. 8, pp. 138–147, 2012.

[43] J. Ruiz-del-Solar and R. Verschae, "Skin detection using neighborhood information," in *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pp. 463–468, 2004.

[44] Q. Zhu, C. Wu, K. Cheng, and Y. Wu, "An adaptive skin model and its application to objectionable image filtering," in *Proceedings of the 12th annual ACM International Conference on Multimedia*, pp. 56–63, 2004.

[45] M. Kawulok, J. Kawulok, and J. Nalepa, "Spatial-based skin detection using discriminative skin-presence features," *Pattern Recognition Letters*, vol. 41, pp. 3–13, 2014.

[46] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[47] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, pp. 259 – 268, 1992.

[48] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *European Conference on Computer Vision*, pp. 720–735, Springer, 2014.

[49] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.