Multiperson Pose Estimation using Thermal and Depth Modalities

Rishab Mehra Meena Chetty John Kamalu
Department of Computer Science, Stanford University
{rishab,mchetty,jkamalu}@cs.stanford.edu

Abstract

Pose Estimation is a high level task that can help us perform other tasks such as activity recognition and risk detection better. In hospitals, senior homes and other sensitive environments, we often do not have access to RGB data due to privacy concerns. In this paper, we explore state of the art results on RGB pose estimation using pose machines, then we create a novel dataset in our partnered Senior Home with only thermal and depth modalities, which have been approved by the Senior Home authorities. We plan on transferring RGB pose models to use only these modalities. We achieve good qualitative results on the Coco Dataset, and the Senior Home Dataset.

1. Introduction

Pose Estimation in general is a widely studied research field [4][11], with some of the biggest machine learning competitions, such as the Coco Keypoints Challenge [10], widely contested. However, Pose Estimation in Hospitals and Senior homes is not studied much. One of the main reasons for this scarcity is privacy concerns in these highly sensitive areas. In this section, we discuss these privacy concerns, and our possible solution to these concerns.

1.1. Privacy Concerns

RGB data is invasive. It can easily perform identity recognition tasks such as face recognition [3]. Techniques such as blurring can be used, but tasks such as person tracking [5] can still be performed and construed as invasive. Further, the original image must be saved somewhere before it can be blurred. Due to these reasons, hospitals and senior home boards are currently against leveraging RGB data. However, even if they do agree to allowing RGB data collection, it requires patients to sign off privacy waivers as done by [1]. This is simply impossible in certain scenarios, such as if the patient is unconscious. Pose estimation can be beneficial in hospitals, however, in order to monitor a patient's condition without having to have people present. Therefore, alternatives are important to explore in order to be able to conduct pose estimation without invading privacy.

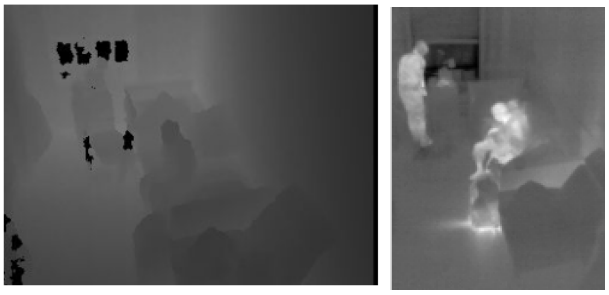

Figure 1. Nearest Neighbor frames from the same room in a Senior Home of depth and thermal modalities respectively

1.2. Depth and Thermal Modalities

Depth and thermal modalities, as seen in Figure 1, are less invasive, and cannot perform tasks such as face recognition. Our goal is to use these modalities to perform state of the art activity detection in Hospitals and Senior homes. In this paper we study how we can maximally utilize these modalities to perform pose estimation, which in turn helps in performing activity detection [7].

2. Related Work

2.1. Pose Estimation in General

Pose estimation is useful for predicting future poses or determining activities. Currently, most research involving pose estimation uses RGB data. In Shih-En Wei et al., the authors create a convolutional pose machine where each stage builds on 2D belief maps of body part locations from the previous stage [11]. In Zhe Cao et al., the authors couple body part confidence maps and affinity fields that measure the relationship between parts in a 2-branch convolutional neural network to predict poses with a mean Average Precision that is 8.5% better than previous models [4][11]. The input images for both of these models are RBG images, where full facial identification is possible and thus a privacy concern.

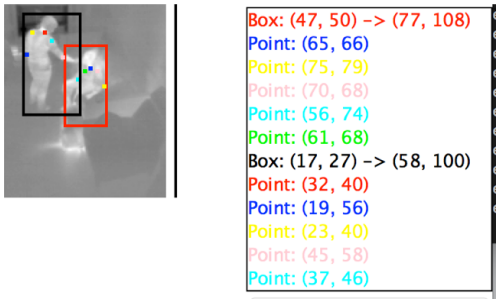

Figure 2. Thermal image labeled with our labeling tool. For each person, we first have the bounding box, and then joints in the order: Neck, L-Elbow, L-Shoulder, R-Elbow, R-Shoulder

2.2. Pose Estimation in the Clinical Setting

More and more, the utility of accurate pose estimation is being introduced into medical and clinical settings, despite the strict privacy requirements.

In Kadkhodamohammadi et al. [8], the authors explore the use of pose estimation of surgical and clinical teams in the operating room (OR), with beneficial applications ranging from performance assessment to collision avoidance. The authors propose a 3D, part-based model that makes use of RGB-D imaging. The model computes the pairwise confidence scores between all parts and locations and then assigns a pairwise weight between all parts, from which pose is reconstructed. Among the authors' most significant accomplishments is the extent to which they were able to optimize a 3-D state space.

In Achilles et al. [2], the authors propose a model to estimate the pose of immobile or bedridden patients as well as patients often covered by an occluding blanket. The model takes depth video as input, and for each frame predicts all joint locations with a compound convolutional/recurrent network architecture. In LSTM form, these predictions are then passed as input to the subsequent frame's RNN componenent to control for temporal consistency.

3. Data Set

3.1. Data Description

Thanks to our partnership with Onlock, we collected two days of Data at a Senior Home in San Francisco. In two rooms of the senior home, we collected thermal data at 8 frames/second and depth data at 24 frames/second. Since depth is collected at a faster framerate, for every thermal image we find the nearest neighbor depth image, and discard the rest of the depth images.

3.2. Data Collection and Labeling

We created a data labelling tool for pose by modifying an existing tool for labeling object detection data [9]. For thermal images, we labeled the bounding box of the person and 5 joints (Neck, L-Elbow, L-Shoulder, R-Elbow, R-Shoulder) as seen in Figure 2. We used the nearest neighbor depth image as a feature, without any labeling, since we believe that pose is clearer in thermal images.

We labelled 1000 thermal images, and found the corresponding nearest neighbor depth images. We used 700 images for our training set, 100 for our validation set, and 200 for our test set.

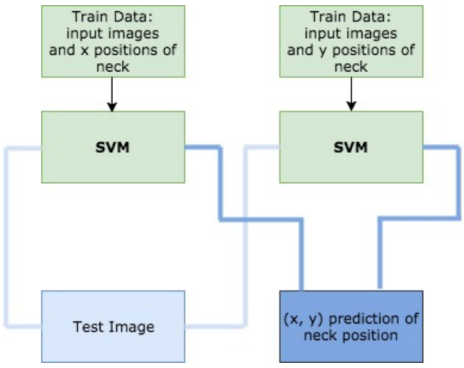

Figure 3. Pipeline of the dual SVM baseline model.

4. Methods

4.1. SVM

For our baseline, we created an SVM model for predicting the position of the neck joint of a person. We had 2 SVM classifiers - one for the x coordinate of the neck, and one for the y coordinate of the neck. As shown in Figure 3, our model predicts the position of the neck joint in an image by running the image through both classifiers and constructing a 2-D coordinate from the corresponding x and y predictions.

4.2. Convolution Pose Machines with Confidence Maps and Affinity Fields

We use a modified version of the model stated in [4], as it is the state of the art pose estimation model. The model is a significantly improved version of the original pose machines [11] as it uses both confidence maps and part affinity fields over multiple stages to predict pose.

The model's confidence value $c(j, l)$ is defined as the relative likelihood that joint j appear at location l. The confidence map C_j is defined as the 2-D cartesian representation of confidence values $c(j, l)$ for all coordinate locations l such that $C_j \in \mathbb{R}^{w \times h}$. The set S is defined as the set of confidence maps C_j for all joints j.

The Part Affinity Fields (PAFs) used in this model map relationships between body parts. These fields are 2-D vector fields that represent relational positions of body parts in

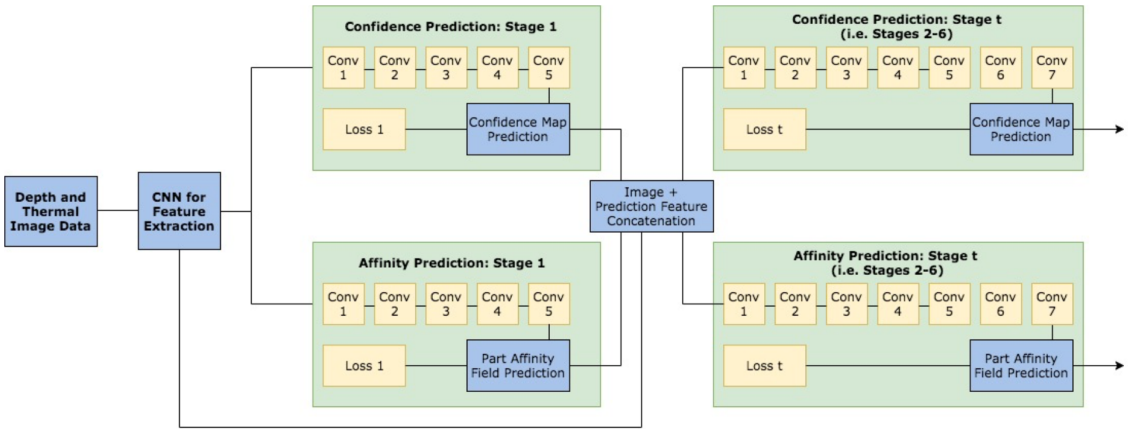

Figure 4. Full 7 stage model , extracting image features at the first stage, and refining part affinity fields, and confidence maps in the remaining 6 stages.

images. If the model is estimating the positions of n limbs in an image, affinity field A_j corresponds to the association of limb j with other body parts in the image where $j \in \{1, ..., n\}$.

The model architecture proposed in Cao et al. [4], and which we use here starts off by extracting features from a VGG-19 model. These features are then passed through a sequence of branching stages wherein one branch predicts/refines the set of confidence maps S and the other branch predicts/refines the set of part affinity fields L. At any stage s_i for all $i > 1$, both branches take as input the concatenation of the original feature mapping F extracted from the image as well as the output of both branches from the immediately preceding stage. Each branch consists of a convolutions neural network, the structure of which can be independently configured during hyperparameter tuning.

At the end of each stage of the model, an L2 loss function is applied to each branch measuring the error between the branch's predicted and ground truth confidence maps or part affinity fields. Both losses are weighted using binary masks such that if an annotation is missing at a particular location, that location's loss will count toward the total loss. The total loss at a stage is the sum of both losses.

Note that the ground truth confidence maps and ground truth part affinity fields are not predefined – the model is only given bounding boxes for each person and coordinate locations for each joint in the image. For this reason, the following ground truth definitions for confidence mappings and part affinity fields as defined in Cao et al. [4] are used.

Let the confidence mapping $C_{j,p}(l)$ represent the likelihood that the j-th joint of person p is found at location i. We express $C_{j,p}(l)$ as a Gaussian distribution parameterized by the distance between l and $l_{j,p}$ the ground truth location for the j-th joint of person p:

$$C_{j,p}(l) = exp(-\frac{\|l - l_{j,p}\|^2}{\sigma^2})$$

To generate the ground truth distribution for joint j independent of person p, we define $C_j(l)$ as the max over p of all confidence maps $C_{j,p}(l)$.

Let the part affinity field $L_{c,p}(l)$ be defined as the ground truth unit vector pointing in the direction of $limb(c, p)$, the c-th limb of person p, *iff* the location l lies within the space occupied by $limb(c, p)$ and *zero* otherwise. To generate the ground truth part affinity field for limb c independent of person p, we define $L_c(l)$ as the average over all $L_{c,p}(l) \neq 0$.

After the last stage of the model, we perform non-maximum suppression on the resultant confidence map to pinpoint potential body parts of interests, which are ideally joints. For each limb, we look at its respective affinity field and calculate the corresponding integral along the line connecting detected joints, and then compare the result to the true limb.

The model predicts confidence maps and affinity fields of size (15, 20) because of the convolutional transformations, which we then upscale to the size of the image using the cv2 package. Then we calculate the loss between the confidence maps and affinity fields at every stage, versus their ground truth.

5. Experiments and Results

5.1. Coco Dataset using Pose Machines [4]

We trained a 7 stage pose machine using the published codebase in [4] with both confidence maps and part affinity fields model on the 2017 Coco Dataset [10], which is an object detection and segmentation dataset including 250,000

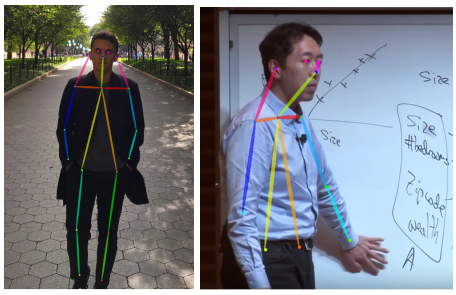

Figure 5. Qualitative results on images of one of our team members, John, and Andrew Ng when using our model trained on the Coco 2017 Dataset.

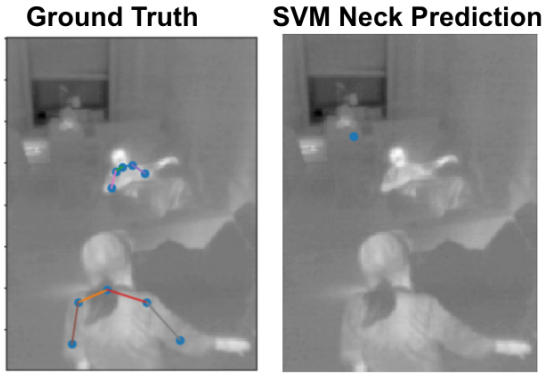

Figure 6. Qualitative result for the neck join using the SVM model.

people and respective keypoint annotations. The model trained till 50 epochs and then the loss flattened out. Qualitative results can be seen in Figure 3. The hyperparameters we used were: weight decay was set to 6e-4, momentum was set to 0.95, batch size was set to 60, learning rate was set to 5e-5, and step size was set to 0.3 every 10 epochs.

5.2. Senior Home Dataset Using SVM

On average, the SVM model predicted the position of the neck joint within 30 pixels of the true position. Figure 6 shows a labelled ground truth image and our baseline model's prediction of a single neck joint. While demonstrably better than those of a random model, these predictions are too inaccurate to be considered useful.

5.3. Senior Home Dataset Using using Adapted Pose Machines [4]

We attempted to optimize by varying such hyperparamters as learning rate, convolution filter sizes, and number of iterations for thermal and depth data. The loss represents the cross entropy loss averaged over the 12 losses (2 for each stage - one for the predicted confidence map, and one for the predicted affinity field). The hyperparameters we used were as follows: weight decay was set to 1e-7, momentum was set to 0.95, batch size was set to 30, learning

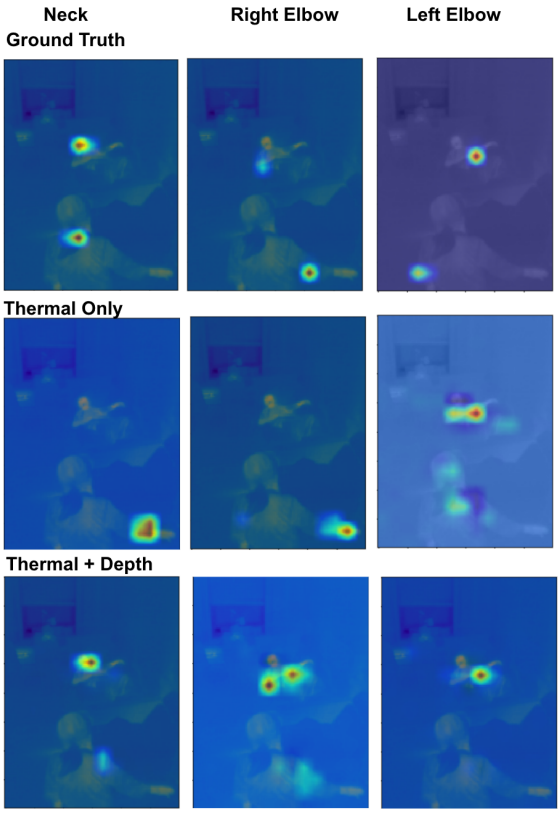

Figure 7. Qualitative confidence maps on an image with 2 people. Ground truth results from using thermal only and from using thermal+depth can be seen.

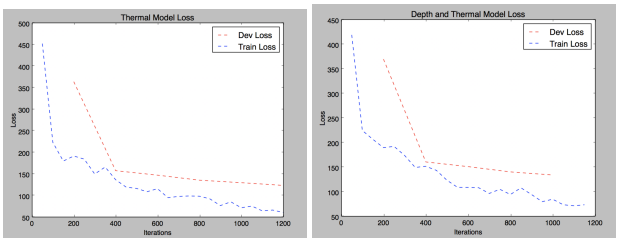

Figure 8. Loss graphs for thermal only vs thermal depth modalities using the full model. The plot for the dev set is not smooth because we calculated it only once per 200 iterations.

rate was set to 1e-3, and step size was set to 0.5 every 10 epochs.

We performed experiments using only thermal data (3 channels of thermal) and using both thermal and depth data (2 channels of thermal and 1 channel of depth). Note that the ground truth labels were available only for thermal data in either case.

Loss graphs for the training and dev set can be seen in Figure 8. Test loss can be seen in Table 1.

The qualitative results can be seen in Figure 7. As we can see, the thermal-depth predictions are generally better than the thermal-only predictions. Thermal-only misclassi-

Stage	Thermal Only	Thermal+Depth
Stage 1	19.2	19.1
Stage 2	17.1	18.2
Stage 3	18.2	16.6
Stage 4	17.5	15.8
Stage 5	17.6	15.3
Stage 6	17.2	14.5
Total	106.8	99.5

Table 1. Average Loss on test set at the various stages of the model using thermal images only and using thermal + depth. Loss at every stage is the sum of the loss of the affinity field and confidence map.

fies the neck, while thermal-depth gives good prediction for the necks. Thermal-only finds the right elbow for 1 person, while thermal-depth nearly finds the right elbow for both (as well as the left elbow, a false positive). For the left elbow, thermal-depth perfectly predicts the elbow of the person sitting on the chair, but misses the person standing.

To improve our results, we need more data, since the original Coco Dataset for which this model was made had over 100,000 images, while our dataset has only 1000 images. We don't need 100,000 images, since our data has less variation, but we still do need a few thousand more to get ideal results.

6. Future Work

6.1. Deconvolution

Currently, we get confidence maps and affinity fields of resolution 15 by 20, and then we upscale the them to 120 by 160 (size of image) to compare to the ground truth confidence maps. In the future, we want to try upscaling using deconvolutions (instead of simply scaling using concatenation), and see how the higher resolution confidence maps affect the results.

6.2. Incorporating YOLOv2 for Person Detection

We plan to train YOLOv2 [6] on bounding boxes for person detection, and then run the pose model only in windows where YOLOv2 detects a person. This way, joints will be predicted based on the location of people rather than solely on the constraints of the image.

6.3. 3D Pose Machines and 2D Pose Machines with LSTMs

If we are able to get high MAP on the Senior Home Data Set quickly, we will also work on 3D pose machines for videos. These 3D pose machines will use 3D convolutions instead of 2D convolutions, and use the temporal information to improve frame level pose.

We also want to compare 3D pose machines to 2D pose machines with LSTMs, to see which ones are better for predicting pose.

7. Acknowledgements

We thank the members of the Stanford Vision and Learning Lab, particularly Serena Yeung, for allowing us to use the Senior Home Data for the project, and providing us with GPU resources to train the models. We also thank Onlock for allowing us to collect data at their Senior homes. Finally, we thank Albert Haque and Sanyam Mehra for their useful comments throughout the quarter.

8. Contributions

8.1. Rishab Mehra

I helped with the getting the data for the Senior Home Dataset, creating the dataset with nearest neighbor thermal and depth images, setting up the Pose Machine model from [4] in Pytorch, running experiments on the model, and labeling thermal images.

8.2. Meena Chetty

I helped with interpreting the Pose Machine model, labeling thermal images, combining thermal and depth image data, pre-processing and post-processing model data and results, and data visualization.

8.3. John Kamalu

I helped with interpreting the Pose Machine model, modifying the labeling tool, labeling thermal images, implementing the SVM baseline, and data visualization.

References

[1] Abdolrahim, A. Kadkhodamohammadi, M. Gangi, N. Mathelin, and Padoy. Articulated clinician detection using 3d pictorial structures on rgb-d data. In *CVPR*, 2016.

[2] F. Achilles, A.-E. Ichim, H. Coskun, F. Tombari, S. Noachtar, and N. Navab. Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications. pages 491–499, 2016.

[3] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[5] Eldar, M. Insafutdinov, L. Andriluka, S. Pishchulin, E. Tang, B. Levinkov, B. Andres, and Schiele. Arttrack: Articulated multi-person tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Joseph, A. Redmon, and Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, 2016.

[7] Jun, A. Liu, D. Shahroudy, G. Xu, and Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *CVPR*, 2016.

[8] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy. Articulated Clinician Detection Using 3D Pictorial Structures on RGB-D Data. 2016.

[9] puzzledqs. Bbox-label-tool. `https:https://github.com/puzzledqs/BBox-Label-Tool`, 2014.

[10] Tsung-Yi, M. Lin, S. Maire, L. Belongie, R. Bourdev, J. Girshick, P. Hays, D. Perona, L. Ramanan, P. Zitnick, and Dollr. Microsoft coco: Common objects in context. In *CVPR*, 2014.

[11] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.