Monocular Real-time Hand Shape and Motion Capture using Multi-modal data

1 Introduction

Architecture: IKnet (inverse kinematics) + DetNet. DetNet produces 3D predictions which IKnet uses to regressing joint rotations
Contributions

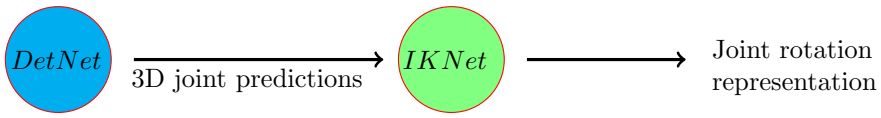
1. A new learning based approach for monocular hand shape and motion capture, which enables the joint usage of 2D and 3D annotated image data as well as stand alone motion capture data
2. An inverse kinematics network that maps 3D joint predictions to the more fundamental representation of joint angles in a single feed forward pass and that allows joint training with both positional and rotational supervision

2 Monocular RGB Images

Does not require depth images (hence monocular) and does not suffer the drawbacks of depth images (sunlight, high power consumption, proximity to sensor)

3 Method

1. Retrieve shape of hand by fitting a hand model to the 3D joint predictions (DetNet)
2. Inverse kinematics (IKnet) takes 3D joint predictions and convert them into a joint rotation representation in an end-to-end manner



4 Joint detection with DetNet

DetNet works on a single RGB image and outputs root-relative scale-normalized 3D hand joint prediction as well as 2D joint predictions in the image space. Has 3 stages: feature extractor, 2D detector, and a 3D detector

4.1 Feature Extractor

Architecture mainly from ResNet50, weights initialized with Xavier initialization. Input image is 128× 128 vector and output is a feature volume F of dimension 32× 32× 256.

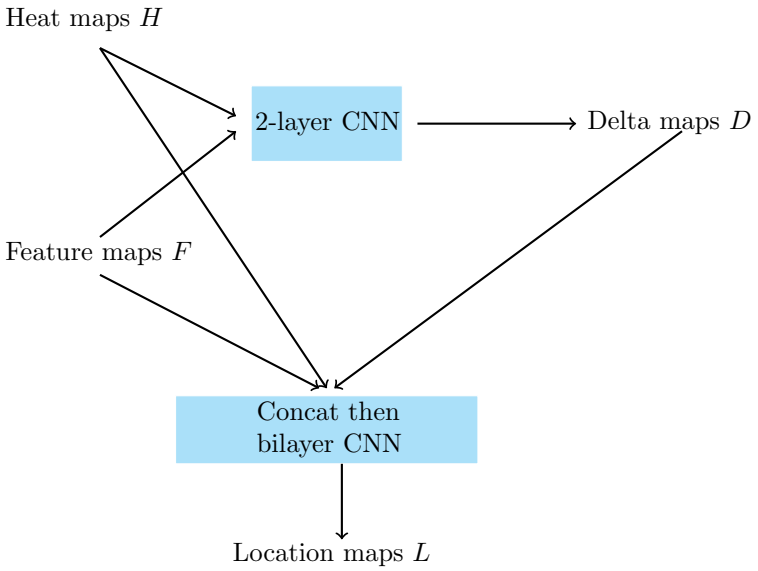
4.2 2D detector

Compact 2-layer CNN that takes feature volume F and outputs 21 heat maps H_j corresponding to the $J = 21$ joints. A pixel in H_j encodes the confidence of that pixel being covered by joint j. Heat maps used for 2D pose estimation, and can be trained using 2D label internet images.

4.3 3D detector

Takes the feature maps F and the heat maps H, and estimates 3D hand joint positions in the form of 21 location maps L. L_j has the same 2D resolution as H_j, and each pixel in L_j encodes joint j's 3D coordinates.

Delta maps encode the bones, and D_b encodes the orientation of bone b, represented by a 3D vector from one joint to the next joint.



4.4 Joint-j location Lookup

To find the location of joint j look the pixels inside the location map L_j and get the pixel with the highest heat value inside H_j. Metacarpophalangeal joint is the root join, and the bone from this joint to the wrist is defined as the reference bone.

5 Loss terms

$$\mathcal{L}_{heat} + \mathcal{L}_{loc} + \mathcal{L}_{delta} + \mathcal{L}_{reg}$$

The loss consists of a regularizer, a loss function for the heat map, a loss function for the location maps, and a loss function for the delta maps.

$$\mathcal{L}_{heat} = ||H^{GT} - H||_F^2$$

the Frobenius norm of the difference from ground truth. To generate the ground truths H_j^{GT} we smooth H_j^{GT} with a Gaussian filter centered at the 2D annotation using a standard deviation of $\sigma = 1$.
For 3D supervised learning we have two terms

$$\mathcal{L}_{loc} = ||H^{GT} \odot (L^{GT} - L)||_F^2$$

and

$$\mathcal{L}_{delta} = ||H^{GT} \odot (D^{GT} - D)||_F^2$$

which measure the difference between ground truth and predicted location maps L and delta maps D, respectively. Since we are only interested in the maxima of the heat maps, we use the element wise matrix product with the heat map (pixels are weighted by the values of the heat map).

6 Hand model and shape estimation

MANO is the mesh model animated by IKnet. MANO can be deformed and posed by the shape parameters $\beta \in \mathbb{R}^{10}$ and pose parameters $\theta \in \mathbb{R}^{21\times 3}$. β controls the shape of the global shape of the hand and θ represents the joint rotations. Together, they allow deforming the mean template $\bar{T} \in \mathbb{R}^{V\times 3}$, where V denotes the number of vertices. The deformed (pre-posed) template is given as

$$\mathcal{T}(\beta, \theta) = \bar{T} + \mathcal{B}_s(\beta) + \mathcal{B}_p(\theta)$$

where $\mathcal{B}_s(\beta)$ is the shape blendshape and $\mathcal{B}_p(\theta)$ is the pose blendshape. The final posed model is given as

$$\mathcal{M}(\theta, \beta) = W(\mathcal{T}(\theta, \beta), \theta, \mathcal{W}, \mathcal{J}(\theta)) \in \mathbb{R}^{V\times 3}$$

where $W(\cdot)$ is a standard linear blend skimming function that takes the deformed template mesh $\mathcal{T}(\beta, \theta)$, pose parameters θ, skinning weights $\mathcal{W}$, and posed joint locations $\mathcal{J}(\theta)$.

6.1 Shape estimation

To find the shape of the hand (as well as the pose described in the paragraph above) we need to estimate the shape parameters β of the MANO model (using the joint position predictions). The hand shape parameters, β, are obtained by minimizing error (summed over bones b)

$$E(\beta) = \sum_b ||\frac{l_b(\beta)}{l_{ref}(\beta)} - l_b^{pred}||_2^2 + \lambda_\beta ||\beta||_2^2$$

1. l_{ref} = length of reference bone in MANO model
2. $l_b(\beta)$ = length of deformed hand model

7 Inverse Kinematics Network IKNet

3D joint locations are not enough to animate hand mesh models, which is useful for CG. CG relies on joint rotations which means we must infer joint rotations from joint locations ⇒ IKnet neural net to infer joint rotations.

7.1 MoCap Data

Data set of pairs of 3D hand joint positions and the corresponding rotation angles. Rotations originally in the axis-angle representation are converted to quaternion representation.

7.2 3DPosData

MoCap data is noiseless which means the IKNet may not be able to generalize for the noisy joint locations from DetNet. DetNet thus produces 3D joint predictions for the training data which is then fed to IKNet.

7.3 IKNet design

7-layer fully-connected neural network with batch normalization and sigmoid activation function except for the last layer with linear activation.

Input is given as $\mathcal{I} = [\mathcal{X}, \mathcal{D}, \mathcal{X}_{ref}, \mathcal{D}_{ref}] \in \mathbb{R}^{4\times J\times 3}$, where $\mathcal{X}$ are the root relative scale-normalized 3D joint positions, $\mathcal{X}_{ref}$ is the same but for rest pose, and $\mathcal{D}$ is the orientation for each bone, with $\mathcal{D}_{ref}$ the same but for rest pose.

Output of IKNet is a quaternion $\hat{\mathcal{Q}} \in \mathbb{R}^{J\times 4}$ describing the global rotation of the