IEEE Journal of Translational Engineering in Health and Medicine

Received 28 October 2018; revised 6 January 2019; accepted 7 January 2019.
Date of publication 14 January 2019; date of current version 29 January 2019.

Digital Object Identifier 10.1109/JTEHM.2019.2892970

In-Bed Pose Estimation: Deep Learning With Shallow Dataset

SHUANGJUN LIU, YU YIN, AND SARAH OSTADABBAS

Augmented Cognition Laboratory, Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02115, USA

CORRESPONDING AUTHOR: S. OSTADABBAS (ostadabbas@ece.neu.edu)

This work was supported by the National Science Foundation (NSF) Award 1755695.

ABSTRACT This paper presents a robust human posture and body parts detection method under a specific application scenario known as in-bed pose estimation. Although the human pose estimation for various computer vision (CV) applications has been studied extensively in the last few decades, the in-bed pose estimation using camera-based vision methods has been ignored by the CV community because it is assumed to be identical to the general purpose pose estimation problems. However, the in-bed pose estimation has its own specialized aspects and comes with specific challenges, including the notable differences in lighting conditions throughout the day and having pose distribution different from the common human surveillance viewpoint. In this paper, we demonstrate that these challenges significantly reduce the effectiveness of the existing general purpose pose estimation models. In order to address the lighting variation challenge, the infrared selective (IRS) image acquisition technique is proposed to provide uniform quality data under various lighting conditions. In addition, to deal with the unconventional pose perspective, a 2- end histogram of oriented gradient (HOG) rectification method is presented. The deep learning framework proves to be the most effective model in human pose estimation; however, the lack of large public dataset for in-bed poses prevents us from using a large network from scratch. In this paper, we explored the idea of employing a pre-trained convolutional neural network (CNN) model trained on large public datasets of general human poses and fine-tuning the model using our own shallow (limited in size and different in perspective and color) in-bed IRS dataset. We developed an IRS imaging system and collected IRS image data from several realistic life-size mannequins in a simulated hospital room environment. A pre-trained CNN called convolutional pose machine (CPM) was fine-tuned for in-bed pose estimation by re-training its specific intermediate layers. Using the HOG rectification method, the pose estimation performance of CPM improved significantly by 26.4% in the probability of correct key-point (PCK) criteria at PCK0.1 compared to the model without such rectification. Even testing with only well aligned in-bed pose images, our fine-tuned model still surpassed the traditionally tuned CNN by another 16.6% increase in pose estimation accuracy.

INDEX TERMS Convolutional neural network (CNN), convolutional pose machine (CPM), histogram of oriented gradient (HOG), in-bed pose estimation, infrared selective (IRS).

I. INTRODUCTION

Human in-bed pose and posture are important health-related metrics with potential values in many medical applications such as sleep monitoring. It is shown that sleeping pose affects the symptoms of many diseases such as sleep apnea [1], pressure ulcers [2], and even carpal tunnel syndrome [3]. Moreover, patients are usually required to maintain specific poses after certain surgeries to get a better recovery result and during pregnancy since certain sleeping postures can cause harm to pregnant women and the fetus. Therefore, long-term monitoring and automatically detecting in-bed poses are of critical interest in healthcare [4].

Currently, besides self-reporting by patients and visual inspection by the caregivers, in-bed pose estimation methods mainly rely on the use of pressure mapping systems. Pouyan *et al.* [5] extracted binary signatures from pressure images obtained from a commercial pressure mat and used a binary pattern matching technique for pose classification. The same group also introduced a Gaussian mixture model (GMM)-based clustering approach for concurrent pose classification and limb identification using pressure data [6]. Pictorial structure model of the body based on both appearance and spatial information was employed to localize the body parts within pressure images in [7]. The authors considered each part of the human body as a vertex in a tree and found how well the appearance of each body part matches its template as well as how far the body parts deviate from their expected respective locations. Finally, the best configuration of body parts was selected by minimizing the total cost. Although pressure mapping based methods are effective at localizing areas of increased pressure and even automatically classifying overall postures [6], the pressure sensing mats are expensive (>$10K) and require frequent maintenance. These obstacles have prevented pressure-based pose monitoring solutions from achieving large-scale popularity.

By contrast, camera-based vision methods for human pose estimation show great advantages including their low cost and ease of maintenance. General purpose human pose estimation has become an active area in computer vision and surveillance research [8], [9]. The methods and algorithms for pose estimation can be categorized into five categories: (i) The classical articulated pose estimation model, which is a pictorial structures model [10], [11]. It employs a tree-structured graphical model to constrain the kinematic relationship between body parts. However, it requires the person to be visible and is prone to errors such as double counting evidence. Some recent works have augmented this structure by embedding flexible mixture of parts (FMP) into the model [12], [13]. (ii) Hierarchical models, which represent the body part in different scale in a hierarchical tree structure, where parts in larger scale can help to localize small body parts [14], [15]. (iii) Non-tree models, which augment the tree structure with additional edges to capture the potential long range relationship between body parts [16], [17]. (iv) Sequential prediction frameworks, which learn the implicit spatial model directly from training process [18], [19]. (v) Deep neural network based method usually in a convolutional neural network (CNN) configuration [20], [21]. A recent CNN-based work, called convolutional pose machine (CPM) employed multi-stage CNN structures to estimate various human poses [22]. The CPM was tested on several well-recognized public datasets and promising results were obtained in estimating general purpose poses.

Although our work focuses on in-bed pose estimation, due to the use of camera for imaging instead of pressure mat, this line of research is categorized under camera-based human pose estimation [23]. It is sensible to assume that pre-trained models on existing datasets of various human poses should be able to address in-bed pose estimation as well. However, it turned out that when it comes to pose monitoring and estimation from individual in sleeping postures, there are significant distinctions between two problems. Since in-bed pose estimation is often based on a long-term monitoring scenario, there will be notable differences in lighting conditions throughout a day (with no light during sleep time), which makes it challenging to keep uniform image quality via classical methods. Moreover, if night vision technology is employed to address this challenge, the color information will be lost. Another difference is on the imaging angle, which for in-bed applications is overview (bird's-eye view) and subject overall orientation will have a different distribution from a common human surveillance viewpoint. For instance, it is possible that human appears upside-down in an overview image, but it is quite rare to see an upside-down human from a side viewpoint. In addition, the similarity between the background (bed sheets) and foreground (human clothing) is magnified in in-bed applications. To the extent of our knowledge, there is no existing work that has addressed these issues. In addition, no specific in-bed human pose dataset has been released to demonstrate and compare the possibilities of employing existing models to serve for in-bed pose estimation.

In this paper, we address the aforementioned challenges and make the following contributions: (i) Developing an infrared selective (IRS) image acquisition method to provide stable quality images under significant illumination variations between day and night. (ii) Improving the pose estimation performance of a pre-trained CPM model from side viewpoint dataset by adding a 2-end histogram of oriented gradient (HOG) orientation rectification method, which improved performance of the existing model over 26.4% on our dataset. (iii) Proposing a fine-tuning strategy for intermediate layers of CPM, which has surpassed the classical model accuracy by 16.6% in detecting in-bed human poses. (iv) Considering practical cases and embedded implementation requirements (e.g. to preserve privacy), an on-demand trigger estimation framework is proposed to reduce computational cost. (v) Building an in-bed human pose dataset with annotation from several realistic life-size mannequins with clothing differing in color and texture in a simulated hospital room via proposed IRS system. The dataset also includes a semi-automated body part annotation tool.

II. METHODS

Most human pose estimation works exclusively address the pose estimation when a human-contained bounding box is given. Instead, our work presents a system level automatic pipeline, which extracts information directly from raw video sequence inputs, while containing all the related preprocessing parts. An overview of our system is presented in Fig. 1. In Section II-A, we first introduce the IRS acquisition method to address the lighting condition variation issue during day and night. Then in Section II-B, we suggest the n-end HOG rectification method to handle the unusual pose distribution from overview angle. Section II-C describes on-demand trigger mechanism, which provides on-demand pose estimation. Finally in Section II-D, an example of general purpose pose estimation models based on deep neural networks is repurposed for in-bed pose estimation.

In particular, we used convolutional pose machine (CPM) as a pre-trained CNN [22]. We also employed a high performance pictorial structure oriented method, called flexible

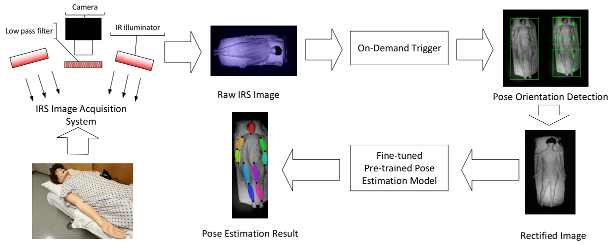


FIGURE 1. Overview of our in-bed human pose estimation system. In-bed images are collected from the proposed IRS system, then based on the system user's demand, pose estimation routine is triggered. Raw images are first preprocessed by a rectification method to get rectified and then fed into a fine-tuned pre-trained pose estimation model to produce pose estimation results.

mixture of parts (FMP) during experimental analysis for estimation accuracy comparison. The rational behind using CPM and FMP is that these two algorithms represent two typical frameworks for pose estimation, one is based on deep learning, and the other is based on the pictorial structure, one of the most classical pose estimation models. In the case of CPM, to deal with the high-volume data requirement issue, a fine-tuning strategy is also suggested, which is based on training only a few specific layers rather than retraining the whole network. Therefore, we were able to evaluate the pose estimation accuracy of both models using our "shallow" in-bed dataset. We chose the term "shallow" to indicate the differences between our IRS in-bed pose data and publicly available general purpose pose data. These differences include limited size of dataset, lack of color information, and irregular orientation and poses that one may take while being in bed.

A. INFRARED SELECTIVE (IRS) IMAGE ACQUISITION SYSTEM

Available datasets for pose estimation are collected under well illuminated environment and the subjects are visible enough to be captured by regular cameras. However, in-bed pose estimation requires to be conducted not only during daytime but also during night time, which means to be functional under a totally dark environment. Night vision cameras are commercially available, however the resultant images are significantly different than images from regular cameras, which raises great challenges to the pose estimation methods.

1) IRS IMAGING SYSTEM IMPLEMENTATION

To address this issue, we developed an IRS image acquisition method, which provides stable quality image under huge illumination variation between day and night. The IRS imaging benefits from the difference between human vision and charge coupled device (CCD) cameras, which show different sensitivity to the same spectrum. CCD cameras capture larger range of spectrum beyond human capability, which makes the visualization possible under dark environment to human. Our system avoids the visible light spectrum, which ranges from 400nm to 700nm, and selects the infrared spectrum ranging from 700nm to 1mm. Different from traditional night vision cameras, which only employ the IR light to enhance the lighting condition during night, we filter out the whole visible light spectrum in order to make the image quality invariant to lighting conditions, thus making robust performance estimation possible. The IRS imaging process and the hardware implementation are shown in Fig. 2a and Fig. 2b, respectively.

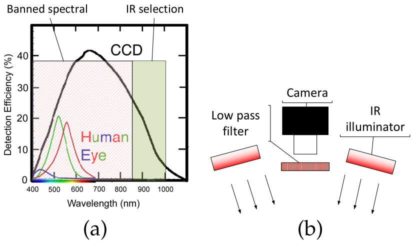


FIGURE 2. Infrared selective (IRS) acquisition method, (a) IRS spectrum [24], (b) IRS hardware diagram.

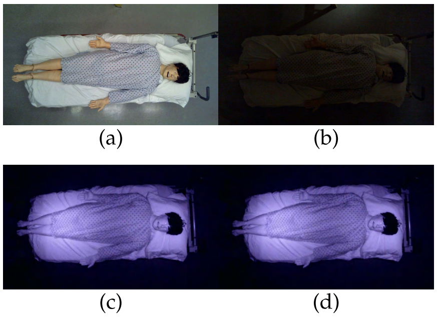


FIGURE 3. Image captured by normal webcam (a) with light on and (b) with light off. The same images captured by IRS imaging system (c) with light on and (d) with light off.

Fig. 3 shows the images captured by IRS system and a comparing pair from a normal webcam. It clearly demonstrates that IRS system provides stable image quality under huge illumination variations. This makes the night monitoring possible without disturbing subjects during sleep. Another advantage of using IRS imaging is it produces high contrast foreground and background, which makes the segmentation easier. In terms of the safety of our IRS imaging system, it is proved that IR light is a non-ionizing radiation, which has insufficient energy to produce any type of damage to human tissue. Most common effect generated by IR is heating [25]. In our case, the visualization radiation is far below the dangerous level due to its low power density.

2) NEW CHALLENGES FROM IRS

IRS provides a way for stable image acquisition for day long monitoring, however the use of the IRS setup results in new challenges. Fig. 3a and Fig. 3b show the images captured from regular cameras with light on and off, respectively. Fig. 3c and Fig. 3d show the images captured by our IRS system the under same conditions. As you can see, the color information is totally lost from this process and the purple color in the image Fig. 3d is resulted form filtering process. To employ existing pose estimation models, we assumed this false color as gray intensity information and replicated this to three channels, what is the standard input format for most pose estimation models. It is shown that the color information is not trivial in pose estimation and its effect on pose estimation accuracy is given in Section III-C.

Moreover, in-bed pose distribution under overview angle will be different from most public datasets collected from regular side viewpoint. Subjects can be commonly upside-down in an overview image because of their in-bed orientation, which is a rare case from a side viewpoint. This difference is also not trivial during estimation process which is shown in both models under our test (Section III-D). One example is shown in Fig. 4 where we employed a pre-trained CPM model to test the pose estimation accuracy of same image in our dataset but with different orientations [22]. The result showed notable differences between the image with portrait orientation and the inverse one.

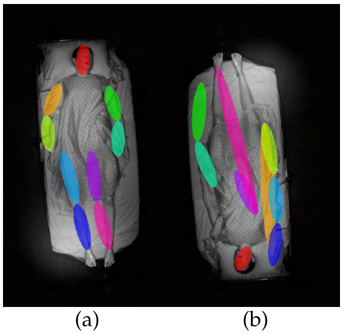


FIGURE 4. Convolutional pose machine (CPM) detection result of same image with different orientations, sleeping position with the head (a) in the top of the image, and (b) in the bottom of the image.

B. IN-BED POSE ORIENTATION DETECTION

Classically, subject orientation problem during pose estimation is handled by data augmentation technique [26], which artificially enlarge the pose dataset using label-preserving transformations [27]. However, this technique often results in an extensive re-training computational time. Assuming that the chosen model is capable of capturing pose information from side view, to utilize the model trained on a large dataset, we present an orientation rectification method to re-align the image to a similar position to training set. In order to employ the pre-trained pose estimation models directly in our application, here we present an n-end HOG rectification method to minimize the image misalignment.

1) BOUNDING BOX DETECTION

We assume under usual home/hospital settings, beds are aligned with one of the walls of the room. In the case of cuboid rooms, this will result in four general categories of in-bed orientations. Suppose the camera is correctly setup to capture images with major axes approximately parallel to the wall orientations. We define these four general in-bed orientations as north, east, south, and west $\{N, E, S, W\}$. The first step to find the general in-bed orientation, is locating the human-contained bounding box in the image. This could be a computationally intensive process over multi-scale extensive search for a common vision task. However in our case, due to IRS imaging, foreground appears with high contrast from the background, which makes the segmentation a straightforward threshold-based algorithm. We further noticed that under IRS, the foreground shows visible edges, in which the bounding box can also be extracted from a classical edge detection algorithm using the 'Sobel' operator. The results of applying these two methods are shown in Fig. 5. When there is no disturbance at surroundings, the edge based bounding box extraction will be more accurate to locate the boundaries. However, threshold based method will be more robust to the noise and a multiple scale search can be employed to improve the results. Information associated with a bounding box are $\mathcal{B} = \{\mathcal{B}_{x_c}, \mathcal{B}_{y_c}, \mathcal{B}_w, \mathcal{B}_h\}$, where $\mathcal{B}_{x_c}, \mathcal{B}_{y_c}$ represent the coordinate of the up-left corner, and $\mathcal{B}_w, \mathcal{B}_h$ represent the width and height of the bounding box, respectively. From the bounding box width and height ratio, in-bed orientations are first categorized into horizontal and vertical ones. To further rectify the orientation, we apply an n-end HOG rectification method as described below.

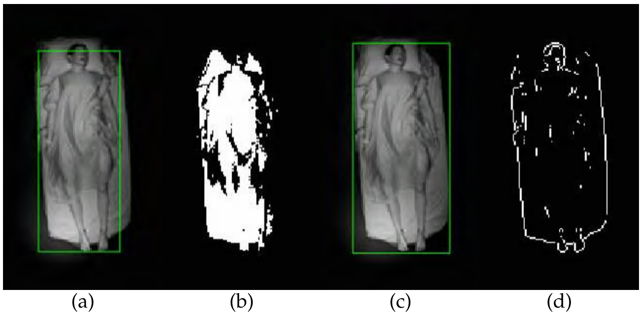


FIGURE 5. Bounding box extraction using (a) threshold method, (b) binary image from thresholding, (c) the edge detection method, (d) edge detected with the 'Sobel' operator.

2) N-END HOG POSE RECTIFICATION METHOD

HOG features were first employed for pedestrian detection [28], which captured the localized portion features by estimating the local gradient orientation statistics. These features show the benefit of being invariant to the geometric and photometric transformations. Since all the horizontally-orientated images can be detected based on the $\mathcal{B}_w/\mathcal{B}_h$ and rotated back into vertical ones, here the classification is between upside-down images vs. portrait ones, all in vertical cases. As upper and lower body parts show clear differences in their overall geometry, we captured information from large scale patches instead of small grids. Therefore, unlike extracting HOG features on dense grid, we only extracted HOG features on sparse locations. To form HOG features in this way, two information is needed. One is HOG descriptor parameters and the other is *interest points'* locations, where HOG operator to be applied at.

For HOG parameters, we employed a 2 × 2 cell structure for each block to capture overall information. The block size is determined by the size of the estimated bounding box as $l_{block} = \min(\bar{\mathcal{B}_w}, \bar{\mathcal{B}_h})$, where $\bar{\mathcal{B}_w}$ and $\bar{\mathcal{B}_h}$ represent the average of width and height of the bounding boxes in images from our IRS dataset, respectively. In practice, the average bounding box information can be achieved by a short period of initial monitoring. Consequently, the cell size is $l_{cell} = l_{block}/2$. For long-term in-bed monitoring applications, once set up, the scale information would stay the same during the monitoring time.

For interesting points' locations, we assumed $\mathcal{B}_w < \mathcal{B}_h$ and the coordination of the first and last interesting points are given as:

$$\begin{aligned} C_{hog}(1) &= (\mathcal{B}_{x_c} + l_{block}/2, \mathcal{B}_{y_c} + l_{block}/2) \\ C_{hog}(n) &= (\mathcal{B}_{x_c} + l_{block}, \mathcal{B}_{y_c} + \mathcal{B}_h - l_{block}/2) \end{aligned} \qquad (1)$$

where n is an integer stands for the total number of interesting points, and $C_{hog}(n)$ is the center of the n-th HOG descriptors. Once the two end interesting point coordination are achieved, other interesting point can be extracted from linear interpolation from them. In our case, we chose $n = 2$ and 2-end HOG features are generated as shown in Fig. 6(a)(b). Extracted HOG features from the interest points are cascaded in top to bottom order into HOG feature vector, f_{2e}. A support vector machine (SVM) model is then employed as binary classifier on extracted HOG features to give prediction result from orientation categories of $\{N, \neq N\}$. We assign this result to an indicator bit_N. Another indicator bit_H comes from the bounding box to show if the subject is horizontal or vertical. The final orientation is decoded from the encoding rule that assigns different bit_H and bit_N to the four orientation categories. This process automatically forms a two-layer decision tree as shown in Algorithm. 1.

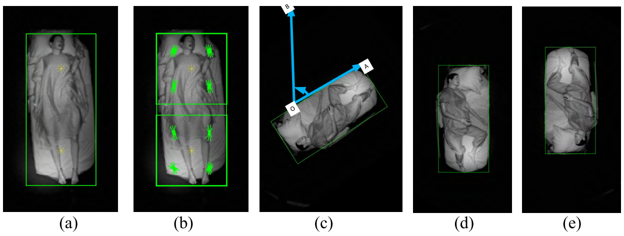


FIGURE 6. 2-end HOG feature extraction and random lying direction rectification by minimal-area encasing rectangle detection, (a) candidate HOG center locations, (b) HOG features extracted from the candidate locations, (c) rectify longer edge to vertical direction, (d) correct rectification, (e) upside down rectification.

3) GENERAL ORIENTATION RECTIFICATION

In the cases that the bed is in a random orientation layout, it is not possible to represent the general lying direction in only 4 categories. Instead, we first employ the minimal-area encasing rectangle algorithm to find a tight bounding box [29]. The principle direction of the bounding box is found from the longer edge and then the image is aligned to the vertical direction as shown in Fig. 6(c)(d). However, this algorithm can only rectify the image to a general vertical layout, yet the upside down case still exists as demonstrated in Fig. 6(e). After this initial rectification, the up or down decision comes out to be a binary classification which falls back to our N-end HOG pose rectification method.

Algorithm 1 2-End HOG Rectification Method

```
Input: Image $I$
Result: General in-bed orientation from $\{N, E, S, W\}$,
        reclined portrait image
Initialization;
Edge detection on $I$, output $I_{bw}$;
Bounding box extraction from $I_{bw}$;
Calculate $\mathcal{B}_w/\mathcal{B}_h$;
if $\mathcal{B}_w/\mathcal{B}_h > 1$ then
    Subject has a horizontal in-bed orientation,
    $bit_H = 1$; Rotate original image to vertical in-bed
    orientation $I$ = Rotate($I$, -90$^o$);
else
    Subject has a vertical in-bed orientation, $bit_H = 0$;
end
2-end HOG extraction to form vector $f_{2e}$;
Get orientation from the SVM classification;
if $N$ then
    $bit_N = 1$;
else
    $bit_N = 0$;
end
Predict in-bed orientation from encoding table.
Rectify image $I$ to $N$ category.
```

C. ON-DEMAND TRIGGER FOR POSE ESTIMATION

Typical applications of in-bed pose estimation are overnight sleep monitoring and long-term monitoring of bed-bound patients. In these cases, human on the bed is often less physically active or even totally immobile. Therefore, we can reasonably hold the following hypothesis: "when the scene is stable, the human pose stays the same." This means that we only need to estimate the pose after each variation in the scene rather than continuously process the video, frame by frame. In this scenario, we propose an on-demand estimation trigger scheme to reduce the computational and power cost of our pose estimation algorithm. This power efficiency is crucial for patient's privacy reasons, since it enables us to build an *in-situ* embedded pose processing system rather than sending all raw videos of the patient during his/her sleep to a base-station for further processing.

Since this process is conducted in an indoor environment, a threshold-based method is used to detect foreground variations. The pose estimation process then is triggered when the scene recover from the variation. Suppose the current state is $\mathcal{S}_{cur} \in \{0, 1\}$ and previous state is $\mathcal{S}_{pre}$, where 1 stands for a dynamic scene and 0 stands for a static one. To get the state value, we make a difference operation by adjacent video frames. If this difference is greater than a threshold, it is assumed to be a dynamic frame, otherwise a static one. When in-bed pose changes, it could be caused by the subject herself or the caregiver. Based on the speed of repositioning, the process possibly contains piece-wise static periods. To suppress this false static state, we employed a backward window $\mathcal{W}_{bf}$ of size N_{bf} to filter the raw state result. The filtered state $\hat{\mathcal{S}}_{cur}$ is 0 only when all states in the backward window show static states, otherwise it is 1. This operation as shown in Algorithm. 2 is designed to favor dynamic states and guarantees a gap between static states if short disturbance occurs between them.

Algorithm 2 On-Demand Pose Estimation Trigger

```
Input: Video stream $I$
Result: Trigger pose estimation process
Initialization;
while new frame do
    Get difference of adjacent frames
    Update $\mathcal{W}_{bf}$
    if $max\ (\mathcal{W}_{bf}) == 1$ then
        $\hat{\mathcal{S}}_{cur} = 1$
    else
        $\hat{\mathcal{S}}_{cur} = 0$
    end
    if $\hat{\mathcal{S}}_{cur} - \hat{\mathcal{S}}_{pre} < 0$ then
        Trigger pose estimation process
    end
    $\hat{\mathcal{S}}_{pre} = \hat{\mathcal{S}}_{cur}$
end
```

D. FINE-TUNING CPM FOR IN-BED POSE ESTIMATION

Even with larger orientation possibilities and full loss of color information, in-bed human poses still share great similarities with ones taken from side views. We believe a well-trained general purpose pose estimation model is still able to capture body parts' features and kinematic constraints between them. In this work, a recent CNN-based pose estimation approach, called convolutional pose machine (CPM) is employed as a pre-trained pose estimation model [22]. CPM employs multi-stage structure to estimate human pose, in which each stage is a multi-layer CNN. Each stage takes in not only the image features, but also previous stage's belief map results as input. The final stage outputs the final estimation results, which are the 14 key joints' coordinates in image domain that include left and right (L/R) ankles, L/R knees, L/R hips, L/R wrists, L/R elbows, L/R shoulders, top of the head and neck. In the original work that introduced CPM [22], CPM with 6 stages has shown promising estimation results on large scale dataset such as MPII [30], LSP [31] and FLIC [32]. However, for a new query image, manual intervention was still required to indicate the exact bounding box of the human in the scene.

Due to the IRS imaging system and 2-end HOG method, our proposed method is able to accurately locate the human-contained bounding box and efficiently rectify the image orientation, which drastically save the cost of extensive search across multi-scale image pyramid. These properties provide a more efficient way to directly apply pre-trained CNN model on an in-bed pose dataset. Furthermore, in order to adapt to the input layer dimension of the pre-trained model, each input image is amplified into three channels, which share the same intensity value.

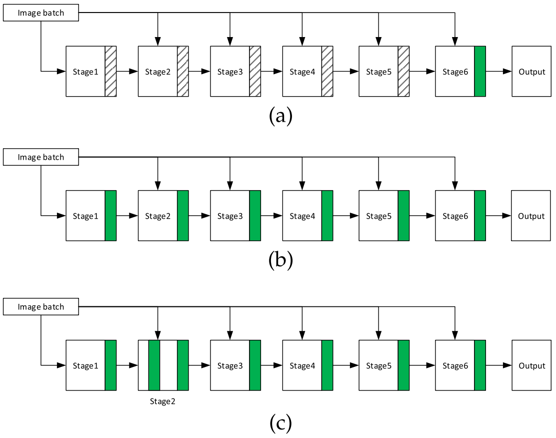


FIGURE 7. Fine-tuning configurations: green block indicates the layers for training (a) MANNE-S6: fine-tuning only the last layer before output, (b) MANNE-AS: fine-tuning last layers of all stages, (c) MANNE-AS-S2C3-#: fine-tuning last layers of all stages as well as the 3rd convolutional layer of stage 2 with # number of iterations.

When available dataset is limited in size, such as our IRS in-bed pose data, it is a golden rule to fine-tune the deep neural network model with only fully connected layers or the last layer [33]. However, based on the multi-stage configuration of the CPM, other fine-tuning approaches can also be applied. In this work, three fine-tuning strategies are proposed, which are illustrated in Fig. 7. First strategy, called MANNE-S6 takes the convention to train the very last layer before output or fully connected layer [33]. Due to the CPM's special configuration with multiple stages, in the second configuration, we train the last layer of each stage, which is called MANNE-AS. We also notice that there is a shared layer in CPM structure, which is the 3rd convolutional layer located in stage 2. Therefore, in third strategy, we further put this layer under training. This strategy is called MANNE-AS-S2C3-200, when it is trained with 200 iterations, and is called MANNE-AS-S2C3-2000, when it is trained with 2000 iterations. More iterations will enhance the probability of capturing more training samples' patterns and tuning the model weights to more representative values. Without any fine-tuning, the pre-trained CPM model using MPII and LSP dataset is called MPII-LSP.

To compare the effectiveness of the deep learning against other non-deep models when our IRS in-bed pose dataset is used, we employed a recently proposed pictorial structure oriented model with flexible mixtures of parts (FMP) [34], which has also shown great general purpose pose estimation performance on small scale human pose datasets such as PARSE [12] and BUFFY [13].

III. EXPERIMENTAL SETUP AND ANALYSIS

A. BUILDING AN IN-BED POSE DATASET

Although there are several public human pose datasets available such as MPII [32], LSP [30], FLIC [31], Buffy [35], they are all mainly from scenes such as sports, TV shows, and other daily activities. None of them provides any specific in-bed poses. To fill this gap, we crafted an image acquisition system based on an IRS configuration and collected IRS data from one male and one female realistic life-size mannequins in a simulated hospital room at the Health Science Department of Northeastern University.

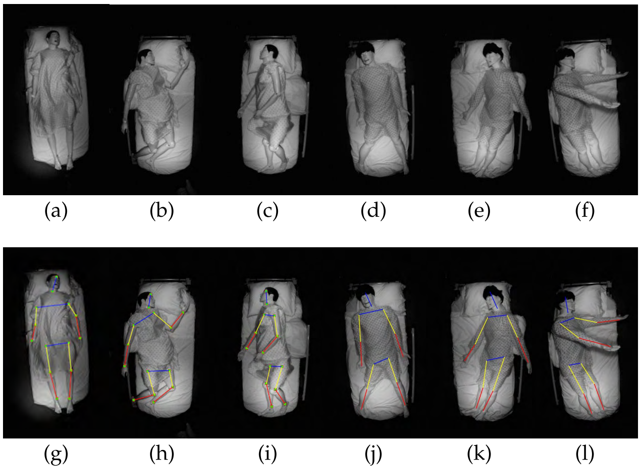


FIGURE 8. Mannequin pose dataset collected in a simulated hospital room. First row images show the raw image collected via IRS system. Second row shows manually annotation pose results of first row images.

Using mannequins gave us the option to collect images from different in-bed postures (supine, left side lying, and right side lying) by changing their poses with high granularity. Limited by the number of available mannequins, we collected data from the mannequins with different clothes, mainly different color/texture hospital gowns. We totally collected 419 poses, some of which are shown in Fig. 8. For comparison purpose, a color edition in-bed pose dataset is also established under the same setting but with an overview normal webcam. Some samples of the colored in-bed pose dataset is shown in Fig. 9. A semi-automated tool for human pose annotation is designed in MATLAB, in which the joint indices follow the LSP convention [30], as shown in Fig. 8(g) to (l). The GUI of this tool (GitHub code available) provides the convenience to label join locations and visibility in a semi-automated way.

B. POSE ESTIMATION PERFORMANCE MEASURE

Throughout the result section, probability of correct keypoint (PCK) criteria is employed for pose estimation performance evaluation, which is the measure of joint localization accuracy [34]. The distance between the estimated joint position and the ground-truth position is compared against a threshold defined as fraction of the person's torso length, where torso length is defined as the distance between person's left shoulder and right hip [30]. For instance, PCK0.1 metric means the estimation is correct when the distance between the estimated joint position and the ground-truth position is less than 10% of the person's torso length. This is usually considered a high precision regime. For the experimental analysis, we illustrate the pose estimation results of different models for the body part categories of total (all body parts), hip, knee, ankle, head, shoulder, elbow, and wrist by combining the estimation results of left and right corresponding limbs.

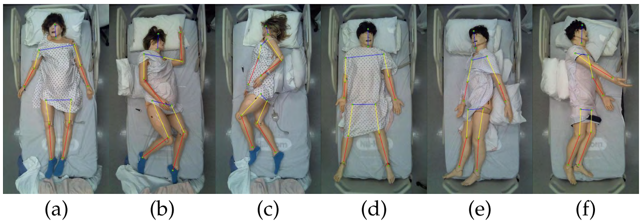
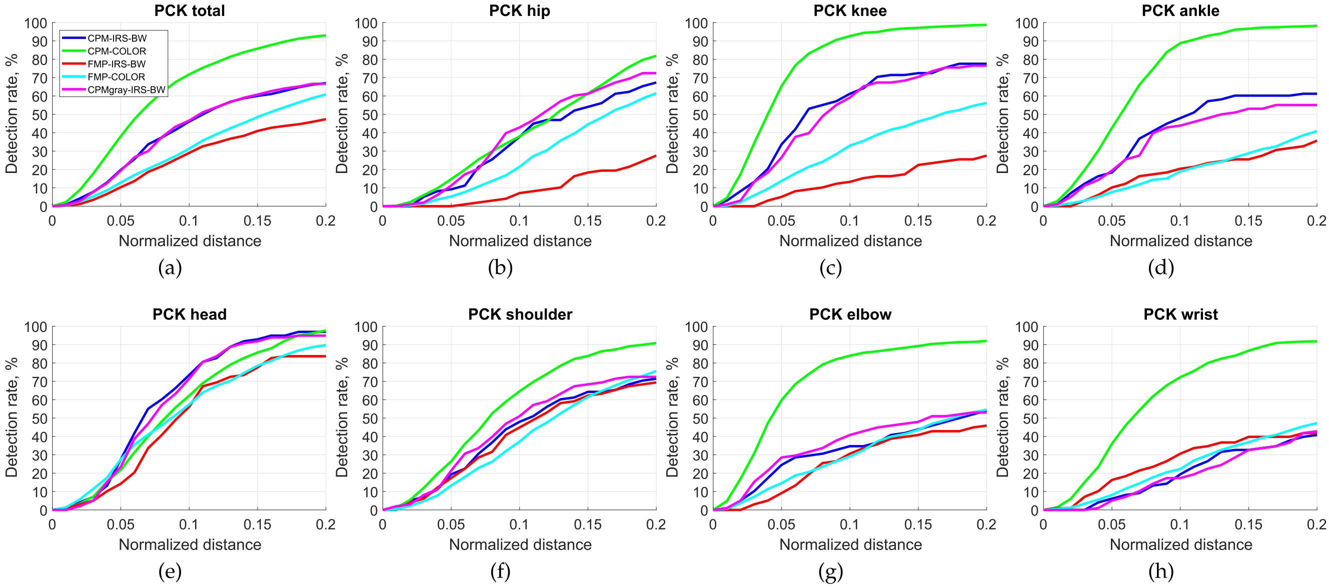


FIGURE 9. Annotated mannequin pose samples collected via webcam system in a simulated hospital room.

C. DOES COLOR INFORMATION MATTER?

One obvious difference between the IRS in-bed dataset and the publicly available general purpose datasets is the loss of color information. To investigate the influence of color loss on pre-trained models, we employed a pre-trained CPM model (trained on MPII and LSP dataset) to estimate poses of our mannequin dataset collected using IRS imaging and a normal webcam, respectively. To exclude the influence of unusual orientation, we only compare these two datasets from portrait image angle. To show the general effect of color loss on other pre-trained models, an FMP model is also evaluated under the same setting.

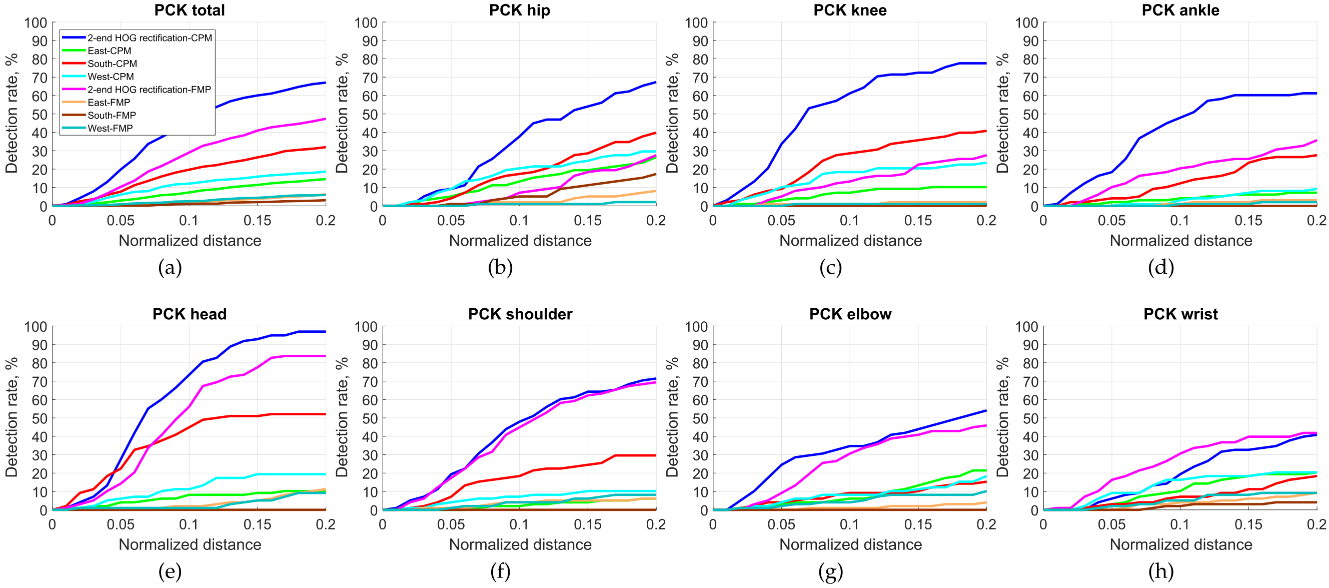
As shown in Fig. 10, both pre-trained models show better result on colored dataset than its black and white (BW) counterpart. Improvements from color information bring much more improvement in CPM than FMP. It shows color information is important in both models and is more helpful in CNN framework. In overall performance, the CPM gives better result. Even its performance on BW edition surpasses the FMP color edition. These results once more clarify our rationale for choosing a CNN based framework as the main pre-trained model for in-bed pose estimation. Another speculation is that whether or not the IRS dataset is just a gray-scale version of RGB camera. To clarify this question, we trained the CPM from scratch with gray-scale versions of the MPII-LSP dataset with 70000 iterations. The test results for this model are also presented in Fig. 10, which shows that the CPM model trained on gray-scale MPII-LSP dataset shows similar performance to the pre-trained CPM on original MPII-LSP dataset, yet not as good as the model fine-tuned



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 10. **Quantitative posture estimation result with different IRS mannequin black and white dataset and webcam mannequin color dataset via MPII-LSP pre-trained CPM model, CPM model pre-trained on gray-scale version of MPII-LSP and FMP model. All images are in the portrait view similar to the general purpose view point common in the MPII and LSP datasets to exclude the in-bed orientation factor.**



(a) (b) (c) (d)

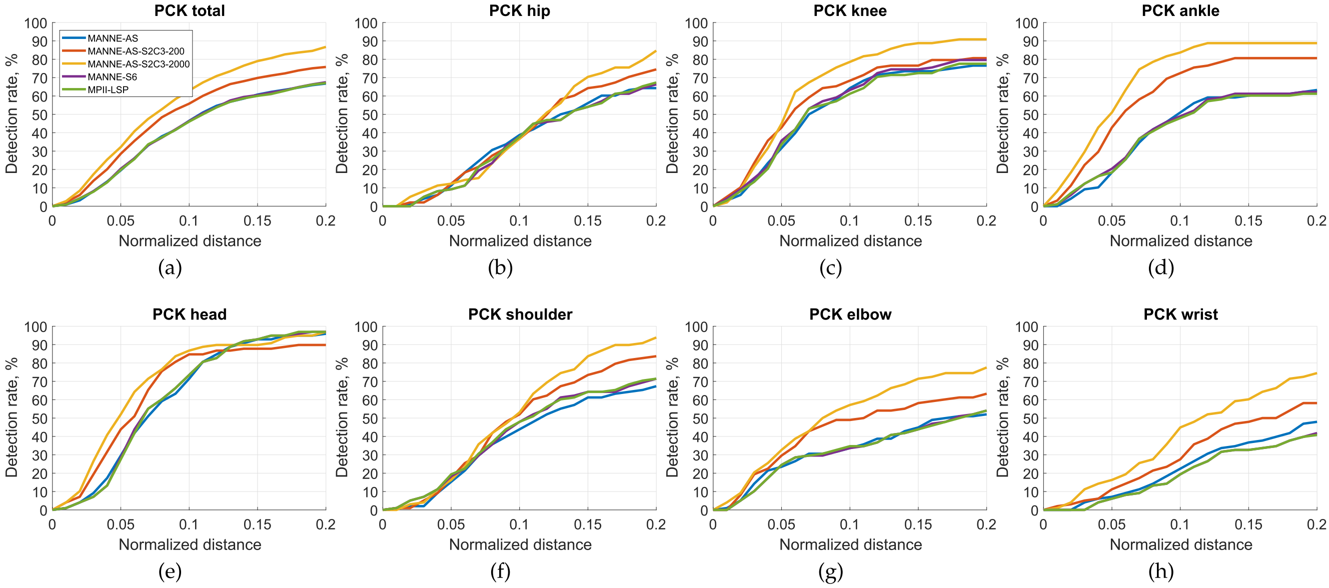
(e) (f) (g) (h)

FIGURE 11. **Quantitative posture estimation result via MPII-LSP pre-trained CPM model on different in-bed orientation images as well as their 2-end HOG rectified version.**

on the IRS version (see Fig. 12). Such difference in pose estimation accuracy could come from the facts that: (1) the pose distribution is different between the humans lying on bed and humans during their daily normal activities, and/or (2) although presented as an intensity image, the IRS image distribution is different from the gray-scale version of its RGB image counterpart. This could caused by the differences in the physics of imaging when using IR vs. visible light.

D. UNUSUAL ORIENTATION HANDLING

To handle the unusual orientation resulted from overview camera angle, a 2-end HOG rectification method was employed. We evaluated the effectiveness of the process in two phases. In the first phase, we tested the accuracy of 2-end HOG orientation detection and rectification method. We augmented our IRS in-bed pose dataset by synthesizing and adding several in-bed orientations in $\{N, E, S, W\}$ general categories for each image. bit_H and bit_N were obtained for



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 12. Quantitative posture estimation result with different fine-tuning strategies as shown in Fig. 7. MPII-LSP stands for the original pre-trained CPM model from MPII and LSP dataset. MANNE-S6 stands for model only fine turned on the last layer of final stage. MANNE-AS stands for the model fine turned with last layers of all stages. MANNE-AS-S2C3-200 and MANNE-AS-S2C3-2000 stand for the model fine turned with last layers of all stages and the 3rd convolutional layer in stage 2 after 200 and 2000 iterations, respectively.

a given image in dataset based on $\mathcal{B}_w/\mathcal{B}_h$ and the results of the SVM classifier as explained in Algorithm. 1. Using a 10-fold cross validation scheme, 99% accuracy in the general orientation detection was achieved. To further evaluate the performance of our proposed method in estimating general lying orientations, we regenerated our test dataset with all possible lying orientations and tested our method on them. As the bounding box realignment can be easily realigned the subject to the near vertical layout, the error only comes from the 2-end HOG alignment process and it was 4% (i.e. 2 cases of alignment error out of 49 random orientation samples).

In the second phase, to further evaluate the pose estimation performance on unusually oriented images (belonging to $\{E, S, W\}$ categories) vs. rectified images (all re-aligned to $\{N\}$ position), we employed a pre-trained CPM model from MPII [30] and LSP [31] dataset and also the flexible of parts (FMP) model [34] as our pose estimation models. We then divided our IRS in-bed pose dataset into two subsets: 370 images for training and 49 for test and used PCK metric for performance evaluation, as suggested in [31]. The estimation performance on images belonging to $\{E, S, W\}$ in-bed orientations categories is compared to the portrait images after 2-end HOG rectification and the results are shown in Fig. 11. These results demonstrate that in-bed orientation significantly affects the pose estimation accuracy and our proposed 2-end HOG rectification method boosts the estimation performance by a large margin for both CNN based and pictorial structure based models. Our method shows promising to act as a generic purpose tool to enhance the performance of pre-trained models for in-bed case.

E. FINE-TUNING OF A DEEP MODEL

To further improve the performance of our chosen neural network model, the MPII-LSP pre-trained CPM, we performed fine-tuning with different configurations as shown in Fig. 7. We trained all three proposed configurations with small iteration (=200) with batch size of 16. Fig. 12 shows the performance of CPM model after different fine-tuning strategies compared to the original pre-trained CPM. Our third fine-tuning configuration, MANNE-AS-S2C3-200, showed the highest estimation performance when compared to the traditional fine-tuning approach. When the iteration number was increased to 2000 for the third configuration, it further improved the estimation results.

In 200 iteration training test, MANNE-S6 does not show improvement over original model, however our proposed strategy, MANNE-AS-S2C3-200 shows clear improvement after 200 iterations in all body parts except the head part. MANNE-AS-S2C3-200 model shows improvement at PCK0.1, however falls behind at PCK0.2. It means the model either gives accurate answer for the head location or drifts far away from the correct location. This may come from the fact that the head part depends more on local image features. This drawback however is resolved after more iterations. Our final fine-tuned model MANNE-AS-S2C3-2000 surpassed the original pre-trained CPM MPII-LSP and also the traditional fine-tuned model MANNE-S6 by nearly 20% at PCK0.2 criterion. One sample of an estimation belief map is shown in Fig. 13. We hypothesize that the success of MANNE-AS-S2C3-2000 is due to the fact that its first 3 layers are reused in all the following stages, which means

TABLE 1. Pose estimation accuracy in PCK0.2 standard using FMP, pre-trained CPM, and our fine-tuned CPM model.

Models	Total	Ankle	Knee	Hip	wrist	Elbow	Shoulder	Head
FMP	60.2	52.0	42.9	51.0	49.0	62.2	78.6	85.7
MPII-LSP CPM	67.1	61.2	77.6	67.3	40.8	54.1	71.4	**96.9**
MANNE-AS-S2C3-2000	**86.7**	**88.8**	**90.8**	**84.7**	**74.5**	**77.6**	**93.9**	**96.9**

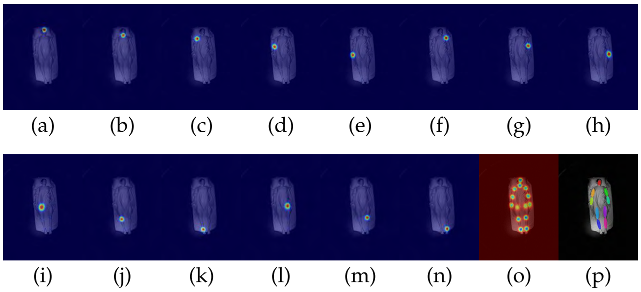


FIGURE 13. Human pose estimation result with our MANNE-AS-S2C3-2000 model. (a)-(n) estimated belief map of head, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, and left ankle, (o) background belief map, and (p) pose visualization.

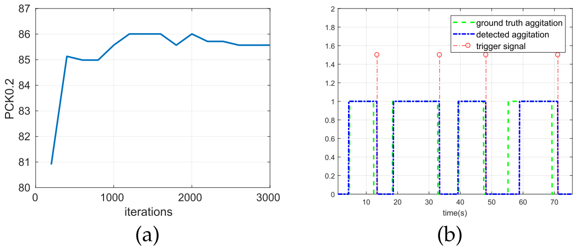


FIGURE 14. PCK0.2 on test set against training iterations and on-demand triggering. (a) PCK0.2 on test set performance against training iterations, (b) on-demand state estimation trigger result.

it has larger influence on the final output, and the outcome performances validate this hypothesis.

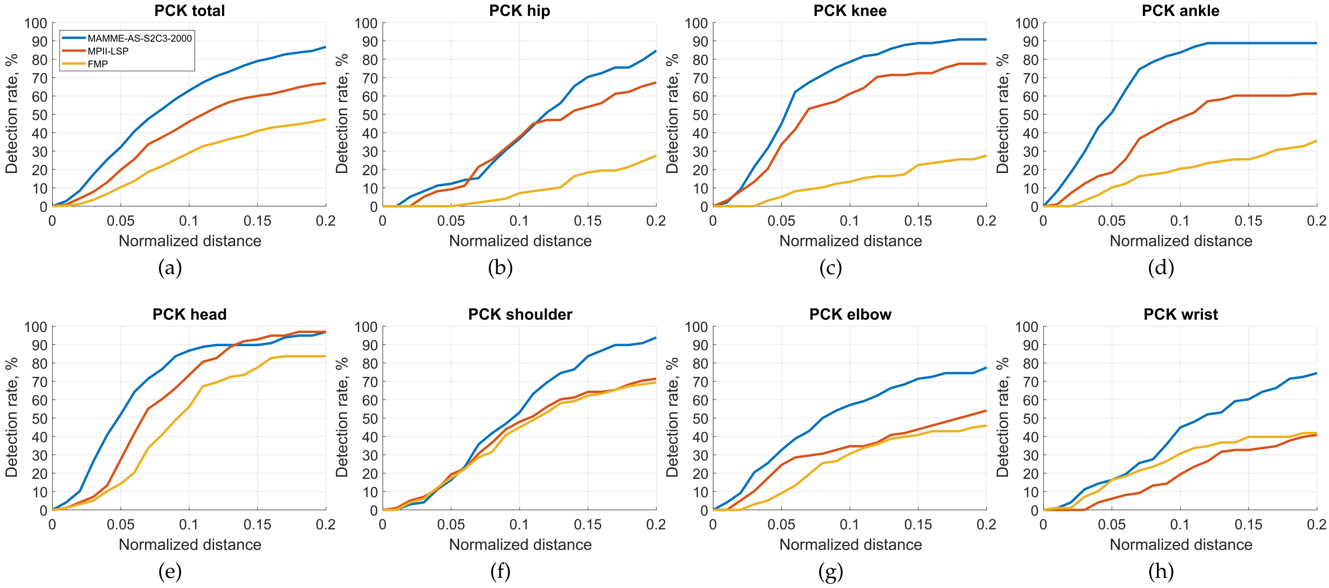
It is worth mentioning that further extending iterations did not improve the accuracy in this experiment. The evaluation of metric PCK0.2 on our test set across iterations from 200 to 3000 is shown in Fig. 14(a). We can see that the accuracy converges to a performance with 1% error range without further improvement over extended training sessions while the model at 2000 iterations shows slightly better than its following successors.

Here, we also present and compare the results of pose estimation using a classical framework against the deep neural network model. We employed a recent augmented pictorial structure based method with flexible mixture of parts (FMP), which showed best pose estimation performance on PARSE dataset [12] at that time and comparable performance to the state-of-the-art non-deep leaning methods [34], [36]. We compared the model performance of our fine-tuned CPM model, pre-trained CPM with MPII-LSP dataset, and the FMP model on orientation rectified IRS dataset to exclude the orientation factors in test result which is shown in Fig. 15. Our fine-tuned model shows advantages in total accuracy across all PCK standards. However surprisingly, trained only on a small dataset, FMP surpasses the CPM performance in all upper body parts' estimation in a high precision regime (PCK0.1) and slightly inferior in the low precision regime (PCK0.2) for head detection. This result once more emphasizes the importance of color information in the CPM model. Instead, FMP essentially employs the HOG features, which highly depend on image gradients. This is the reason that FMP surpasses the pre-trained CPM in several body parts' estimation. For example, the head, shoulder, and elbow show obvious shape features compared to other body parts, which is more easily captured by the HOG descriptors than the color information. The quantitative result of PCK0.2 is shown in Table 1 and our fine-tuned model surpasses the second best model by 19.6%.

F. ON-DEMAND ESTIMATION TRIGGER

To validate the effectiveness of our on-demand trigger pipeline, we video monitored a mannequin on bed via using IRS system. In this video, we mimicked a practical scenario, where hospital bed is moved around and kept in place for a while after each relocation. In this process, mannequin was located in different in-bed orientations as part of the $\{N, E, S, W\}$ general categories, defined in Section II-B. We simulated 4 times relocation in the video and each stable period in between lasted approximately 6–8 seconds, which is enough for our algorithm to distinguish the static states ($\mathcal{S} = 0$) from the dynamic states ($\mathcal{S} = 1$). The pose estimation algorithm is triggered only at each falling edge, when $\mathcal{S}$ transits from 1 to 0, and not frame by frame.

To generate the the ground-truth label for the video, we replayed the video and annotated the start and end points of the dynamic states manually by recording their frame index. Our on-demand trigger method is also applied on this video with backward window $\mathcal{W}_{bf}$ of size $N_{bf} = 30$. As the test video has a frame rate of 11.28 frame/s, this window is approximately 2.66s. Fig. 14(b) shows our state estimation results against the ground-truth and the trigger signal to initiate the estimation pipeline. It shows that our algorithm is successful in triggering the estimation after each dynamic to static state transition. There is a slight lag between our trigger and ground-truth label due to the use of backward window of size 2.66s. In practice, caregivers in nursing homes and hospitals usually perform posture repositioning for pressure re-distribution on a regular basis to prevent bed born complications such as pressure ulcers [37].



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 15. **Quantitative posture estimation result with different pre-trained general purpose pose detection models against our fine-tuned model on the rectified IRS in-bed pose images.**

Considering a recommended 2-hour interval between repositioning, even 10 seconds lag can only result in 0.12% information loss and the loss caused by our lag is much smaller.

IV. DISCUSSION AND FUTURE WORK

In this work, we have presented a comprehensive system to estimate in-bed human poses and address the challenges associated with this specific pose estimation problem. The issue of huge lighting variations for this application is addressed by our proposed IRS imaging system. The image differences between the overview angle used for human in-bed monitoring and the side angle often used in available human pose datasets is handled by our proposed 2-end HOG rectification method, which effectively improve the performance of existing pose estimation models for irregular poses. In CV applications, this issue is usually handled by extensively augmenting the dataset to cover all possible orientations. However, our rectification method avoids the time/memory expense of retraining the whole network by this preprocessing steps. Without a large dataset, retraining a deep neural network from scratch is not feasible. In this paper, we explored the idea of using a shallow (limited in size and different in perspective and color) dataset collected from in-bed poses to fine-tune a CNN, which was pre-trained on general human poses. We showed that classical fine-tuning principle is not always effective and the network architecture matters. For the specific CNN, the CPM model, our proposed fine-tuning model demonstrated clear improvement over the classical one.

The problem of in-bed pose estimation still has other challenges that remain. The main one is the high probability of being covered by a sheet or blanket while on bed. In fact, vision-based methods would no longer be functional in this case. Other sensing modalities may provide other forms of indication for pose inference, however it is less likely to be able to retrieve color information from those modalities. In this respect, this work is also a pilot study for pose estimation under information loss. In future work, we plan to address this issue by employing other sensing modalities to complement vision information. Test on real human data is also anticipated in our next step.

ACKNOWLEDGMENT

The code and data are available at GitHub code and in-bed datasets.

REFERENCES

[1] C. H. Lee, D. K. Kim, S. Y. Kim, C.-S. Rhee, and T.-B. Won, "Changes in site of obstruction in obstructive sleep apnea patients according to sleep position: A DISE study," *Laryngoscope*, vol. 125, no. 1, pp. 248–254, 2015.

[2] J. Black *et al.*, "National pressure ulcer advisory panel's updated pressure ulcer staging system," *Adv. Skin Wound Care*, vol. 20, no. 5, pp. 269–274, 2007.

[3] S. J. McCabe, A. Gupta, D. E. Tate, and J. Myers, "Preferred sleep position on the side is associated with carpal tunnel syndrome," *Hand*, vol. 6, no. 2, pp. 132–137, 2011.

[4] M. Abouzari *et al.*, "The role of postoperative patient posture in the recurrence of traumatic chronic subdural hematoma after burr-hole surgery," *Neurosurgery*, vol. 61, no. 4, pp. 794–797, 2007.

[5] M. B. Pouyan, S. Ostadabbas, M. Farshbaf, R. Yousefi, M. Nourani, and M. D. M. Pompeo, "Continuous eight-posture classification for bed-bound patients," in *Proc. 6th Int. Conf. Biomed. Eng. Informat.*, Dec. 2013, pp. 121–126.

[6] S. Ostadabbas, M. B. Pouyan, M. Nourani, and N. Kehtarnavaz, "In-bed posture classification and limb identification," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2014, pp. 133–136.

[7] J. J. Liu, M.-C. Huang, W. Xu, and M. Sarrafzadeh, "Bodypart localization for pressure ulcer prevention," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 766–769.

[8] S. Liu and S. Ostadabbas, "Inner space preserving generative pose machine," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 718–735.

[9] S. Liu and S. Ostadabbas, "A semi-supervised data augmentation approach using 3D graphical engines," in *Proc. 9th Int. Workshop Hum. Behav. Understand., Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 11130, 2018.

[10] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1014–1021.

[11] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[12] S. Antol, C. L. Zitnick, and D. Parikh, "Zero-shot learning via visual abstraction," in *Proc. ECCV*, 2014, pp. 401–416.

[13] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[14] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 723–730.

[15] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 256–269.

[16] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3041–3048.

[17] L. Karlinsky and S. Ullman, "Using linking features in learning non-parametric part models," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 326–339.

[18] D. Munoz, J. A. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 57–70.

[19] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 82–90.

[20] L. Pishchulin *et al.*, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4929–4937.

[21] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[22] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4724–4732.

[23] S. Liu and S. Ostadabbas, "A vision-based system for in-bed posture tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2017, pp. 1373–1382.

[24] A. Treiman. *Life at the Limits: Earth, Mars, and Beyond*. Accessed: Sep. 2018. [Online]. Available: http://www.lpi.usra.edu/education/fieldtrips/2005/activities/ir_spectrum/

[25] A. Zamanian and C. Hardiman, "Electromagnetic radiation and human health: A review of sources and effects," *High Freq. Electron.*, vol. 4, no. 3, pp. 16–26, 2005.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[27] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3642–3649.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[29] H. Freeman and R. Shapira, "Determining the minimum-area encasing rectangle for an arbitrary closed curve," *Commun. ACM*, vol. 18, no. 7, pp. 409–413, 1975.

[30] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[31] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. BMVC*, vol. 2, 2010, p. 5.

[32] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3674–3681.

[33] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[34] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[35] D. Ramanan, "Learning to parse images of articulated bodies," in *Proc. NIPS*, vol. 1, 2007, pp. 1129–1136.

[36] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1465–1472.

[37] S. Ostadabbas, R. Yousefi, M. Nourani, M. Faezipour, L. Tamil, and M. Pompeo, "A posture scheduling algorithm using constrained shortest path to prevent pressure ulcers," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2011, pp. 327–332.