ESTIMATION OF PARTIALLY OCCLUDED HUMAN JOINTS USING A BAYESIAN APPROACH AND AN APPLICATION OF HUMAN IMAGE INPAINTING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET ANIL DURSUN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2021

Approval of the thesis:

ESTIMATION OF PARTIALLY OCCLUDED HUMAN JOINTS USING A BAYESIAN APPROACH AND AN APPLICATION OF HUMAN IMAGE INPAINTING

submitted by **AHMET ANIL DURSUN** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. Temel Engin Tuncer
Supervisor, **Electrical and Electronics Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Çağatay Candan
Electrical and Electronics Engineering, METU _____

Prof. Dr. Temel Engin Tuncer
Electrical and Electronics Engineering, METU _____

Assoc. Prof. Dr. Sevinç Figen Öktem
Electrical and Electronics Engineering, METU _____

Assist. Prof. Dr. Elif Vural
Electrical and Electronics Engineering, METU _____

Assoc. Prof. Dr. Seniha Esen Yüksel
Electrical and Electronics Engineering, Hacettepe University _____

Date: 01.02.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ahmet Anıl Dursun

Signature :

ABSTRACT

ESTIMATION OF PARTIALLY OCCLUDED HUMAN JOINTS USING A BAYESIAN APPROACH AND AN APPLICATION OF HUMAN IMAGE INPAINTING

Dursun, Ahmet Anıl
M.S., Department of Electrical and Electronics Engineering
Supervisor: Prof. Dr. Temel Engin Tuncer

February 2021, 87 pages

Human pose estimation is a well-known computer vision task that has applications in the fields of surveillance, computerized outfit planners, and video special effects. State-of-the-art pose estimators are based on CNN structures and use visual features obtained from single images. However, occlusions are prevalent problems in the natural scenarios for this task, and the performances of CNN-based estimators degrade significantly under occlusion conditions. In this thesis, a novel Bayesian approach, BAKE: Bayesian Approach for occluded Keypoint Estimation, is presented to estimate the positions of occluded human joints in a given video sequence. This approach uses a well-known CNN-based pose estimator Openpose [1] to detect the visible human joints in a given image and then develops a Bayesian framework to complete missing pose elements. This approach can be evaluated as an alternative for 3D CNN structures in terms of embedding the information in time-dependent event sequences. In our case, the problem is ill-conditioned since it is in general not possible to complete the missing joints for an arbitrary occlusion on the human body accurately. However, it is possible to localize some missing joints in certain regions based on the apriori information on the human skeleton with a certain confidence. This apriori

information is obtained from the previous frames of the video sequence. A statistical human body model is generated by defining the joint length and angle parameters. The parameters of the model are calculated from the non-occluded frames of the video sequence as the local information as well as a human pose database is utilized for obtaining the general joint statistics. Then, on a partially occluded video frame, body length and angle distribution are updated by using the visible joints. These length and angle distributions are used for the estimation of occluded joints. A new confidence score is also proposed. This confidence score is used to develop a hybrid technique which combines the predictions of the Openpose and the proposed method, BAKE. Several experiments are performed to compare the outputs of the Openpose, BAKE, and the hybrid approach. It is shown that BAKE outperforms Openpose in general and the hybrid method generates a slight improvement over the BAKE. In addition to the BAKE method, an inpainting method for the partially occluded human video frames is proposed. In this method, non-occluded images of the target person obtained from the video sequence are used with a 3D body reconstruction algorithm, SMPLify-x [2]. Image patches are transferred from non-occluded images to occluded parts after a matching process and the corresponding results are shown.

Keywords: Occluded human pose estimation, convolutional neural networks, Bayesian inference, statistical modelling, image inpainting

ÖZ

KISMİ KAPANMAYA UĞRAMIŞ İNSAN EKLEMLERİNİN BAYESYEN YAKLAŞIMLA KESTİRİMİ VE İNSAN GÖRÜNTÜSÜ TAMAMLAMA UYGULAMASI

Dursun, Ahmet Anıl
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Temel Engin Tuncer

Şubat 2021 , 87 sayfa

İnsan poz kestirimi güvenlik, kıyafet uygulamaları ve özel video efektleri alanlarında gerçekleştirilmesi önemli bir problemdir. Bu problemin çözümünde tek bir görüntüden faydalanarak görsel öznitelikleri kullanan konvolüsyonel sinir ağları tabanlı metotlar en başarılı yöntemlerdir. Ancak kapanmalar doğal senaryolarda sıkça ortaya çıkmakta ve KSA tabanlı insan poz kestirimi metotlarının performasını ciddi oranda düşürebilmektedir. Bu çalışmada, verilen bir görüntüde kapanmaya uğrayan insan eklemlerinin konumunu kestiren ve özgün bir Bayesyen yaklaşım olan BAKE metodu sunulmaktadır. Bu yaklaşımda görüntüdeki görünür eklemleri bulması için popüler bir kestirim metodu olan Openpose'dan [1] faydalanılmakta, sonrasında ise Bayesyen bir çerçevede kapanmaya uğramış noktaların yerleri kestirilmektedir. Bu metot, zamana bağlı olay dizilerini istatistiksel bir şekilde modelleyebilmesi bakımından 3 boyutlu KSA yapılarına bir alternatif olarak görülebilir. Kapanmaya uğramış eklemlerin kestirimi kötü koşullu bir problem olsa da önsel bilgiden faydalanılabilir ve bu önsel bilgi önceki kapanmaya uğramamış video karelerinden elde edilebilmekte-

dir. Bu kapsamda, eklem uzunlukları ve açıları tanımlanarak istatistiksel bir vücut modeli oluşturulur. İstatistiksel parametreleri hesaplamak amacıyla kapanmaya uğramamış video karelerinden faydalanılarak lokal bilgi, geniş bir insan pozu veri kümesinden faydalanılarak da genel bilgi elde edilir. Sonrasında, kapanmaya uğramış video karesinde görülebilir noktalar kullanılarak uzunluk ve açı dağılımları güncellenir. Kaba kuvvet yöntemiyle de kapanmaya uğramış eklemler için pozisyon kestirimi yapılır. İnsan eklemlerini tamamlama işleminin ne kadar güvenilir yapıldığını ölçen yeni bir güven puanı da ortaya konmuştur. Bu puan Openpose ve BAKE algoritmalarının çıktılarını birleştiren hibrit bir algoritmada da kullanılmıştır. Yapılan deneylerde Openpose, BAKE ve hibrit yaklaşım karşılaştırılmış, BAKE'nin Openpose'dan üstün başarı sergilediği gösterilmiştir. Hibrit metodun da BAKE'nin performansını bir miktar artırdığı gözlemlenmiştir. Ayrıca kısmi kapanma yaşanan insan görüntülerinde tamamlama yapacak bir metot da önerilmektedir. Bu metotta yine videodaki kapanmaya uğramamış görüntülerden faydalanılmakta, ayrıca da görüntüden 3 boyutlu vücut örgüsü oluşturabilen SMPLify-x [2] algoritması kullanılmaktadır. Kapanmaya uğramamış karelerdeki görüntü parçaları kapanma olan karede kapanma yaşanan bölgedeki görüntü parçalarıyla eşleştirilmekte ve eşleştirilmiş görüntü parçaları kapanma yaşanan bölgeye aktarılarak tamamlama işlemi yapılmaktadır. Bu metoda ait sonuçlar da sunulmuştur.

Anahtar Kelimeler: Kapanma durumunda insanda poz kestirimi, konvolüsyonel sinir ağları, Bayesci çıkarım, istatistiksel modelleme, görüntü tamamlama

To my beloved family

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Dr. T. Engin Tuncer, for his support and guidance. He showed me the points that I was wrong with a good criticism while he always encouraged me during this thesis work. I am grateful and glad to have him as the supervisor of my thesis.

I am grateful of my parents, my sister and Çağıl Ezgi Aydemir for their endless support and understanding. They have stood by me during the difficult times.

I would like to thank ASELSAN Inc. for letting me to do my thesis.

Lastly, I am thankful for the Scientific and Technological Research Council of Turkey and TÜBİTAK 2210-A National Scholarship Program for financial support they provided in my graduate studies.

TABLE OF CONTENTS

APPENDICES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	
2D	2 Dimensional
3D	3 Dimensional
CNN	Convolutional Neural Networks
BAKE	Bayesian Approach for occluded Keypoint Estimation
SMPL	Skinned Multi-Person Linear Model
GMM	Gaussian Mixture Model
COCO	Common Objects in Context
HOG	Histogram of oriented gradients
PAF	Part Affinity Field
MPJPE	Mean Per Joint Positional Error
EDM	Euclidean Distance Matrix
SCAPE	Shape Completion and Animation of People
IoU	Intersection over Union
ROC	Receiver Operating Characteristic
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index

CHAPTER 1

INTRODUCTION

Occlusions in visual systems occur when an object in the scene cannot be observed by the camera while it is still in the camera's field of view, spatially. Reasons for occlusions are the interactions of objects in the scene while different parts of an object could also cause occlusions. This second type of occlusions is named self-occlusions since both the occluded and occluder objects are the same. Occlusions are important problems for applications of surveillance [3], human-robot interactions [4], augmented reality [5] etc., since lower-level computer vision tasks such as object detection/segmentation, object tracking, or pose estimation used in these complicated applications suffers from occlusions in general.

It is possible to categorize occlusions depending on their severity [6]. If the object becomes completely unobservable by the camera while it is still in the scene, it is fully occluded. Therefore, the visual information related to it is lost in the image. In case of complete or significant occlusions, object detection, segmentation, or pose estimation from a single image may not be possible. While in the most general scenario, it may not be possible to estimate the positions of occluded objects (occluded human keypoints in our case), temporal information in a video sequence can be used for estimation on the low to mild occluded sequences.

There could be cases where only a portion of the objects are occluded, namely partially occluded cases. In such cases, visual cues of occluded human joints can be obtained from the previous video frames where there is no occlusion. One critical task is how to exploit the information in the local video frames in addition to the global anatomical constraints for the human joints extracted from certain databases. In this thesis, a new approach is presented as a solution. This new method uses the

Bayesian framework which is shown to perform effectively in modeling the local and global information for the human joints. The proposed approach is compared with the state-of-the-art human pose estimation CNN network, Openpose[1], results. In this thesis, an additional problem is also investigated. This is the inpainting of partially occluded human images in video sequences. In this case, only the local information from the previous frames is used with a 3D human pose estimation algorithm, SMPLify-x[2].

1.1 Motivation and Problem Definition

2D Human pose estimation under partial occlusions is an important computer vision task especially for applications of surveillance [7] and computerized outfit planners [8]. In this problem, it is desired to detect the partially occluded body part locations of the target person in an image. In video surveillance applications, there are scenarios where the 2D joint positions of the tracked person are desired to be detected in the occluded region. In computerized outfit planners, a target person is tried to be dressed up with predefined clothes, or the clothes are desired to be obtained from images of a person who wears them. For images in wild, dressing up a person or observing the clothing of a person becomes difficult due to different problems where one of them is occlusion. Thus, it is desired to estimate the pose of a person in partially occluded scenarios for such applications [9, 10].

Completion of occluded keypoints is an ill-conditioned problem in general since the information of the occluded keypoint is lost and reconstruction is ambiguous. In order to recover the lost information related to the occluded part, apriori information is required such as the physical shape of the human body. However, human body motion has a high amount of variety. Considering it as a rigid object and inferring about the occluded regions is not feasible in general. Therefore, a model is required which utilizes both anatomical constraints and motion information of the target person in order to estimate the 2D locations of occluded joints. In this work, a Bayesian framework is proposed which uses a database and previous non-occluded frames of video in order to embed the anatomical and kinematic relations of the human body joints. Human pose estimation under partial occlusions task is performed with the

help of the proposed model.

A method for inpainting the occluded regions of the target human utilizing computer graphics is also proposed in this thesis. Such an application could be used in video special effects and restoration of damaged videos [11]. This method reconstructs the 3D body structure of the human in non-occluded and partially occluded frames and transfers the previously observed body patches to the occluded region. This is a different approach than popular inpainting methods [12] which do not consider the complex appearance of human images.

1.2 Proposed Methods and Models

There are two methods presented in this thesis. The first one is the Bayesian pose estimation for partial occlusions. This method first uses a base model for 2D pose estimation and a foreground extraction method to find the visible and occluded joints. In this study, Openpose[1], a popular neural network-based pose estimator, is used as the base pose estimator with a Gaussian Mixture Model foreground extraction algorithm [13]. Such a classification of joints or keypoints, allows us to find the positions of occluded ones with the information obtained from the visible ones. A simple 13-keypoint skeleton human body model is used throughout the study. Model parameters are defined as the normalized lengths and angles obtained from this 13-keypoint model. These parameters let us have shift and scale invariance through the process and they are assumed to have a joint Gaussian distribution. COCO dataset [14] and frames, where the target person is not occluded, are used for the estimation of the mean vector and covariance matrix of this joint Gaussian distribution. In mean and covariance parameters, COCO contributes with the global information while non-occluded frames contribute with the local information obtained from the target person in the video sequence.

After obtaining the joint distribution of angle and length parameters, the occluded keypoints are ordered with a priority scheme which determines the proximity of the visible keypoints with occluded ones. In the order of most prior to less prior, approximate positional probability distributions of occluded keypoints are estimated.

In order to estimate approximate positional probability distributions, a derivation using the joint distribution of length and angle parameters is given. Then, positional probability distributions are estimated by sampling the derived approximate probability mass function in possible search regions which are specific to each occluded keypoint and found by investigating the maximum probabilistic length values of occluded keypoints with visible ones. Maximum likely positions of occluded keypoints in the obtained positional distributions become the estimates of the proposed Bayesian approach. The joint-wise confidence scores are also obtained from the positional distributions.

The proposed approach also gives a new overall body confidence score output which implies how accurately the pose of a given partially occluded body can be recovered. In addition, a hybrid method that utilizes both the Openpose and the Bayesian approach is given. This method makes predictions in accordance with the confidence ratios of the two methods' predictions.

The second method is an inpainting application for the occluded regions of human images. 3D body meshes for both non-occluded and partially occluded frames are reconstructed to this end. In the reconstruction process, 2D keypoint estimates of the partially occluded body are utilized as the input of a 3D body reconstruction algorithm, SMPLify-x [2] in our case. SMPLify-x algorithm simply tries to fit a predefined 3D SMPL [15] body mesh's pose and shape parameters to given 2D joint coordinates. After the reconstruction step, 3D body meshes are rendered on the video frames. Corresponding occluded region of 3D body mesh is known with the help of a foreground extraction algorithm again. Then, the process is transferring the non-occluded image patches to the corresponding occluded region with the required warping operations, briefly.

The notation is as follows. Vectors are shown with lowercase boldface letters, $\mathbf{x} \in \mathbb{R}^{\mathrm{n}}$ throughout the paper where matrices are shown with uppercase boldface letters, $\mathbf{A} \in \mathbb{R}^{\mathrm{nxm}}$. Sets are represented with uppercase letters, such as the set B. Moreover, $f_{\mathbf{x}}(\mathbf{x})$ is used for a probability density function on continuous valued variable $\mathbf{x}$, while $p_{\mathbf{y}}(\mathbf{y})$ represents probability mass function defined for discrete valued variable $\mathbf{y}$. Multivariate Gaussian distribution is denoted as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is the mean vector

and Σ is the covariance matrix.

1.3 Thesis Contributions

Our contributions for different problems are as follows:

- Problem 1: Estimation of occluded keypoints.
  - A 13-keypoint skeleton body model is proposed.
  - A simple Bayesian model which employs both local and global information is proposed.
  - A Gaussian formulation with cylindrical parameters to encode 2D human body pose is proposed.
  - An approximation for probability distribution of missing joints on Cartesian pixel coordinates by using cylindrical parameters of the statistical model is proposed.
  - An overall body pose predictability score is proposed.
  - Joint-wise confidence score is proposed.
  - A hybrid method is proposed to improve the proposed Bayesian technique.
- Problem 2: Inpainting of occluded human body parts.
  - A new method for inpainting the occluded human body parts in video sequences is presented.

1.4 The Outline of the Thesis

In Chapter 2, background information related with human pose estimation problem and 3D body reconstruction literature will be mentioned. For the human pose estimation task, used body models in the literature will be explained. Popular pose estimation approaches and algorithms will be introduced. Among those popular methods,

Openpose is used as a base pose estimation algorithm in this study, so the structure of the Openpose will be explained. Moreover, strategies in the literature for performance improvement under partial occlusion conditions will be given. For 3D body reconstruction, created 3D human body models in the literature will be introduced.

In Chapter 3, an occlusion detection method is proposed which is used further in the methods of both Bayesian approach for occluded keypoint estimation and human image inpainting. A comparison is performed between the proposed occlusion detection scores in Section 3.2. This chapter is an auxiliary part of the main work of this thesis.

In Chapter 4, the proposed Bayesian pose estimation method under partial occlusion conditions will be explained. The mathematical formulation and details of the algorithm will be given. Confidence score outputs of the algorithm will be explained in Section 4.1.6 and the Hybrid approach which uses the outputs of both Openpose and proposed Bayesian approach is introduced in Section 4.1.7. The experimental results will be given in Section 4.2.

In Chapter 5, the inpainting application of partially occluded human images will be given. The inpainting process for the partially occluded frames using the non-occluded video frames will be explained. Experiments related to the inpainting application will be given at the end of this chapter.

Discussion and possible future work related with the thesis will be given in Chapter 6.

In Appendix A, a derivation for the positional distributions of occluded keypoints is given. A derivation for the total body length distribution is also given in Appendix B. Test frames related to the BAKE update weight experiment are shown in Appendix C. In Appendix D, inpainting results of two video sequences are shown for the BAKE-Openpose inpainting comparison experiment. Lastly, the results of the candidate set selection experiment are shown in Appendix E.

CHAPTER 2

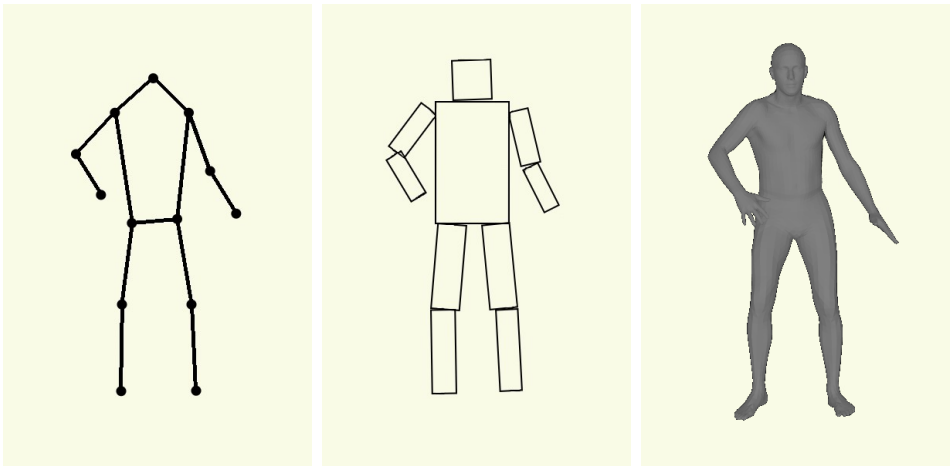
HUMAN POSE ESTIMATION

Human pose estimation is an important problem which founds wide-spread applications in the literature. In 2D, it is defined as the detection of the body part locations in a given image with respect to a graphical body model. In this section, previous works on human pose estimation are presented. In addition, 3D human pose modeling is discussed.

2.1 Previous Works on Human Pose Estimation

Human pose estimation has many applications in virtual reality, human-computer interaction, video surveillance, medical assistance, etc. [16] A predefined body model is required for pose estimation applications. In general, there are three types of body models that exist in the literature [17, 16]: Skeleton-based, cardboard-based, and volume-based models as shown in Figure 2.1. Skeleton-based models consist of a set of joints and the connections between joints correspond to related body parts. Used skeleton models vary in terms of included joints in different algorithms [18, 1]. In cardboard-based body models, body parts are represented with rectangular blocks and in volume-based body models, 3D mesh models are used.

In the Bayesian occluded pose estimation part of this study, a 13-keypoint skeleton-based body model is utilized. The keypoints in the model are nose, left/right shoulder, left/right elbow, left/right wrist, left/right hip, left/right knee, and left/right ankle. This body model includes the joints that can act independently, so it simplifies our analysis by including only the major keypoints. Openpose [1], the base network used in this study, uses a 25-keypoint skeleton-model and the COCO dataset [14] uses a 17-



(a) Skeleton-based (b) Cardboard-based (c) Volume-based

Figure 2.1: Body models used in the human pose estimation problem

keypoint model. They commonly include eyes and ears as extra keypoints, however, in 3D, relative positions of eyes and ears with respect to nose keypoint can be assumed as constant when the orientation of the head is known. Openpose also includes 3 extra keypoints per foot (heel, big toe, and small toe keypoints), the mid-hip keypoint and neck keypoint different than our model. A similar case is also valid for extra foot keypoints as in face keypoints, so it could be assumed that their relative position to ankles in 3D does not change when the leg orientation is known. Furthermore, mid-hip and neck keypoints are at the middle of right/left hip and shoulder keypoints. Therefore, it is not necessary for our analysis to insert these keypoints in the used skeleton-model.

2.1.1 2D Human Pose Estimation Methods

It is useful to mention the general human pose estimation algorithms briefly, before the specific problem of pose estimation under partial occlusions, since we adapt to use a popular pose estimator Openpose. It is a widely studied problem for the past 2 decades. There are 2 main approaches for the pose estimation problem in 2D: bottom-up and top-down approaches. In bottom-up architectures, location of each keypoint is predicted using the visual cues and further, those keypoints are linked with each

other for assigning each keypoint to a single individual with different strategies. In top-down approaches, firstly the individuals are detected with some object detectors, and then, the keypoint locations are tried to be found inside the detection boxes for each individual. Both frameworks have some advantages and disadvantages as stated in [16, 19]. Note that this study focuses on single person pose estimation problem, so the used video sequences contain only one human. Thus, we do not consider the problem of assigning found keypoints to different individuals in the image.

Before the increasing popularity of neural networks, hand-crafted visual features like HOG features are used for estimating 2D keypoint locations in images. Such methods are summarized in [17]. After the success of convolutional neural networks in object classification and detection problems, it is started to be used in the pose estimation problem and state-of-the-art methods are generally based on CNN structures [18, 20, 1, 21]. Among those, Openpose is a widely known open-source algorithm. It is also a CNN-based architecture for multi-person 2D keypoint estimation and used in this study. It includes the concatenation of CNN blocks, which consists of two sub-blocks. The first one is responsible for predicting PAFs (part affinity fields) and the second sub-block predicts the confidence maps (spatial distribution of each keypoint). PAFs (Part affinity fields) are the attractive point of Openpose where they imply the relationships between different keypoints, in other words, the body limbs. Such relations are useful for predicting unseen body joints since the anatomical structure of the human body is learned and utilized in this way. The concatenation of PAF and confidence map structures provides the capability of capturing features in a hierarchical order. Therefore, Openpose can successfully extract the visual features in images and is good at estimating the visible keypoints.

2.1.2 Performance Measure for 2D Human Pose Estimation

- Mean Per Joint Position Error in Pixels (MPJPE) [22]:
  MPJPE is the mean of the Euclidean distances of predicted joints and ground truth joints. It is originally used in [23] for 3D human pose estimation in millimeters. However, it could also be used for pixel coordinates as in [22]. MPJPE

can be expressed as given below.

$$MPJPE = \frac{1}{N_j} \sum_{m=1}^{N_j} \|\mathbf{p_m^{pred}} - \mathbf{p_m^{gt}}\|_2 \tag{2.1}$$

where N_j is the number of joints, $\mathbf{p_m^{pred}}$ and $\mathbf{p_m^{gt}}$ are the predicted and ground truth position vectors of m^{th} joint, respectively.

2.1.3 Methods for Human Pose Estimation under Partial Occlusion

State-of-the-art neural network based methods like Openpose are good at estimating the visible keypoints, however, when partial occlusions emerge, their performances decrease significantly. For Openpose, PAFs make it possible to estimate the partial occlusions, however, it does not focus on completing the missing human joints. It works on single images and does not use the person-specific kinematic and anatomical information which may be hidden in the previous image sequence.

In order to overcome the problems encountered when partial occlusions exist, data augmentation strategies are utilized. In [9], a two-stage structure that uses occlusion-aware bounding boxes is proposed. Since the method includes a segmentation block in the architecture, it is proposed to utilize person detection datasets. In [24], a data augmentation strategy that creates occlusions synthetically on available pose estimation datasets is used for training the CNN network. In [25], for dealing with self occlusions, existing datasets are enriched by using the 3D cylinder man model. In [26], a new dataset is introduced for occlusions that emerged from multi-person interactions. Although the above strategies improved the obtained results, they do not insert the person-specific anatomical information like body ratios into the models. They tend to give general estimations in case of partial occlusions.

There are some methods that employ extra information sources such as anatomical constraints or temporal information into their models for partially occluded human pose estimation. In [25, 27], temporal information is imported to the model via an architecture capable of taking multiple frames as input and imposing optical flow constraints. However, these methods have short-term memory and long duration occlusions cause an increase in errors. Although joint angle limits are used in [28],

importing anatomical constraints without temporal information is also problematic. In [10, 29], sparse formulations are used through matrix recovery methods for reconstructing the occluded 3D keypoints. Depth information is also used for 3D human pose estimation in [30, 31]. In some 3D human pose estimation approaches, Euclidean distance matrices (EDMs) for keypoint pairs are used [32, 33] for the recovery of occluded keypoints. The above methods usually use general constraints for the estimation of occluded joints which returns reasonable results only for certain cases.

The proposed Bayesian approach utilizes both general and local data in contrast to mentioned strategies. The local data provides temporal information and person-specific anatomic constraints to the model which is more suitable to complete the occluded human joints in different scenarios.

2.2 3D Human Body Models

In the application of inpainting partially occluded human images, we propose to use 3D body models, so it is useful to mention them. In general, used models are triangle meshes. These meshes consist of a certain number of vertices and connection patterns. Set of vertices can be defined as,

$$G = \{g_1, ..., g_n\} \tag{2.2}$$

The connection patterns include the edges and triangles or triangular faces. Edges, e_i, are the lines between vertices while triangles, f_i, are formed by the edges. For a defined 3D mesh, all possible edges and triangles are not used, instead sets of edges, ϵ, and triangles, F, are predefined as,

$$\epsilon = \{e_1, ..., e_k\}, e_i \in GxG \tag{2.3}$$

$$F = \{f_1, ..., f_m\}, f_i \in GxGxG \tag{2.4}$$

with certain number of elements.

There are different works in the literature to model the 3D human body mathematically. Creating realistic 3D human bodies in a variety of body poses with different body shapes and rigging the body meshes with a skeletal structure in order to animate them are some of the aims [34]. SCAPE [35] and SMPL [15] are popular 3D body models used in the computer graphics literature. They both utilize triangular meshes and learn parameters from 3D scans of different people in a variety of body poses. SCAPE models 3D bodies with pose and shape parameters by the deformation of triangles from a template body, while SMPL models them with vertex displacements. In this study, SMPL body model is used for the inpainting application, since it is a more accurate model than SCAPE as stated in [15]. Template triangular mesh of SMPL body which contains 6890 vertices with 13776 triangles is given in Figure 2.2. SMPL models the vertex positions using given pose parameters, given shape parameters, and the learned model parameters.

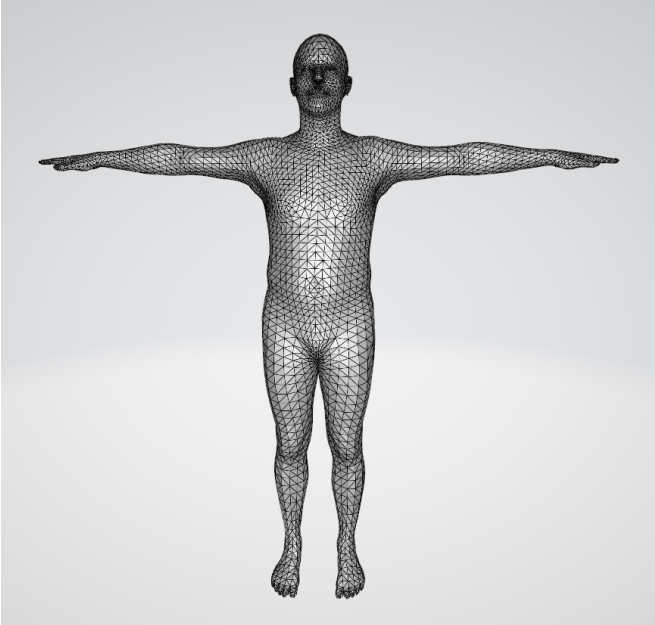


Figure 2.2: SMPL 3D body template constructed by triangles

CHAPTER 3

OCCLUSION DETECTION

Occlusion detection is the first step of the proposed methods. It can be defined as the detection of frames where the target person is occluded on a given video sequence. In the proposed human image inpainting method, it is required to detect which part of the body is occluded if occlusions exist. Moreover, in the Bayesian occluded pose estimation method, visible and occluded keypoints have to be determined. In experiments of both methods, it is assumed that the used video sequences are stationary, the scenes are static where only one person is moving and there are available frames where the target person is not in the scene, for simplicity. Furthermore, occlusions in video sequences are created with black boxes in order to perform controlled experiments and three types of occlusions are assumed to exist: upper body occlusions, middle body occlusions, and lower body occlusions. An example video frame based on the given scenario is shown in Figure 3.1. Then, an occlusion detection strategy is proposed to detect the occurrence of occlusion, the type of occlusion, and the occluded keypoints on a given frame. Note that any occlusion detection and segmentation strategy could also be used as a prior step of the BAKE or the inpainting method. The proposed occlusion detection strategy performs well enough in the test scenarios and it is also expected to work on natural occlusion cases on stationary videos.

3.1 Proposed Method for Occlusion Detection on Monocular Stationary Video

A foreground extraction algorithm and Openpose [1] are utilized collaboratively for the occlusion detection strategy. Keypoint estimates of Openpose are found and reduced to the 13-keypoint model where the keypoints were given in Figure 3.2. Using

Figure 3.1: An example video frame based on the given scenario

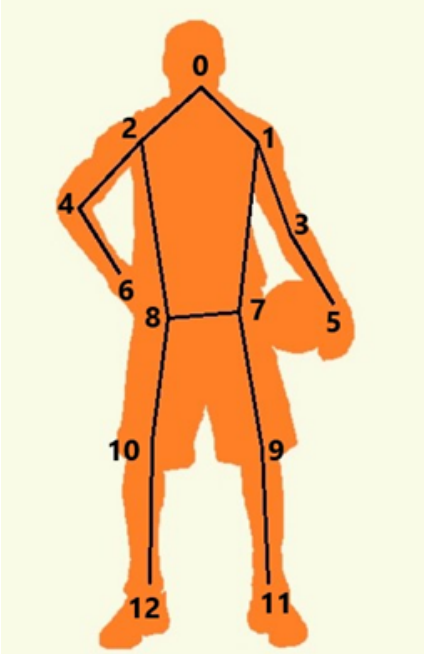


Figure 3.2: 13-keypoint skeleton body model

the Openpose estimates, the detection box is defined as the smallest box that contains the reduced Openpose estimates.

Then, a foreground mask that gives the pixel regions of moving foreground objects is required. The detection box contains the information of the whole body while the foreground mask gives information of only the visible parts of the human body in the video sequence. A Gaussian Mixture Model (GMM) based foreground extraction

algorithm is employed for this purpose which is given in [13]. This method estimates a GMM probability density function for each pixel in a recursive update manner by using the frames where the target person is out of the scene. Pixel values varying at the static background image are modeled in this way. Then, at inference, if the corresponding likelihood of a pixel is smaller than a predefined threshold, it is identified as a foreground pixel, since it does not fit the trained background model.

Obtained foreground mask with GMM based background subtraction method, $\mathbf{M}_0$, is defined as,

$$\mathbf{M}_0[m,n] = \begin{cases} 1, & \text{if (m,n) corresponds to a foreground pixel} \\ 0, & \text{if (m,n) corresponds to a background pixel} \end{cases} \tag{3.1}$$

and an example is shown in Figure 3.3a for the given frame in Figure 3.1, In order to focus only on the region where the target person exists, a larger detection box is defined. This box is the expansion of the detection box with a certain number of pixels in each direction with the top left corner coordinates, x_0^{ld}, y_0^{ld} and bottom right corner coordinates, x_1^{ld}, y_1^{ld}. Outside of the larger detection box is taken as the background region. In this way, limited foreground area, M_1 is obtained as,

$$\mathbf{M}_1[m,n] = \begin{cases} \mathbf{M}_0[m,n], & \text{if } x_0^{ld} < m < x_1^{ld} \text{ and } y_0^{ld} < n < y_1^{ld} \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

where the example frame is given in Figure 3.3b. However, there could be still noisy foreground estimates inside the larger detection box. Furthermore, the foreground human body consists of separate connected parts as the output of the background subtraction method. In order to clean the noise and unify the foreground human silhouette inside the larger detection box, morphological operations, i.e., opening and closing are used. Opening is applied first to filter out small undesired foreground regions. Obtained mask, $\mathbf{M}_2$ is given as,

$$\mathbf{M}_2[m,n] = \mathbf{M}_1[m,n] \circ \mathbf{K}_o[m,n] = (\mathbf{M}_1[m,n] \ominus \mathbf{K}_o[m,n]) \oplus \mathbf{K}_o[m,n] \tag{3.3}$$

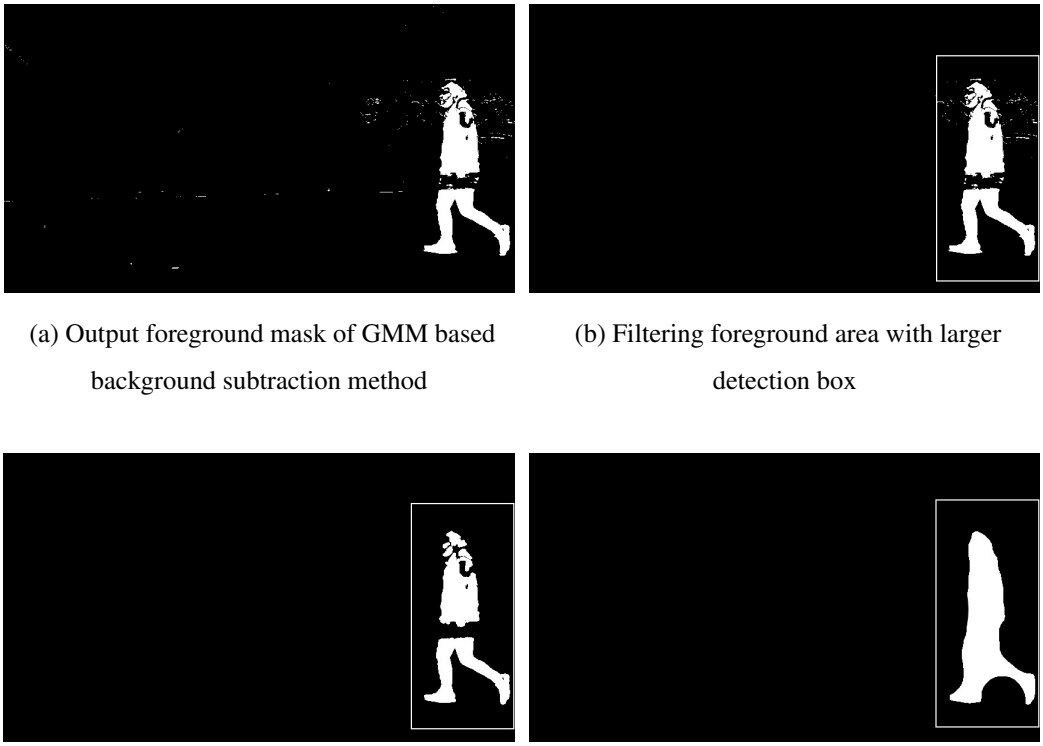
where $\circ$ is the opening operation, $\ominus$ is the morphological erosion operation, $\oplus$ is the morphological dilation operation and $\mathbf{K}_o$ is the structural element used in the opening process. Opening is applied on Figure 3.3b and the result is shown in Figure 3.3c. Then, in order to connect the separated parts of the resultant foreground human body, closing operation is applied and the final foreground mask, $\mathbf{M}_3$ is obtained as,

$$\mathbf{M}_3[m,n] = \mathbf{M}_2[m,n] \bullet \mathbf{K}_c[m,n] = (\mathbf{M}_2[m,n] \oplus \mathbf{K}_c[m,n]) \ominus \mathbf{K}_c[m,n] \quad (3.4)$$

where $\bullet$ is the closing operation and $\mathbf{K}_c$ is the structural element used in closing. The result of the closing operation on the Figure 3.3c, which will be mentioned as the resultant foreground human silhouette for further parts, is given in Figure 3.3d. Used structural elements for opening and closing operations are defined as circular elements with different sizes. Sizes of structural elements are determined by investigating the target video. Different sized elements could be chosen depending on the video. Furthermore, in $\mathbf{M}_3$, the smallest box containing the foreground area is defined as the foreground box and used in further steps.

The existence of the target person in the frame is assumed to be known apriori. Then, Algorithm 1 is used to detect the occlusion and its type. In Step 1 of Algorithm 1, predictions of Openpose are investigated for video frames where the target is in the scene. If the head keypoint, $\mathbf{p}_0$ is not predicted by Openpose while at least one foot keypoint ($\mathbf{p_{11}}$ or $\mathbf{p_{12}}$) is predicted, the occlusion exists at the top part of the body. For the opposite case where Openpose predicts $\mathbf{p}_0$ but could not predict any foot keypoint, ($\mathbf{p_{11}}$ or $\mathbf{p_{12}}$), it is concluded that the occlusion occurs at the bottom part. If $\mathbf{p}_0$ is detected with at least one foot keypoint, we conclude that Openpose can predict the whole body. Then, the resultant foreground human silhouette is investigated for further steps. In step 2, the existence of mid-body occlusions is searched by connected component labeling. If there are multiple connected components in the resultant foreground human silhouette, it is concluded that the occlusion exists in the middle part of the target human body as shown in Figure 3.4.

If there is only one connected component, an area ratio score for extracting the occlusion information from the foreground box and detection box is proposed to be used in



(a) Output foreground mask of GMM based background subtraction method

(b) Filtering foreground area with larger detection box

(c) Result of opening operation after using the larger detection box for filtering

(d) Result of closing operation after opening operation

Figure 3.3: Process to obtain the resultant foreground human silhouette with proposed steps

step 3. Thresholding using this score gives the result of whether the occlusion exists or not on the target person. Then, the type of occlusion is detected by investigating the centroid of foreground box position with the detection box. If the centroid is closer to the upper side of the detection box, it means the occlusion is at the lower part of the body and for the opposite condition, occlusion is at the upper part. There are two alternatives for the score to detect occlusions. The first one is intersection over union (IoU) which is the ratio of intersected area to the unified area of the boxes. The second one is the intersection area of the boxes over the detection box area (IoBox). The rationale of choosing such a metric is that detected keypoints stay inside the resultant foreground human silhouette generally for non-occluded cases, so the IoBox becomes one, while the IoU score becomes less than one in general since the foreground box and detection boxes do not fit exactly. When occlusions exist, both scores become similar. Therefore, the second metric, IoBox, is more discriminative for this task as

Algorithm 1 Occlusion Detection on Target Person

Step 1: Investigate the detection of head and foot keypoints by Openpose.

if ($\mathbf{p}_0$ is detected) AND ($\mathbf{p}_{11}$ and $\mathbf{p}_{12}$ are not detected) **then**

OCC = True, OCC TYPE = Bottom

else if ($\mathbf{p}_0$ is not detected) AND ($\mathbf{p}_{11}$ or $\mathbf{p}_{12}$ is detected) **then**

OCC = True, OCC TYPE = Top

else if ($\mathbf{p}_0$ is detected) AND ($\mathbf{p}_{11}$ or $\mathbf{p}_{12}$ is detected) **then**

Step 2: Investigate if the mid part of the body occluded using the number of foreground connected components, N_{fg}.

if $N_{fg} > 1$ **then**

OCC = True, OCC TYPE = Mid

else

Step 3: Investigate the area ratio score ($S_{fg,det}$) with threshold, $T_{fg,det}$, y coordinates of foreground box center, y_c^{fg}, upper side of detection box, y_0^{det} and lower side of detection box, y_1^{det} to detect occlusion.

if ($S_{fg,det} < T_{fg,det}$) AND ($|y_0^{det} - y_c^{fg}| > |y_1^{det} - y_c^{fg}|$) **then**

OCC = True, OCC TYPE = Top

else if ($S_{fg,det} < T_{fg,det}$) AND ($|y_0^{det} - y_c^{fg}| \leq |y_1^{det} - y_c^{fg}|$) **then**

OCC = True, OCC TYPE = Bottom

else

OCC = False

end if

end if

end if

the experiments also showed.

If an occlusion is detected on the target person, the occluded keypoints have to be identified for the proposed Bayesian approach, BAKE. Algorithm 2 summarizes the used strategy. All of the keypoints in the defined body model are investigated to this end as stated in Algorithm 2. If a keypoint is predicted by Openpose and its predicted position corresponds to a foreground pixel, it is classified as a visible keypoint. If Openpose could not predict a joint or the predicted joint stays in the background area, that joint is classified as an occluded one.

(a) Detected keypoints by Openpose on mid occlusion occurred frame

(b) Image masked with foreground mask. Blue box is larger the detection box, purple box is the detection box and cyan boxes are the foreground boxes which contain foreground connected components

Figure 3.4: Example case where mid body occlusion occurs

Occurrence of occlusion with its type (upper body, mid-body, and lower body occlusions) and the occluded keypoints are detected using the given procedure. In the following part, occlusion detection scores IoU and IoBox will be compared.

3.2 Comparison of IoU and IoBox Scores for Occlusion Detection

IoU and IoBox are the proposed area ratio scores for the occlusion detection procedure. If the existence of occlusion could not found by steps 1 and 2 of Algorithm 1, the area ratio score is compared with a threshold as given in step 3. In order to compare the performance of these scores, several experiments are done. For a given video sequence, synthetic occlusions for the upper and lower body are generated in four different severity levels for both. Example frames for generated upper occlusions are given in Figure

Occluded and non-occluded frames are labeled in all video sequences. Then, detection results are obtained for both IoU and IoBox scores by iterating the threshold values. In order to compare results, receiver operating characteristic (ROC) curves for all levels of upper and lower occlusions are drawn by using the obtained results. The vertical axis of the ROC curve is the true positive rate. It is the ratio of the num-

Algorithm 2 Identification of Visible and Occluded Keypoints

Identify the set of occluded keypoints, OC, and the set of visible keypoints, VIS, for Bayesian occluded pose estimation with the resultant foreground human silhouette, M_3, if an occlusion occurs.

if OCC == True **then**
  $VIS = \emptyset, OC = \emptyset$
  for $i = \{0, 1, .., 12\}$ **do**
    if $\mathbf{p}_i = (p_i^0, p_i^1)$ is detected **then**
      if $\mathbf{M}_3(p_i^0, p_i^1) = 1$ **then**
        $VIS \leftarrow VIS \cup \{i\}$
      else
        $OC \leftarrow OC \cup \{i\}$
      end if
    else
      $OC \leftarrow OC \cup \{i\}$
    end if
  end for
else
  $VIS = \{0, 1, ..., 12\}, OC = \emptyset$
end if

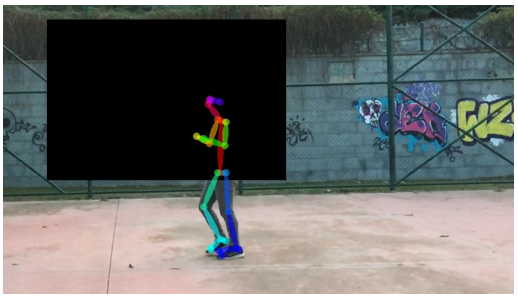
ber of accurately detected occluded frames over the number of all occluded frames. The horizontal axis of the ROC curve is the false positive rate which is the ratio of the number of false alarms (frames detected as occluded but actually not occluded) over the number of all non-occluded frames. In a ROC curve, it is desired to pick a threshold that gives a result close to 1 in true positive rate and 0 in false positive rate which corresponds to detecting all occluded frames accurately while not detecting non-occluded frames as occluded. Obtained ROC curves for the upper occlusion scenario are given in Figure 3.6 while curves for the lower occlusion scenario are given in Figure 3.7.

When the results are compared for IoBox and IoU, IoBox is a better score for both upper and lower occlusions, since the areas under the IoBox ROC curves are larger than the corresponding IoU areas. Furthermore, as the occlusion severity level increases,

(a) Example frame for level 1 upper occlusion

(b) Example frame for level 2 upper occlusion

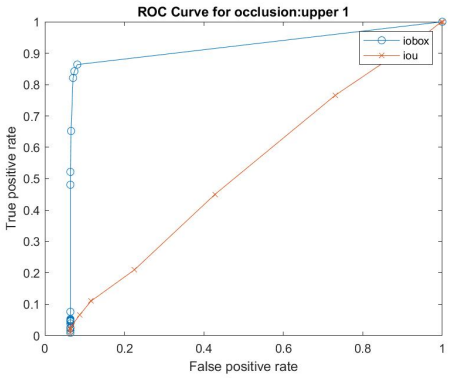
(c) Example frame for level 3 upper occlusion



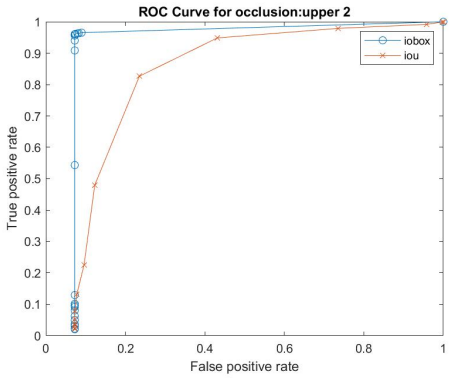
(d) Example frame for level 4 upper occlusion

Figure 3.5: Example frames for different levels of upper body occlusions

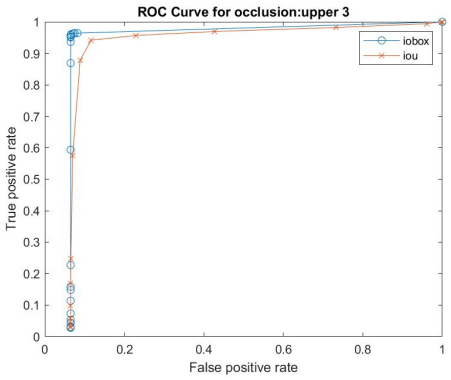
detection results get better as seen from ROC curves, since the area under curves also increases for both scores. As the results of the experiments, IoBox score is used in other videos for the occlusion detection task.



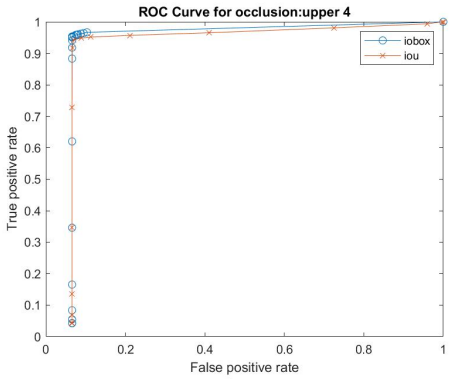
(a) ROC Curve for level 1 upper occlusion



(b) ROC Curve for level 2 upper occlusion

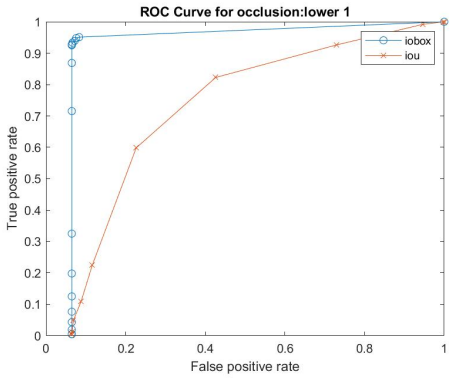


(c) ROC Curve for level 3 upper occlusion

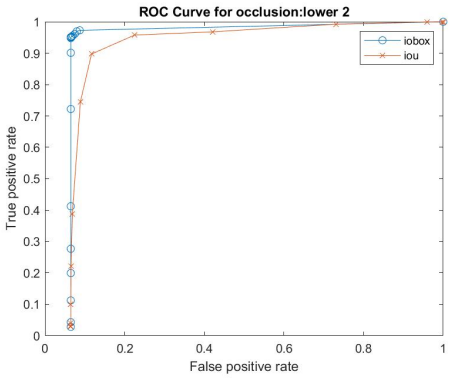


(d) ROC Curve for level 4 upper occlusion

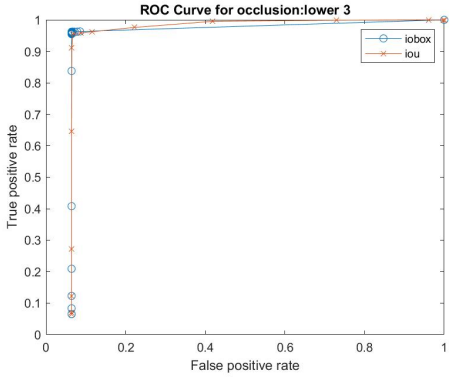
Figure 3.6: ROC Curves for different levels of upper body occlusions



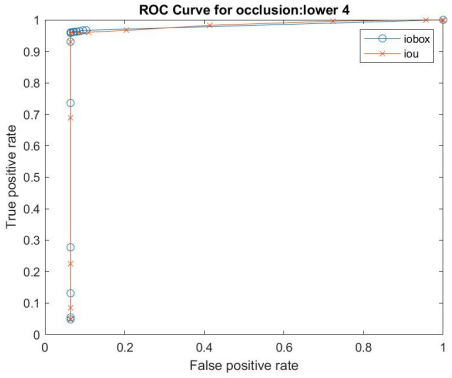
(a) ROC Curve for level 1 lower occlusion



(b) ROC Curve for level 2 lower occlusion



(c) ROC Curve for level 3 lower occlusion



(d) ROC Curve for level 4 lower occlusion

Figure 3.7: ROC Curves for different levels of lower body occlusions

CHAPTER 4

BAYESIAN POSE ESTIMATION UNDER PARTIAL OCCLUSIONS

A Bayesian approach for the estimation of partially occluded keypoints is presented in this chapter. The method is explained in detail in the following part and experimental results are presented in Section 4.2.

4.1 Bayesian Approach for Occluded Keypoint Estimation

On partially occluded video frames, pose estimation algorithms that use visual features are not effective. A strategy to embed the motion character and anatomical features of the target person is required when partial occlusions occur, since visual features are not observable. In this thesis, a Bayesian framework is proposed to estimate occluded keypoints by using the visible ones. A well-known CNN-based pose estimator Openpose [1] is used to detect visible keypoints and using the statistical information obtained from the COCO dataset [14] and non-occluded video frames, estimation of occluded keypoints are performed. In the following part, used body model and the statistical model parameters will be introduced. Then, how the estimation process is performed for a single occluded keypoint will be explained.

4.1.1 Body Model and Parameters of The Method

A skeleton body model is used which contains 13-keypoints in this method. These keypoints are the major body keypoints that can represent a human body and motion effectively as stated in Section 2.1. In Figure 4.1, used skeleton model is shown with black points and lines.

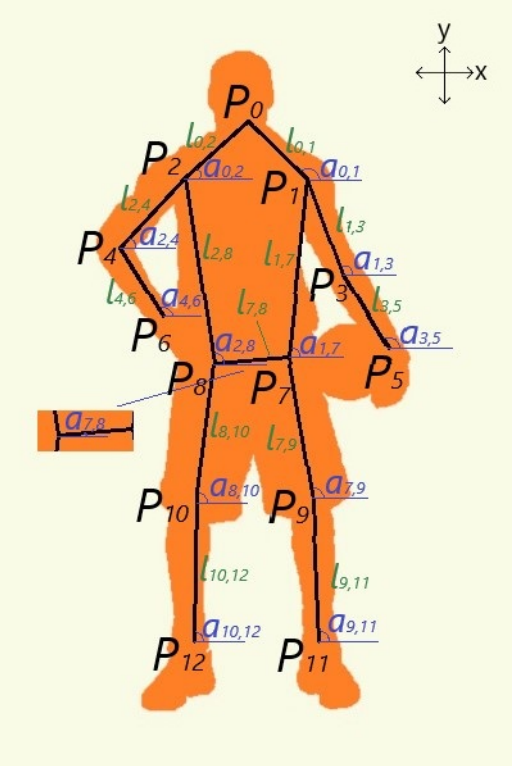


Figure 4.1: 13-keypoint skeleton human body model used in this paper. Keypoints are numbered from $\mathbf{p}_0$ to $\mathbf{p}_{12}$. Lengths between the keypoints are represented by $l_{k,n}$. Angle for each keypoint pair is represented by $a_{k,n}$.

Using the 13-keypoint skeleton model, length parameters, $l_{k,n}$, and angle parameters, $a_{k,n}$ where $k = 0, ..., 11,\ n = 1, ..., 12,\ k < n$ are defined. These parameters are obtained from the combinations of all keypoint pairs. For the 13-keypoint model, $\binom{13}{2} = 78$ length and angle parameters exist. In Figure 4.1, some of these parameters are shown. The length values are defined as the Euclidean distances in pixels normalized by a total body length parameter which is given as,

$$l_{k,n} = ||\mathbf{p_k} - \mathbf{p_n}||_2/T, \quad k = 0, ..., 11, \quad n = 1, ..., 12, \quad k < n \tag{4.1}$$

where $\mathbf{p}_k$ is the position vector of k^{th} keypoint in pixel coordinates and the total body length, T, is defined as the sum of segment lengths in 13-skeleton body model and given as,

$$T = ||\mathbf{p_0} - \mathbf{p_1}||_2 + ||\mathbf{p_0} - \mathbf{p_2}||_2 + \\ ||\mathbf{p_1} - \mathbf{p_3}||_2 + ||\mathbf{p_1} - \mathbf{p_7}||_2 + ... + \\ ||\mathbf{p_{10}} - \mathbf{p_{12}}||_2 + ||\mathbf{p_9} - \mathbf{p_{11}}||_2 \quad (4.2)$$

Note that such a normalization provides scale invariance for the length parameters. Then, the angle parameter, $a_{k,n}$ is defined as the angle between the x coordinate axis and the vector $\mathbf{p}_k - \mathbf{p}_n$ in the counterclockwise direction. It is defined as,

$$a_{k,n} = \begin{cases} cos^{-1}(\mathbf{u_{k,n}(0)}), & \text{if } \mathbf{u_{k,n}(1)} \geq 0 \\ 2\pi - cos^{-1}(\mathbf{u_{k,n}(0)}), & \text{if } \mathbf{u_{k,n}(1)} < 0 \end{cases} \\ k = 0, ..., 11, \quad n = 1, ..., 12, \quad k < n \quad (4.3)$$

where $\mathbf{u}_{k,n}$ is also defined as the 2-dimensional unit vector given as,

$$\mathbf{u_{k,n}} = \frac{\mathbf{p_k} - \mathbf{p_n}}{||\mathbf{p_k} - \mathbf{p_n}||_2} \quad (4.4)$$

Then, the parameter vector $\mathbf{x} = [l_{0,1},\ l_{0,2}, ..., l_{11,12},\ a_{0,1},\ a_{0,2}, ...a_{11,12}]$ is the vector which consists of all the length and angle parameters.

4.1.2 Positional Distribution Estimation of a Single Occluded Keypoint

Parameter vector $\mathbf{x}$ is used for the positional distribution estimation of a target occluded keypoint, $\mathbf{p}_k$, given a partially occluded body annotation and total body length, T. Note that the visible and occluded keypoints are assumed to be known and grouped with the sets VIS and OC. Then, $\mathbf{x}$ is a random vector for our case and it is assumed that $\mathbf{x}$ has a multivariate Gaussian probability density function, $f_{\mathbf{x}}(\mathbf{x})$, which is given as,

$$f_{\mathbf{x}}(\mathbf{x}) = N(\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x}) \quad (4.5)$$

$\boldsymbol{\mu_x}$ is the mean vector and $\boldsymbol{\Sigma_x}$ is the covariance matrix for $\mathbf{x}$. These terms are estimated through an update procedure by using the global and local estimates. The global estimates are obtained from the COCO database [14] and they are denoted as $\boldsymbol{\mu_{gl}}$ and $\boldsymbol{\Sigma_{gl}}$. The local variables $\boldsymbol{\mu_{lc}}$ and $\boldsymbol{\Sigma_{lc}}$ are obtained from the statistics extracted from the non-occluded frames of the evaluated video sequence. Then, the update equation for $\boldsymbol{\mu_x}$ and $\boldsymbol{\Sigma_x}$ are given as,

$$\boldsymbol{\mu_x} = (1-\alpha)\boldsymbol{\mu_{gl}} + \alpha\boldsymbol{\mu_{lc}} \tag{4.6}$$

$$\boldsymbol{\Sigma_x} = (1-\alpha)\boldsymbol{\Sigma_{gl}} + \alpha\boldsymbol{\Sigma_{lc}} \tag{4.7}$$

where $\alpha < 1$ is a scalar.

The above equations allow us to find the mean and covariance terms. In our case, some of the joints are occluded. Hence, the parameters in $\mathbf{x}$ corresponding to these joints are not observable. In order to estimate these parameters, a Bayesian framework is proposed. For this purpose, the parameters in $\mathbf{x}$ are separated in two parts, namely known parameters, $\mathbf{x_k}$, and unknown parameters, $\mathbf{x_u}$, whose parameters are obtained from the Bayesian theorem as given in (4.9) and (4.10). The statistics for $\mathbf{x}$ can be rewritten as shown below in terms of $\mathbf{x_k}$ and $\mathbf{x_u}$.

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_k} \\ \mathbf{x_u} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x_k}} \\ \boldsymbol{\mu}_{\mathbf{x_u}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x_k}} & \boldsymbol{\Sigma}_{\mathbf{x_k x_u}} \\ \boldsymbol{\Sigma}_{\mathbf{x_u x_k}} & \boldsymbol{\Sigma}_{\mathbf{x_u}} \end{pmatrix} \right) \tag{4.8}$$

Then, the statistics for the parameters of the occluded joints, $\mathbf{x_u}$, can be obtained as,

$$\boldsymbol{\mu}_{\mathbf{x_u}|\mathbf{x_k}=\tilde{\mathbf{x}}_\mathbf{k}} = \boldsymbol{\mu}_{\mathbf{x_u}} + \boldsymbol{\Sigma}_{\mathbf{x_u x_k}}\boldsymbol{\Sigma}_{\mathbf{x_k}}^{-1}(\tilde{\mathbf{x}}_\mathbf{k} - \boldsymbol{\mu}_{\mathbf{x_k}}) \tag{4.9}$$

$$\boldsymbol{\Sigma}_{\mathbf{x_u}|\mathbf{x_k}=\tilde{\mathbf{x}}_\mathbf{k}} = \boldsymbol{\Sigma}_{\mathbf{x_u}} - \boldsymbol{\Sigma}_{\mathbf{x_u x_k}}\boldsymbol{\Sigma}_{\mathbf{x_k}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x_k x_u}} \tag{4.10}$$

where $\tilde{\mathbf{x}}_\mathbf{k}$ is a realization of $\mathbf{x_k}$.

Our target is to find the position vector of the occluded keypoint using the statistics given in (4.9) and (4.10). Let $\mathbf{p_k}$ be the position vector of an occluded keypoint. We

need to find the probability mass function for this occluded keypoint by considering a continuous probability density function and the related continuous position vector, $\mathbf{p_k^c}$. The probability mass function for a given total body length can be written as,

$$p_{\mathbf{p_k}|T}(\mathbf{p_k}|T) = P(\|\mathbf{p_k} - \mathbf{p_k^c}\|_2 < \Delta_p) \tag{4.11}$$

where $2\Delta_p$ is the scalar step size of discrete valued coordinate vector in x and y directions.

In order to find a simple expression for the above probability, we need to consider only the length and angle parameters for the single occluded joint alone. These parameters for $\mathbf{p_k}$ are denoted by $\mathbf{x_r^k}$. Figure 4.2 is an explanation for this point. In Figure 4.2, red dots indicate occluded keypoints. Yellow dot is also an occluded keypoint where we would like to calculate the position vector, namely $\mathbf{p_k}$. Black dots correspond to non-occluded keypoints. Pink lines represent the unknown length parameters. Since we need to consider only the related parameters for finding the $\mathbf{p_k}$, in Figure 4.2b, blue lines are shown as the unknown segments for only the single occluded keypoint. The angle and length parameters are kept in the vector $\mathbf{x_r^k}$. Therefore, $\mathbf{x_r^k}$ can be defined as the random vector of related variables with the occluded keypoint with a probability density function $f_{\mathbf{x_r^k}}(\mathbf{x_r^k})$ which is marginalized from $f_{\mathbf{x_u}|\tilde{\mathbf{x}}_\mathbf{k}}(\mathbf{x_u}|\tilde{\mathbf{x}}_\mathbf{k})$. Using the derivation in Appendix A, the probability mass function for the occluded keypoint position $\mathbf{p_k}$ given total body length can be written approximately as,

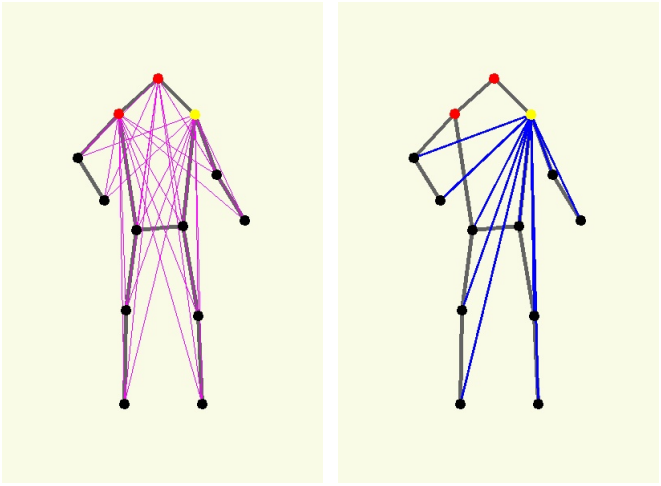
$$p_{\mathbf{p_k}|T}(\mathbf{p_k}|T) \approx \beta_1 f_{\mathbf{x_r^k}}(\tilde{\mathbf{x}}_\mathbf{r}^k) \prod_{i \in VIS} \Delta_{a_{k,i}} \tag{4.12}$$

where $\tilde{\mathbf{x}}_\mathbf{r}^k$ is the realization of $\mathbf{x_r^k}$ for given $\mathbf{p_k}$, VIS is the given visible joints set and $\Delta_{a_{k,i}}$ is given as,

$$\Delta_{a_{k,i}} = 2tan^{-1}(\frac{\Delta_p}{\|\mathbf{p_k} - \mathbf{p_i}\|_2}) \tag{4.13}$$

β_1 in (4.12) is a normalization constant.

Equation (4.12) is highly nonlinear and it is decided to be evaluated in the given image frame for each pixel coordinate. In order to decrease the computational complexity,



(a) Illustration of unknown variables

(b) Illustration of related variables

Figure 4.2: Examples of unknown and related variables for the positional distribution estimation process of an occluded keypoint. In both figures, black dots are the visible, red dots are the occluded keypoints and the yellow dot is the occluded keypoint under consideration. The gray lines form the defined body model. In (a), keypoint pairs that form the unknown variables are shown with purple lines while in (b) keypoint pairs that form the related variables are shown with blue lines.

the search region for the occluded keypoint is limited by a square region of $2T(\mu+3\sigma)$ x $2T(\mu+3\sigma)$ where μ is the mean value of the length between the occluded keypoint and the non-occluded keypoint and σ is the standard deviation for the same parameter. Multiple square regions are determined by considering different segments between the occluded keypoint and the non-occluded keypoints. The intersection of these square regions is taken as the final search region. In Figure 4.3, this process is shown.

Note that, total body length can be estimated from the image sequence at the beginning of the process where there is no occlusion. However, since the position of the person might change after the occlusion, total body length might also change. The calculations are performed in pixel dimension. Hence, such a change might generate substantial total body length differences. Therefore, we need to estimate the probability mass function of total body length, $p_T(T)$. This estimation is discussed in the following part.

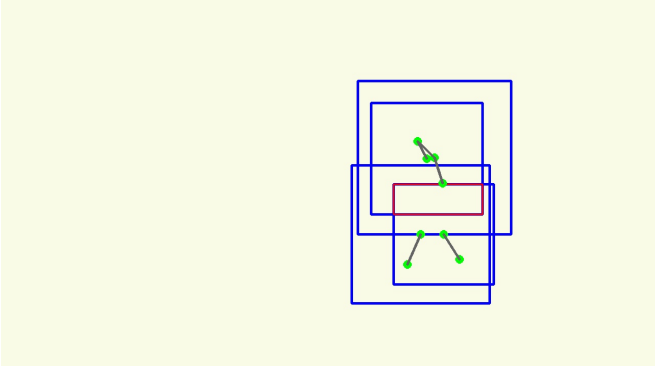


Figure 4.3: A possible search region for the right hip keypoint, ($\mathbf{p}_8$), for a partially occluded body on the whole image, is given. Green dots represent the visible keypoints. Blue boxes are obtained through the visible keypoints neighboring the occluded keypoint. The intersection of blue boxes is represented as a red box which is the possible search region.

4.1.3 Estimation of Total Body Length Probability Distribution

The total body length parameter, T given in (4.2), is an important scalar used for the scale-invariant solution of the occluded keypoint estimation. When some of the keypoints are occluded, we present a Bayesian approach to estimate the probability mass function of T. This approach is based on an approximate formulation given in Appendix B.

First, the general joint distribution for the parameter vector, $\mathbf{x}$, namely $f_{\mathbf{x}}(\mathbf{x})$ is marginalized by using the vector of known parameters, $\mathbf{x_k}$ and $f_{\mathbf{x_k}}(\mathbf{x_k})$ is obtained. Let $\mathbf{a_k}$ and $\mathbf{l_k}$ be the random vectors of known angles and lengths. Their relationship with $\mathbf{x_k}$ is given as,

$$\mathbf{x_k} = \begin{pmatrix} \mathbf{a_k} \\ \mathbf{l_k} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_{a_k} \\ \boldsymbol{\mu}_{l_k} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{a_k}} & \boldsymbol{\Sigma}_{\mathbf{a_k l_k}} \\ \boldsymbol{\Sigma}_{\mathbf{l_k a_k}} & \boldsymbol{\Sigma}_{\mathbf{l_k}} \end{pmatrix} \right) \tag{4.14}$$

Let $\tilde{\mathbf{a}}_\mathbf{k}$ be the known angle vector for a particular observation. The conditional mean $\boldsymbol{\mu}_{\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k}}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k}}$ terms can be written under the Gaussian assumption,

i.e.,

$$\boldsymbol{\mu}_{\mathbf{l_k}|\mathbf{a_k}=\tilde{\mathbf{a}}_\mathbf{k}} = \boldsymbol{\mu}_{\mathbf{l_k}} + \boldsymbol{\Sigma}_{\mathbf{l_k a_k}} \boldsymbol{\Sigma}_{\mathbf{a_k}}^{-1} (\tilde{\mathbf{a}}_\mathbf{k} - \boldsymbol{\mu}_{a_k}) \tag{4.15}$$

$$\boldsymbol{\Sigma}_{\mathbf{l_k}|\mathbf{a_k}=\tilde{\mathbf{a}}_\mathbf{k}} = \boldsymbol{\Sigma}_{\mathbf{l_k}} - \boldsymbol{\Sigma}_{\mathbf{l_k a_k}} \boldsymbol{\Sigma}_{\mathbf{a_k}}^{-1} \boldsymbol{\Sigma}_{\mathbf{a_k l_k}} \tag{4.16}$$

The Gaussian conditional distribution $f_{\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k}}(\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k})$ with its parameters in (4.15) and (4.16) is used to estimate the probability mass function of total body length parameter, T, by considering the non-normalized known length vector observation, $\tilde{\mathbf{l}}_\mathbf{k}$. This is achieved through an approximation given in Appendix B where probability mass function for T is given as,

$$p_T(T) \approx \frac{\beta_2}{T^2} f_{\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k}}((\tilde{\mathbf{l}}_\mathbf{k}/T)|\tilde{\mathbf{a}}_\mathbf{k}) \tag{4.17}$$

where β_2 is a normalization constant. Note that β_2 is selected such that sum of mass function in (4.17) is unity.

4.1.4 Estimation of the Occluded Keypoint Position

In this part, the estimation of the occluded keypoint position is discussed. This estimation is achieved by using probability mass function in (4.17) as well as the conditional probability mass function in (4.12). For this purpose, (4.17) is sampled for different values of T. For each of these T values, occluded keypoint probability mass function is estimated as,

$$p_{\mathbf{p_k}}(\mathbf{p_k}) = \sum_{n=1}^{N} p_{\mathbf{p_k}|T_n}(\mathbf{p_k}|T_n) p_T(T_n) \tag{4.18}$$

where it is assumed that T is sampled by N points, namely T_n, $n = 1, ..., N$. In Figure 4.4, probability mass function in (4.17) is shown for different values of T. Note that this type of characteristics can be sampled with a sufficiently small number of values centered around the peak. Hence, the computational complexity can be reduced by

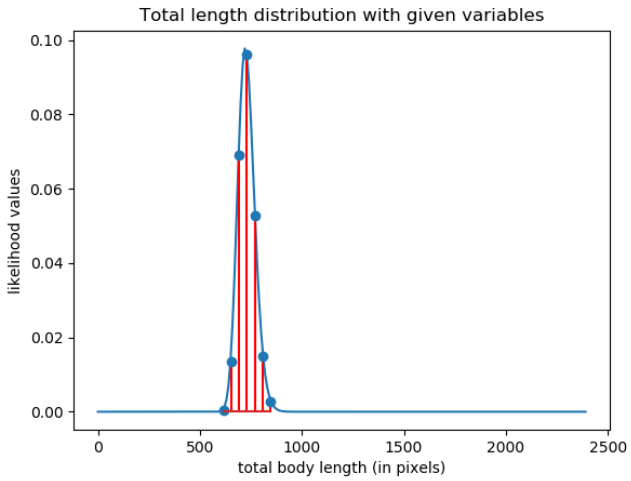


Figure 4.4: Illustration of total body length sampling process for a given partially occluded body

selecting an appropriate value for N. Note that this process ignores the small values of $p_T(T)$.

Then, the estimation of the keypoint position $\mathbf{p_k}$ is given as a maximization problem as below,

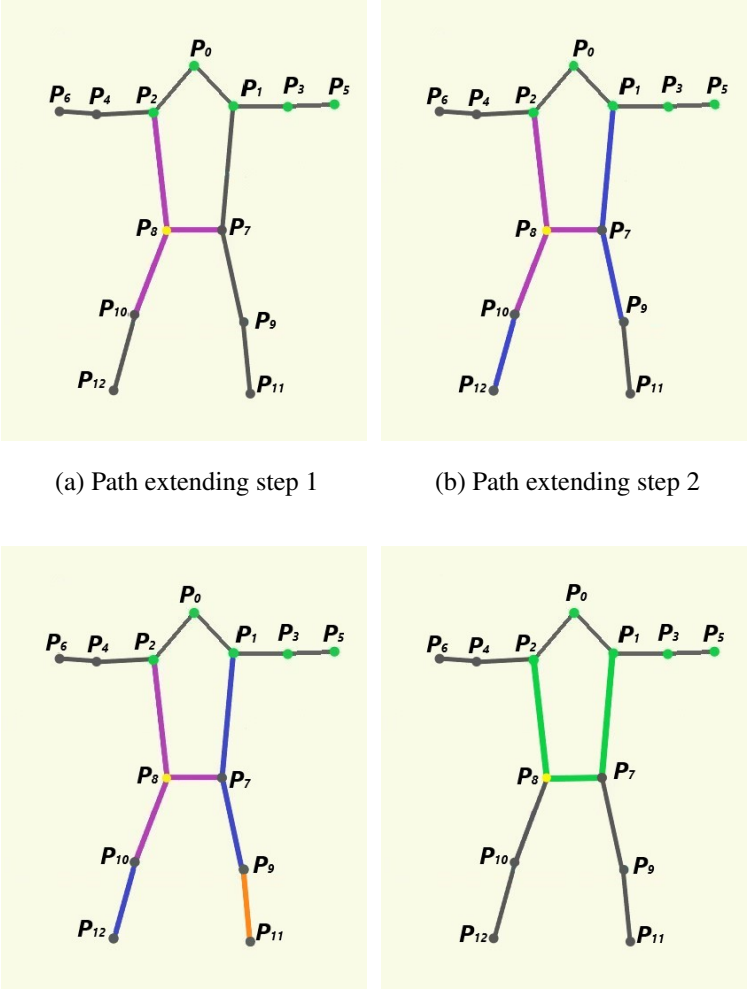
$$\mathbf{p_k^*} = \arg\max_{\mathbf{p_k}} p_{\mathbf{p_k}}(\mathbf{p_k}) \tag{4.19}$$

4.1.5 Solution Paths for Multiple Occluded Keypoints

In the previous part, solution for a single occluded keypoint is presented. In practice, the position of several occluded keypoints should be estimated. In this process, it is advantageous to use the positions of estimated keypoints for the estimation of remaining occluded keypoints. This requires a solution path estimation process for the best solution. We have considered two different strategies for the solution path estimation. It turns out that while there could be different techniques for this purpose, their solution paths, as well as their estimation errors, are not significantly different. While this is so, we selected an algorithm which is more simple and with a better estimation error. In the following part, this algorithm is explained.

The proposed approach first selects the "most suitable" occluded keypoint for estimation. This corresponds to a keypoint which has the largest number of non-occluded keypoints in its first immediate neighborhood, second immediate neighborhood, etc. in order to identify a score which gives an estimate for processing quality. Once this keypoint is found, then the solution path for the keypoint is estimated. The estimated occluded keypoint is regarded as a visible keypoint for further steps of keypoint selection.

Solution path is used to determine the related keypoints with the occluded keypoint using the 13-keypoint skeleton body model in Figure 4.1. If the solution path of an occluded keypoint includes previously estimated keypoints, those estimated keypoints are used as visible keypoints in the solution. This solution path is found by using a path extending strategy where neighboring keypoints are traced up to the end keypoints. End keypoints are defined as the $\mathbf{p}_0$, $\mathbf{p}_5$, $\mathbf{p}_6$, $\mathbf{p}_{11}$ and $\mathbf{p}_{12}$ respectively. If a visible (non-occluded) neighbor is reached, that path is closed. If no visible keypoint is reached, that path is removed from the solution. In Figure 4.5, an example of the path extending for the solution path of the right hip ($\mathbf{p}_8$) is given where green points represent the visible joints and the yellow one is the occluded keypoint under consideration. In Figure 4.5a, the first possible paths from $\mathbf{p}_8$ are shown with purple lines. One of the paths ends at a visible point $\mathbf{p}_2$ while the others continue to extend. In Figure 4.5b, the possible paths for the second step are shown with blue lines where another path reaches a visible joint, $\mathbf{p}_1$. In Figure 4.5c, the only possible path remaining is shown with an orange line. Note that, the paths reaching joints $\mathbf{p}_{11}$ and $\mathbf{p}_{12}$ are removed from the solution since they do not reach a visible joint. The resulting solution path is given in Figure 4.5d with green lines. The solution path includes $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_7$ points as shown in Figure 4.5d. In Section 2-B, the process of finding possible search regions to obtain the positional distribution of an occluded keypoint is given. In Figure 4.6, the rectangle denotes the final search region by applying this process. Then, the distribution for the keypoint position is estimated by using Equation (4.18) and this distribution is shown inside the rectangle region as a purple point cloud for the keypoint $\mathbf{p}_8$.



(a) Path extending step 1

(b) Path extending step 2

(c) Path extending step 3

(d) Solution path of right hip joint

Figure 4.5: Illustration of path extending steps: (a) - (c) and solution path: (d) for right hip joint ($\mathbf{p}_8$)

4.1.6 Pose Predictability Score Estimation

In this part, predictability scores for the overall body and individual keypoints are presented. Overall body predictability score is used to give a measure of how good the occluded keypoints are estimated. This confidence score can be seen as a measure to evaluate the overall estimation quality. The estimation of the keypoint predictability score is especially important to develop a hybrid technique presented in the next section.

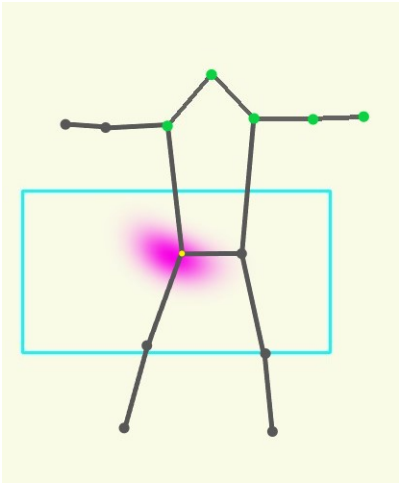


Figure 4.6: The positional probability distribution of the right hip ($\mathbf{p}_8$) for the occlusion pattern given in Figure 6. Green points are the visible keypoints, the cyan box represents the possible search region, and purple is used to represent pixels with high probability values. The yellow point is the ground truth position of $\mathbf{p}_8$.

While there are some examples in the literature for the estimation of confidence score [1], the expression for these are not completely characterizing the probability distribution of the parameter. In our case, a different confidence score is presented. This is based on the fact that the worst confidence is obtained when the positional distribution of the keypoint is uniform. Therefore, a normalized confidence score with respect to uniform distribution can better represent the statistical character of the occluded keypoint. The proposed confidence score, c_k^{BAKE}, is given as,

$$c_k^{BAKE} = b_2 log(\frac{var_{uni}}{var(\mathbf{p_k})}) \tag{4.20}$$

where $var(\mathbf{p_k})$ is the estimated variance of the occluded keypoint position using the found probability mass function, $p_{\mathbf{p_k}}(\mathbf{p_k})$, and b_2 is a constant for normalization. var_{uni} is the variance of uniform position distribution of the occluded keypoint and given as,

$$\begin{aligned} var_{uni} &= \sum_{u=1}^{W}\sum_{v=1}^{H}\frac{1}{WH}((u-\frac{W}{2})^2+(v-\frac{H}{2})^2) \\ &= \frac{(H^2+W^2+4)}{12} \end{aligned} \tag{4.21}$$

where H is the height and W is the width of the image.

Then, the total body confidence score is given as,

$$c_{tot} = \frac{\sum_{k \in OC} c_k^{BAKE}}{|OC|} \tag{4.22}$$

which is the mean of occluded keypoint confidences also proposed in [36].

4.1.7 Hybrid Prediction Approach of Occluded Joints

Bayesian approach for occluded keypoint estimation (BAKE), uses local and global information extracted from video sequence under consideration as well as COCO dataset [14]. This allows us to combine both long-term and short-term information into the estimation procedure. In contrast, Openpose [1] has only limited capacity for this purpose. Nevertheless, both approaches can estimate the occluded keypoints. While there are cases where Openpose fails to estimate some occluded keypoints, its performance in general sets a baseline for comparison. We have done several simulations to compare BAKE and Openpose. It turns out that while BAKE has a better performance than Openpose in general, there is no guarantee that this will be so for all cases. In order to take advantage of both techniques, a hybrid approach, B-OP, is presented. The combination of these two techniques is based on the confidence scores of occluded keypoints obtained by the two techniques. The position estimate for the hybrid technique can be written as,

$$\mathbf{p_k^{B-OP}} = \gamma_k \mathbf{p_k^{BAKE}} + (1 - \gamma_k)\mathbf{p_k^{OP}} \tag{4.23}$$

where $\mathbf{p_k^{BAKE}}$ is the position estimate for BAKE and $\mathbf{p_k^{OP}}$ is the position estimate of Openpose respectively. γ_k is a weighting term obtained from the confidence scores of both techniques and is given as,

$$\gamma_k = \frac{c_k^{BAKE}}{c_k^{BAKE} + c_k^{OP}} \tag{4.24}$$

In (4.24), the confidence scores for the individual methods are consistent within their

techniques but they have different scales. In order to use (4.24), we need to equalize the scale difference between confidence scores of both methods. This is achieved by considering a validation video sequence and an affine relationship for the error and the confidence score. Validation video sequence is an annotated video sequence where the ground truth, occluded and non-occluded keypoints are well defined. The affine relationship between the keypoint distance error with respect to ground truth and the confidence score is given as,

$$e = \eta_{k_0} c + \eta_{k_1} \tag{4.25}$$

where η_{k_0} and η_{k_1} are affine weights for the predictions of Openpose or BAKE for the k^{th} joint, e is the joint's prediction error, c is the confidence score for the joint.

The validation video sequence is used to estimate the η weights for the 13-keypoint skeleton body model for both Openpose and BAKE where a least-squares solution is utilized. For each of 13 joints, different sets of coefficients should be found. The distributions of errors in pixels and confidence scores for the Openpose and BAKE are obtained from the validation set. Equation (4.25) can be written in matrix form for a single keypoint as,

$$\mathbf{A_k} \boldsymbol{\eta_k} = \mathbf{e_k} \tag{4.26}$$

where $\boldsymbol{\eta_k}$ is the weight vector $[\eta_{k_0}\ \eta_{k_1}]^T$. In (4.26), N_vx2 matrix $\mathbf{A}_k = [\mathbf{c}_k \mathbf{1}]$ is composed of confidence score vector $\mathbf{c}_k$ corresponding to the scores estimated for the joint k in the validation set with N_v images. $\mathbf{1}$ is a vector of ones. $\mathbf{e}_k$ is the vector of distance errors for the k^{th} joint in the validation set. The least squares solution for (4.26) can be written as,

$$\boldsymbol{\eta_k} = (\mathbf{A_k}^T \mathbf{A_k})^{-1} \mathbf{A_k}^T \mathbf{e_k} \tag{4.27}$$

The above solution for the weights of affine relations should be computed for both Openpose and BAKE. Then, the distance errors for both Openpose and BAKE can be

calculated by using Equation (4.25). In order to have equivalence between confidence measures, the error for both cases should be the same as shown below,

$$\begin{aligned} e &= \eta_{k_0}^{OP} c_k^{OP} + \eta_{k_1}^{OP} \\ &= \eta_{k_0}^{BAKE} c_k^{BAKE} + \eta_{k_1}^{BAKE} \end{aligned} \tag{4.28}$$

The relation between two confidence measures can be written as,

$$c_k^{OP} = \frac{\eta_{k_0}^{BAKE}}{\eta_{k_0}^{OP}} c_k^{BAKE} + \frac{\eta_{k_1}^{BAKE} - \eta_{k_1}^{OP}}{\eta_{k_0}^{OP}} \tag{4.29}$$

in terms of the estimated relation weights $\eta_{k_0}^{OP}$, $\eta_{k_1}^{OP}$, $\eta_{k_0}^{BAKE}$ and $\eta_{k_1}^{BAKE}$. The above equation can be used to obtain a confidence score $\hat{c}_k^{BAKE}$ which is compatible with the c_k^{OP} measure. This is given as,

$$\hat{c}_k^{BAKE} = z_k c_k^{BAKE} + q_k \tag{4.30}$$

where z_k and q_k are the first and second coefficients in Equation (4.29).

Using $\hat{c}_k^{BAKE}$ in (4.30), we compute γ_k in (4.24). Then, the keypoint estimate for the hybrid method is obtained by using (4.23).

4.2 Experimental Results

Currently, a common database for the pose estimation under partial occlusion is not available. Previous works in the literature [9, 10, 33], use their synthetic data for evaluation. We also generated a synthetic dataset in order to compare the performances of BAKE, Openpose, and B-OP (hybrid technique in section 4.1.7). The process for data generation is similar to [9, 10, 33]. Selected frames from three image sequences are shown in Figure 4.7. These three video sequences are used as ground truth. Examples of possible synthetic occlusions for the image sequence 1 are shown in Figure 4.8. Three different occlusion scenarios are considered, namely, lower, upper, and middle occlusions respectively. For lower occlusions we consider three different occlusion levels with occlusion severity levels labeled as l1, l2 and l3 respectively, as

shown in Figure 4.8a, 4.8b and 4.8c. Similarly, three different upper occlusions are considered with severity described as u1, u2 and u3 respectively as shown in Figure 4.8d, 4.8e and 4.8f. There are only two middle occlusions, m1 and m2 as shown in Figure 4.8g and 4.8h.

Given the occluded video sequence as in Figure 4.8, the occluded and visible keypoints are determined using Algorithm 2 in Section 3.1 for each frame. Then, the occluded keypoints are estimated using the BAKE, Openpose, and B-OP algorithms as outlined in Section 4.1. The performances of these three techniques are compared using mean per joint positional error (MPJPE) in terms of pixels as it is given in [22]. Expression for MPJPE is given in Equation (2.1). For occluded pose estimation problem, it is also useful to rewrite it as,

$$MPJPE = \frac{1}{N_{OC}} \sum_{m \in OC} E_m \tag{4.31}$$

in terms of E_m where E_m is the distance error between the true and estimated occluded keypoint position in pixels and is given as,

$$E_m = \|\mathbf{p_m^{pred}} - \mathbf{p_m^{gt}}\|_2 \tag{4.32}$$

In this equation $\mathbf{p_m^{pred}}$ is the estimated keypoint position, $\mathbf{p_m^{gt}}$ is the ground truth for the same keypoint.

The summation is over all occluded keypoints in the set OC where there are N_{OC} occluded keypoints in total. When there are several frames for evaluation, the average MPJPE score should be evaluated which is given as,

$$MPJPE_{av} = \frac{1}{N_{fr}} \sum_{m=1}^{N_{fr}} MPJPE_m \tag{4.33}$$

where $MPJPE_m$ is the MPJPE score as in (4.31) for the m^{th} frame. N_{fr} is the number of test frames.

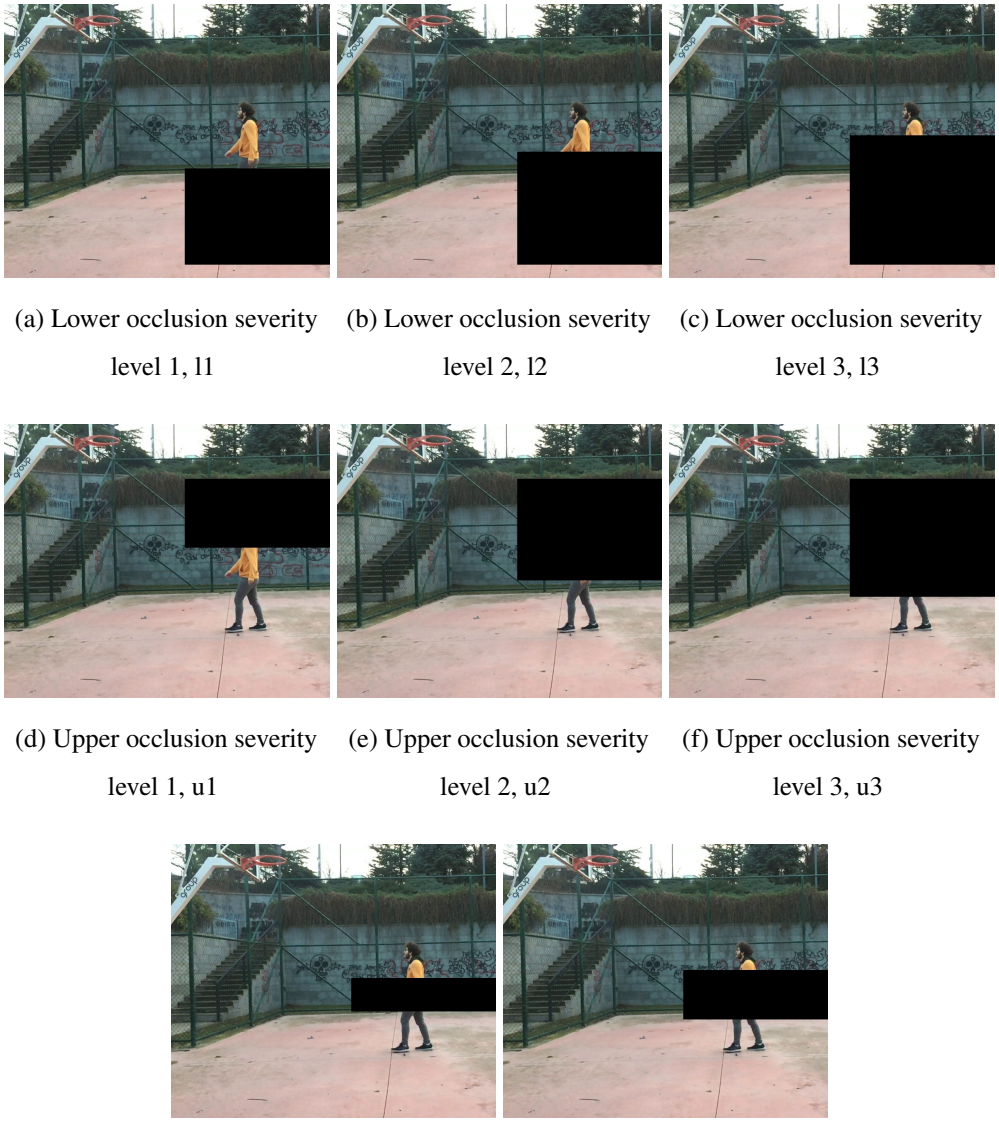
In order to see the effect of using a total body length value distribution instead of using only the maximum probable T value, an experiment is set. In sequence 1, average

(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

Figure 4.7: Example frames in test sequences. row (a) is from sequence 1, row (b) is from sequence 2, and row (c) is from sequence 3.



(a) Lower occlusion severity level 1, l1

(b) Lower occlusion severity level 2, l2

(c) Lower occlusion severity level 3, l3

(d) Upper occlusion severity level 1, u1

(e) Upper occlusion severity level 2, u2

(f) Upper occlusion severity level 3, u3

(g) Middle occlusion severity level 1, m1

(h) Middle occlusion severity level 2, m2

Figure 4.8: Examples of synthetically occluded frames in test sequence 1. Row (a) is lower occlusions, row (b) is upper occlusions and row (c) is middle occlusions.

MPJPE errors of each occlusion case for both multiple T values and single T value cases are found for this purpose. The results are shown in Figure 4.9. In general, using multiple T values decreases the error. When the occlusion severity decreases, using single T values also results in similar errors with multiple T values.

Then, the ordering strategy is investigated. For different occlusion patterns on the

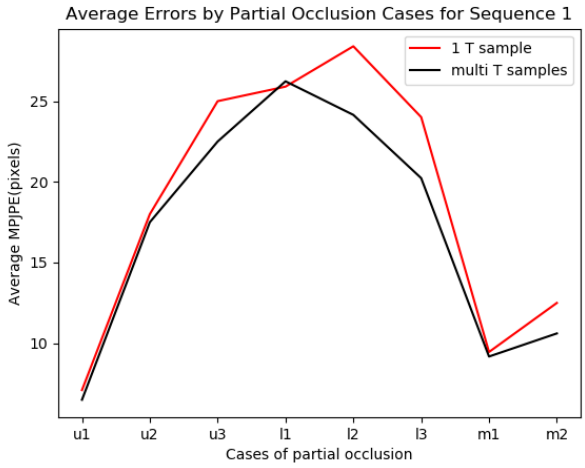


Figure 4.9: Average MPJPE results of BAKE for different strategies to use total body length T. Horizontal axis corresponds to occlusion cases where u is for upper, l is for lower and m is for mid occlusions. Vertical axis is the average MPJPE of test frames for each occlusion case.

occluded body as shown in Figure 4.10, two different ordering strategies are examined. The occlusion patterns are chosen so that the ordering strategies give different keypoint selection orders. Method 1 is the used strategy mentioned in section 4.1.5. Method 2 also utilizes the occluded neighboring keypoints. It selects the "most suitable" occluded keypoint by considering the largest number of first immediate non-occluded neighbors - occluded neighbors, second immediate non-occluded neighbors - occluded neighbors, etc.

Results are given in Figure 4.11 for this experiment. Using different ordering strategies does not affect the obtained average MPJPE errors. Therefore, the simpler strategy Method 1 is chosen to be followed.

The Bayesian approach proposed in this paper uses global and local information and combines them in Equations (4.6) and (4.7). This is an important and effective step in the proposed method. When the action for the person is steady, local information is sufficient to represent the motion statistics. On the other hand, when the action is changing in time, global information extracted from the COCO database must be used to capture the motion statistics since the local information obtained from the non-occluded previous frames does not reflect the motion statistics. These points are

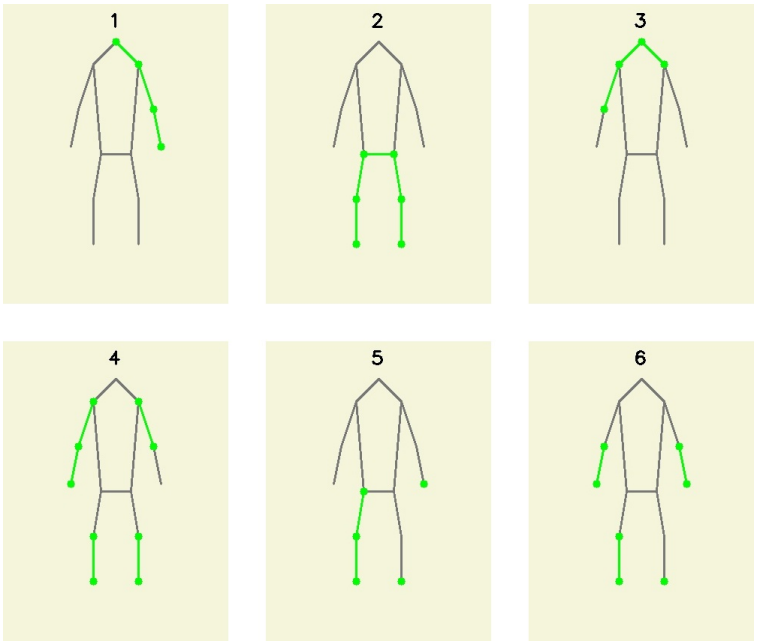


Figure 4.10: Occlusion patterns for the ordering strategy test. Green points and lines represent the hypothetical visible points where the gray lines represent unknown parts.

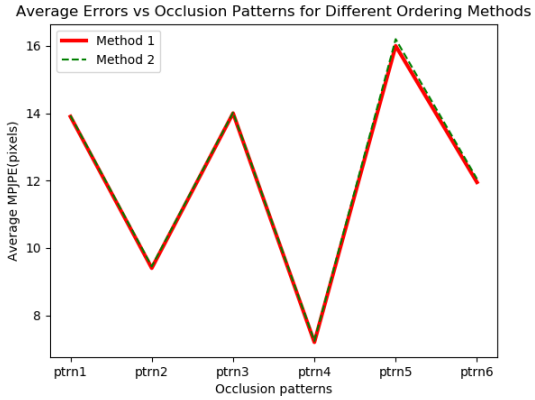


Figure 4.11: Average MPJPE results of BAKE for different occluded keypoint selection ordering strategies. Horizontal axis corresponds to occlusion patterns. Vertical axis is the average MPJPE of test frames for each occlusion case.

reflected by the first experiment which includes 350 non-occluded frames of a steady motion followed up with a similar sequence of 14 occluded frames and then, the last 14 occluded frames involve a different action than the non-occluded frames. These two sequences are shown in Figures C.1a and C.1b in Appendix (C). The proposed

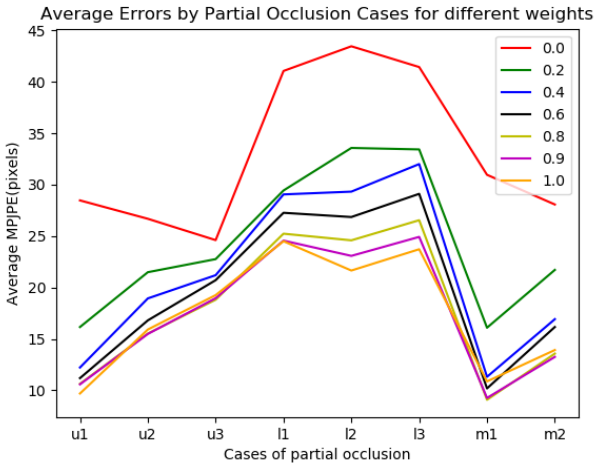


Figure 4.12: Average MPJPE results of BAKE for the first 14 frames of test sequence where walking motion is performed with different update weights: 0.2, 0.4, 0.6, 0.8, 0.9, and 1. Horizontal axis corresponds to occlusion cases where u is for upper, l is for lower and m is for mid occlusions. Vertical axis is the average MPJPE of test frames for each occlusion case.

method, BAKE, requires the estimation of 156 parameters for the 13-keypoint skeleton model. 350 non-occluded frames are used to estimate the local parameters as in Equations (4.6) and (4.7). Then, the BAKE algorithm is applied for the next 14 occluded frames to obtain MPJPE values for each occlusion case. The result is shown in Figure 4.12. Horizontal axis shows the occlusion cases with different severities as it is described in Figure 4.8. Since this 14 frame test sequence has a similar action as the previous 350 frames, the best MPJPE scores are obtained for $\alpha \in [0.8 - 1.0]$. This shows that local information obtained from the previous 350 frames is sufficient to estimate the occluded keypoints for this sequence. Then, the next 14 occluded frames, which have a different action than the 350 frames, are considered. Figure 4.13 shows the result for this sequence. As it is seen from this figure, the best MPJPE scores are obtained for small values of α. This shows that local information is not sufficient and global information should be included as in Equations (4.6) and (4.7). When we consider the first and second 14 frame sequences together, the result that is shown in Figure 4.14 is obtained. In this case, the evaluated sequence contains a mixture of actions and the best MPJPE scores are obtained for $\alpha \in [0.2 - 0.6]$.

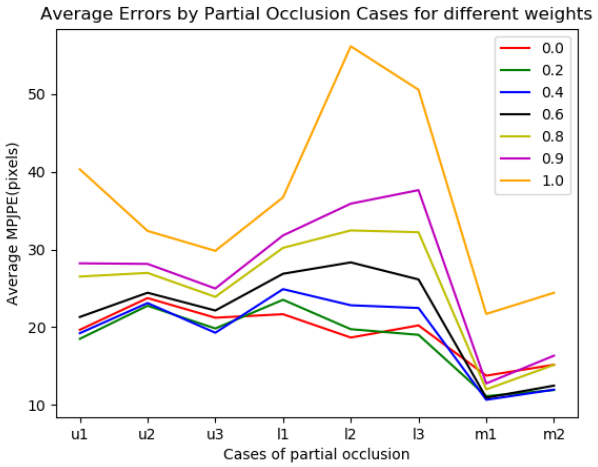


Figure 4.13: Average MPJPE results of BAKE for the second 14 frames of test sequence where a different motion is performed than the 350 non-occluded frames with different update weights: 0.2, 0.4, 0.6, 0.8, 0.9, and 1. Horizontal axis corresponds to occlusion cases where u is for upper, l is for lower and m is for mid occlusions. Vertical axis is the average MPJPE of test frames for each occlusion case.

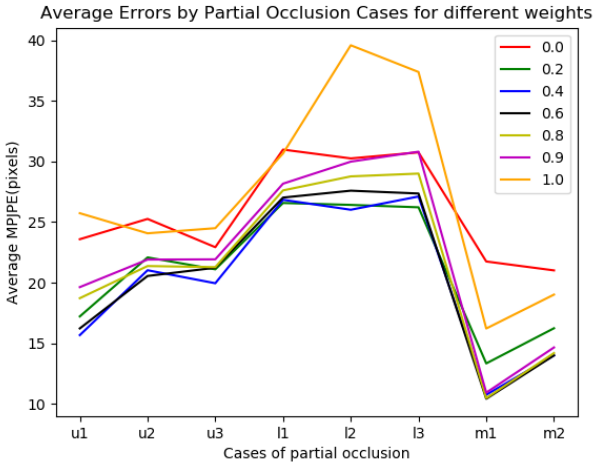


Figure 4.14: Average MPJPE results of BAKE for the whole 28 frames of test sequence with different update weights: 0.2, 0.4, 0.6, 0.8, 0.9, and 1. Horizontal axis corresponds to occlusion cases where u is for upper, l is for lower and m is for mid occlusions. Vertical axis is the average MPJPE of test frames for each occlusion case.

The two proposed methods, BAKE and hybrid B-OP, are compared with the Openpose [1] by using three different video sequences as discussed in Figure 4.7. Figure

4.15 shows the average MPJPE scores for each sequence and method. As it is seen from this figure, the proposed method, BAKE, performs significantly better than the Openpose for all three sequences. The hybrid B-OP method can slightly improve the performance of BAKE in certain cases. Overall, this figure shows that the proposed Bayesian approach is a good technique to include long-term and short-term information in an effective framework compared to the CNN approach in [1].

In Figure 4.16, BAKE and Openpose are compared for a single occlusion. In Figure 4.16a, it is shown that Openpose cannot estimate the occluded keypoints. On the other hand, BAKE estimates the keypoints as shown in Figure 4.16b. In the evaluation of these two algorithms, only the occluded keypoints where Openpose returns predictions are considered. Note that BAKE method always gives the predictions for the occluded keypoints thanks to the Bayesian approach.

In Figure 4.17, the results of the process for obtaining an equivalent confidence score expressed in Equation (4.30) for the BAKE algorithm is shown for joint 0 (nose). The orange dots represent the confidence score - MSE relation for the BAKE algorithm. The orange line is the LS best fit for these points. When these points are mapped into an equivalent score as in Equation (4.30), green dots are obtained. The best LS fit for green points is the green line which matches perfectly with the best fit for the Openpose which is shown with a blue line. This shows that the hybrid technique can be constructed appropriately by using the confidence score of Openpose and the modified BAKE confidence score.

In Figure 4.18, keypoint confidence score, c_{BAKE}, in (4.20) is compared with the Openpose in terms of keypoint estimation error, E_m, in Equation (4.32) for all the occluded keypoints in all the frames of image sequence 1. As it is seen in this figure, confidence scores for BAKE and Openpose closely match and E_m decreases as the confidence increases. In Figure 4.19, total body confidence score, c_{tot}, and the corresponding occluded body estimation error, MPJPE, is shown for a video sequence. As the body confidence score in (4.22) increases the estimation error MPJPE decreases. Hence, this shows that the total body confidence score in (4.22) accurately represents the predictability of the occluded keypoints in the body.

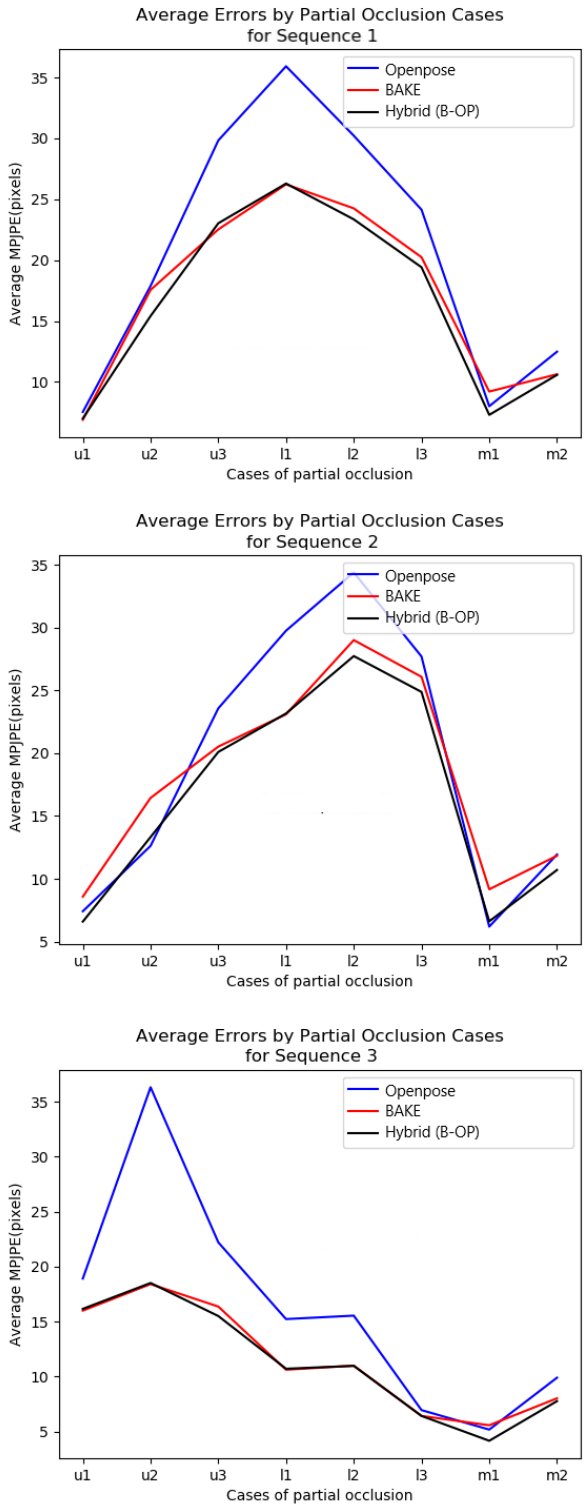


Figure 4.15: Results of the obtained tests are given for sequences 1, 2 and 3 from top to bottom respectively. Horizontal axis corresponds to occlusion cases where u is for upper, l is for lower and m is for mid occlusions. Vertical axis is the average MPJPE of all test frames for each occlusion case.

(a) Prediction of Openpose

(b) Prediction of BAKE

Figure 4.16: Example case where Openpose could not predict the locations of all occluded keypoints while the proposed approach does.

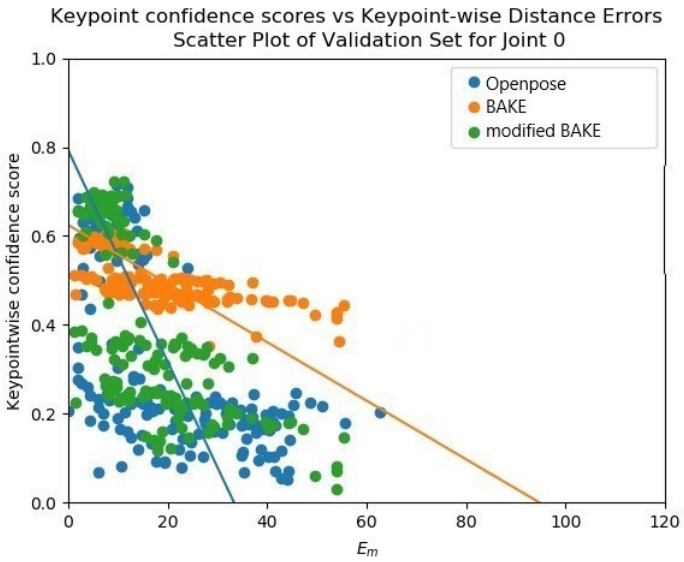


Figure 4.17: Example distribution of validation set for the joint 0 (nose). Orange line represents the relation of confidence scores vs. errors of proposed method via $\eta_{0_0}^{BAKE}$ and $\eta_{0_1}^{BAKE}$ coefficients while blue line represents relation of confidence scores vs. errors of Openpose via $\eta_{0_0}^{OP}$ and $\eta_{0_1}^{OP}$ coefficients.

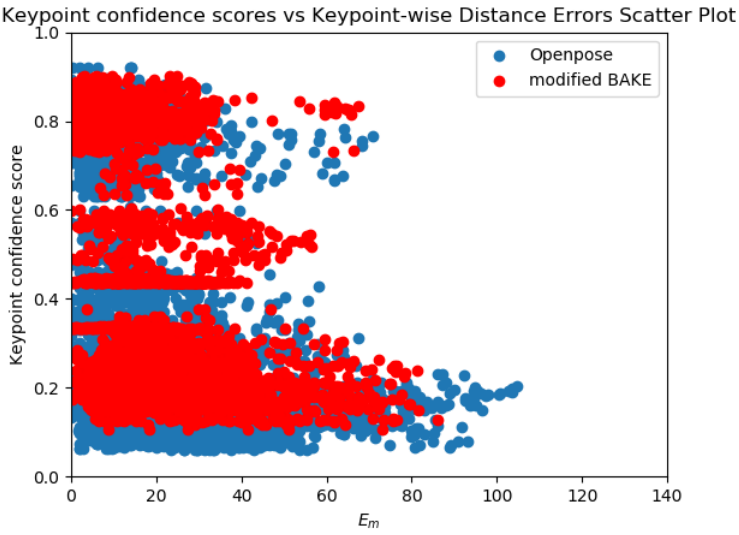


Figure 4.18: Scatter plots of keypoint-wise confidence scores vs. keypoint-wise distance errors, E_m, of BAKE and Openpose for all the occluded keypoints in sequence 1.

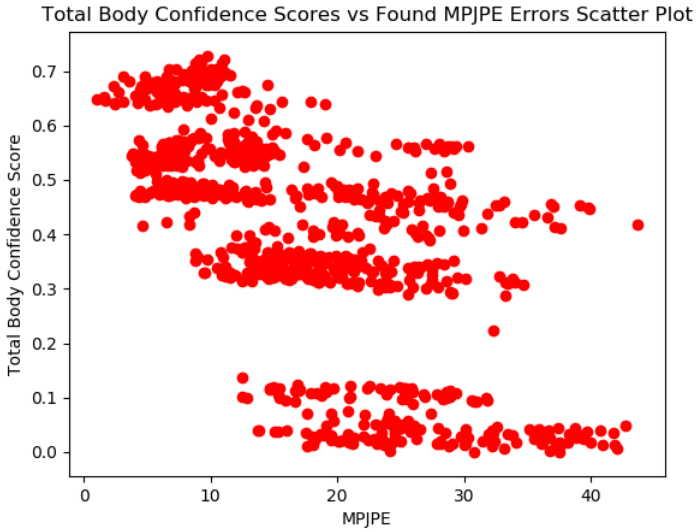


Figure 4.19: Scatter plot of total body predictability scores vs. MPJPE errors of BAKE for sequence 1.

CHAPTER 5

HUMAN IMAGE INPAINTING APPLICATION ON VIDEO FOR PARTIALLY OCCLUDED FRAMES

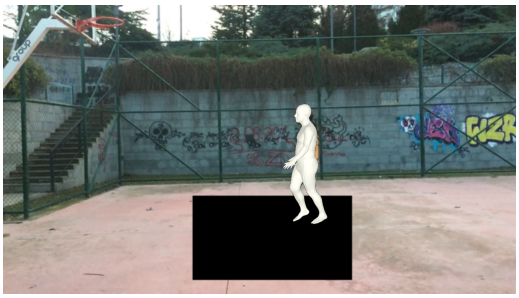
Image inpainting is the process of filling the missing or corrupted regions of an image [37]. In this section, a new method for inpainting the partially occluded human images in a video sequence is proposed. In this problem, focus is only inpainting the occluded human parts. Occluded background regions can be recovered with the mentioned methods in [11, 12]. The proposed method employs a human pose estimation method to complete the missing joints. Then, a 3D body reconstruction method is used to replace the occluded part with image patches extracted from the previous frames. In the proposed method, human pose estimation is performed by using the proposed BAKE method. The results of this are also compared with the case where Openpose is used. Several experiments are done in order to show the effectiveness of the proposed approach. The details of the proposed approach are presented in the following part.

5.1 Inpainting Using the Non-Occluded Candidate Images

In order to inpaint the occluded parts of human images, occlusion detection and segmentation processes are required. The occlusion detection strategy in Section 3.1 is utilized for this purpose in the further parts of this section. After detecting the occluded frames and occluded regions of them, human image inpainting can be applied by using the non-occluded frames. In order to inpaint a partially occluded body image, a set of candidate non-occluded images is given. Then, 3D human body reconstruction from images is proposed to be used for the inpainting process. The target person is reconstructed in both the occluded frame and candidate non-occluded

frames. In non-occluded frames, Openpose [1] and SMPLify-x [2] algorithms are used together. Openpose estimates the 2D keypoint locations of the target person. SMPLify-x uses the keypoint estimates and defines a camera in the 3D scene with an approximate focal length value. It estimates the unknown camera translation and global orientation of the body first. Then, since it estimated the approximate position of the body with respect to the camera, it tries to fit the projected 2D positions of 3D SMPL body joints [15] to Openpose keypoint estimates by optimizing the body shape and pose parameters over a re-projection loss [2]. It outputs the 3D vertex coordinates of the SMPL body in camera coordinates. In target partially occluded frame, using the Openpose and the proposed Bayesian approach, 2D keypoint locations are estimated and similarly, SMPLify-x estimates the 3D SMPL body. In this work, an open-source python library, Pyrender, is used for rendering the 3D SMPL body onto the target images. Frames containing the estimated and projected SMPL bodies for an occluded and non-occluded frame are given in Figure 5.1.

(a) Estimated 3D SMPL body on a non-occluded frame



(b) Estimated 3D SMPL body on an occluded frame

Figure 5.1: Estimated and rendered 3D SMPL bodies in example frames

As previously mentioned, SMPL bodies include 6890 vertices with 13776 triangles. Let G and F are the sets of vertices and triangles for a given SMPL body, respectively. The goal of the application is to transfer the triangular image patches in non-occluded frames to the occluded region of the partially occluded frame. For a given frame with a projected SMPL body, some triangles can be observed by the camera while some of them cannot. The unobservable triangles are the ones that stay at the backside of the 3D body with respect to the camera. It is desired to identify the observable triangles in each non-occluded frame and also in the occluded frame using the SMPL bodies. For

a given camera and given 3D object in the scene, there are different methods to find visible 3D surface points in computer graphics [38] which is a must in the rendering process. During the rendering, Pyrender also finds the visible 3D surface points, however, it does not identify directly the visible triangles of the 3D body. Then, we developed a strategy to determine the visible triangles of the 3D SMPL body for a given frame by using the depth image rendered by Pyrender, $\mathbf{I}_D$, where an example is given in Figure 5.2.

Figure 5.2: An example depth image of an SMPL body. Only the depth information of the SMPL body exists. The closest point to the camera is shown with black while the farthest point is shown with white.

Note that, only the depth information of 3D SMPL body exists in this image. Using the depth image and vertices of SMPL body, we find the visible vertices and then, associate them with triangles to identify visible triangles. In order to find visible vertices, 3D coordinates of vertices are converted to pixel coordinates by applying perspective projection from 3D camera coordinates to 2D image plane as,

$$\mathbf{p}_{proj}^{i} = \mathbf{M}_{proj}\mathbf{p}_{cam}^{i} \tag{5.1}$$

where $\mathbf{M}_{Proj}$ is the 3x4 projection matrix defined by SMPLify-x and $\mathbf{p}_{cam}^{i}$ is the i^{th}

vertex position in camera coordinates with the homogeneous form which is given as,

$$\mathbf{p}_{cam}^{i} = \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} \tag{5.2}$$

Then, the pixel coordinates of i^{th} vertex, u^i and v^i are calculated by using the vector $\mathbf{p}_{proj}^{i}$ as,

$$\mathbf{p}_{proj}^{i} = \begin{bmatrix} x_i' \\ y_i' \\ z_i' \end{bmatrix}, \quad u^i = \frac{x_i'}{z_i'}, \quad v^i = \frac{y_i'}{z_i'} \tag{5.3}$$

After finding the vertex pixel coordinates, the procedure for the identification of visible triangles is applied. This procedure is given in Algorithm 3. In step 1 of Algorithm 3, the set of visible SMPL vertices, G^V, is found by checking the depth values in the depth image and each vertex. Then, in step 2, elements of f_i are investigated whether they are in the set G^V or not. The set F^V is identified for estimated SMPL bodies in this way.

Algorithm 3 Identification of Visible SMPL Triangles

$G^V = \emptyset$, $F^V = \emptyset$

Step 1: For each vertex, check the equality of vertex depth value, z^i, to the value in the depth image's corresponding pixel, $\mathbf{I}_D(u^i, v^i)$

for $i = \{1, 2, .., 6890\}$ **do**

  if $z^i == \mathbf{I}_D(u^i, v^i)$ **then**

    $G^V \leftarrow G^V \cup \{g_i\}$

  end if

end for

Step 2: Note that elements of F are subsets of G with three elements. For each element f_i of F, check whether the elements of f_i, g_{i_1}, g_{i_2}, g_{i_3}, are also elements of G^V.

for $i = \{1, 2, .., 13776\}$ **do**

  if $g_{i_1} \in G^V$ AND $g_{i_2} \in G^V$ AND $g_{i_3} \in G^V$ **then**

    $F^V \leftarrow F^V \cup \{f_i\}$

  end if

end for

In the estimated SMPL body of the partially occluded frame, some of the triangles are not visible due to the occlusion. It is necessary to identify those, so they could be inpainted with the corresponding visible triangles in candidate frames. Foreground mask of partially occluded frame, $\mathbf{M}_3$ given in Section 3.1, is used for the identification of occluded triangles. Algorithm 4 explains the process of finding the occluded triangles. In step 1, found visible vertices of occluded frame are investigated with foreground mask, $\mathbf{M}_3$. If the corresponding vertex pixel is not in the foreground region, it is concluded that it is not seen in the scene, so it is occluded. In Figure 5.3, occluded vertices of the frame given in Figure 5.1b are shown as the output of step 1. In step 2, set of occluded vertices, G^O, is investigated to find set of occluded triangles, F^O. If any vertex of the triangle f_i is occluded, it is also concluded as an occluded triangle.

Then, occluded triangles in the partially occluded frame are known, so the inpainting can be done by triangular image patch transferring from the candidate non-occluded

Figure 5.3: Occluded (red) and visible (white) vertices of the frame given in Figure 4.4 (b) as the output of the step 1 of Algorithm 4

Algorithm 4 Identification of Occluded SMPL triangles on Partially Occluded Frame

$G^O = \emptyset$, $F^O = \emptyset$

Step 1: For each element of visible vertex set in occluded frame, G_O^V, check whether the corresponding pixel stays in the foreground or background area.

for each $g_i \in G_O^V$ **do**

  if $\mathrm{M}_3(u^i, v^i) == 0$ **then**

    $G^O \leftarrow G^O \cup \{g_i\}$

  end if

end for

Step 2: For each element, f_i of visible triangle set in occluded frame, F_O^V, check whether the elements of f_i; g_{i_1}, g_{i_2}, g_{i_3}, are also elements of G^O.

for each $f_i = \{g_{i_1}, g_{i_2}, g_{i_3}\} \in F_O^V$ **do**

  if $g_{i_1} \in G^O$ OR $g_{i_2} \in G^O$ OR $g_{i_3} \in G^O$ **then**

    $F^O \leftarrow F^O \cup \{f_i\}$

  end if

end for

frames. However, it is required to decide the correspondences of occluded triangles and candidate frames for transferring patches. Algorithm 5 explains the process of relating occluded triangles with candidate frames. According to Algorithm 5, numbers of common triangles between the sets of visible triangles in candidate frames and oc-

cluded triangles in the partially occluded frame are compared. The candidate with the largest number of commonly visible triangles is chosen as the first frame to be used in inpainting. Common triangles are decided to be transferred from this candidate frame to the occluded region. Then, for the non-associated occluded triangles, other candidates are investigated similarly. Hence, the process continues recursively until all the candidates are used or all the occluded triangles are associated with candidates for image patch transfer.

Algorithm 5 Associating Occluded SMPL Triangles with Candidate Non-occluded Frames for Image Patch Transferring

Set of used candidates, $C_u = \emptyset$, set of not used candidates, $C_n = \{c_1, c_2, ..., c_{N_c}\}$ where there are N_c candidate frames.

Set of occluded triangles which are not associated yet with candidate frames, $F_n^O = F^O$

Set of ordered candidates for image patch transferring, $O = \emptyset$

while ($|C_n| \neq 0$) and ($|F_n^O| \neq 0$) **do**

  for each $c_k \in C_n$ **do**

    Let F_{c_k} be the set of common triangles between set of occluded triangles not associated yet, F_n^O, and set of visible triangles in k^{th} candidate frame, F_k^V.

    $F_{c_k} = \emptyset$

    for each $f_i \in F_n^O$ **do**

      if $f_i \in F_k^V$ **then**

        $F_{c_k} \leftarrow F_{c_k} \cup \{f_i\}$

      end if

    end for

  end for

  $k* = \arg\max_k (|F_{c_k}|)$

  $F_n^O \leftarrow F_n^O \setminus F_{c_k}$

  $C_n \leftarrow C_n \setminus \{c_k\}$

  $O \leftarrow O \cup \{k\}$

end while

Then, each of the associated occluded triangles is transferred by the candidate frames

according to the ordered candidate set, O, and common triangle sets, F_{c_k}. Image patch transferring is performed through affine transformations of triangular image patches. Affine transformation from a source pixel coordinates, (u_t, v_t), to target pixel coordinates, (u_s, v_s), is given as,

$$\begin{bmatrix} u_t \\ v_t \\ 1 \end{bmatrix} = \mathbf{T} \begin{bmatrix} u_s \\ v_s \\ 1 \end{bmatrix} \tag{5.4}$$

where $\mathbf{T}$ is the 2x3 transformation matrix given as,

$$\mathbf{T} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \tag{5.5}$$

where a_{13} and a_{23} are the coefficients responsible for translation. a_{11}, a_{12}, a_{21} and a_{22} are the coefficients responsible for scaling and rotation operations. For the occluded triangle, f_i with vertices g_{i_1}, g_{i_2} and g_{i_3}, let the pixel coordinates of vertices in occluded frame be $(u_o^{i_1}, v_o^{i_1})$, $(u_o^{i_2}, v_o^{i_2})$ and $(u_o^{i_3}, v_o^{i_3})$, while the pixel coordinates of vertices in associated candidate frame be $(u_c^{i_1}, v_c^{i_1})$, $(u_c^{i_2}, v_c^{i_2})$ and $(u_c^{i_3}, v_c^{i_3})$. Then the coefficients of affine transformation matrix for the occluded triangle f_i is found as,

$$\begin{bmatrix} a_{11}^i \\ a_{12}^i \\ a_{13}^i \\ a_{21}^i \\ a_{22}^i \\ a_{23}^i \end{bmatrix} = \mathbf{D}_i^{-1} \begin{bmatrix} u_c^{i_1} \\ u_c^{i_2} \\ u_c^{i_3} \\ v_c^{i_1} \\ v_c^{i_2} \\ v_c^{i_3} \end{bmatrix} \tag{5.6}$$

where $\mathbf{D}_i$ is the matrix which consists of the candidate frame vertex coordinates and

given as,

$$\mathbf{D}_i = \begin{bmatrix} u_o^{i_1} & u_o^{i_1} & 1 & 0 & 0 & 0 \\ u_o^{i_2} & u_o^{i_2} & 1 & 0 & 0 & 0 \\ u_o^{i_3} & u_o^{i_3} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & u_o^{i_1} & u_o^{i_1} & 1 \\ 0 & 0 & 0 & u_o^{i_2} & u_o^{i_2} & 1 \\ 0 & 0 & 0 & u_o^{i_3} & u_o^{i_3} & 1 \end{bmatrix} \tag{5.7}$$

For the found transformation matrix of the occluded triangle f_i, the pixels in the triangular image patch of the occluded frame are transformed to the coordinates of the associated candidate frame. The aim of transforming pixel coordinates from occluded to candidate frame is to obtain the value of occluded pixels from the candidate. Then, the pixel is inpainted in the occluded frame. However, the transformed coordinate is real-valued instead of integer pixel coordinates. Therefore, bilinear interpolation is performed using the transformed pixel coordinates in the candidate frame. The bilinear interpolation operation for the transformed coordinates, (u'_o, v'_o), around the four corner pixels; (u_b, v_b), $((u_b + 1), v_b)$, $(u_b, (v_b + 1))$ and $((u_b + 1), (v_b + 1))$, is given as,

$$\begin{aligned} \mathbf{I}_o(u_o, v_o) = (u'_o - u_b)(v'_o - v_b)\mathbf{I}_{c_k}((u_b + 1), (v_b + 1)) \\ + (u_b + 1 - u'_o)(v'_o - v_b)\mathbf{I}_{c_k}(u_b, (v_b + 1)) \\ + (u'_o - u_b)(v_b + 1 - v'_o)\mathbf{I}_{c_k}((u_b + 1), v_b) \\ + (u_b + 1 - u'_o)(v_b + 1 - v'_o)\mathbf{I}_{c_k}(u_b, v_b) \end{aligned} \tag{5.8}$$

where $\mathbf{I}_{\mathbf{c_k}}$ is the candidate frame associated with the occluded triangle, f_i, and $\mathbf{I}_o$ is the occluded frame. In Figure 5.4, corresponding positions of given points in (5.8) are shown. If there are occluded triangles left unmatched with candidate frames, interpolation is applied again to inpaint the pixels in these triangles. Delaunay triangulation [39] and linear barycentric [40] interpolation is used to this end with the open-source

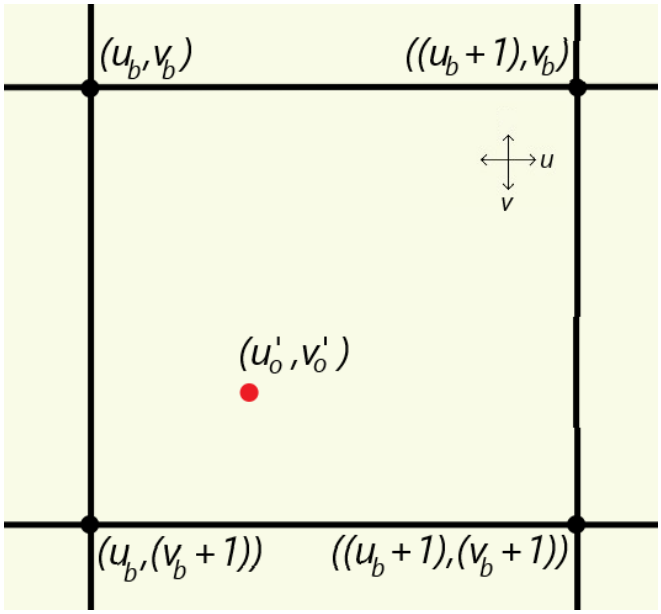


Figure 5.4: Illustration of coordinates for bilinear interpolation

Figure 5.5: Result of inpainting procedure for the partially occluded frame.

software library Scipy [41]. In Figure 5.5, occluded image given in Figure 5.1b is inpainted using the outlined method.

In the following part, experimental results related to inpainting method will be presented.

5.2 Experiments Related with Human Body Inpainting

In order to show the effectiveness of the proposed inpainting method, several experiments are done. In these experiments, occlusion detection on the test videos is done using the proposed method given in Section 3.1. The foreground mask, M_3, in Section 3.1 is also used to detect occluded vertices and triangles.

5.2.1 Comparison for the Effects of BAKE and Openpose in the Inpainting Application

In this section, inpainting results of occluded video frames from different video sequences will be shown. The performance of the application is directly related to the base method SMPLify-x [2]. SMPLify-x obtains the anatomical information via the found 2D keypoints of the target person. However, obtained 2D pose elements do not contain enough information about the body shape in general. Moreover, SMPL bodies do not have a clothing layer. Thus, the clothing information of the target person is lost. Therefore, the performance of the application is limited inherently due to the limitations of SMPLify-x. Nevertheless, the performance of the application can be improved if the used keypoints for the occluded frame get more accurate. In this test, we tested BAKE and Openpose methods in terms of correctly predicting the occluded keypoints. 3D SMPL bodies reconstructed for the occluded frames with both Openpose and BAKE. Ground truth 3D body is also reconstructed for these frames. Then, in order to measure the similarity of the reconstructed and projected SMPL bodies with the ground truth SMPL body, the average of Euclidean distances between the corresponding vertices of reconstructed body and ground truth body in the occluded part, E_v, is used which is given as,

$$E_v = \frac{1}{N_o} \sum_{g \in G^O} \|\mathbf{v_g^{pred}} - \mathbf{v_g^{gt}}\|_2 \tag{5.9}$$

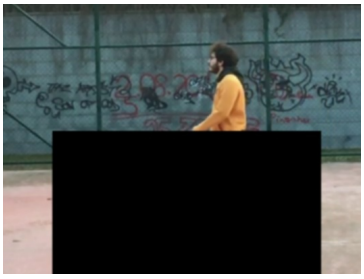
where G^O is the set of occluded vertices of the projected SMPL body in the occluded frame with N_o elements. $\mathbf{v_g^{pred}}$ is the pixel coordinate vector of vertex g in the occluded frame while $\mathbf{v_g^{gt}}$ is the pixel coordinate vector of SMPL body projected to the

original frame where synthetic occlusion is not inserted. Two video sequences are utilized in this experiment. Example frames for the occluded frames are given with their corresponding ground truth images in Figure 5.6 for the video sequence 1.

Projected 3D bodies which are constructed from original image and occluded image for the frame given in Figure 5.6a and 5.6b are shown in Figure 5.7. Vertices in the set G^O for the test frame in Figure 5.7 are shown in Figure 5.8. In Figure 5.8a, ground truth vertices are shown with the vertices reconstructed by the occluded keypoint estimates of BAKE while in Figure 5.8b, ground truth vertices are shown with the vertices reconstructed by the occluded keypoint estimates of Openpose. One-to-one correspondence of the given vertices is known, so the average of Euclidean distances between the vertices in the occluded region, E_v, can be calculated for both reconstructions with BAKE and Openpose.

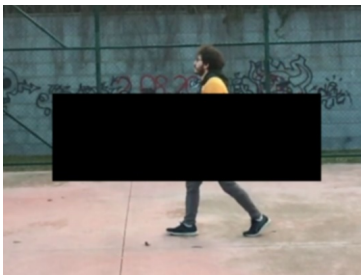
For both sequences, 4 test frames are used for each occlusion pattern: lower occlusion, middle occlusion, and upper occlusion. Comparison results are given in Figure 5.9. The vertical axis is the average Euclidean distance error while the horizontal axis corresponds to each test frame where l, m, and u are used for representing lower, middle, and upper, respectively. Furthermore, all inpainted test frames of the two test sequences are shown in Appendix D.

In lower and upper occlusion conditions, reconstructed bodies with occluded keypoint estimates of BAKE are closer to the ground truth bodies than Openpose. In such occlusion cases, BAKE gives more accurate estimates since it has the capability of embedding the motion information. In occlusions that occurred on the middle part of the body, Openpose might also be able to reconstruct accurate results as much as BAKE, since the upper and lower parts of the human body are visible and Openpose can localize the middle part using the information of upper and lower parts.



(a) A test frame where lower part is occluded

(b) Corresponding original frame



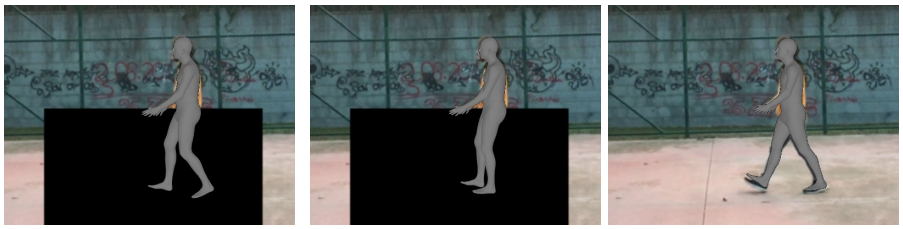
(c) A test frame where mid part is occluded

(d) Corresponding original frame

(e) A test frame where upper part is occluded

(f) Corresponding original frame

Figure 5.6: Occluded test frames and corresponding original frames for sequence 1

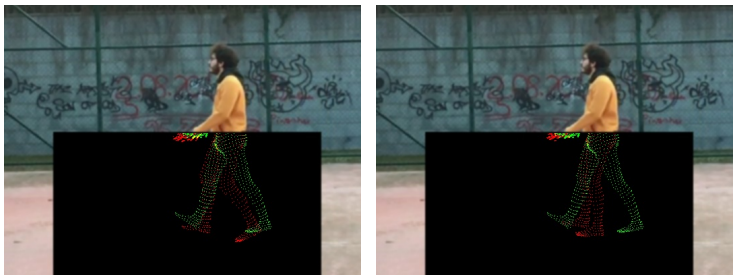


(a) Projected 3D SMPL body where the occluded keypoint estimations of BAKE are used in reconstruction

(b) Projected 3D SMPL body where the occluded keypoint estimations of Openpose are used in reconstruction

(c) Projected 3D SMPL body on the original frame

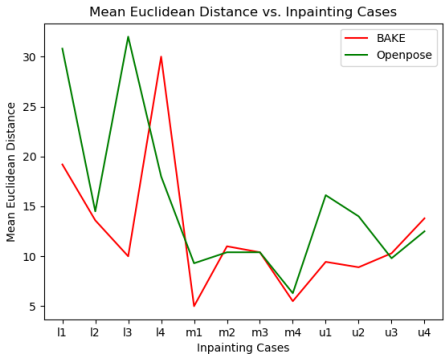
Figure 5.7: Projected SMPL bodies on occluded frame using BAKE and Openpose and the ground truth SMPL body on original frame



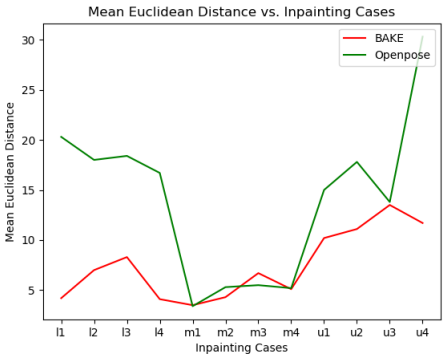
(a) Vertices of ground truth body and the reconstructed body in occluded frame by using BAKE

(b) Vertices of ground truth body and the reconstructed body in occluded frame by using Openpose

Figure 5.8: Vertices of ground truth and predicted bodies in the occluded region. Red corresponds to the vertices of the predicted body while green corresponds to the vertices of the ground truth body. Since the correspondence of each vertex between two images is known, E_v can be used to measure the similarity of the reconstructed SMPL body in the occluded frame with the ground truth body in the original frame.



(a) E_v vs Inpainting Cases for video sequence 1



(b) E_v vs Inpainting Cases for video sequence 2

Figure 5.9: Obtained E_v errors for given video sequences

5.2.2 Effect of Chosen Candidate Images

In this experiment, the effect of chosen candidate images for the inpainting process is examined. Two candidate image sets are used from video sequence 2 for this purpose. Candidate set 1 and 2 are shown in Figure 5.10 and Figure 5.11.

The occluded test frames for sequence 2, which are used in the comparison of BAKE-Openpose in Section 5.2.1, are also used in this experiment. 3D reconstructions of bodies are done using the occluded keypoint estimates of BAKE for these test frames. Then, the inpainting is performed using the candidate sets 1 and 2 separately. In Figure 5.12, an inpainted test frame with the given candidate sets and the ground truth of the frame are shown.

In order to measure the similarity of the inpainted images with the ground truth, PSNR [42] and SSIM [42, 43] metrics are used. PSNR is the peak signal-to-noise ratio and for 8-bit images, it is given as,

$$PSNR = 20log_{10}(\frac{255}{MSE}) \tag{5.10}$$

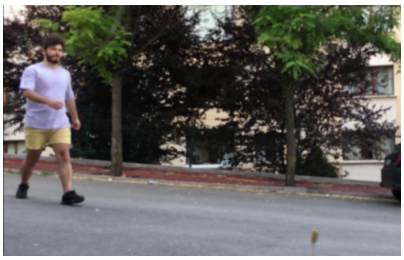
(a) Candidate 1

(b) Candidate 2

(c) Candidate 3

(d) Candidate 4

Figure 5.10: Candidate Images Set 1 for Inpainting



(a) Candidate 1

(b) Candidate 2

(c) Candidate 3

(d) Candidate 4

Figure 5.11: Candidate Images Set 2 for Inpainting

(a) Inpainted image with candidate set 1

(b) Inpainted image with candidate set 2

(c) Ground truth

Figure 5.12: Inpainting results with candidate sets 1 and 2 for a test frame and its ground truth

where MSE is given as,

$$MSE = \frac{\sum_{n}^{N} \sum_{m}^{M} (\mathbf{I_1}(m,n) - \mathbf{I_2}(m,n))^2}{M * N} \tag{5.11}$$

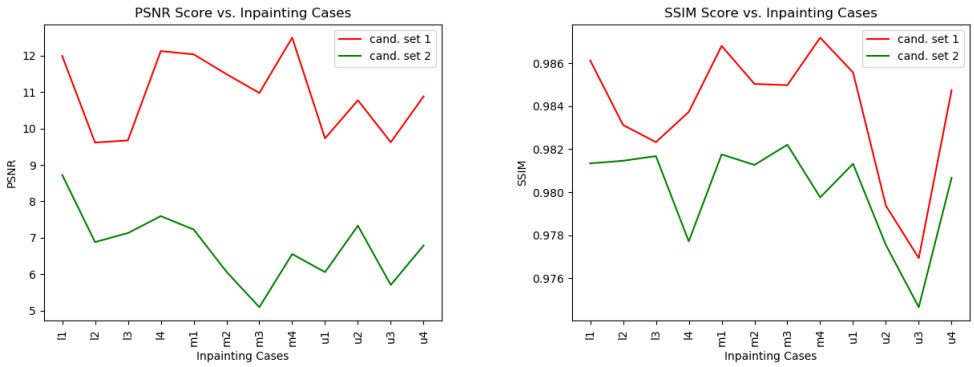
for the images $\mathbf{I_1}$ and $\mathbf{I_2}$.

SSIM [42, 43] is a metric more adapted to the human visual perception system since it compares images on three different aspects: luminance, contrast, and structure. SSIM metric is given as,

$$SSIM = \frac{(2\mu_{I_1}\mu_{I_2} + C_1)(2\sigma_{I_1,I_2} + C_2)}{(\mu_{I_1}^2 + \mu_{I_2}^2 + C_1)(\sigma_{I_1}^2 + \sigma_{I_2}^2 + C_2)} \tag{5.12}$$

where the parameters μ_{I_1} and μ_{I_2} are the mean parameters while σ_{I_1} and σ_{I_2} are the variance parameters of the images $\mathbf{I_1}$ and $\mathbf{I_2}$, respectively. σ_{I_1,I_2} is the covariance parameter between between two images. C_1 and C_2 are small constants to avoid instability when $(\mu_{I_1}^2 + \mu_{I_2}^2)$ or $(\sigma_{I_1}^2 + \sigma_{I_2}^2)$ are very close to zero, respectively [43].

In the occluded regions, similarity of the intersecting regions of ground truth body silhouettes and predicted body silhouettes are investigated for the comparison. Results of the experiment are given in Figure 5.13. Furthermore, all of the inpainted frames related to the candidate test are shown in Appendix E.



(a) PSNR scores for each test frame

(b) SSIM scores for each test frame

Figure 5.13: PSNR and SSIM results of the inpainting application using candidate sets 1 and 2

Both sets give successful inpainting results, especially when the SSIM scores are considered. The smallest SSIM score is around 0.975 which means the occluded region in inpainted images is very similar to the corresponding region of the original image indeed. Nevertheless, using candidate set 1 resulted in better SSIM and PSNR scores than using candidate set 2 in all test frames. The main reason for that result is the body pose angle of the test frames is closer to the candidate set 1 than candidate set 2. There are more common triangles between an occluded test frame and a candidate frame of set 1 than a candidate frame of set 2. This provides the ability to inpaint more area of the occluded region using only one candidate frame, so a smooth inpainted region is obtained. Additionally, there are many triangles on the occluded region of test images that are not visible in any image of the candidates in set 2. Those unmatched regions are interpolated during the inpainting, however, as the interpolated regions are getting larger, the inpainting performance is degraded. Lastly, the illumination of the frames in candidate sets 1 and 2 are different due to the body pose angle differences and the illumination of the candidate set 1 is closer to test frames than candidate set 2.

CHAPTER 6

CONCLUSIONS

A Bayesian approach, BAKE, is proposed for 2D human pose estimation under partial occlusions as a part of this thesis work. BAKE uses the statistical information extracted from the local frames as well as a database which represents the global information for the human body. This Bayesian framework is integrated with the CNN network structure to embed the characteristics of time-dependent human pose features to estimate occluded keypoints. In this sense, BAKE can be seen as an alternative to 3D CNN structures with the advantage of convenient training and computational complexity. It is shown that local statistics is important to model the stationary body motion whereas global statistics is useful to model the changes in the motion. The robustness of the proposed action-specific distribution update strategy is shown through several experiments in terms of the MPJPE performance metric. It is shown that BAKE performs significantly better than Openpose which sometimes fails to give estimates for certain occluded keypoints. A novel keypoint confidence score, as well as a total body confidence score, are presented. It is shown that the proposed confidence scores accurately represent the predictability of the occluded keypoints. A hybrid method is also presented to further improve the performance of the BAKE algorithm. The proposed hybrid technique slightly improves the BAKE performance. Several experiments and results are presented to show the performance of the proposed approaches in natural video sequences.

In the second part of this work, an inpainting method for the partially occluded human images is proposed. This method employs 2D pose estimation under partial occlusions and 3D body reconstruction. Using the non-occluded frames and corresponding 3D SMPL bodies, an image patch transfer procedure is explained for inpainting the

occluded regions of a target human image. Through several experiments, realistic results are obtained in different video sequences, and the improvement on the problem of occluded pose estimation with BAKE is also adapted for this application. Openpose and BAKE are compared for the 3D body reconstruction step and it is shown that 3D bodies reconstructed with BAKE keypoints outperform Openpose in general. Even though this method can be directly used in applications of video special effects and it is inspiring for the computerized outfit planners, there are problems related to the 3D reconstruction process which creates 3D bodies only using the 2D human keypoints. Even in the cases where 2D pose elements are accurate, the reconstructed occluded bodies do not exactly match with the target person due to the differences in terms of body shape between the real and the reconstructed body. Furthermore, clothing of the target person is missing in the reconstructed body.

As future work, the occlusion detection algorithm can be improved for natural occlusion patterns. In this work, occlusions were modeled by black boxes and the proposed occlusion detection method performed well for these cases. It is also expected to work on natural scenarios. Therefore it can be employed for such scenarios to measure its performance. For the occluded keypoint estimation, the proposed method can be tested to estimate occluded keypoints in cases where more complex motions are realized. In order to improve the human image inpainting results, a 3D body reconstruction algorithm which emphasizes the anatomical properties and the clothing of the target person can be utilized.

REFERENCES

[1] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021.

[2] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] H. Chandel and S. Vatta, "Occlusion detection and handling: A review," *International Journal of Computer Applications*, vol. 120, pp. 33–38, 06 2015.

[4] A. Casalino, S. Guzman, A. Maria Zanchettin, and P. Rocco, "Human pose estimation in presence of occlusion using depth camera sensors, in human-robot coexistence scenarios," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–7, 2018.

[5] T. Yuan, G. Tao, and W. Cheng, "Real-time occlusion handling in augmented reality based on an object tracking approach," *Sensors*, vol. 10, 04 2010.

[6] L. Beng Yong, L. Liew, W. Cheah, and Y. Wang, "Occlusion handling in videos object tracking: A survey," *IOP Conference Series: Earth and Environmental Science*, vol. 18, 01 2014.

[7] D. L. D. C, "Video surveillance systems–a survey," *International Journal of Computer Science Issues*, vol. 8, 07 2011.

[8] S. Tran, M. Du, S. Chanda, R. Manmatha, and C. Taylor, "Searching for apparel products from images in the wild," *CoRR*, vol. abs/1907.02244, 2019.

[9] P. S. R. Kishore, S. Das, P. S. Mukherjee, and U. Bhattacharya, "Cluenet : A deep framework for occluded pedestrian pose estimation," in *BMVC*, 2019.

[10] J.-B. Huang and M.-H. Yang, "Estimating human pose from occluded images," *Computer Vision – ACCV 2009 Lecture Notes in Computer Science*, p. 48–60, 2010.

[11] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting of occluding and occluded objects," in *IEEE International Conference on Image Processing 2005*, vol. 2, pp. II–69, 2005.

[12] S. A. Chavan and N. M. Choudhari, "Various approaches for video inpainting: A survey," in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pp. 1–5, 2019.

[13] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," *Proceedings of IEEE Conf. Computer Vision Patt. Recog, vol. 2*, vol. 2, 01 2007.

[14] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, pp. 248:1–248:16, Oct. 2015.

[16] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 03 2020.

[17] W. Gong, X. Zhang, J. González, A. Sobral, T. Bouwmans, C. Tu, and E. hadi Zahzah, "Human pose estimation from monocular images: A comprehensive survey," *Sensors (Basel, Switzerland)*, vol. 16, 2016.

[18] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, p. 34–50, 2016.

[19] R. Jena, "Out of the box: A combined approach for handling occlusion in human pose estimation," *CoRR*, vol. abs/1904.11157, 2019.

[20] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net++: Multi-person 2d and 3d pose detection in natural images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 01 2019.

[21] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016.

[22] A. Kadkhodamohammadi and N. Padoy, "A generalizable approach for multi-view 3d human pose regression," *CoRR*, vol. abs/1804.10462, 2018.

[23] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.

[24] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3d human pose estimation to occlusion?," *CoRR*, vol. abs/1808.09316, 2018.

[25] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[26] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3d body pose estimation from monocular RGB input," *CoRR*, vol. abs/1712.03453, 2017.

[27] B. Tekin, X. Sun, X. Wang, V. Lepetit, and P. Fua, "Predicting people's 3d poses from short sequences," *CoRR*, 04 2015.

[28] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1446–1455, 2015.

[29] D. Bautembach, I. Oikonomidis, and A. Argyros, "Filling the joints: Completion and recovery of incomplete 3d human poses," *Technologies*, vol. 6, no. 4, p. 97, 2018.

[30] U. Rafi, J. Gall, and B. Leibe, "A semantic occlusion model for human pose estimation from a single depth image," pp. 67–74, 06 2015.

[31] 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," pp. 491–499, 10 2016.

[32] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] X. Guo and Y. Dai, "Occluded joints recovery in 3d human pose estimation based on distance matrix," *CoRR*, vol. abs/1807.11147, 2018.

[34] A. Yiannakides, A. Aristidou, and Y. Chrysanthou, "Real-time 3d human pose and motion reconstruction from monocular rgb videos," *Computer Animation and Virtual Worlds*, vol. 30, 05 2019.

[35] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," *ACM Trans. Graph*, vol. 24, pp. 408–416, 2005.

[36] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[37] B. Furht, ed., *Image Inpainting*, pp. 315–316. Boston, MA: Springer US, 2006.

[38] J. Bittner and P. Wonka, "Visibility in computer graphics," *Environment and Planning B: Planning and Design*, vol. 30, pp. 729–755, 09 2003.

[39] C. Barber, D. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software*, vol. 22, 02 1998.

[40] K. Hormann, "Barycentric interpolation," in *Approximation Theory XIV: San Antonio 2013* (G. E. Fasshauer and L. L. Schumaker, eds.), (Cham), pp. 197–218, Springer International Publishing, 2014.

[41] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python," 2001–.

[42] A. Horé and D. Ziou, "Image quality metrics: Psnr vs. ssim," pp. 2366–2369, 08 2010.

[43] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

APPENDIX A

ESTIMATION OF APPROXIMATE PROBABILITY MASS FUNCTION OF OCCLUDED KEYPOINT

Let $2\Delta_p$ be the step size of discrete valued coordinate vector $\mathbf{p_k}$ while $\mathbf{p_k^c}$ is the continuous valued coordinate vector. Conditional probability mass function of occluded keypoint for a given total body length, $p_{\mathbf{p_k}|T}(\mathbf{p_k}|T)$ can be defined as,

$$p_{\mathbf{p_k}|T}(\mathbf{p_k}|T) = P(\|\mathbf{p_k} - \mathbf{p_k^c}\|_2 < \Delta_p) \tag{A.1}$$

It is defined as a circular region around $\mathbf{p_k}$ with radius Δ_p for analytical simplicity instead of a square grid. The given probability value equals the result of integral given as,

$$P(\|\mathbf{p_k} - \mathbf{p_k^c}\|_2 < \Delta_p) = \int_{a^L_{k,v_0}}^{a^U_{k,v_0}} \dots \int_{a^L_{k,v_K}}^{a^U_{k,v_K}} \int_{l^L_{k,v_0}}^{l^U_{k,v_0}} \dots \int_{l^L_{k,v_K}}^{l^U_{k,v_K}} f_{\mathbf{x_r^k}}(\mathbf{x_r^k}) \, da_{k,v_0} \dots da_{k,v_K} \, dl_{k,v_0} \dots dl_{k,v_K} \tag{A.2}$$

where v_i is the i^{th} visible joint and a^L_{k,v_i}, a^U_{k,v_i}, l^L_{k,v_i} and l^U_{k,v_i} are the lower and upper limit functions for related angle and length variables respectively. Note that it is assumed there are K visible joints.

l^L_{k,v_i} and l^U_{k,v_i} can be defined as,

$$l^L_{k,v_i} = \frac{\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2 - \Delta_p}{T} \tag{A.3}$$

$$l^U_{k,v_i} = \frac{\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2 + \Delta_p}{T} \tag{A.4}$$

In order to approximate the above integral in a simpler form, a^L_{k,v_i} and a^U_{k,v_i} parameters have to be defined independent of l_{k,v_i}, so that the circular region $\|\mathbf{p_k} - \mathbf{p^c_k}\|_2 < \Delta_p$ is needed to be converted to an annular sector defined by l^L_{k,v_i}, l^U_{k,v_i}, a^L_{k,v_i} and a^U_{k,v_i} parameters as shown in Figure A.1. Therefore, the angle limits a^L_{k,v_i} and a^U_{k,v_i} are defined as,

$$a^L_{k,v_i} = \tilde{a}_{k,v_i} - tan^{-1}(\frac{\Delta_p}{\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2}) \tag{A.5}$$

$$a^U_{k,v_i} = \tilde{a}_{k,v_i} + tan^{-1}(\frac{\Delta_p}{\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2}) \tag{A.6}$$

where $\tilde{a}_{k,v_i}$ is the value of angle variable defined with points $\mathbf{p_k}$ and $\mathbf{p_{v_i}}$ using the Equations (4.4) and (4.3).

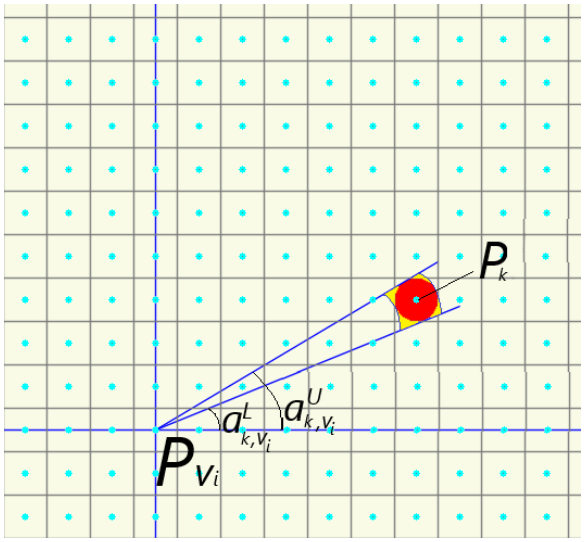


Figure A.1: Illustration of deviation of unit area $\|\mathbf{p_k} - \mathbf{p^c_k}\|_2 < \Delta_p$ for an example point. Red area is the original region. p_{v_i} is the i^{th} visible point and it is set to the center of horizontal and vertical axes for illustration. By defining the angle limits a^L_{k,v_i} and a^U_{k,v_i}, this region expands through the yellow region.

Defining a^L_{k,v_i} and a^U_{k,v_i} in this way causes a deviation in the integral area as shown in

Figure A.1 such that the area changes with the distance $\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2$. The expression for the annular sector area is given as,

$$area = 4\Delta_p \|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2 \, tan^{-1}\left(\frac{\Delta_p}{\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2}\right)$$
$$= 4\nu\Delta_p^2 \, tan^{-1}(1/\nu) \quad \text{(A.7)}$$

where ν is also given as,

$$\nu = \frac{\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2}{\Delta_p}, \;\; \nu \geq 2 \quad \text{(A.8)}$$

However, this deviation is bounded as shown in Figure A.2, the plot of this area with respect to $\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2$ distance. Note that the smallest $\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2$ distance could be $2\Delta_p$ when $\mathbf{p_k}$ is different than p_{v_i}, so the plot on Figure A.2 starts from the distance $2\Delta_p$.

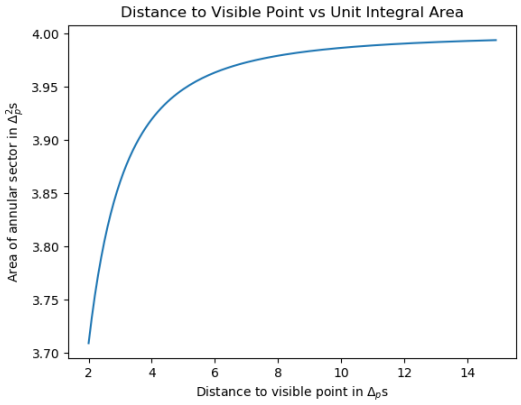


Figure A.2: Plot of distance to visible point vs. angular sector area

Since the newly defined area in Figure A.1 is too small, the integral given above can be approximated in a multiplicative form as,

$$p_{\mathbf{p_k}|T}(\mathbf{p_k}|T)$$

$$\approx \Delta_{a_{k,v_0}} ... \Delta_{a_{k,v_K}} \Delta_{l_{k,v_0}} ... \Delta_{l_{k,v_K}} f_{\mathbf{x_r^k}}(\tilde{\mathbf{x}}_\mathbf{r}^k)$$

$$= \beta_1 f_{\mathbf{x_r^k}}(\tilde{\mathbf{x}}_\mathbf{r}^k) \prod_{v_i \in VIS} \Delta_{a_{k,v_i}} \quad \text{(A.9)}$$

where $\Delta_{a_{k,v_i}}$ and $\Delta_{l_{k,v_i}}$ are also given as,

$$\Delta_{a_{k,v_i}} = a_{k,v_i}^U - a_{k,v_i}^L = 2tan^{-1}(\frac{\Delta_p}{\|\mathbf{p_k} - \mathbf{p_{v_i}}\|_2}) \quad \text{(A.10)}$$

$$\Delta_{l_{k,v_i}} = \frac{l_{k,v_i}^U - l_{k,v_i}^L = 2\Delta_p}{T} \quad \text{(A.11)}$$

As given in Equation (A.11), $\Delta_{l_{k,v_i}}$ is independent of the keypoint position, $\mathbf{p_k}$, so multiplications of $\Delta_{l_{k,v_i}}$ terms in Equation (A.9) is combined in β_1 term. Furthermore, for given total body length T, given visible joints set, VIS with positions of visible joints and a given position of the occluded joint, value of $\mathbf{x_r^k}$ can be calculated with Equations (4.1), (4.4) and (4.3) as the realized value vector $\tilde{\mathbf{x}}_\mathbf{r}^k$.

APPENDIX B

ESTIMATION OF APPROXIMATE PROBABILITY MASS FUNCTION OF TOTAL BODY LENGTH

Let T be the discrete-valued total body length variable with $2\Delta_T$ steps while T^c is the continuous-valued variable and ω equals to $1/T^c$. Probability mass function of discrete valued total body length, $p_T(T)$ can be defined as,

$$p_T(T) = P(T - \Delta_T < T^c < T + \Delta_T) \tag{B.1}$$

Given probability value can be calculated through the integral from ω_L to ω_U as,

$$P(T - \Delta_T < T^c < T + \Delta_T) = \int_{\omega_L}^{\omega_U} f_{\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k}}(\omega\tilde{\mathbf{l}}_\mathbf{k}|\tilde{\mathbf{a}}_\mathbf{k})\, d\omega \tag{B.2}$$

where ω_L and ω_U are also given as,

$$\omega_L = \frac{1}{T + \Delta_T} \tag{B.3}$$

$$\omega_U = \frac{1}{T - \Delta_T} \tag{B.4}$$

The interval $(1/(T+\Delta_T)\,,\, 1/(T-\Delta_T))$ becomes too small as the value of T increases. Therefore, the integral given above can be approximated as,

$$p_T(T) \approx \Delta_\omega f_{\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k}}((\frac{\omega_U + \omega_L}{2}\tilde{\mathbf{l}}_\mathbf{k})|\tilde{\mathbf{a}}_\mathbf{k}) \tag{B.5}$$

where Δ_ω is given as,

$$\Delta_\omega = \omega_U - \omega_L = \frac{2\Delta_T}{T^2 - \Delta_T^2} \tag{B.6}$$

As $T \gg \Delta_T$, $(\omega_U + \omega_L)/2$ and Δ_ω can also be approximated as,

$$T \gg \Delta_T \rightarrow \frac{\omega_U + \omega_L}{2} = \frac{1}{T}, \quad \Delta_\omega = \frac{2\Delta_T}{T^2} \tag{B.7}$$

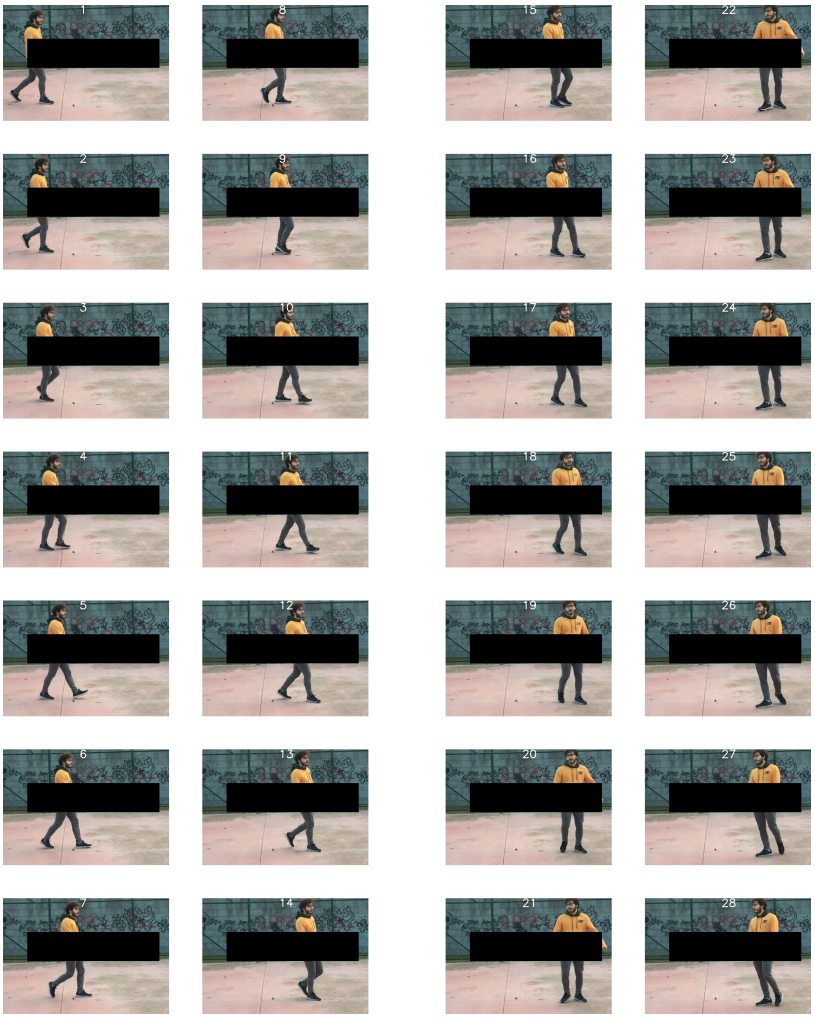
Finally, the expression in Equation (B.5) simplifies to,

$$p_T(T) \approx \frac{\beta_2}{T^2} f_{\mathbf{l_k}|\tilde{\mathbf{a}}_\mathbf{k}}((\tilde{\mathbf{l}}_\mathbf{k}/T)|\tilde{\mathbf{a}}_\mathbf{k}) \tag{B.8}$$

where β_2 is the normalization constant.

APPENDIX C

TEST FRAMES RELATED WITH UPDATE WEIGHT TEST



(a) First 14 test frames (b) Second 14 test frames

Figure C.1: Test frames related to the update weight test. The first 14 frames are the continuation of the walking sequence while a different motion character than the non-occluded frames exists in the second 14 frames

APPENDIX D

INPAINTING RESULTS FOR DIFFERENT CANDIDATE SETS

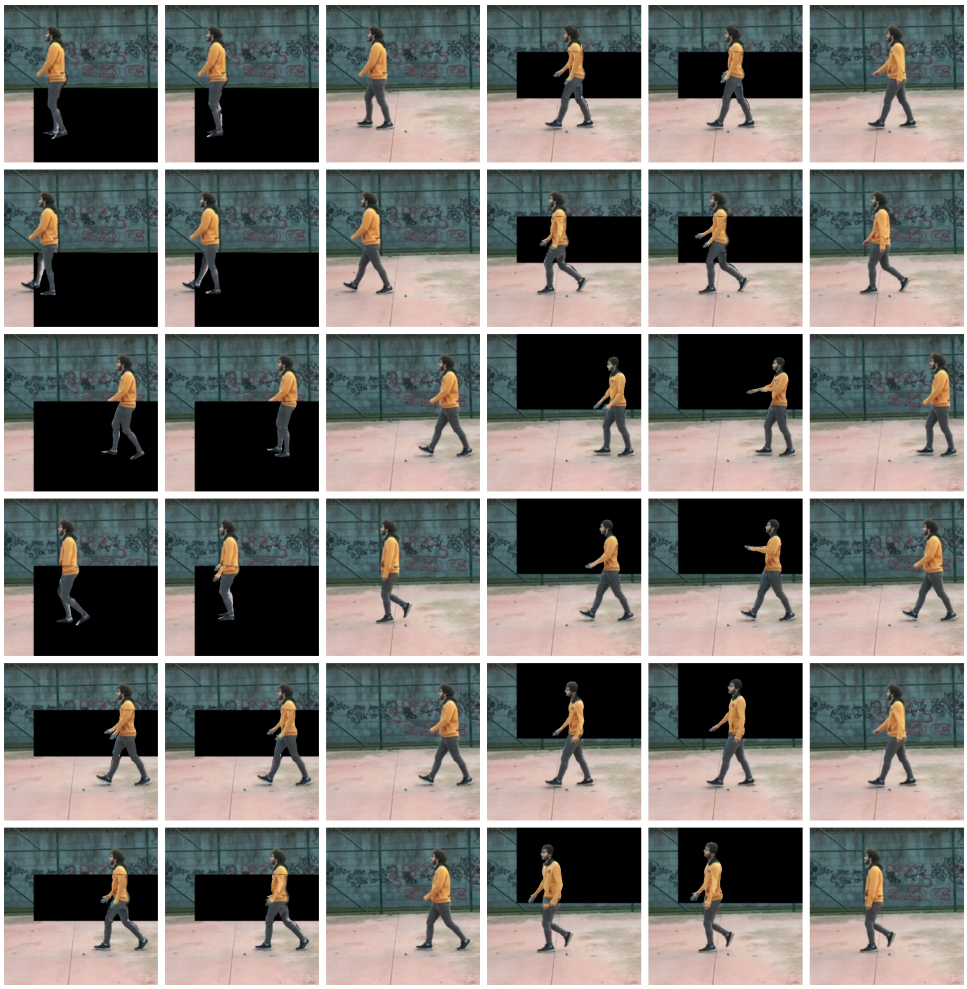


Figure D.1: Inpainting results of video sequence 1 related to the BAKE-Openpose comparison experiment. Columns 1 and 4 correspond to inpainting results with occluded keypoint estimates of BAKE while columns 2 and 5 correspond to inpainting results with occluded keypoint estimates of Openpose. Columns 3 and 6 are the ground truths.

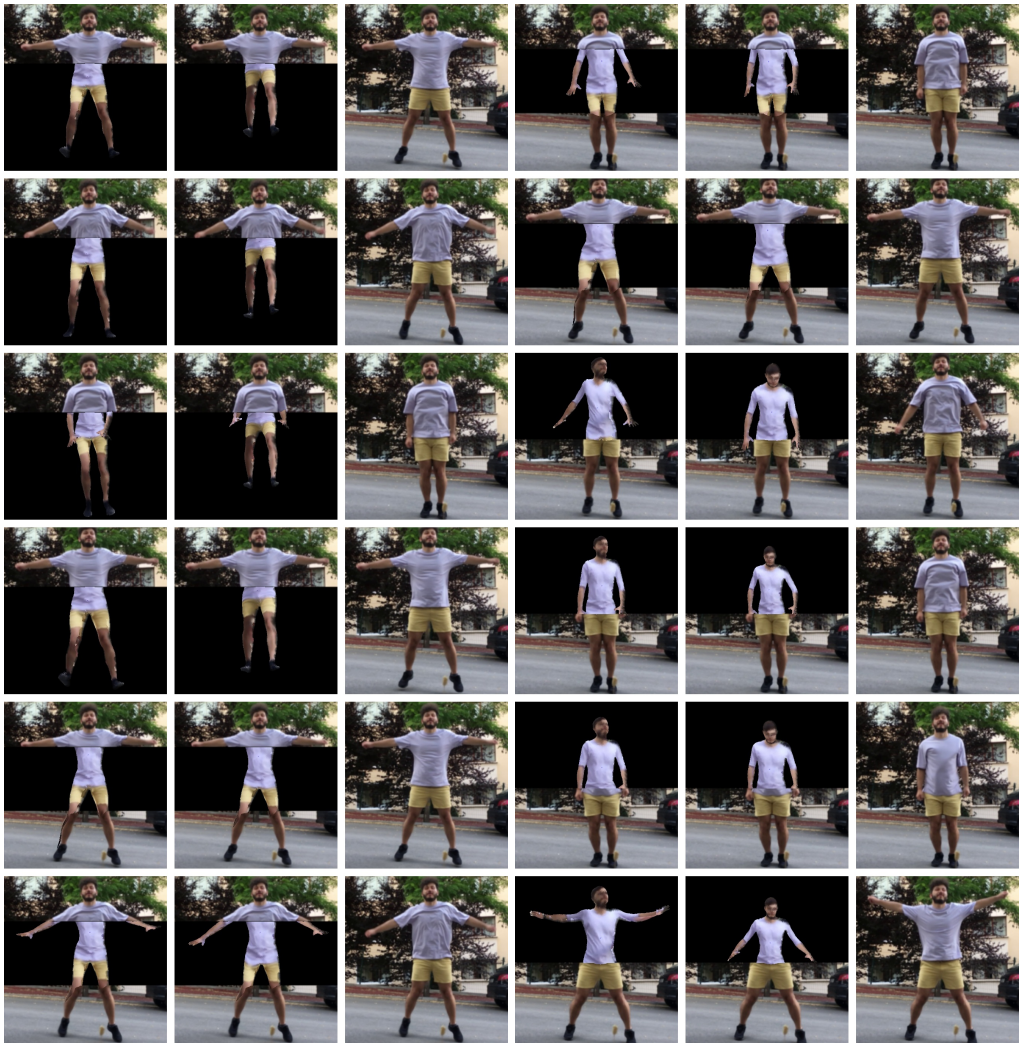


Figure D.2: Inpainting results of video sequence 2 related to the BAKE-Openpose comparison experiment. Columns 1 and 4 correspond to inpainting results with occluded keypoint estimates of BAKE while columns 2 and 5 correspond to inpainting results with occluded keypoint estimates of Openpose. Columns 3 and 6 are the ground truths.

APPENDIX E

INPAINTING RESULTS FOR DIFFERENT CANDIDATE SETS

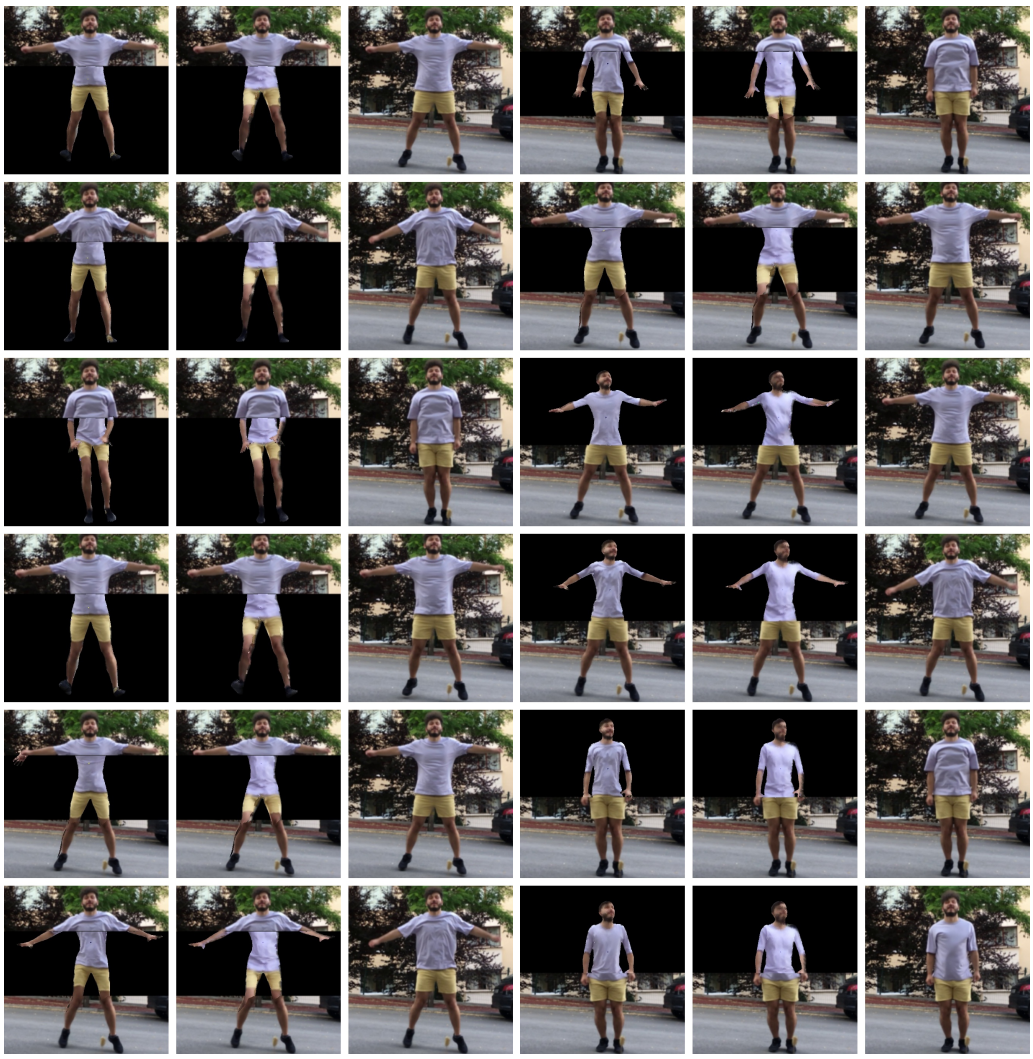


Figure E.1: Inpainting results of video sequence 2 related to the candidate selection experiment. Columns 1 and 4 correspond to inpainting results with candidate set 1 while columns 2 and 5 correspond to inpainting results with candidate set 2. Columns 3 and 6 are the ground truths.