Hand Pose Estimation Using Deep Stereovision and Markov-chain Monte Carlo

Rilwan Remilekun Basaru
City, University of London
London, United Kingdom
Remilekun.basaru.1@city.ac.uk

Chris Child, Eduardo Alonso, Greg Slabaugh
City, University of London
London, United Kingdom
C.Child@city.ac.uk
E.Alonso@city.ac.uk
Gregory.Slabaugh.1@city.ac.uk

Abstract

Hand pose is emerging as an important interface for human-computer interaction. The problem of hand pose estimation from passive stereo inputs has received less attention in the literature compared to active depth sensors. This paper seeks to address this gap by presenting a data-driven method to estimate a hand pose from a stereoscopic camera input, by introducing a stochastic approach to propose potential depth solutions to the observed stereo capture and evaluate these proposals using two convolutional neural networks (CNNs). The first CNN, configured in a Siamese network architecture, evaluates how consistent the proposed depth solution is to the observed stereo capture. The second CNN estimates a hand pose given the proposed depth. Unlike sequential approaches that reconstruct pose from a known depth, our method jointly optimizes the hand pose and depth estimation through Markov-chain Monte Carlo (MCMC) sampling. This way, pose estimation can correct for errors in depth estimation, and vice versa. Experimental results using an inexpensive stereo camera show that the proposed system more accurately measures pose better than competing methods.

1. Introduction

The problem of tracking articulated objects has attracted increasing attention in the field of computer vision, as it provides a natural method of Human Computer Interaction (HCI) [9], [10]. Inference of the pose and gesture of the human hand is an important challenge in this area. Active vision approaches for hand pose estimation using depth sensors such as Leap Motion and Kinect have made considerable progress in recent years. These cameras actively dissipate electromagnetic waves into the scene, probing how far each point in the field of view is away from the imaging device. While active vision techniques provide good shape information and robustness to clutter, they present several limitations, including: large energy consumption, a poor form factor, less accurate near distance coverage, and poor outdoor usage.

In contrast, in this paper we explore the use of passive vision for the estimation of hand pose using a stereovision system composed of adjacent RGB cameras. Such a camera

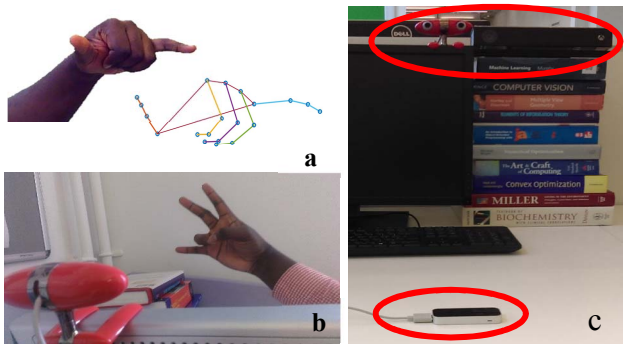

Figure 1: *Using an inexpensive stereo camera RGB images of the hand from two perspectives are captured to regress for hand point 3D position (a). The proposed technique can use a stereo rig system to estimate hand articulation and pose (b). We capture training dataset using RGBD, stereo and hand gesture detection devices (c).*

rig does not project light into the scene, and therefore has complementary advantages to depth imaging, including less energy consumption. However, hand pose estimation in this context is a more challenging computer vision problem, one that has received less attention in the literature. We address this gap by proposing a novel framework that combines jointly optimal depth and hand pose estimation in a unified framework using Markov-chain Monte Carlo (MCMC) sampling and deep learning. Our research is motivated by the possibility of estimating articulation with the input of stereo cameras from an egocentric, stereoscopic perspective. We are inspired by human vision, which can efficiently discern articulations and perform tracking activities with passive, binocular input. As our experiments show, our approach is compatible with inexpensive stereo vision systems, such as the rig shown in Figure 1, to produce robust hand pose inference. The proposed technique also relies on a robust hand segmentation procedure. We do not address hand segmentation in this paper as there is a large body of literature on this subject (see, for example, [1], [21]).

1.1. Contribution

Unlike several approaches to pose estimation from stereo capture that explicitly recover disparity before regressing for the pose in a sequential manner we present a joint optimization approach that is robust against potential errors

in the depth estimation. Thus, this reduces the burden on the pose estimation framework to be robust against erroneous depth recovery. The consequence of our approach is that we iteratively revise for errors in depth proposal. This allows for simultaneous correction of proposed depth estimation and the resulting pose estimation to jointly optimize the likelihood of the depth and hand pose estimation given the stereo input.

Lastly, unlike the work in [14], which utilizes a state-of-the-art tracking method that is sensitive to erroneous initialization and anatomical hand size as discussed in [17], we propose a semi-generative approach that is experimentally proven to work on different sizes and tones of hand without pre-calibration.

The rest of the paper is structured as follows: the next section presents a general survey of related work. Section 3 presents a detailed description of our methodology while Section 4 elaborates on the details of our implementation of the proposed technique. Experiments and results are discussed in Section 5 and we conclude in Section 6.

2. Related Work

Unlike active depth camera based input, less work has been performed on stereo-based passive camera input for hand pose/gesture recognition. Techniques proposed to address stereo based hand pose estimation are largely grouped into two main categories, namely: depth map based and non-depth map. Depth map based methods assume that the mapping between the stereo input and hand pose is strongly based on disparity information being a hidden variable. This is largely influenced by the recent success in robust hand tracking and pose estimation from depth images. These techniques attempt to recover dense or at least a semi-dense depth image before applying state of the art depth based pose estimation. An example of this is [2], where a robust technique that focuses on depth recovery of hand pose is presented, specifically with the aim of later using it for hand pose estimation. [14] also proposed using recovered disparity for pose estimation. It utilizes an Adaptive GMM segmentation [19] to localize the hand skin region before recovering disparity based on stereo matches. Using the estimated hand skin region, it refines the disparity image recovered by constraining the disparity from proposed stereo matches. Finally, hand segmentation is further applied to the final disparity and [18] is used to track hand poses based on the recovered disparity image. A key drawback in this approach is that it assumes that the stereo algorithm will recover disparity/depth with same consistency and accuracy. This is not always the case particularly with a low-quality stereo camera like the one used in this paper. An erroneous disparity recovery will yield a wrong pose.

On the other hand, non-depth based approaches, while still exploiting parallax information, do not attempt to explicitly extract a depth map of the scene. This is typified by the approach presented in [15]. Here a generative hand model approach is used to optimize the appropriate hand pose that yields stereo color consistency between the two cameras. Like most model-driven approaches in hand pose recovery, it does not require the tedious procedure of establishing a robust dataset. However, the approach does require an explicit definition of the anatomical size and hand pose constraint for the skinned model. Also, because of the method's temporal dependency, it is sensitive to the initialization of the pose. Another example is [3]. Here, the pose estimation was preceded by first extracting the hand contour in both images in the stereo pair before matching points along contour in one image to those in the other using dynamic time warping. This allows for the reconstruction of a 3-D contour of the hand, used to establish hand contour tracking for subsequent finger tracking. Again, this approach is sensitive to the starting point selection to determine which pair of points on the contours serve as a seed to subsequent correspondence matching. Nonetheless, this only results in an aggregative tracking of the finger and pose, not providing a dense estimation of the spatial position of the other joints of the hand for a complete hand gesture/pose estimation.

Recently, CNNs have become a prevalent computer vision tool especially in stereo matching and pose estimation from depth images. The work in [7], implements a Siamese network to discriminate between similar and dissimilar patches from stereo pair. The work of [20] and [17] present the use of a CNN to regress for a heat map that indicates the likelihood that a joint will be at a 3D location. Unlike CNNs, Markov-chain Monte Carlo (MCMC) has been an ever-popular machine learning tool. It allows for the sampling in very high dimensional space with no analytical estimation of the probability of such space. Previously, MCMC has been used explicitly in data association and detection [4], [12], [13] and [16]. Inspired by these references, we apply MCMC to stochastically propose depth images that are tested against observed stereo information and prior probability to estimate the hand pose.

3. Methodology

In hand pose estimation, we aim to regress for the spatial location of the different hand joints given a pair of images from a stereo capture of the hand. In this work, we recognize the success of depth data in non-rigid body pose estimation, hence we aim to exploit this as a hidden variable between a stereo image input variable and the spatial pose output. To this end, we conceptualize our problem to jointly solving for two variables: the depth image and the spatial pose of hand joints.

3.1. Stereo-Depth-Pose

For a given stereo image pair, $\boldsymbol{S}$ of a scene of a hand pose, $\boldsymbol{H}$ with a depth image, $\boldsymbol{D}$, we assume that the hand pose

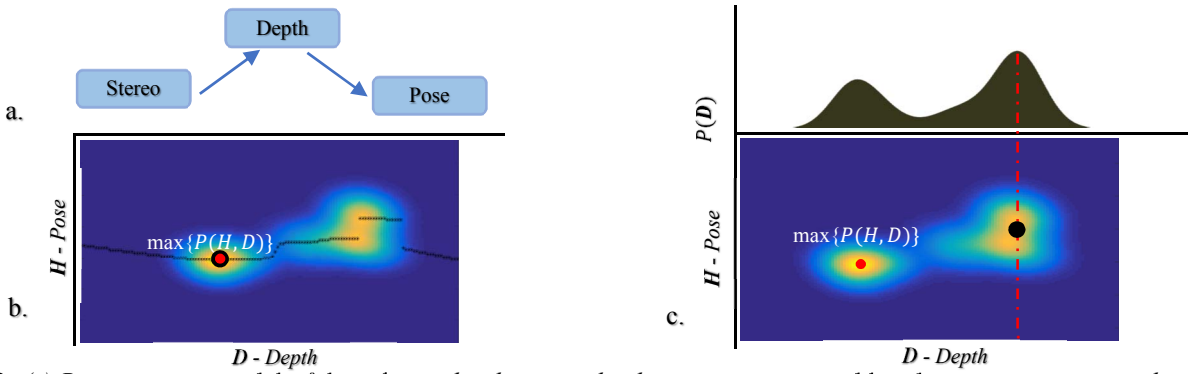

Figure 2: *(a) Bayesian tree model of the relationship between depth, stereo images and hand pose in our proposed model. (c) Illustrates a conventional approach to estimating pose from stereo capture, where the optimum depth is first resolved and then used to factorize the joint probability to identify the maximizing pose. (b) Our approach on the other stochastically propose potential depth solution (along the black line) and then we establish a maximizing pose. This will guarantee identifying the joint maximum point (illustrated with the red dot) with enough depth proposals.*

induces a depth surface, that in turn induces the detected stereo image in a Bayesian tree model. See Figure 2a. Our goal is then reduced to establishing the pose, $\boldsymbol{H}^*$ and depth, $\boldsymbol{D}^*$ values that maximize the posterior distribution of $\boldsymbol{H}$ and $\boldsymbol{D}$ given an observed stereo image pair, $\boldsymbol{S}$.

$$\boldsymbol{H}^*, \boldsymbol{D}^* = \underset{H,D}{\operatorname{argmax}} P(\boldsymbol{H}, \boldsymbol{D}|\boldsymbol{S}) \quad (1)$$

Following from our Bayesian tree model, we assume that $\boldsymbol{H}$ and $\boldsymbol{S}$ are conditionally independent, given $\boldsymbol{D}$. This implies

$$P(\boldsymbol{S}, \boldsymbol{H}|\boldsymbol{D}) = P(\boldsymbol{S}|\boldsymbol{D})P(\boldsymbol{H}|\boldsymbol{D}) \quad (2)$$

and

$$P(\boldsymbol{S}|\boldsymbol{H}, \boldsymbol{D}) = P(\boldsymbol{S}|\boldsymbol{D}). \quad (3)$$

From Bayes' theorem, we can infer that

$$P(\boldsymbol{S}|\boldsymbol{H}, \boldsymbol{D}) = \frac{P(\boldsymbol{H}, \boldsymbol{D}|\boldsymbol{S})P(\boldsymbol{S})}{P(\boldsymbol{H}, \boldsymbol{D})}, \quad (4)$$

and that given Eq. 3 and Eq. 4 we have that

$$P(\boldsymbol{H}, \boldsymbol{D}|\boldsymbol{S}) = \frac{P(\boldsymbol{S}|\boldsymbol{D})\, P(\boldsymbol{H}, \boldsymbol{D})}{P(\boldsymbol{S})}. \quad (5)$$

Note $P(\boldsymbol{H}, \boldsymbol{D}) = P(\boldsymbol{H}|\boldsymbol{D})P(\boldsymbol{D})$, then from Eq. 5 we have that

$$P(\boldsymbol{H}, \boldsymbol{D}|\boldsymbol{S}) = \frac{P(\boldsymbol{S}|\boldsymbol{D})\, P(\boldsymbol{H}|\boldsymbol{D})P(\boldsymbol{D})}{P(\boldsymbol{S})} \quad (6)$$

and that Eq. 1 can be represented as

$$\boldsymbol{H}^*, \boldsymbol{D}^* = \underset{H,D}{\operatorname{argmax}} P(\boldsymbol{S}|\boldsymbol{D})\, P(\boldsymbol{H}|\boldsymbol{D})P(\boldsymbol{D}). \quad (7)$$

The posterior joint probability of $\boldsymbol{H}$ and $\boldsymbol{D}$ yields a very high dimensional space. An intuitive solution to this joint probability will be to first determine the depth image, $\boldsymbol{D}^*$ that best describes the observed stereo image pair, $\boldsymbol{s}$,

$$\boldsymbol{D}^* = \underset{\boldsymbol{D}}{\operatorname{argmax}}\, P(\boldsymbol{S} = \boldsymbol{s}|\, \boldsymbol{D}) \quad (8)$$

before using $\boldsymbol{D}^*$ to resolve for the corresponding pose,

$$\boldsymbol{H}^* = \underset{\boldsymbol{H}}{\operatorname{argmax}}\, P(\boldsymbol{H}|\boldsymbol{D}^*)P(\boldsymbol{D}^*). \quad (9)$$

This is the approach of several papers on hand pose estimation from stereo capture, including [2], [3] and [14]. Here the aim was to first establish a robust depth image given a stereo image capture that can then be used to predict the hand pose. However, this does not fully optimize the pose-depth joint probability space. This is because it assumes that the depth that maximizes $P(\boldsymbol{D})$ coincide with the point (i.e. the pose and depth image) that maximizes in the pose-depth joint distribution. This is not always the case. Consider Figure 2c, where a hypothetical joint distribution between $\boldsymbol{H}$ and $\boldsymbol{D}$ is presented for a given stereo image pair. The maximum probability is indicated with the red dot. First marginalize along $\boldsymbol{H}$ for the depth probability,

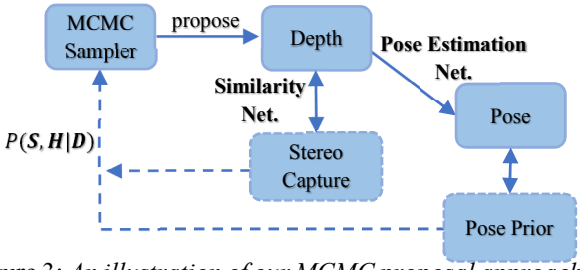

Figure 3: *An illustration of our MCMC proposal approach. The probability of the proposed depth and the recovered pose is used to inform the next depth proposal.*

$P(\boldsymbol{D}) = \sum_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{D})$ to identify $\boldsymbol{D}^*$(analogous to resolving for a robust depth from a given a stereo image pair as in Eq. 8). $\boldsymbol{H}^*$ is then determined by maximizing $P(\boldsymbol{H}|\boldsymbol{D}^*)$ - illustrated with the red dotted line (analogous to Eq. 9). Note how the optimized maximum does not coincide with the joint maximum. Secondly, it assumes that the depth image resolved from the stereo image is fully correct or else even more robust and complex pose estimation from depth techniques will be required to handle erroneous depth recovery. Inspired by [4] we take a different approach. We search for the optimum $\boldsymbol{D}^*$ along the manifold described by the optimum $\boldsymbol{H}$ for all potential depth images, as in

$$\boldsymbol{D}^* = \underset{\boldsymbol{D}}{\operatorname{argmax}}\, [\phi_H\, P(\boldsymbol{S}|\boldsymbol{D})], \quad (10)$$

where

$$\phi_H = \max_{\boldsymbol{H}} \{ P(\boldsymbol{H}|\boldsymbol{D})P(\boldsymbol{D}) \}. \quad (11)$$

and in turn compute $\boldsymbol{H}^*$ using Eq.9. Note the effect of this

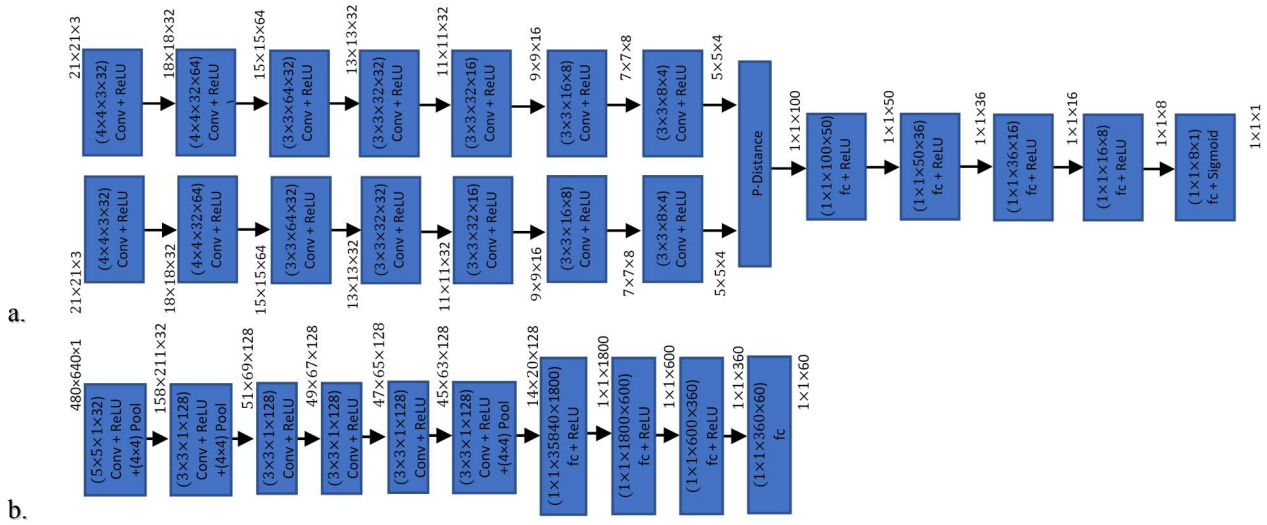

Figure 4: *Structure of two CNN used. (a) a Siamese Network used as a similarity measure between two potentially matching square patches of pixels, (b) illustrates the structure used for discriminatively regressing for pose given a depth image.*

has shown in Figure 2b, where the manifold is illustrated with the black line. Thus, we iteratively travel through the high dimensional space of the depth and the pose, by proposing a depth and evaluating for the Eq. 6 in search for a maximum.

3.2. Probability of observed stereo image given proposed depth

To efficiently propose a depth image, first we segment the reference stereo image[1] into superpixels using SLIC [5]. We represent a hand depth image with a vector $\boldsymbol{d}$ of the depth values of all the superpixels that lie within the hand region. Henceforth, we will refer to this vector as the depth configuration vector. For a proposed depth image, we have

$$P(\boldsymbol{S}|\boldsymbol{D}) = \log\left\{\prod_{j}^{J} P(\boldsymbol{S}|d_j)\right\} = \sum_{j}^{J} log\left[P(\boldsymbol{S}|d_j)\right], \quad (12)$$

where there are J hand superpixels. We model the probability of a stereo image pair given the depth the j^{th} superpixel, $P(\boldsymbol{S}|d_j)$ as the re-projection affinity of the proposed d_j. For a proposed depth, we use the intrinsic and extrinsic parameters of the stereo rig, to re-project pixels in the reference stereo image plane onto the corresponding image plane, before computing affinity. We quantify the quality of a proposed depth based on how re-projected superpixel matches the original superpixel. Hence, we have that for stereo image pair with superpixel, x_j in the left image with a centroid pixel position, $\begin{bmatrix} x_L^j \\ y_L^j \end{bmatrix}$ and a proposed depth d_j,

$$P(\boldsymbol{S}|d_j) = C\left(I_L(x_L^j, y_L^j), I_R(x_R^{*j}, y_R^{*j})\right) \quad (13)$$

where $C(,)$ is a window based matching cost function that gives a measure of affinity and

$$\begin{bmatrix} x_R^* \\ y_R^* \end{bmatrix} = F\left(\begin{bmatrix} x_L \\ y_L \end{bmatrix}, d\right) = d\begin{bmatrix} x_L \\ y_L \\ 1 \end{bmatrix} \boldsymbol{P}_L^{-1}[\boldsymbol{R}|\boldsymbol{t}]\boldsymbol{P}_R. \quad (14)$$

Here $\boldsymbol{P}_L$ and $\boldsymbol{P}_R$ are the projection matrices of the left and right stereo camera pair and $\boldsymbol{R}$ and $\boldsymbol{t}$ are the relative extrinsic matrix and vector respectively – established for the stereo camera using [6]. We represent $C(,)$, as a Siamese network, as in [7]. The first subnet consists of a pair of layers, each composed of convolution followed by ReLU, as shown in Figure 4a. This is followed by the P-Distance layer that computes the square distance of each feature vector in one of the pair of subnet to the other. Finally followed by four fully connected (fc) and then ReLU layers, and then a fully connected then sigmoid layer. The output of the sigmoid layer is the similarity score, $C(,)$. Hence the probability of the observed stereo image, $\boldsymbol{S}$ given a proposed depth configuration, $P(\boldsymbol{S}|\boldsymbol{D})$ is modelled as the similarity of the disparity correspondence resolved from the proposed depth.

3.3. Probability of pose conditioned on depth

The second component is the probability of the pose, $\boldsymbol{H}$ given depth, $\boldsymbol{D}$. Note that the ultimate task is to establish ϕ_H, where we redefine it as

$$\phi_H = P\left(\boldsymbol{H} = \underset{\boldsymbol{H}}{\operatorname{argmax}}\{P(\boldsymbol{H}|\boldsymbol{D})\}\right)P(\boldsymbol{D}), \quad (15)$$

such that $P(\boldsymbol{H} = \boldsymbol{h})$ is the probability of unique pose, $\boldsymbol{h}$ based on the hand pose prior distribution. Hence, we apply a discriminative model that resolves for pose, $\widehat{\boldsymbol{H}}$ given $\boldsymbol{D}$.

[1] The reference stereo image is one of the two images in the pair such that each pixel in the reference image, we seek a correspondence in the other image. Hence a resulting disparity image registers perfectly with the reference stereo image.

We then assume that the discriminatively resolved pose, $\widehat{\boldsymbol{H}}$ is the pose that maximizes the posterior, $\underset{\boldsymbol{H}}{\operatorname{argmax}} P(\boldsymbol{H}|\boldsymbol{D})$ and that $P(\boldsymbol{H}=\widehat{\boldsymbol{H}})$ is the maximum posterior probability, $\max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{D})$. The discriminative model used here is also a CNN. We refer to this CNN as the pose-estimation network. This CNN takes a single channel depth image (from the proposed depth configuration) and outputs a $3*K$-dimensional vector that represents the 3D spatial coordinates of all K joints that describe a hand pose. So, in effect, for a given depth image, the pose-estimation network computes a single pose. ϕ_H is the product of the probability of the estimated pose (based on the pose prior, $P(\boldsymbol{H})$) and the probability of the given depth image (based on the depth image prior, $P(\boldsymbol{D})$). Both priors are described in the following subsection. The structure of the pose-estimation network is illustrated in Figure 4b. This consists of six convolutional layers (each followed with a ReLU and three also with a Pooling layer) followed by four fully connected layers (each followed with a ReLU layer except the last). The output of the final fully connected indicates the joint positions.

3.4. Prior over Depth and Pose

Pose: Let $\boldsymbol{h}$ denote the hand pose vector in a $3*K$-dimensional space $\boldsymbol{V}$. To establish a pose prior over the hand, we add a constraint that resolved joint configuration should be a member of a subspace, $\boldsymbol{W} \subset \boldsymbol{V}$. We establish a criterion for $\boldsymbol{W}$, based on the components that spans the poses in our prior dataset. Applying principal component analysis (PCA) on the prior dataset of potential hand poses, the N most significant components were established, $\boldsymbol{E} = [\boldsymbol{e}_1, \dots, \boldsymbol{e}_N]$ where $N \ll 3*K$. We then apply a constraint that a newly resolved pose $\boldsymbol{h}'$ should be represented by a linear combination of the established component, $\boldsymbol{h}' - \boldsymbol{\mu} \approx \sum_i^N a_i \boldsymbol{e}_i$. Where $\boldsymbol{\mu}$ and a denote the mean pose of all joint configurations in the prior dataset and a scalar value respectively. To this end, we established the probability of a resolved pose $\boldsymbol{h}'$ as

$$P(\boldsymbol{H}=\boldsymbol{h}') = e^{\,||\boldsymbol{E}\boldsymbol{a}^*+\boldsymbol{\mu}-\boldsymbol{h}'||} \tag{16}$$

where

$$\boldsymbol{a}^* = \boldsymbol{E}^+(\boldsymbol{h}'-\boldsymbol{\mu}) \tag{17}$$

where $||.||$ denotes the l^2-norm and $\boldsymbol{E}^+$ is the pseudo-inverse of the $\boldsymbol{E}$. $\boldsymbol{a}^*$ is then the least square estimation to the coefficients of the components that yields $\boldsymbol{h}'$ under a linear combination. We then use the exponentiated Euclidean distance between this linear combination of components and $\boldsymbol{h}'$ as a measure of prior probability. In effect, a 3D joint configuration (pose) that is like those in the dataset will be more accurately mapped onto $\boldsymbol{W}$ and re-mapped back.

Depth: Using the hand region segmentation, the Euclidean distance between the mean hand pixel position in both images of the stereo pair is used to estimate the general distance of the hand to the camera, using the baseline and focal lengths of the stereo rig. The prior of over depth at all superpixels in the scene is modelled with a Gaussian, with a mean as the estimated general distance, R and an arbitrary standard deviation, σ, as in $P(\boldsymbol{D}=\boldsymbol{d}) = \sum_i \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(d_i-R)^2}{2}}$.

3.5. Metropolis-Hastings Algorithm

Algorithm: Joint Depth and Pose Estimation using the Metropolis-Hastings Algorithm

Input: $\boldsymbol{S}$;
Output: $\boldsymbol{H}^*$;
Initialize $\boldsymbol{D}^{(0)}$, $\boldsymbol{H}^{(0)} = \underset{\boldsymbol{H}}{arg\max}\, P(\boldsymbol{H}|\boldsymbol{D}^{(0)})$;
Let $\boldsymbol{D}^* = \boldsymbol{D}^{(0)}, \boldsymbol{H}^* = \boldsymbol{H}^{(0)}$;
for $i = 0$ *to* $L-1$ **do**
  Sample $u \sim U_{[0,1]}$;
  Sample $\boldsymbol{D}' \sim q(\boldsymbol{D}'|\boldsymbol{D}^{(i)})$;
  if $log(u) < log(\alpha(\boldsymbol{D}'|\boldsymbol{D}^{(i)}))$ **then**
    $\boldsymbol{D}^{(i+1)} = \boldsymbol{D}'$; $\boldsymbol{H}^{(i+1)} = \underset{\boldsymbol{H}}{arg\max}\, P(\boldsymbol{H}|\boldsymbol{D}^{(i+1)})$;
  else
    $\boldsymbol{D}^{(i+1)} = \boldsymbol{D}^{(i)}$; $\boldsymbol{H}^{(i+1)} = \boldsymbol{H}^{(i)}$;
  end
  if $P(\boldsymbol{H}^{(i+1)}, \boldsymbol{D}^{(i+1)}|\boldsymbol{S}) > P(\boldsymbol{H}^*, \boldsymbol{D}^*|\boldsymbol{S})$ **then**
    $\boldsymbol{D}^* = \boldsymbol{D}^{(i+1)}$; $\boldsymbol{H}^* = \boldsymbol{H}^{(i+1)}$;
  end
end

To achieve an informed framework for proposing depth images (configuration), we exploit Markov-chain Monte Carlo. We iteratively propose a new depth configuration, $\boldsymbol{D}'$ conditioned on the previous proposal, $q(\boldsymbol{D}', \boldsymbol{D}^{(i)})$. We implement this distribution by randomly perturbing the elements of the depth vector $\boldsymbol{d}^{(i)}$ that describes $\boldsymbol{D}^{(i)}$. Hence the probability of the newly proposed depth vector, $\boldsymbol{d}'$ is conditioned on the previous depth proposal, $\boldsymbol{d}^{(i)}$. We then evaluate for the acceptance ratio, α, where

$$\alpha(\boldsymbol{D}', \boldsymbol{D}^{(i)}) = min\left\{1, \frac{P(\boldsymbol{D}', \boldsymbol{H}'|\boldsymbol{S})}{P(\boldsymbol{D}^{(i)}, \boldsymbol{H}^{(i)}|\boldsymbol{S})}\right\} \tag{18}$$

and $\boldsymbol{H}' = \max_{H} P(\boldsymbol{H}, \boldsymbol{D}')$. Note that we ignore the ratio of the probability of proposing a particular $\boldsymbol{D}'$ given $\boldsymbol{D}^{(i)}$, $q(\boldsymbol{D}', \boldsymbol{D}^{(i)})$ and the reverse, as these are equal and hence cancel out. If the acceptance ratio is higher than u, a sample between 0 and 1, the proposed depth configuration and the corresponding maximizing pose are considered as a potential candidate for solution. Hence the higher the probability of the newly proposed depth configuration (relative to the previous proposal) the more likely it would be accepted as a potential solution. We evaluate all potential solutions by maximizing for Eq. 6. See above for the pseudo-code of the Metropolis-Hastings Algorithm. The effect from this is that we evaluate for the depth configuration that is most consistent with the observed stereo capture and that yields the more probable pose, from a sample set with a distribution that is consistent with the solution. This is because the MCMC samples the depth

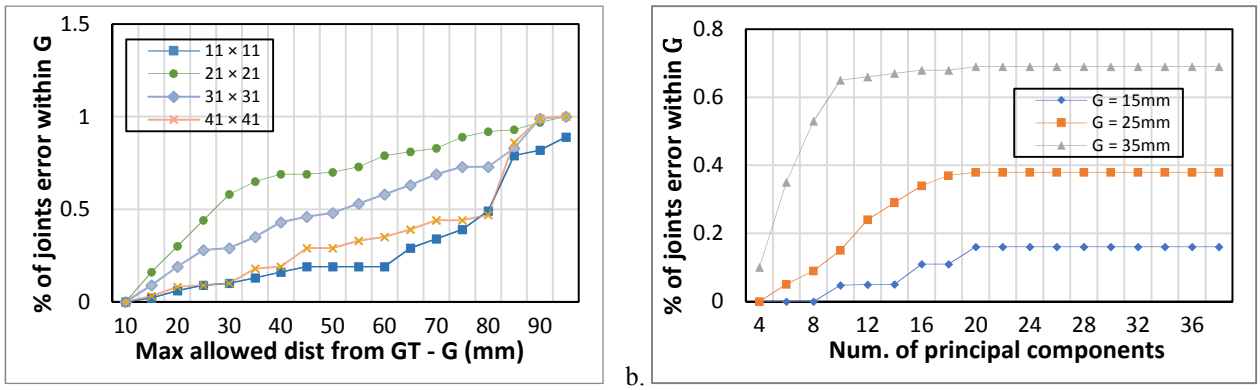

Figure 5: *Evaluating the significance of the window size and the number of components used for spanning the prior pose space. (a) The graph illustrates the percentage of joint pose prediction with a margin of error, G for different window sizes. (b) Shows the percentage of correctly predicted joint position as different number of components are used.*

configuration based on the same criteria, i.e. the depth that are more consistent with the stereo capture and that yields more probable pose are proposed more.

4. Implementation details

Both the pose-estimation and similarity networks were implemented using the VLFeat MatConvNet [8] and trained on a NVIDIA Titan X GPU with 6GB memory.

Similarity Network: This CNN was trained with the learning rate of 0.001. We ran 10 epochs, reducing the learning rate by 10% every epoch. The decay weight and momentum was set as 0.0005 and 0.09 respectively. Like [7], we train the similarity network to map a pair of window regions, $< I_L(\boldsymbol{p}),\ I_R(\boldsymbol{q}) >$ from the left and right stereo pair to a cost, c. For each superpixel, a square window region centered on its centroid pixel is considered. We base this on a hinge loss, $max(0, g + c_- + c_+)$, where g, c_- and c_+ are the margin, output of the CNN from a non-matching input window patch pair and the output of the CNN from a matching input window patch pair. We establish matching pair windows by reprojection based on the camera parameters of the stereo cameras and the ground truth depth at the superpixel. We set the value of g to 0.2.

Pose-Estimation Network: The pose-estimation network has a significantly greater number of weights due to the larger input image. This explains the need of the pooling layers absent in the similarity network. We train this CNN with a learning rate of 0.00001 for 150 epochs. Decay weight and momentum were set as 0.005 and 0.09 respectively. We train under a mean squared error between the output vector and the ground truth pose vector.

The prediction phase of the entire framework for a frame of stereo images under 200 MCMC proposals will took 360 seconds. See Figure 3 for the entire framework. Here proposed depth is evaluated with the Similarity Network and simultaneously used to recover pose using the Pose Estimation Framework that is evaluated against the Pose prior.

5. Experiment and Results

We present a proof of concept by evaluating the performance of the proposed technique. The approach was validated experimentally, presenting both qualitative (Figure 7) and quantitative (Figure 6a) results. Four main comparisons were made, these include: pose estimation prediction made from single shot depth recovery, estimation made without our pose prior; estimation made using proposal in [2]; and estimation made using depth acquired using active RGBD camera sensor. The results were quantitatively appraised for accuracy by computing the percentage of correctly predicted joint position, $\frac{\sum_{p \epsilon N} F\{|\boldsymbol{s}_p^{GT} - \boldsymbol{s}_p| < G\}}{N}$, where $\boldsymbol{s}_p^{GT}$ and $\boldsymbol{s}_p$ are the ground truth and the predicted 3D joint position of all joints, p in the testing dataset; $F\{\}$ is a function that returns 1 for *true* input and 0 otherwise; and N is the to total number of joints evaluated (across all the frames). We also computed the mean distance error, $\frac{1}{N}\sum_{p \epsilon N} |\boldsymbol{s}_p^{GT} - \boldsymbol{s}_p|$ to quantitatively evaluate the performance of the test.

5.1. Dataset

To establish a database of strong registration between the triplet of data: stereo, depth and pose, acquisition was carried out on the stereo camera, a RGBD camera, and an off-the-shelf hand pose detector. The RGBD and stereo cameras were almost adjacently positioned with the pose detector positioned perpendicularly as shown in Figure 1c. Using camera calibration [6], depth data from an RGBD sensor was registered to the left image of the RGB pair. It suffices that the spatial position of the hand pose detector relative to the stereo camera is unchanged during capture of training data. To train the similarity network, a binary class dataset was to be created with matching pairs of image patches (from the left and right stereo image) considered as a positive class and non-matching considered otherwise. In the case of the pose-estimation network.

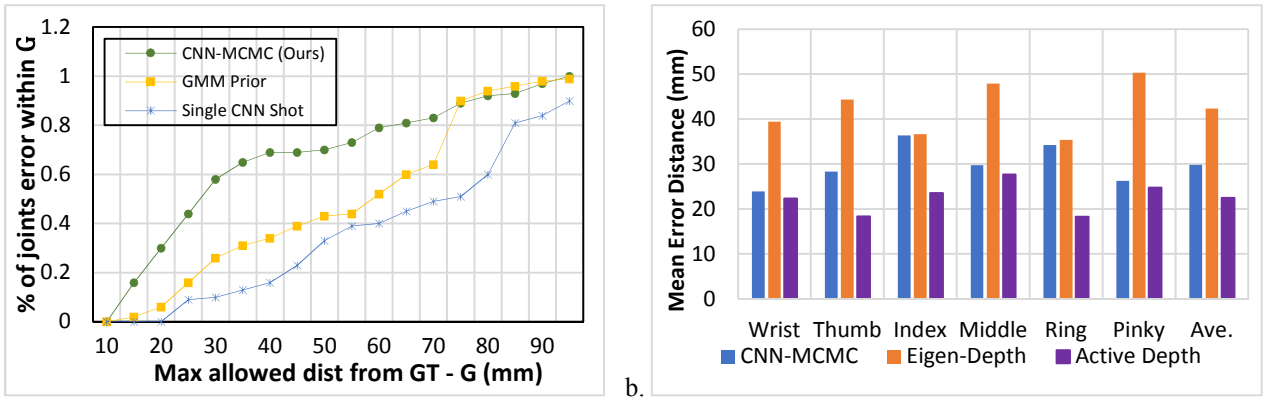

Figure 6: *A baseline comparison of our approach. (a) The graph illustrates the percentage of accurately predicted joint pose prediction (within a margin of error), for our approach in comparison to the single shot depth estimation and to GMM based prior. (b) Bar chart showing the mean joint position error per finger for our approach, the work proposed in [2] and RGBD camera based pose estimation.*

Data was captured from 12 participants (12,000 stereo pairs in total) of different skin tone, hand size and gender. Data from two participants was reserved for testing, and the remaining data (from the other ten participants) was used for training in a cross validation manner. SLIC segmentation was applied to all reference stereo images, producing approximately 300 superpixels per image. Note that only a fraction of these 300 superpixels are hand region superpixels. The amount of hand superpixels (ranging approximately from 30 to 60 per image capture) depends on the distance from camera and the size of the hand. All in all, about 540,000 patches where used in training the similarity network. Each hand pose is represented by 20 joints i.e. $K = 20$. These included the wrist; the thumb (fingertip, distal and intermediate); the index, middle, ring and pinky finger (each with a fingertip, distal, intermediate and proximal joint).

5.2. Baseline Comparison

To optimize the performance of our proposed technique we experimented with two significant parameters. These includes the window size (of stereo comparison) and number of components used to store pose prior information. The window size determines the size of the input stereo pair regions that is fed into the similarity network for comparison and subsequently, the number of the weights of the similarity network. From Figure 5, one can identify a gradual improvement in the accuracy as the size of the window reduces. 41×41, 31×31 and 21×21 window sizes yielded a 18.23%, 35.54% and 65.218% of accurately predicted joint position within an error of 35 mm, respectively. This trend stops when a window size of 11 × 11 window is applied, resulting in 13% accurate predictions (see Figure 5a). A second parameter was the number of components used. Recall from Eq. 16 and 17 that from the $3 * K$ components only N are used. The significance of the number of components used, is also presented in Figure 5. Figure 5b illustrates the increase in the percentage of accurate joint prediction as the number of components increases, however this improvement in prediction performance stops after 10 to 18 of the most significant components have been used.

As well as the parameter evaluation, two baseline comparisons were made. The first was predicting the pose using a single shot depth estimation, and the second was predicting pose without the pose prior.

Single-shot depth recovery: For a given stereo capture, we evaluate all potential matching pixels along the epipolar line on the corresponding stereo pair under the Similarity Network and apply a greedy search approach to establish a disparity image. We then apply the pose-estimation network to directly estimate for the pose. Figure 6a validates our hypothesis presented in Section 3.1. The superiority of our jointly optimal, iterative depth proposal is apparent here, particularly at lower error thresholds. The ability to continuously reevaluate the depth solution whilst resolving for pose contributes to this performance. In fact, there is a 389.8% more correctly predicted joint positions (within a 35mm error margin) when our approach is taken in comparison to the single shot approach. Although this superiority diminishes as the error threshold increases, our iterative approach produces a more accurate hand pose estimation from stereo capture. The qualitative results in Figure 7 (4^{th} row) corroborates this result, as better pose estimation is achieved with our approach in comparison – particularly in the first, fourth and fifth columns.

GMM prior: Another component of our derivation is the pose prior. We evaluate the effectiveness of our PCA based approach by comparing it against a GMM (Gaussian Mixture Model) based approach. For this we apply an expectation maximization to establish a $3 * K$ dimensional GMM model that represents the probability of a pose (as in [11]). We experiment to establish the optimum component. We present the performance of this approach in Figure 6a.

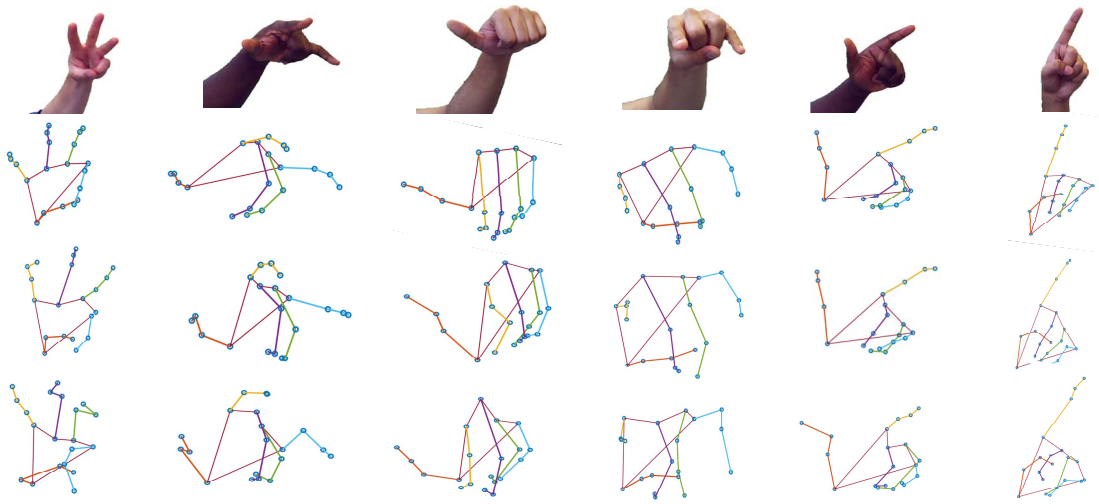

Figure 7: *Qualitative results of pose estimation using real stereo captured poses. The reference image of the stereo pair is shown in the 1st row. The results from our full technique are presented in the 2nd row. The 3rd row shows the pose estimation result from using our method but with a GMM pose prior while 4th row shows result from using the single-shot CNN.*

Again, results show the significance of the PCA based model, with our approach producing 109.6% more correctly predicted joint positions (within a 35mm error margin). This is largely owed to the first identifying the highly discriminating components in the pose subspace before establishing a prior model. This superiority is shown in Figure 7 (3rd row), particularly in the first, third and sixth columns. Our PCA based approach better constraints for a more realistic hand pose.

5.3. Comparison against [2]

To further validate our work against published literature we evaluate performance of our work to the work proposed in [2]. As introduced in Section 2, [2] regresses for robust hand depth estimation using eigen leaf node based variant of a regression forest. The paper motivates its approach with depth recovery specifically for hand pose estimation. To evaluate this, we applied the pose-estimation network to directly regress for pose from the recovered depth using the approach in [2]. We present, the performance in Figure 6b. Again, like the single shot approach this approach performs significantly less than our joint optimization approach. On average our approach preforms 29.55% better than the proposal in [2] (29.80 mm to 42.32mm error). This corroborates the significance of jointly optimizing for both pose and depth. The single shot approach assumes a high-quality depth prediction and will yield a poor result when the preceding depth estimation is poor.

5.4. Comparison against Active Depth Sensor

To evaluate the significance of the work done in the general context of gesture recognition, we compare the accuracy of the pose estimation prediction made to pose estimation made from depth image acquired from the RGBD camera. Again, we apply pose estimation using the pose-estimation network. Figure 6b presents the evaluative comparison. Compared to our approach, the RGBD based pose prediction was relatively more accurate in predicting thumb, the index and ring finger joints. This is due to large variance in their 3D position across the training and testing dataset. Across all five fingers, the mean joint position error of estimated pose from the RGBD depth image is 21.99mm, this is only 9.304mm lower than the mean joint position error of our technique (30.802mm). Considering the low-quality nature of the stereo camera used the proposed approach exhibits robustness against inconsistency and noise in stereo capture to an extent that it is on par with pose estimation made from an active depth sensor. This is significant, has it shows potential of overcoming the drawbacks of RGBD discussed in Section 1 without a significant drop in the accuracy of pose estimation.

6. Conclusion

In this work, we present a novel approach to pose estimation from stereo capture by proposing a MCMC-CNN approach of joint optimization. We have shown experimentally, that our joint optimization approach outperforms the conventional single shot depth estimation approach. For future work, we aim to propose a closed form solution to the estimation of the depth configuration by establishing a parametric relationship between the depth configuration and the stereo cost. This will allow for the parallelizing the CNN execution in a single run to achieve a real-time pose estimation.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] Hasan, M.M.,and Mishra, P.K. (2012). Superior Skin Color Model using Multiple of Gaussian Mixture Model. In the British Journal of Science, Volume 6 (1).

[2] Basaru, R., Alonso, E., Child, C., and Slabaugh, G., (2016). HandyDepth: Example-based Stereoscopic Hand Depth Estimation using Eigen Leaf Node Features. In Proc. of the IWSSIP International Conference. Bratislava, Slovakia.

[3] Romero J., Kragic D., Kyrki V., and Argyros A., (2008). Dynamic Time Warping for Binocular Hand Tracking and Reconstruction. In Proc. of the ICRA, Pasadena, California USA.

[4] Collins R., and Carr P., (2016) Hybrid Stochastic/Deterministic Optimization for Tracking Sports Players and Pedestrians. In Proc. of the ECCV International Conference, Amsterdam, Netherlands.

[5] Achanta R., Shaji A., Smith K., Lucchi A., Fua P., and Susstrunk S., (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods, In the IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 34 (11).

[6] Zhang, Z. (1999). Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In Proc. of the ICCV. Corfu, Greece.

[7] Žbontar J., and LeCun Y., (2016) Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches, arXiv preprint:1510.05970, 2016.

[8] http://www.vlfeat.org/matconvnet/ [Accessed 10th May 2017]

[9] https://www.spectacles.com/ [Accessed 17th February 2017]

[10] https://www.oculus.com/ [Accessed 17th May 2017]

[11] Burke, M. and Lasenby J. (2014) Single Camera Pose Estimation using Bayesian Filtering and Kinect Motion Priors. arXiv preprint: 1405.5047v2, 2014.

[12] Oh, S., Russell, S., and Sastry, S., (2009) Markov Chain Monte Carlo Data Association for Multitarget Tracking. In the journal of Transactions on Automatic Control, Volume 54(3).

[13] Brau, E., Barnard, K., Palanivelu, R., Dunatunga, D., Tsukamoto, T., and Lee, P. (2011) A Generative Statistical Model for Tracking Multiple Smooth Trajectories. In Proc. of the CVPR, Colorado Springs, Colorado, USA.

[14] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. 3d Hand Pose Tracking and Estimation Using Stereo Matching. arXiv preprint: 1610.07214, 2016.

[15] Argyros, A. (2017) Back to RGB: 3D tracking of hands and hand-object interactions based on short-baseline stereo. arXiv preprint: 1705.05301, 2017.

[16] Ge, W., and Collins, R., (2008) Multi-target Data Association by Tracklets with Unsupervised Parameter Estimation. In Proc. of the BMVC, Leeds, UK.

[17] Ge, L., Liang, H., Yuan, J., and Thalmann, D., (2016) Robust 3D Hand Pose Estimation in Single Depth Images: from Single-view CNN to Multi-View CNNs. In Proc. of the CVPR, Las Vegas, Nevada, USA.

[18] Sun X., Wei Y., Liang S., Tang X., and Sun J., (2015). Cascaded Hand Pose Regression. In Proc. of the CVPR, Boston, Massachusetts, USA.

[19] Zivkovic Z., (2004) Improved Adaptive Gaussian Mixture Model for Background Subtraction in Pattern Recognition. In Proc. of the ICPR, Cambridge, UK.

[20] Toshev A. and Szegedy C., (2014) DeepPose: Human Pose Estimation via Deep Neural Networks. In Proc. of the CVPR, Columbus, Ohio, USA.

[21] Phung, S., Bouzerdoum, A., and Chai, D., (2005). Skin Segmentation Using Color Pixel Classification: Analysis and Comparison. In the IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 27 (1).