UNIVERSITY OF CALIFORNIA SAN DIEGO

Continuous Behavior Acquisition in Clinical Environments

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Electrical Engineering
(Intelligent Systems, Robotics, and Control)

by

Kenny J. Chen

Committee in charge:

Professor Vikash Gilja, Chair
Professor Nikolay Atanasov
Professor Michael Yip

2018

The Thesis of Kenny J. Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California San Diego

2018

EPIGRAPH

The only way to swim fast is to swim fast.

Coach Ray Wong

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Vikash Gilja for his constant support and guidance throughout all these years. I am deeply indebted to him for the many opportunities he has provided me with since my undergraduate years and for his commitment to my growth in scientific research and communication. Vikash's passion for science and teamwork is admiring, and I hope that I can one day be as inspirational as he has been to me. I would also like to thank Michael Yip and Nikolay Atanasov for serving on my committee and for our many conversations on the latest news in robotics, and I hope that I can one day help advance the field of robotics just as they are doing right now. I am also thankful for all my past and present labmates in TNEL who have always been supportive of me, including Aashish Patel, Akin Omigbodun, Daril Brown, John Hermiz, Kevin Moses, Nathan Gong, Nick Rogers, Paolo Gabriel, Stephen Estrin, Tejaswy Pailla, Venkatesh Elango, and Wahab Alasfour. I am specifically thankful for my mentor Paolo Gabriel and his continuous guidance and patience ever since he took me under his wing. I attribute my growth as a researcher to his mentorship, and I could not have asked for a better mentor than him.

I am also thankful for my parents and brother Kellen Chen for their unconditional love and unwavering support in pursuing my dreams. They have always been there for me during both the highs and the lows in my life, and my appreciation for them is something that words cannot express. Finally, I am thankful for my BudEEs Alejandro Cervantes, Billy Zeng, Bronson Arucan, Jeanette Nguyen, Kevin Le, Sam Vineyard, and Thomas An for their amazing friendship and for all of our crazy times both in and out of the classroom. My success in the ECE B.S./M.S. program at UC San Diego would not have been possible without their love and support these past five years, and I will always remember our (oddly fun) late-night study sessions at Geisel for midterms and finals.

The chapters of this thesis consist of published and submitted conference/journal manuscripts. The thesis author was the primary investigator and author of these papers.

- Chapter 4, in part, has been submitted for publication of the material as it may appear in K. Chen,

P. Gabriel, A. Alasfour, C. Gong, W.K. Doyle, O. Devinsky, D. Friedman, P. Dugan, L. Melloni, T. Thesen, D. Gonda, S. Sattar, S. Wang, and V. Gilja, "Patient-specific pose estimation in clinical environments," *IEEE Journal of Translational Engineering in Health and Medicine*, 2018. Chapter 4, in part, is also a reprint of the material as it appears in K. Chen, P. Gabriel, A. Alasfour, W.K. Doyle, O. Devinsky, D. Friedman, T. Thesen, and V. Gilja, "Patient-specific pose estimation in a clinical environment," *SoCal Machine Learning Symposium*, 2017.

- Chapter 5, in part, has been submitted for publication of the material as it may appear in K. Chen, P. Gabriel, A. Alasfour, C. Gong, W.K. Doyle, O. Devinsky, D. Friedman, P. Dugan, L. Melloni, T. Thesen, D. Gonda, S. Sattar, S. Wang, and V. Gilja, "Patient-specific pose estimation in clinical environments," *IEEE Journal of Translational Engineering in Health and Medicine*, 2018.

ABSTRACT OF THE THESIS

Continuous Behavior Acquisition in Clinical Environments

by

Kenny J. Chen

Master of Science in Electrical Engineering

(Intelligent Systems, Robotics, and Control)

University of California San Diego, 2018

Professor Vikash Gilja, Chair

Continuous behavioral labels of hospital patients provide quantitative data that can be informative for both research studies and clinical applications. An analysis of neural correlates to natural behavioral labels extracted from pose estimates, for example, could enable more robust brain-machine prostheses for those with limb loss or motor impairment. Automated patient motion analysis could also provide additional insight to clinicians during motor scoring assessments and seizure classification for a better-informed diagnosis. Likewise, continuous patient safety monitoring enabled by posture annotations could detect potential bed falls or other injury risks and quickly alert nurses to administer preventive measures. Such scenarios rely on consistent and accurate patient posture tracking in clinical environments.

While many existing pose estimation frameworks are effective when subjects are located in uncluttered settings, clinical environments can provide several visual challenges that these general frameworks are not calibrated for. In this thesis, we propose a semi-automated approach for improving upper-body pose estimation in noisy clinical environments, whereby we adapt and build around an existing joint tracking framework to improve its robustness to environmental uncertainties. The proposed framework uses subject-specific convolutional neural network (CNN) models trained on a subset of a patient's RGB video recording chosen to maximize the feature variance of each joint. Furthermore, by compensating for scene lighting changes and by refining the predicted joint trajectories through a Kalman filter with fitted noise parameters, the expanded framework can yield more consistent posture annotations in these settings when compared to general methods. The perspectives gained from this work provide better insight in developing a practical pose estimation framework for researchers and clinicians in these environments.

Chapter 1

Introduction

1.1 Motivation

Accurate patient joint tracking and posture estimates provide quantitative data that can be experimentally and clinically informative. Upper-body annotations for long-term continuous video of patients in the epilepsy monitoring unit (EMU), for example, can be used to further explore the relationship between neural activity and unconstrained human movement when combined with a neural recording system [1, 2]. Analysis of neural correlates to behavioral labels extracted from long duration naturalistic datasets collected in the hospital could then provide a pathway for more robust brain-computer interfaces (BCI's). These include assistive robotic arms [3–5] and neural prostheses [6, 7] for those with limb loss or total paralysis. Alternatively, posture annotations can be used to objectively score patient motor capabilities to enhance current subjective assessments. For instance, the Unified Parkinson's disease rating scale (UPDRS) [8] is the current standard for evaluating the severity of motor impairment associated with Parkinson's disease, but it involves a qualitative evaluation by interview and clinical observation. The outcome of this process is limited to the clinician's interpretation during examination and can be inconsistent between evaluators. Combining such assessments with additional insight from objective motion analysis could help improve the efficacy of treatment protocols. Other motor scoring assessments (e.g., BOT-2 [9], FMA [10], MAS [11])

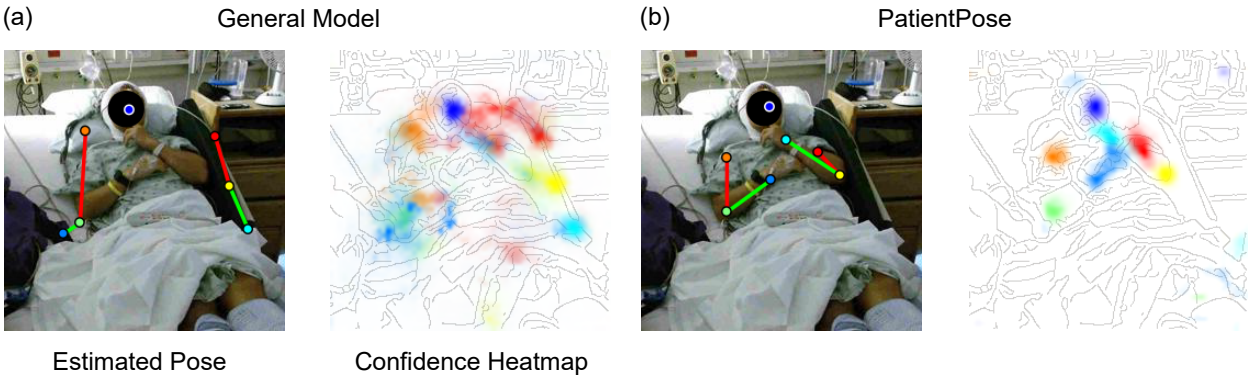

Figure 1.1: **Comparison of pose estimation models.** Upper-body posture annotations and their corresponding probability heatmaps using (a) a prepackaged model and (b) our patient-specific model. Both models were developed using the Caffe-Heatmap architecture. Our proposed framework accounts for variability in clinical environments to improve pose estimates, and can be more confident and accurate than generalized methods. Subject 1 is depicted in these images.

would benefit similarly.

Several studies in automated motor scoring incorporate wearable devices (such as inertial measurement units [12, 13] and accelerometers [14–16]) that collect kinematic data of subject appendages, but may risk complications from prolonged wear of physical sensors [17]. These systems can be complemented with less invasive video-based tracking methods that supplant physical sensors when they are temporarily removed for relief. Additionally, for patients who are unable to wear such sensors due to injuries at the wrists or at other attachment areas, video-based joint tracking and behavioral labeling can create a nonintrusive means to monitor their well-beings.

To further motivate this work, other uses of patient pose estimates include enhanced seizure classification and automated patient safety monitoring. Current methods for classifying epileptic seizure types depend on a manual inspection of seizure semiologies [18], and an analysis of patient movements during an episode could provide clinicians with further intuition on which seizure treatment protocol to select. In addition, preventive measures for accidental bed falls could be administered on a shorter notice, since automated patient video monitoring can alert nurses of impending bed falls faster than current weight sensors. Video safety monitoring can also analyze overall patient motility during an extended period of stay and

automatically notify nurses to prevent pressure ulcers or other complications from developing. This would allow nurses to attend more patients who require manual care by freeing their time from constant check-ups.

To this end, we introduce PatientPose, an adaptation of Caffe-Heatmap [19] for semi-automated pose estimation in clinical environments. Our additions to the existing pose estimation framework include three key elements that enable more accurate and consistent patient posture tracking than before:

(i) A preprocessing step to accommodate for the frequent scene lighting changes found in hospital rooms.

(ii) A training technique that targets separate convolutional neural network (CNN) models specifically to each patient to capture the high variance of postures a subject can realize during their hospital stay.

(iii) A Kalman filter with tuned noise parameters which refines the predicted joint trajectories.

We show that for three subjects recorded in two research clinics, the extended system provides an increase in tracking performance when compared to two state-of-the-art generalized frameworks (Figure 1.1).

1.2 Related Work

The importance and potential impact of human pose estimation is supported by the substantial history of research in this field. Recent work in computer vision [19–27] suggests using deep convolutional neural networks (CNN's) to automatically estimate joint locations in long-term recording sessions. Toshev and Szegedy [25] were the first to use CNN's for human pose estimation and regressed joint coordinates directly from a cascade of deep CNN regressors. More recently, Pfister *et al.* [19] instead regressed confidence *heatmaps* for the joint positions of each input frame and improved estimates by aligning and pooling heatmaps with neighboring frames. This framework was then extended by Charles *et al.* [26] who recursively filtered and processed the estimates for even better labels. Cao *et al.* [27] later used a two-branch multi-stage CNN architecture to encode the location and orientation of body parts into a set of 2D vector fields and achieved real-time multi-person pose estimation.

While the aforementioned frameworks were developed independently from subject and environment and intend to be all-inclusive solutions, they often do not generalize well to patients within clinical environments. Previous works on improving pose estimation performance in these complex environments take advantage of a wide range of available sensors [28–33]. Achilles *et al.* [28] used a single depth camera to regress joint coordinates specifically for body tracking under blanket occlusion, and Liu *et al.* [29] relied on a novel infrared image acquisition technique using a bird's-eye view in order to monitor patient sleeping postures. Belagiannis *et al.* [30] combined information from multiple RGB cameras to track surgeons and medical staff in operating rooms, and Kadkhodamohammadi *et al.* [31] improved upon pose estimation in operating rooms by using depth sensors in tandem with multiple RGB cameras. Chaaraoui *et al.* [32] also used a multi-camera setup but for vision-based monitoring and action recognition by learning subject activity patterns from estimated silhouettes. However, none have attempted to extract high-quality joint estimates to track freely-behaving patients in hospitals across hours of data using a single RGB camera. Capturing RGB video is trivial with the current state of consumer technology, and to our knowledge this work is the first to create a pose estimation framework that specifically targets subjects in clinical environments using only one angle of recorded RGB video. Additionally, the proposed extensions to Pfister *et al.*'s Caffe-Heatmap [19] do not modify the original framework's central CNN architecture and could potentially be adopted to improve other general pose estimators, and our framework is capable of a real-time implementation after a patient's initial training procedure.

1.3 Thesis Overview

The remainder of this thesis consists of five chapters, and is organized as follows:

In Chapter 2, we introduce our data acquisition system and the different modalities we captured during each recording session. We outline our novel dataset we recorded of several consenting patients in different research clinics, and we also detail the various clinical considerations and concerns during the data

acquisition phase and our solutions to address them.

In Chapter 3, we provide a brief overview of two state-of-the-art generalized pose estimation frameworks that we used as benchmarks against our framework.

In Chapter 4, we present our three main augmentations to an existing pose estimation framework that enable more consistent posture tracking in clinical environments. Each component serves a specific purpose, and we provide our reasons for implementing each one into our pipeline.

In Chapter 5, we analyze each component to convince the reader that our techniques are reasonable for improving patient pose estimation in noisy clinical settings. Additionally, we explain our evaluation criteria and provide pose estimation accuracy comparisons between our framework and the two benchmark general frameworks. These comparisons used selected test sets of patient data for three subjects recorded in various clinical monitoring units.

In Chapter 6, we summarize the contributions made in this thesis and discuss the trade-offs between our patient-specific framework and general pose estimation frameworks. In particular, we note that our framework serves a different purpose than generalized methods and we discuss who the target user is for each one. We also consider several potential directions of future work and refer the reader to the appendix for preliminary results on two that are particularly interesting.

Chapter 2

Systems and Data Acquisition

2.1 Subject Recording and Dataset Description

In this work, we conducted our experiments using a subset of a novel dataset. Three patients with intractable epilepsy were enrolled according to protocols approved by the Institutional Review Board (IRB) at the New York University (NYU) Langone Comprehensive Epilepsy Center and the Rady Children's Hospital (RCH), San Diego, Pediatric Epilepsy Center. Video was recorded using a Microsoft Kinect v2 during each patient's stay, targeting 1–2 days post-implant of electrodes when the subjects were expected to be most active. Video was recorded using multiple modalities (i.e., RGB, depth, infrared), but only the RGB images were considered for this study. Specific details regarding the duration of each subject's recording session and the number of frames used for framework training/evaluation are provided in Table 2.1, and a sketch of the setup can be seen in Figure 2.1. Note that the Kinect v2 RGB camera samples at either 15 or 30 frames-per-second (fps) depending on room luminance and horizontally flips all images when saving to disk. Our data acquisition system was fit onto a custom-built mount that stood 5 feet tall and was placed about 20 degrees to the left of Subjects 1 and 3 (S1 and S3) and 45 degrees to the left of Subject 2 (S2).

Although the work presented in this thesis only considers RGB video for three subjects, we note that our novel dataset also includes other modalities recorded from four adult subjects and five adolescent

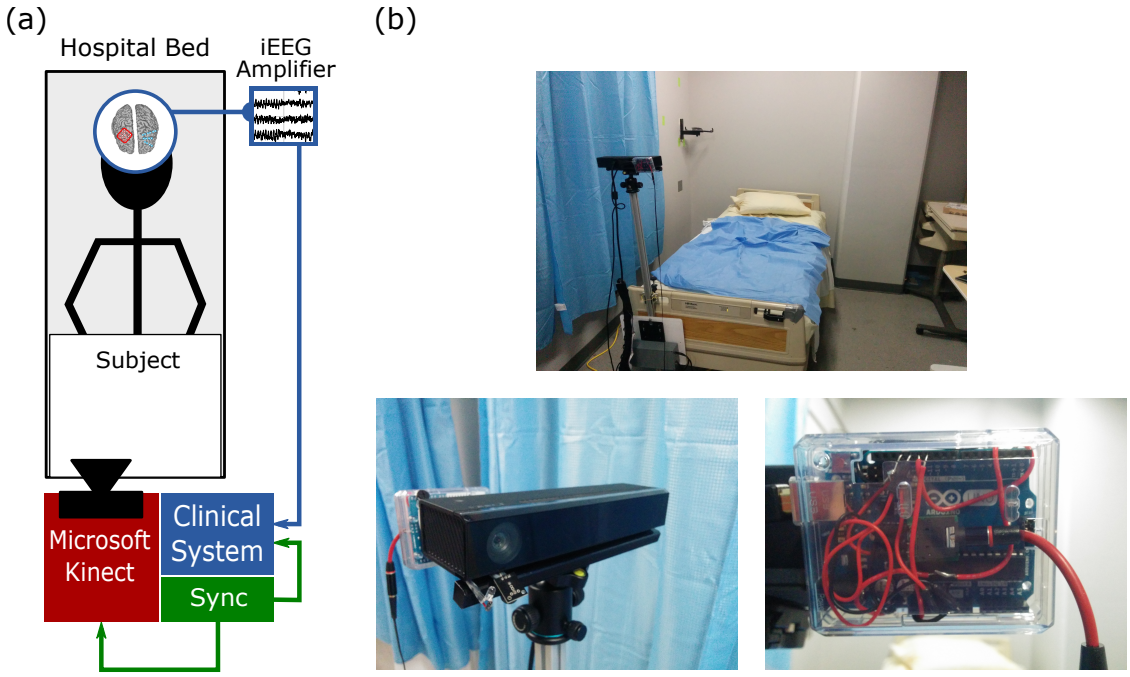

Figure 2.1: **Data acquisition system.** Our data acquisition system was deployed to each hospital using a similar setup as seen in (a) the cartoon schematic and (b, top) our mock hospital room. Multiple modalities were captured during each patient recording session, including RGB/depth/infrared video by a Microsoft Kinect (b, bottom left) and neural activity by cortical implants connected to an intracranial electroencephalography (iEEG) amplifier. The modalities were then time-aligned using an Arduino microcontroller (b, bottom right).

Table 2.1: **Dataset summary.** Subject datasets varied in recording duration and number of frames, but the number of training frames used per dataset was consistent across all subjects (2,000 for CNN model training and 500 for Kalman filter parameter training). Note that the Kinect automatically recorded S1's dataset at 15fps and S2/S3's dataset at 30fps due to a difference in room luminance.

Subject	**Study ID**	**Hours**	**Number of Frames**		
			Total	*Training*	*Testing*
S1	NY531	1.8	94,470	2,500	3,000
S2	RCH1	5.8	625,127	2,500	1,000
S3	RCH3	22.2	2,399,469	2,500	1,000

subjects across more than 250 collective hours in the two hospitals. These subjects were implanted with either electrocorticography (ECoG) or stereoelectroencephalography (sEEG) electrodes for electrophysiological monitoring of epileptic seizures and other complications they may have had. In addition to the modalities recorded by the Kinect, the signals captured by these electrodes were also collected to create a dataset that can enable a wide range of potential studies in neurophysiology.

(a) (b)

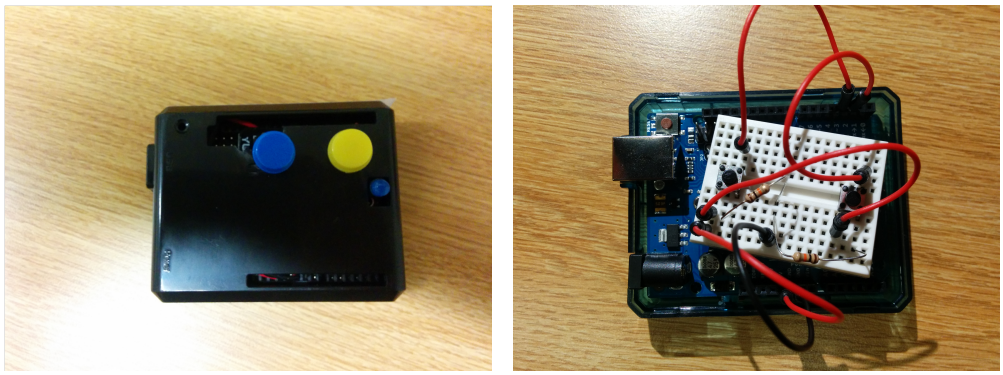

Figure 2.2: **Prototype of the opt-out device.** (a) A front-view of our prototype for a patient's opt-out device, and (b) the internals of the device showing its simple two-button circuit and the Arduino Uno. Patients were given two options to temporarily opt-out of data collection during a continuous recording session: one which paused data collection for 15 minutes (blue button) and one which erased the previous 15 minutes of data recording (yellow button). Another version of this device featuring a studier construction was created prior to the system's deployment, but a photo of it was not taken.

2.2 Time-Aligning Recorded Modalities

Data stemming from multiple sources (i.e., Microsoft Kinect and neural acquisition system) needed to be synchronized to ensure that information between the sources were relevant at each time step. Although the RGB, depth, and infrared modalities were automatically aligned via the Kinect's internal clock, synchronization between patient video and acquired neural signals was not as trivial, and relying on wall-clock time resulted in poor data alignment. As part of our data acquisition system, we used an Arduino Uno as an intermediate device that sent audio/visual cues to the Kinect through a speaker and an LED, and pulses to the neural acquisition system through a 3.5mm connection (Figure 2.1). These triggers were then used to calculate the offset in time between the sources for a better temporal alignment of data.

2.3 Patient Recording Opt-Out

As part of our agreement for recording patients in each hospital, we were required to create a device that temporarily paused the data acquisition system during a continuous recording session. Patients were

often bound to their hospital beds and could not move without nurse assistance, and therefore occasionally had private moments they did not wish to be recorded. To provide patients with an option of pausing the system during these sensitive moments, we built an additional "opt-out" device using another Arduino Uno microcontroller (Figure 2.2). This device included two buttons that allowed patients to opt-out of recording: one which paused data collection for 15 minutes and one which deleted the previous 15 minutes of data recording. An LED on the device also indicated when the 15 minute pause was nearing its end.

2.4 Handling Sensitive Data

Patient data collected from hospitals required careful data handling procedures to ensure full compliance with hospitals confidentiality agreements. In addition to our contractual bindings for upholding patient anonymity, we were also obligated as researchers to responsibly manage and store sensitive (and potentially identifying) patient data. Therefore, we stored our collected datasets onto encrypted and password-protected external hard drives to prevent access from unauthorized persons. Furthermore, video recorded from the Kinect was anonymized prior to visualizations, and only those who were authorized could view the de-anonymized images.

Chapter 3

General Pose Estimation

General pose estimation frameworks are effective when subjects are located in uncluttered settings, but they can be unreliable when applied to noisy environments such as epilepsy monitoring and intensive care units. Such locations present several visual challenges that these generic frameworks do not account for, including variance in lighting conditions throughout a recording session, non-subject (e.g., clinician, nurse, visitor) interferences, and environmental occlusions (e.g., bed blankets, head wrapping, hospital gown). As a result, estimated joint locations of patient video recorded in the hospital using all-inclusive pose estimators can be inconsistent in its predictions and may not generalize well to clinical environments.

Two state-of-the-art generic pose estimation frameworks include Pfister *et al.*'s Caffe-Heatmap [19] and Cao *et al.*'s OpenPose [27]. Although these frameworks both aim to predict subject joint positions, they are vastly different in their approaches. Caffe-Heatmap estimates joint coordinates by first regressing confidence heatmaps of positions using a pretrained model, and then aligning neighboring heatmaps with optical flow to take advantage of temporal information. OpenPose, however, encodes the location and orientation of body parts into a set of 2D vector fields and uses nonparametric representation to learn the association of body parts to individuals. While these algorithms are quite impressive in their ability to work on data in many different environments, we found that their predictions were sensitive to the visual challenges within clinical environments; this motivated the work presented in this thesis.

Chapter 4

Patient-Specific Pose Estimation

To calibrate general pose estimation frameworks for clinical environments, we developed three key add-ons to address the visual challenges found within these settings. In this chapter, we present these three augmentations in detail and discuss the challenges that each component tries to compensate for. In particular, we present a preprocessing step that accounts for shifts in lighting conditions, a high-quality training strategy that captures a patient's diverse collection of postures, and a Kalman filter with trained noise parameters that better predicts partially occluded joints. These methods were built around the Caffe-Heatmap architecture [19], and a complete pipeline of our framework can be seen in Figure 4.1.

4.1 Image Preprocessing

4.1.1 Cropping

RGB frames were originally recorded at a resolution of 1920x1080 pixels in width and height, but then cropped and resized to 256x256 pixels centered around the patient for memory efficiency during GPU training. The location of cropping was manually selected once per patient dataset and used to crop all frames within that same set.

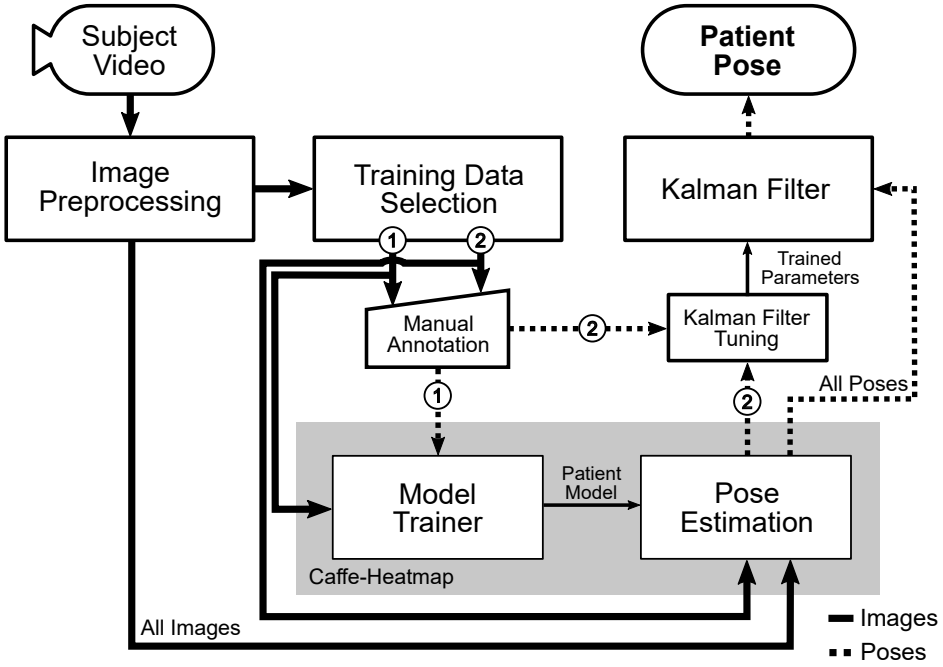

Figure 4.1: **Pipeline of framework.** The proposed framework extends Caffe-Heatmap to improve pose estimation of patient video recorded in clinical environments. Prior to estimation, a new patient-specific CNN model is trained using a subset of preprocessed video frames that maximize feature variance (①). This model is then used to estimate the joint positions of the same patient from additional video, which are then refined using a Kalman filter with noise parameters trained using another subset of preprocessed frames (②). This work used 2,000 frames for ① and 500 frames for ②.

4.1.2 Scene Lighting Normalization

Hospital rooms can fluctuate in lighting conditions as a result of its dynamic atmosphere. Clinicians, nurses, and visitors frequently enter and exit the patient room and often block sources of light from entering the recording camera's lense, resulting in video frames with inconsistent brightness. In addition, for long duration recording sessions, natural sunlight entering from a nearby window can vary in intensity as the sun rises and sets, thereby affecting image consistency. To account for these fluctuations, image brightness was normalized by first transforming each frame to the Hue-Saturation-Value (HSV) color space and then applying contrast-limited adaptive histogram equalization (CLAHE) [34] onto the value layer with an 8x8 tile size. Regions with similar surroundings (e.g., bed sheets) were susceptible to noise amplification when normalized using global or regular adaptive equalization [35], and CLAHE limited the amount those regions could increase in contrast (Figure 4.2).

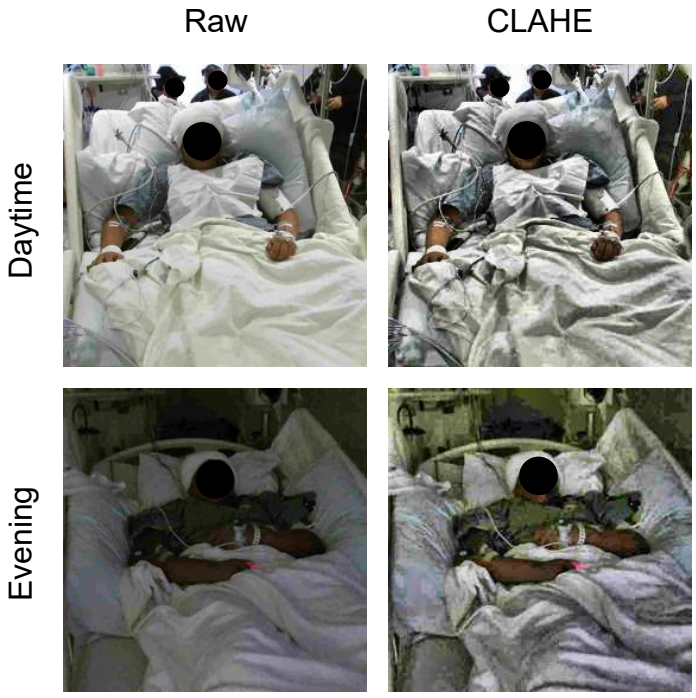

Figure 4.2: **Scene lighting normalization.** Raw patient images (left) during daytime (top) and evening (bottom) were significantly different in lighting conditions for the same clinic. However, after applying contrast-limited histogram equalization (CLAHE, right), the frames across a dataset became more consistent in brightness. Subject 3 is depicted in these images.

4.2 Convolutional Neural Network Models

4.2.1 Motivation

Prepackaged convolutional neural network (CNN) models trained using movie or video frames work well against other generic pose estimation datasets (e.g., BBC Pose [36], Common Objects in Context (COCO) [37], Frames Labeled in Cinema (FLIC) [38], MPII Human Pose [39]), but can be less reliable when applied to videos of hospital patients due to various challenges unique to the clinical setting. Therefore, we trained a separate CNN model for each of our subjects using an extracted subset of frames held out from the test set. These high-quality training sets were designed to capture the wide range of postures the corresponding patient may naturally take on throughout a recording session.

4.2.2 High-Quality Training Dataset

To maximize posture diversity and therefore feature variance in a patient's training data, frames were selected from both movement and non-movement periods. This was accomplished by first applying the Gunnar-Farnebäck dense optical flow algorithm [40] onto the raw RGB video of the same patient to calculate the average magnitude of scene movement between adjacent frames. A threshold on this average flow empirically set to 0.15 pixels per frame then partitioned patient RGB video into periods of movement and idleness. Afterwards, a subset of frames was uniformly sampled from the segmented video such that frames drawn from movement and rest periods were distributed $70\%/30\%$. Using this strategy, 2,000 frames for model training were selected across the entire span of each patient's dataset which captured different postures the patient may take on during their stay. Frames with significant patient occlusions were excluded.

4.2.3 Patient-Specific Model Training

Ground truth (x, y) coordinates of the seven joints (i.e., head, left/right hands, elbows, and shoulders) were manually marked for each training set using a custom labeling script. A CNN model was then trained for each patient using the Caffe-Heatmap model training architecture [19] on an NVIDIA GeForce GTX 1080 Ti with the annotated images. One million iterations of batch size 14 were used with a learning rate of 10^{-8} and momentum of 0.95, and each iteration took approximately 0.75 seconds for a total of nine days of training per model. The resulting models learned features specific to each patient through the high-quality training set. Using the same hardware and configurations, training a generic model on the FLIC dataset with about 4,500 frames would span around twelve days.

4.2.4 Inference via Patient-Specific Model

To enable easy adoption of our augmentations onto other pose estimators, we did not directly modify the Caffe-Heatmap framework. We therefore treated it as a black box, with the inputs as the patient-

specific Caffe [41] model and N number of frames, and an output of seven 256x256 confidence heatmaps for each frame. Each joint location was then taken to be at the *argmax* of its corresponding heatmap, resulting in a 2x7xN structure of $(x, y) \in [0, 256]$ joint coordinates. For each frame, inference spanned $\sim$0.03 seconds when using the same NVIDIA GeForce GTX 1080 Ti (compared to $\sim$10 seconds per frame on an Intel Xeon CPU E5-2630), enabling the potential for a real-time implementation. Specific details of the Caffe-Heatmap architecture can be found in [19].

4.3 Kalman Filter

4.3.1 Motivation

Joint locations estimated by a patient-specific CNN model were generally reasonable, but occasionally contained jitter or large jumps when a patient moved quickly or was occluded. Therefore, a standard Kalman filter [42] was used as a post-processing step to leverage the temporal information found between frames in order to refine any noisy measurements. The Kalman filter consists of two primary components (a state transition function and a measurement function) that model the underlying physics of a system to predict its state over time, making it an appropriate choice for denoising estimated joint trajectories. In addition, we chose to use a Kalman filter (as opposed to a non-causal Kalman smoother [43]) to preserve our framework's potential to be implemented in real-time due to the filter's causality.

4.3.2 General Equations

The Kalman filter is a recursive two-step process which iteratively predicts a system's next state using past information and a predefined model, then updates its predictions using external sensor measurements. These two functions are defined by the linear state transition and measurement matrices A and H. In addition, the estimated state μ_t at each t^{th} iteration is accompanied with a covariance Σ_t that measures the

accuracy of the estimate to the true state. In the Kalman filter's prediction step, we have:

$$\hat{\mu}_t = A\mu_{t-1}\,,$$

$$\hat{\Sigma}_t = A\Sigma_{t-1}A^\top + Q\,,$$

where the hat indicates that these values are purely estimates by the filter without considering any outside measurements yet. The Q term above is the covariance of the process noise that captures the error between the transition model and the true dynamics of the system, and it is assumed to be Gaussian distributed in this work. In the update step, we have:

$$K_t = \hat{\Sigma}_t H^\top (H\hat{\Sigma}_t H^\top + R)^{-1}\,,$$

$$\mu_t = \hat{\mu}_t + K_t(z_t - H\hat{\mu}_t)\,,$$

$$\Sigma_t = \hat{\Sigma}_t - K_t H\hat{\Sigma}_t\,,$$

where K_t is the Kalman gain that adjusts the next predicted state μ_t and covariance Σ_t depending on the accuracy of the model. In this step, external sensor measurements z_t provide the filter with additional information on the system's next possible state, and the R term captures the noise in these measurements (also assumed to be Gaussian). Complete derivations of these equations can be found in [44–46].

In this work, we used a constant velocity model [47] to define the state transition and measurement matricies A and H in these equations, and we assumed independent movement between the seven joints. Therefore, we ran a separate Kalman filter on each joint, in which the 4D state estimates μ_t contained a joint's (x,y) pixel position and velocity at time t. Additionally, we used the (x,y) coordinates provided by a patient's CNN model as the external z_t measurements to update the filter's predictions on the system's state. These Kalman filter equations recursively computed a next-best-guess on a joint's position using the predefined constant velocity model and pose estimates by the CNN model.

4.3.3 Learning the Noise Parameters

The Q and R process and measurement noise covariances are critical components to the Kalman filter that model unforeseen perturbations on the system. In the context of this work, the Q term captures how erroneous the constant velocity model is to the real dynamics of a patient and the R term captures the variability in the CNN's pose estimates to the true positions. However, these matrices are frequently difficult to estimate and are often constructed using prior knowledge of the problem, tediously tuned by hand, or assumed to be independent between variables for convenience. Abbeel *et al.* [48] demonstrated that Q and R can be learned by maximizing the joint likelihood between the states and the measurements of a training dataset. In other words, for T number of training datapoints, the optimal parameters Q^j_{MLE} and R^j_{MLE} for each j^{th} patient joint can be formulated as:

$$\left\langle Q^j_{MLE}, R^j_{MLE} \right\rangle = \arg\max_{Q,R} \log p(x^j_{0:T}, z^j_{0:T}) \,. \tag{4.1}$$

Here, the joint probability distribution between the sequence of ground truth states $x^j_{0:T}$ and CNN pose estimates $z^j_{0:T}$ is:

$$p(x^j_{0:T}, z^j_{0:T}) = p(x^j_0) \prod_{t=1}^{T} p(x^j_t \,|\, x^j_{t-1}) \prod_{t=0}^{T} p(z^j_t \,|\, x^j_t) \tag{4.2}$$

with some prior $p(x^j_0)$, where the Gaussian motion and observation models are:

$$p(x^j_t \,|\, x^j_{t-1}) = \mathcal{N}(x^j_t; A x^j_{t-1}, Q) \,, \tag{4.3}$$

$$p(z^j_t \,|\, x^j_t) = \mathcal{N}(z^j_t; H x^j_t, R) \,. \tag{4.4}$$

Substituting (4.2), (4.3), and (4.4) into (4.1) results in:

$$Q^j_{MLE} = \arg\max_{Q} \Big[-T \log |2\pi Q| - \sum_{t=1}^{T} (x^j_t - A x^j_{t-1})^\top Q^{-1} (x^j_t - A x^j_{t-1}) \Big] \,,$$

$$R^j_{MLE} = \arg\max_{R} \Big[-(T+1) \log |2\pi R| - \sum_{t=0}^{T} (z^j_t - H x^j_t)^\top R^{-1} (z^j_t - H x^j_t) \Big] \,.$$

Finally, by computing the closed form solutions, we have the following optimal equations:

$$Q^j_{MLE} = \frac{1}{T}\sum_{t=1}^{T}(x^j_t - Ax^j_{t-1})(x^j_t - Ax^j_{t-1})^\top, \tag{4.5}$$

$$R^j_{MLE} = \frac{1}{T+1}\sum_{t=0}^{T}(z^j_t - Hx^j_t)(z^j_t - Hx^j_t)^\top. \tag{4.6}$$

4.3.4 Patient-Specific Noise Parameter Training

Equations (4.5) and (4.6) require ground truth states x^j_t and CNN pose estimates z^j_t from a set of training data for each patient joint. In addition, because (4.5) depends on states at times t and $t-1$, the training joints must be continuous over time. Therefore, in an attempt to capture the process variability across the entire span of a patient dataset, we constructed a "semi-continuous" training subset using the following steps. First, we segmented a patient's video into periods of movement and idleness using the same optical flow method as described in *Section 4.2.1*. Afterwards, we extracted the first 10 frames of a movement period for 50 periods chosen uniformly across the span of the patient's video. This resulted in a set of 500 "semi-continuous" frames (50 discontinuous movement periods of 10 continuous frames each) for each patient which we used to train patient-specific noise parameters. Occluded segments, movements less than 10 frames in length, and frames used for evaluation were excluded during period selection. After extraction, these 500 frames were manually annotated for ground truth joint positions. To obtain the ground truth states x^j_t, joints were assumed to have zero initial velocity at the start of each movement period, and the remaining velocity values were calculated as the difference in pixel position between adjacent frames within the same period. The frames of each segment were then sent through Caffe-Heatmap's pose estimator along with the corresponding patient-specific model to obtain z^j_t.

To calculate a patient's set of measurement noise covariances R_*, we directly implemented (4.6) for each joint such that $R^j_* = R^j_{MLE}$ using all 500 training datapoints. Equation (4.6) only depends on values at time t, and therefore its training set need not be continuous. However, because (4.5) depends on values at times t and $t-1$, we first calculated a separate $Q^{j,m}_{MLE}$ for each m^{th} movement period of length $T = 10$

frames in the semi-continuous set using (4.5). The resulting $M = 50$ matrices per joint were the covariances that maximized the data likelihood in their corresponding movement sequence. A joint's process noise parameter Q_*^j was then taken to be the average of these covariances, such that:

$$Q_*^j = \frac{1}{M} \sum_{m=1}^{M} Q_{MLE}^{j,m} \,. \tag{4.7}$$

These calculated parameters $Q_* \in \mathbb{R}^{4x4x7}$ and $R_* \in \mathbb{R}^{2x2x7}$ for each patient modeled any unforeseen perturbations on the system throughout a patient's dataset at runtime of the filter.

This chapter, in part, has been submitted for publication of the material as it may appear in K. Chen, P. Gabriel, A. Alasfour, C. Gong, W.K. Doyle, O. Devinsky, D. Friedman, P. Dugan, L. Melloni, T. Thesen, D. Gonda, S. Sattar, S. Wang, and V. Gilja, "Patient-specific pose estimation in clinical environments," *IEEE Journal of Translational Engineering in Health and Medicine*, 2018. This chapter, in part, is also a reprint of the material as it appears in K. Chen, P. Gabriel, A. Alasfour, W.K. Doyle, O. Devinsky, D. Friedman, T. Thesen, and V. Gilja, "Patient-specific pose estimation in a clinical environment," *SoCal Machine Learning Symposium*, 2017. The thesis author was the primary investigator and author of this material. Reprinted with permission.

Chapter 5

Pose Estimation Performance

In this chapter, we first provide an analysis of each component to convince the reader that our techniques are reasonable for improving pose estimation in clinical environments. Then, to validate our methods as a whole, we present pose estimation accuracy comparisons between our framework and two state-of-the-art general frameworks using selected test sets of patient data for three subjects recorded in various clinical monitoring units. In this part, we define our criteria for evaluating our framework and show that it is a reasonable approach for measuring performance, and we provide quantitative results at various spatial tolerances from the ground truth for each method. A representative demo video of patient pose estimation can be viewed at *https://youtu.be/c3DZ5ojPa9k*.[1]

5.1 Analysis of Components

5.1.1 Scene Lighting Normalization

After normalizing image brightness by applying CLAHE onto the value layer of each HSV frame, we observed a significant reduction in scene lighting variance throughout each patient dataset. This reduction is depicted by the histograms of the mean V-channel magnitude for each frame in the S2 dataset using

[1] Video frames have been blurred for patient confidentiality

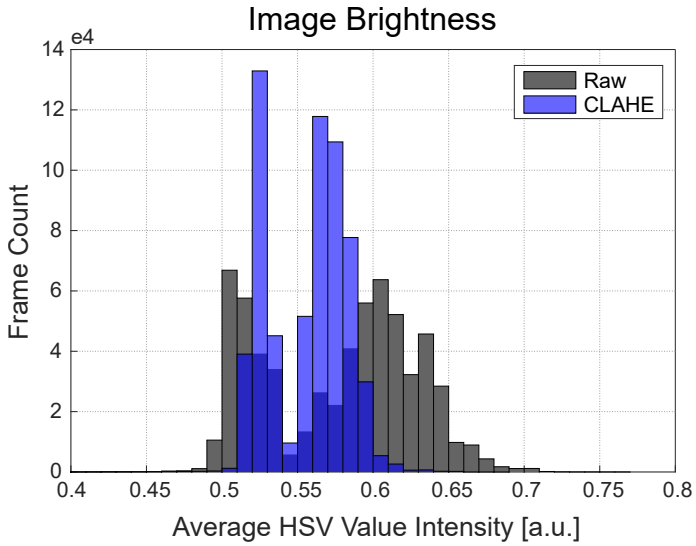

Figure 5.1: **Average brightness before and after normalization.** Overlaid distributions of the average image brightness before (gray) and after (blue) CLAHE confirm that the lighting conditions across images become more similar after equalization. Average brightness of a frame was measured by taking the mean of a frame's value-layer intensity. Each histogram used the entire Subject 2 dataset (N = 625,127 frames).

0.005 bin widths before and after lighting normalization (Figure 5.1). The value-layer in the HSV color space corresponds to image brightness, and therefore the lower histogram variance after normalization indicates a higher similarity in lighting conditions within the patient dataset than before. This translates into joint features that are more likely to be consistent in visibility.

5.1.2 High-Quality Training

To establish that our high-quality training strategy can capture a large variety of postures within a patient dataset, we compared against another manually annotated subset defined as the first 15 minutes of frames for the same patient ($\sim$13.5k frames at 15 fps). Patients were observed to engage in different postures depending on the time of day (e.g., upright vs. rest), and we therefore inferred that frames extracted using our "high-variance" (HV) training strategy would contain greater posture diversity than frames within this "low-variance" (LV) set. To investigate this, we first defined each posture as a 14-dimensional vector containing the (x,y) pixel coordinates for each of the seven joints. These vectors were then projected down to 2-dimensional space using t-distributed stochastic neighbor embedding (t-SNE) [49] for a graphical

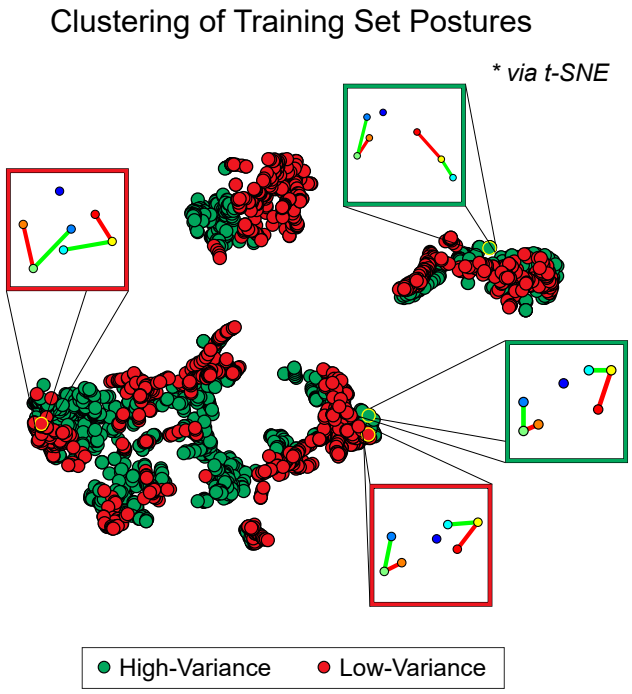

Figure 5.2: **Visualization of posture varieties.** Manually annotated postures from two training strategies were projected onto a 2D space using t-SNE to provide a graphical intuition of the various poses included in each set. "High-variance" training frames were selected from periods of movement and rest across the entire span of the patient dataset, whereas "low-variance" training frames were the first 15 minutes of recording. Note that the "high-variance" set initially contained twice as many unique postures than "low-variance," but was uniformly downsampled to prevent bias in the projection. Points within the same cluster resemble similar postures.

intuition of the posture coverage between the two strategies. Only unique datapoints were considered in this analysis, and we observed that the HV set initially contained twice as many unique postures than the LV set. Therefore, prior to t-SNE dimensionality reduction, we uniformly sampled the HV set to ensure an equal number of datapoints that would have otherwise biased the t-SNE manifold towards the more represented HV postures. In addition, the two sets were concatenated prior to projection to ensure compatibility in the output space. In this analysis, the exact t-SNE algorithm was implemented with a standard Euclidean distance metric for 1,000 iterations at a perplexity of 50 and a learning rate of 500, and the two subsets were derived from S1.

The results after projecting the 14D postures onto a 2D space (Figure 5.2) represent a low-dimensional clustering of different patient postures extracted from the two training strategies. Each color-coded data-

point represents a unique set of seven joint coordinates, and points within the same cluster resemble similar postures. Despite downsampling the HV set to match the size of the LV set for an unbiased projection, the HV set still visually occupies a larger area in the projected space. This suggests that our high-quality training strategy can capture a diverse collection of patient postures. In addition, the spread of the HV datapoints encompasses nearly all of the LV points, indicating that there may be little to no trade-off between posture diversity and coverage quality when extracting training data from the entire span of frames. Therefore, our high-quality training strategy constructs a more informative training set when compared to frames extracted from a limited window of time, and can provide the CNN architecture with more representations of each joint to train on.

5.1.3 Kalman Filter with Trained Noise Parameters

Immediate joint coordinates estimated by Caffe-Heatmap using a patient-specific CNN model are still subject to inconsistencies during periods of quick movements or patient occlusions. However, a Kalman filter with trained noise parameters refines these predictions and reduces the jitter and noise within estimated paths. Prior to optimizing S1's left hand in the testing data, the original trajectory demonstrated reasonable tracking with an average error of 10.42 ± 5.85 pixels from the ground truth. In contrast, a denoised trajectory using patient-specific parameters followed the true path more closely at an average error of 8.23 ± 5.19 pixels and exhibited less jitter at sharp turns. This is illustrated in Figure 5.3 which shows a segment of S1's left hand trajectory using the different Q and R noise parameters. For reference, using stock constant velocity parameters resulted in an average error of 9.94 ± 6.34 pixels within the same test set, and these observations were consistent throughout all joints and subjects.

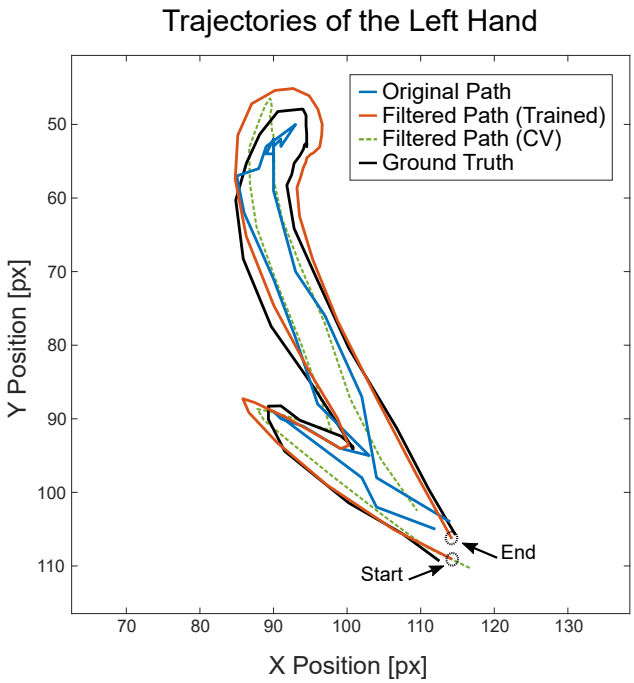

Figure 5.3: **Comparison of Kalman filter trajectories.** Estimated (x, y) coordinates of S1's left hand during an example segment of movement before (blue) and after (orange) using a Kalman filter with fitted noise parameters, as compared to the ground truth (black). A filtered path using constant velocity (CV) noise parameters (green) is also provided for reference. Across all testing data for Subject 1's left hand, the trained Kalman filter produced a lower average prediction error of 8.23 $\pm$5.19 pixels from the ground truth, compared to the original path's error of 10.42 $\pm$5.85 pixels.

5.2 Comparison to General Pose Estimation

5.2.1 Evaluation Criteria

Performance was measured using the Euclidean distance of estimated joint coordinates against an additional set of manually annotated frames held out from the training set for each patient. These frames were chosen for their variety in postures, fluctuations in lighting conditions, and occasional nurse appearances. For each patient test set, we compared our framework's pose estimation performance against two state-of-the-art generalized frameworks by evaluating joint estimates from each method at distances between 0 and 30 pixels (px) from the ground truth. These methods included Caffe-Heatmap by Pfister *et al.* [19] trained on FLIC and OpenPose by Cao *et al.* [27] trained on COCO.[2] Figure 5.4a provides a spatial

[2]Models were provided out-of-box by their respective authors

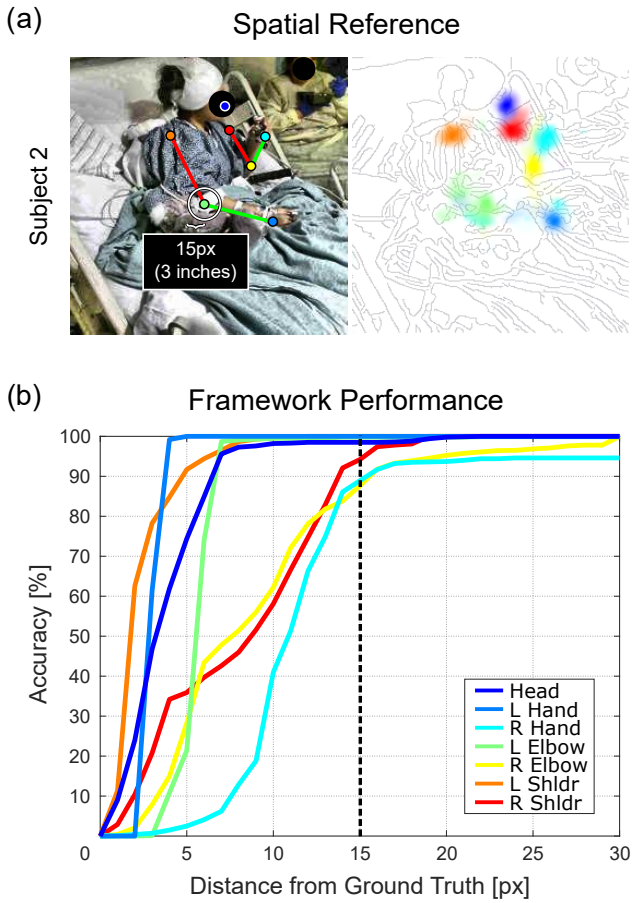

Figure 5.4: **Spatial reference and Subject 2 accuracy curves.** (a) Example skeleton and heatmap with a 15-pixel (3-inch) radius for spatial reference, and (b) Subject 2 accuracy curves between 0 and 30 pixel tolerances from the ground truth.

reference of 15 pixels (approximately 3 inches) and Figure 5.4b shows an example plot of joint accuracies at varying tolerances.

5.2.2 Pose Estimation Results

At a tolerance of 15 pixels, our framework was more accurate than Caffe-Heatmap by $42.4 \pm 8.3\%$ and OpenPose by $11.4 \pm 3.9\%$ on average across our three patient test sets (Table 5.1). Patient hands and elbows were typically the most challenging joints to estimate for every method, but we saw more consistent tracking in these categories using our framework. Figure 5.5 shows a complete performance comparison against the two generalized methods for all three subjects. With Subjects 1 and 2, we observed

Table 5.1: **Pose estimation accuracy rates at 15 pixels.** Subject pose estimation accuracy rates [%] at a 15-pixel tolerance for each category of joints averaged between the left and right body parts. The average and standard deviation was calculated using the corresponding accuracies of that method for each subject.

Subject 1

Method	Head	Hands	Elbows	Shoulders	Average
CH-FLIC	95.6	0.4	22.1	30.7	49.8 $\pm$41.0
OpenPose	99.4	37.3	93.4	88.4	83.7 $\pm$28.6
Ours	**99.9**	**87.8**	**99.1**	**95.6**	**96.5 $\pm$5.6**

Subject 2

Method	Head	Hands	Elbows	Shoulders	Average
CH-FLIC	78.4	2.0	16.6	48.8	49.2 $\pm$34.1
OpenPose	**99.4**	49.4	69.2	92.9	82.3 $\pm$23.1
Ours	98.5	**94.5**	**93.7**	**97.1**	**96.8 $\pm$2.2**

Subject 3

Method	Head	Hands	Elbows	Shoulders	Average
CH-FLIC	82.0	38.9	23.9	12.0	49.7 $\pm$30.9
OpenPose	97.9	48.0	52.5	79.5	75.6 $\pm$23.5
Ours	**99.4**	**60.1**	**73.0**	**80.2**	**82.5 $\pm$16.4**

a considerable improvement on tracking performance for all seven joints, and our framework labeled at least 80% of frames for any joint within 15 pixels from the ground truth. In addition, our framework provided $\sim$50% more hand annotations at this tolerance when compared to OpenPose for these two subjects.

In contrast to the test sets for Subjects 1 and 2, Subject 3's chosen test set contained more frequent hand occlusions in which Subject 3 often placed their left and right hands behind their head during rest. This decreased the overall tracking consistency across all methods for these two joints. However, our framework still on average provided $22.0 \pm 9.5\%$ and $9.8 \pm 6.8\%$ more hand labels than Caffe-Heatmap and OpenPose, respectively. For Subject 3's elbows, the second most challenging category, we saw an overall increase in performance by $38.2 \pm 17.7\%$ against Caffe-Heatmap and $11.3 \pm 5.7\%$ against OpenPose when using our framework. This suggests that our framework can be more consistent within reasonable spatial tolerances for particularly noisy segments of video compared to generic methods.

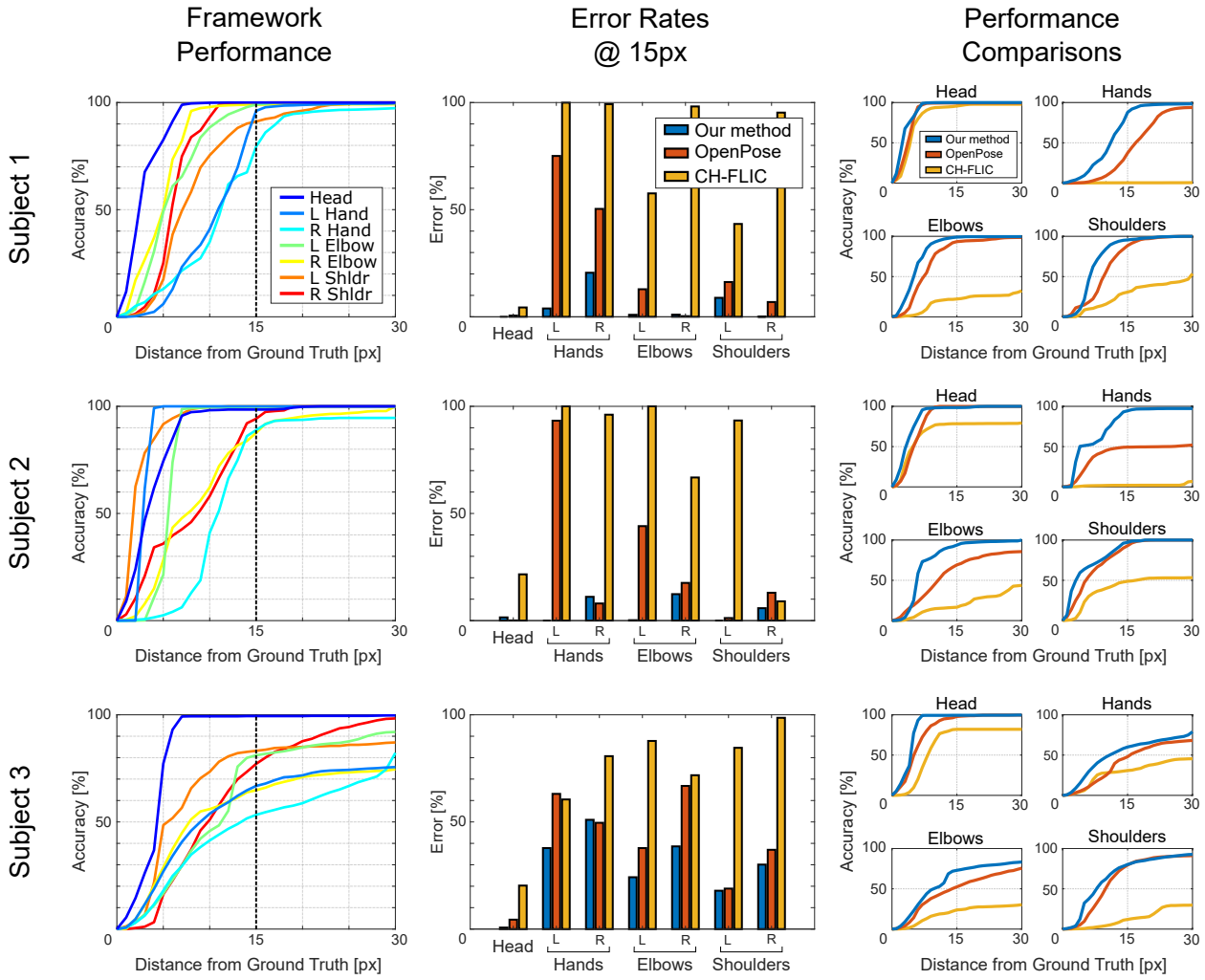

Figure 5.5: **Pose estimation comparison.** Performance of our proposed method as compared to OpenPose and Caffe-Heatmap with FLIC (CH-FLIC) for each subject is shown above. The left column provides the accuracies of each joint at various tolerances from the ground truth using our framework, and the middle shows the error rates at 15 pixels. The right column compares the accuracies of each category of joints averaged between the left and right body parts.

This chapter, in part, has been submitted for publication of the material as it may appear in K. Chen, P. Gabriel, A. Alasfour, C. Gong, W.K. Doyle, O. Devinsky, D. Friedman, P. Dugan, L. Melloni, T. Thesen, D. Gonda, S. Sattar, S. Wang, and V. Gilja, "Patient-specific pose estimation in clinical environments," *IEEE Journal of Translational Engineering in Health and Medicine*, 2018. The thesis author was the primary investigator and author of this material. Reprinted with permission.

Chapter 6

Conclusion

In this thesis, we presented several extensions onto an existing pose estimation framework to improve posture tracking in clinical environments. By extracting images from periods of movement and idleness across the entire span of a patient's dataset, we can construct a subset of training frames that captures a diverse collection of postures for a patient-specific CNN model. Furthermore, by accounting for the frequent lighting changes often found in these environments and by refining the predicted trajectories through a Kalman filter with trained noise parameters, our framework can provide more reliable annotations on a patient's pose in these settings when compared to general pose estimation frameworks.

Our framework relies solely on low-resolution RGB images to be implemented and therefore can be used by anyone with a means of recording RGB video. In addition, our augmentations can be potentially adopted to improve other pose estimators, and our framework is capable of running in real-time after training as a consequence of the Kalman filter's causality. We have open-sourced our standalone PatientPose toolbox,[1] and we encourage others to use our framework for their own experimental or clinical studies or to apply and build upon our methods. However, we suggest that the trade-off between PatientPose and generalized frameworks should be considered before use. In particular, although we have demonstrated the potential to substantially improve posture estimation quality with our add-ons, we note that our framework's

[1] *https://github.com/TNEL-UCSD/PatientPose*

upfront cost of labeling and training a separate CNN model for each patient is greater. Frameworks that are built independent from subjects and environments are often prepackaged with general models that can be applied to patient data right away without any additional work, and therefore may be the preferred choice for those seeking an immediate solution. However, for others who require a more custom approach which can result in a higher consistency and accuracy of pose annotations in these environments, we encourage them to look into PatientPose as a means to extend beyond current general methods.

Work in this field will continue to expand with the increasing desire for automated behavioral labels, since these labels can be informative for both research studies and clinical applications. Analysis of neural correlates to natural behaviors extracted from pose estimates, for example, could enable more robust brain-computer interfaces that would assist those with motor disabilities. In addition, automated patient tracking can provide a way to monitor patient safety, and could improve current motor scoring assessments, overall patient management, and the effectiveness of treatment protocols. Such studies and applications all seek to improve the quality of our health care.

6.1 Future Work

The trade-offs between general frameworks and PatientPose directly motivate future work that could explore the use of insights from generalized frameworks in order to reduce the upfront efforts per patient, or to develop a framework for "hospital-specific" models that generalize across patients within the same hospital. In particular, hospitals frequently rotate patients in and out of rooms, and new patients that require automated monitoring may arrive faster than researchers or clinicians can create patient-specific CNN models. To address this, a framework for "hospital-specific" models could be developed using the insights gained from this thesis. Hospitals may have features that are specific to their environments (e.g., bed sheets, hospital gowns, room decor), and a CNN model that is generalized across all patients within the same hospital could provide similar tracking performance by capturing those features within the hospital

model. In addition, as new patients arrive, a researcher or clinician could manually annotate a subset of new patient data and retrain their hospital's CNN model in order to expand its encapsulation of hospital-specific features. While a deeper study on the potential trade-offs will need to be made before drawing any conclusions, such a method that casts a wider net could provide a more practical "HospitalPose" framework for research studies and clinical applications within the same clinical environment.

Another direction of potential future work could take advantage of the Microsoft Kinect's depth sensor for an additional dimension of patient posture information. This depth information could either be used in tandem with our PatientPose framework as a post-processing step to refine predicted joints coordinates (e.g., background subtraction), or be incorporated within the model training process using a modified CNN architecture that instead regresses 3D joint confidence heatmaps to provide (x, y, z) joint coordinates relative to the camera. Outside of this thesis, we have already looked into improving PatientPose with depth information, and we demonstrate the viability of depth-based tracking in Appendix A by showing preliminary results of patient hand tracking using only data captured by the Kinect's depth sensor. These results suggest that a combination of the two modalities could further improve patient tracking performance.

Finally, the work presented in this thesis will be used by my colleagues to conduct research studies between neural correlates and natural human behaviors. More specifically, we are interested in exploring the possibility of decoding left and right hand movements from neural activity recorded from cortical implants. Our previous methods for labeling left and right movements in patient video used optical flow on predefined regions to detect flow magnitude; however, false-positives were often triggered by outside influences such as camera bumps, shifts in lighting, environmental objects, and non-subject interferences. Posture annotations from recorded patient video can extract a greater level of detail in patient behaviors as opposed to optical flow, and we show the viability of pose estimates for such an application in Appendix A by generating preliminary results of behavior extraction via an actogram of patient hand movements.

Appendix A

Preliminary Results of Future Work

A.1 Depth-Based Tracking

In the work presented in this thesis, we only considered RGB frames in order to build around Caffe-Heatmap to improve pose estimation performance in clinical environments. This was done to create an extended framework with modular components that other researchers could pick and choose from to improve their own pose estimation framework, rather than alter the core architecture. However, we note that relying only on RGB frames limits our framework's potential, and we are continually seeking to improve upon our framework's tracking performance. One possible direction of future work includes taking advantage of other recorded modalities in addition to RGB frames, such as depth and infrared images. In particular, depth images provide an extra dimension of information that RGB frames lack and could enable access to techniques such as background subtraction to improve patient detection, and infrared images could be used in place of CLAHE to track patient postures in dim lighting conditions or complete darkness. Although we have yet to work with infrared data, our preliminary results for depth-based hand tracking suggest that using depth data in tandem with PatientPose could potentially improve the performance of posture tracking in clinical environments even further (Figure A.1). Our results also serve as a proof-of-concept for patient hand tracking with only a depth sensor.

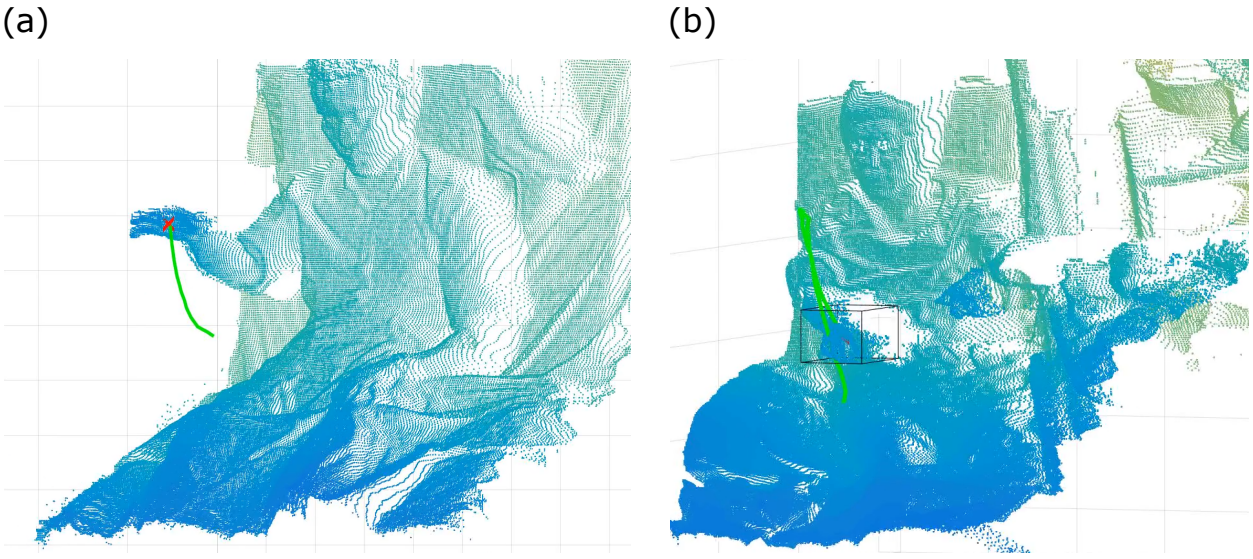

Figure A.1: **Depth-based hand tracking.** Depth frames were first converted to point clouds to work in 3-dimensional space. We show preliminary results of depth-based hand tracking using (a) data of a colleague captured in a mock hospital room and (b) data of Subject 1 captured in the NYU epilepsy monitoring unit.

To track the position of a subject's hand using a depth sensor, frames were first converted to 3D point clouds. The Microsoft Kinect v2 records depth images at a dimension of 512x424, and the value at each (x,y) index indicates the distance from the sensor to an unobstructed object (in arbitrary units). With this, depth images were projected onto a 3D world frame using MATLAB's *Computer Vision System* toolbox to create the point clouds. A bounding box was then manually selected in front of the subject for a desired region of tracking, and the position of a hand was taken to be the average (x,y,z) coordinates of the points within the selected area. This method was tested on data recorded in both a mock hospital room and an academic epilepsy monitoring unit (Figure A.1). While this preliminary work is simple in its idea, our results motivate future work that could incorporate depth information into the PatientPose framework.

A.2 Patient Actograms from Pose Estimates

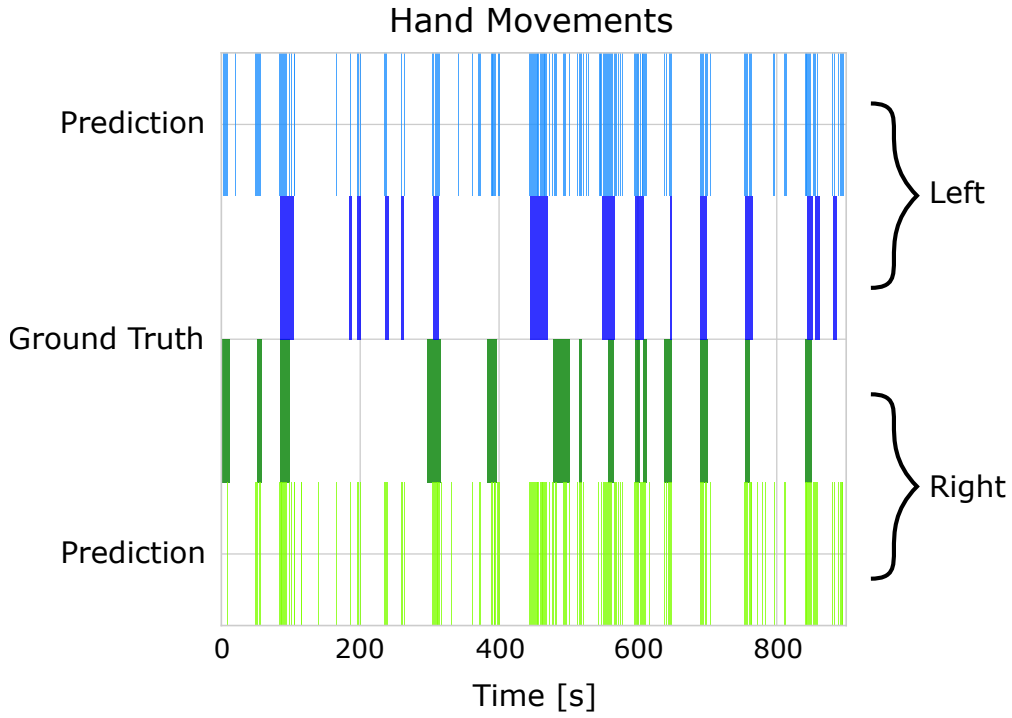

Figure A.2: **Actogram of left and right hands.** Preliminary results of detecting the onset of patient left and right hand movements by using a threshold on extracted pose estimates. This motivates future work on extracting high-level contextual labels of patient behavior from pose estimates for research studies and clinical applications. This 900 second subset was derived from Subject 1's dataset.

Estimated joint coordinates of continuous patient video can provide a rich amount of information regarding patient movements and behaviors, and in this preliminary work we used extracted pose estimates from a 900 second subset of patient video to detect the onset of left and right hand movement. Using a threshold similar to the one described by Wang *et al.* [2] on the pose estimates, we then compared the predicted onsets of hand movement against manually obtained ground truth labels through an actogram (Figure A.2). The results above suggest that extracting a deeper context of patient behavior is possible with posture annotations, and future work could improve the accuracy of such movement detections. In particular, some of my lab colleagues are currently experimenting with long short-term memory (LSTM) networks to increase this prediction accuracy and have seen encouraging preliminary results.

References

[1] Paolo Gabriel, Werner K. Doyle, Orrin Devinsky, Daniel Friedman, Thomas Thesen, and Vikash Gilja. Neural correlates to automatic behavior estimations from RGB-D video in epilepsy unit. In *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016.

[2] Nancy Xin Ru Wang, Ali Farhadi, Rajesh Rao, and Bingni Brunton. AJILE movement prediction: Multimodal deep learning for natural human neural recordings and video. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[3] Jose M Carmena, Michail A Lebedev, Roy E Crist, David M O'Doherty, Miguel A Nicolelis, and Others. Learning to control a brain-machine interface for reaching and grasping by primates. *Public Library of Science (PLOS) Biology*, 1:193–208, 2003.

[4] Meel Velliste, Sagi Perel, Chance Spalding, Andrew S. Whitford, and Andrew B. Schwartz. Cortical control of a prosthetic arm for self-feeding. *Nature*, 458:1098–1101, 2008.

[5] John K. Chapin, Karen A. Moxon, Ronald S. Markowitz, and Miguel A. L. Nicolelis. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience*, 2:664–670, 1999.

[6] Vikash Gilja, Paul Nuyujukian, Cindy A. Chestek, John P. Cunningham, Byron M. Yu, Joline M. Fan, Mark M. Churchland, Matthew T. Kaufman, Jonathan C. Kao, Stephen I. Ryu, and Krishna V. Shenoy. A high-performance neural prosthesis enabled by control algorithm design. *Nature Neuroscience*, 15:1752–1758, 2012.

[7] Leigh R. Hochberg, Mijial D. Serruya, Gerhard M. Friehs, Jon A. Mukand, Maryam Saleh, Abraham H. Caplan, Almut Branner, David Chen, Richard D. Penn, and John P. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442:164–171, 2006.

[8] Christopher C. Goetz. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations. *Movement Disorders*, 18, 2003.

[9] Jean Crosetto Deitz, Deborah Kartin, and Kay Kopp. Review of the Bruininks-Oseretsky test of motor proficiency, second edition (BOT-2). *Physical & Occupational Therapy In Pediatrics*, 27:87–102, 2007.

[10] Axel R Fugl-Meyer, Lisbeth Jaasko, Ingegerd Leyman, Sigyn Olsson, and Solveig Steglind. The post-stroke hemiplegic patient: A method for evaluation of physical performance. *Scandinavian Journal of Rehabilitation Medicine*, 7:13–31, 1975.

[11] Janet H. Carr, Roberta B. Shepherd, Lena Nordholm, and Denise Lynne. Investigation of a new motor assessment scale for stroke patients. *Physical Therapy*, 65(2):175–180, 1985.

[12] Christina Strohrmann, Rob Labruyere, Corinna N. Gerber, Hubertus J. van Hedel, Bert Arnrich, and Gerhard Troster. Monitoring motor capacity changes of children during rehabilitation using body-worn sensors. *Journal of Neuroengineering and Rehabilitation*, 10:83, jul 2013.

[13] Avinash Parnandi, Eric Wade, and Maja J. Matarić. Motor function assessment using wearable inertial sensors. *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, aug 2010.

[14] Dheeraj Kumar, Jayavardhana Gubbi, Bernard Yan, and Marimuthu Palaniswami. Motor recovery monitoring in post acute stroke patients using wireless accelerometer and cross-correlation. *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013.

[15] Jayavardhana Gubbi, Aravinda S. Rao, Kun Fang, Bernard Yan, and Marimuthu Palaniswami. Motor recovery monitoring using acceleration measurements in post acute stroke patients. *BioMedical Engineering OnLine*, 12, apr 2013.

[16] Alfredo Lucas, John Hermiz, Jamie LaBuzetta, Yevgeniy Arabadzhi, Navaz Karanjia, and Vikash Gilja. Use of accelerometry to monitor motor impairment to unilaterally impaired stroke patients. *Unpublished*.

[17] Michael Schukat, David McCaldin, Kejia Wang, Guenter Schreier, Nigel H. Lovell, Michael Marschollek, and Stephen J. Redmond. Unintended consequences of wearable sensor use in healthcare. *International Medical Informatics Association (IMIA) Yearbook of Medical Informatics*, pages 73–86, 2016.

[18] Alberto Verrotti, Pasquale Striano, Vincenzo Belcastro, Sara Matricardi, Maria Pia Villa, and Pasquale Parisi. Migralepsy and related conditions: Advances in pathophysiology and classification, 2011.

[19] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[20] Rob Fergus, George Williams, Ian Spiro, Christoph Bregler, and Graham W. Taylor. Pose-sensitive embedding by nonlinear NCA regression. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2010.

[21] Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W. Taylor, and Christoph Bregler. Learning human pose estimation features with convolutional networks. In *Proceedings of the International Conference on Learning Representations*, 2014.

[22] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. MoDeep: A deep learning framework using motion features for human pose estimation. In *Proceedings of the Asian Conference on Computer Vision*, 2015.

[23] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014.

[24] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann Lecun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. *Computing Research Repository (CoRR)*, 2013.

[26] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. Personalizing human video pose estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, nov 2016.

[27] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, nov 2017.

[28] Felix Achilles, Alexandru Eugen Ichim, Huseyin Coskun, Federico Tombari, Soheyl Noachtar, and Nassir Navab. Patient MoCap: Human pose estimation under blanket occlusion for hospital monitoring applications. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2016.

[29] Shuangjun Liu, Yu Yin, and Sarah Ostadabbas. In-bed pose estimation: Deep learning with shallow dataset. In *arXiv Preprint*, nov 2017.

[30] Vasileios Belagiannis, Xinchao Wang, Horesh Ben Shitrit, Kiyoshi Hashimoto, Ralf Stauder, Yoshimitsu Aoki, Michael Kranzfelder, Armin Schneider, Pascal Fua, Slobodan Ilic, Hubertus Feussner, and Nassir Navab. Parsing human skeletons in an operating room. *Machine Vision and Applications*, 27, 2016.

[31] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. A multi-view RGB-D approach for human pose estimation in operating rooms. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, jan 2017.

[32] Alexandros Chaaraoui, José Padilla-López, Francisco Ferrández-Pastor, Mario Nieto-Hidalgo, and Francisco Flórez-Revuelta. A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. *Sensors*, 14(5):8895–8925, 2014.

[33] Francis Seung-hyun Baek. Autonomous patient safety assessment from depth camera based video analysis. *Dissertation at UC San Diego*, 2016.

[34] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems*, pages 474–485. 1994.

[35] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart Ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39:355–368, 1987.

[36] James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. Automatic and efficient human pose estimation for sign language videos. *International Journal of Computer Vision*, 110:70–90, 2014.

[37] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, may 2014.

[38] Ben Sapp and Ben Taskar. MODEC: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[39] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[40] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA)*, 2003.

[41] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014.

[42] Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35, 1960.

[43] Aleksandr Aravkin, James V. Burke, Lennart Ljung, Aurelie Lozano, and Gianluigi Pillonetto. Generalized Kalman smoothing: Modeling and algorithms. *Automatica*, 86:63–86, 2017.

[44] Byron M Yu and Krishna V Shenoy. Derivation of Kalman filtering and smoothing equations. Technical report, Stanford University, 2004.

[45] Ramsey Faragher. Understanding the basis of the Kalman filter via a simple and intuitive derivation. *IEEE Signal Processing Magazine*, 29:128–132, 2012.

[46] Greg Welch and Gary Bishop. An introduction to the Kalman filter. In *Proceedings of the ACM Special Interest Group on Computer Graphics (SIGGRAPH)*, 2001.

[47] X. Rong Li and Vesselin P. Jilkov. Survey of maneuvering target tracking. Part I: Dynamic models. *IEEE Transactions on Aerospace and Electronic Systems*, 2003.

[48] Pieter Abbeel, Adam Coates, Michael Montemerlo, Andrew Y. Ng, and Sebastian Thrun. Discriminative training of Kalman filters. In *Proceedings of Robotics: Science and Systems I*, 2005.

[49] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.