A Unified Approach for Occlusion Tolerant 3D Facial Pose Capture and Gaze Estimation using MocapNETs

Ammar Qammaz
Computer Science Department
University of Crete, Heraklion, Crete, Greece
and
Institute of Computer Science, FORTH
Heraklion, Crete, Greece
ammarkov@ics.forth.gr

Antonis A. Argyros
Computer Science Department
University of Crete, Heraklion, Crete, Greece
and
Institute of Computer Science, FORTH
Heraklion, Crete, Greece
argyros@ics.forth.gr

Abstract

We tackle the challenging problems of 3D facial capture, head pose and gaze estimation. We do so by extending MocapNET, a highly effective deep learning motion capture framework. By leveraging state-of-the-art RGB/2D joint estimators, the proposed network ensemble converts 2D facial keypoints into a real-time 3D Bio-Vision Hierarchy (BVH) skeleton in an end-to-end fashion, incorporating inverse kinematics computations. Our approach achieves satisfactory performance on benchmark datasets and also architecturally excels in challenging scenarios with significant facial occlusions. Moreover, it runs in real-time on CPU, which makes it an ideal choice for applications requiring low-latency interactions. Overall, our unified approach for facial capture, head pose and gaze estimation provides a robust solution for capturing facial expressions and visual focus, with huge potential in HCI and AR/VR applications. Notably, our approach is naturally integrable with MocapNETs for 3D human body and hands pose estimation, offering one of the few state-of-the-art unified approaches that enable holistic recovery of 3D information regarding human gaze, face, upper/lower body, hands, and feet.

1. Introduction

Facial pose estimation and gaze detection are indispensable in various computer vision applications, including automotive safety, assistive technologies, collaborative robotics, human-computer interaction, smart homes and AR/VR. The rise of neural networks has significantly influenced the development of methods for monocular gaze estimation [13], head pose estimation [4], and facial capture [74]. In line with the remarkable advancements seen this year, such as the release of ChatGPT, previously distinct areas of study in 3D human perception are converging [47, 66, 53, 72, 70, 68, 23]. The field of computer vision research is witnessing a trend of integrating diverse concepts under a unified subsymbolic knowledge base [32].

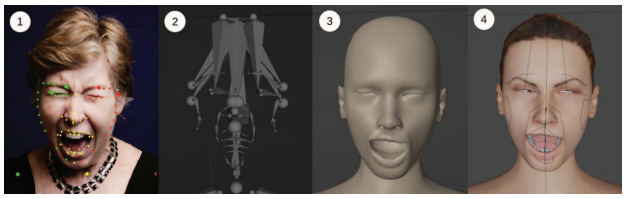

Figure 1. (1) Given an RGB image, we extract 68 2D Multi-PIE [28]/IBUG [54] facial landmarks, and 2 iris positions. We encode the points as enhanced Normalized Signed Rotation Matrices (eNSRMs) [51] and feed them as input to our neural network that directly emits a (2) Bio Vision Hierarchy (BVH) [40] skeleton, including inverse kinematics that can (3) be directly used to animate the vertices of a skinned facial model or (4) with the help of a 3D renderering program like Blender [8], a realistic virtual avatar.

Motivated by this trend, we propose (see Figure 1) a novel approach that extends the highly effective MocapNET deep learning motion capture framework [1, 49, 50, 51, 48] to now accurately estimate 3D head pose, facial capture, and gaze detection. MocapNET was originally designed for human body pose estimation, offering low computational complexity and an extendable Bio Vision Hierarchy (BVH) [40] motion capture output format compatible with various 3D graphics engines and editors. It has demonstrated occlusion tolerance [50], has been extended to 3D hand pose estimation [51], and was proven to be suitable for low-resource devices like mobile platforms [48].

Building upon this foundation, the work presented in this paper tackles 3D gaze and facial estimation, Novelties of the method include leveraging facial symmetry and being occlusion tolerant by design. We introduce a series

of novel additions to the baseline like incorporating Sobol sequences [57] and integrated BVH rendering sample generation among others, to overcome the encountered challenges and to achieve our goal. Moreover, this work completes the puzzle of transcending the boundaries for total human capture using MocapNETs since now 3D gaze, face, upper/lower body, hands, and feet can all be addressed within the same unified framework.

Summarizing our contributions:

1. To the best of our knowledge, this is the first work to directly regress facial 3D Bio-Vision-Hierarchy (BVH) [40] motion capture data in an end-to-end fashion from monocular RGB/2D data.

2. The proposed approach runs on CPU-only systems with state of the art computational performance.

3. Combined with upper body [50], feet and hand [51] MocapNETs, it yields a total body capture solution.

Our source code is publicly available on GitHub [2], serving as a resource to bolster further research.

2. Related Work

Due to the broad scope of related work we will attempt to address them on a per topic basis and subsequently identify similar methods to the one we propose in the holistic total capture scenario. The methods can be broadly split into two major categories, 1-stage and 2-stage methods. 1-stage methods directly regress the output from RGB observations. 2-stage methods first estimate 2D landmarks from RGB, and then regress the target based on this 2D data.

3D Head Pose Estimation: A recent survey [4] summarizes 3D head pose estimation research. The problem is long standing with early approaches using RGB-D cameras and stochastic optimization [45]. The advent of neural networks yielded seminal works like Liu et al. [36] that regress roll, pitch and yaw in 1 stage using convolutional neural networks (CNNs), while methods like [61] using 2D landmark-based face alignment to improve accuracy. Our method has conceptual similarities to ASMNET [21] which is a 2-stage method that uses a statistical shape model to match 2D observations to facial structure and thus regress its pose.

3D Gaze Estimation: Gaze estimation methods focus primarily on the eyes and typically employ head mounted RGB cameras. A recent survey [13] provides a detailed study of the topic. Influential 1-stage methods include ITracker [35] which paired a CNN with an RNN for temporal refinement, MPIIGaze [69] which uses a CNN to regress 3D gaze from a single eye image, achieving high accuracy and robustness to variations in head poses and illumination conditions. Our method, however, falls in the 2-stage category with notable methods that follow this approach being EyeNet [64] that operates on infrared images of the eye from a head mounted camera. Methods that also handle head pose include [44] for full views of the body far from the camera. RT-GENE [22] achieves real-time performance and demonstrated accurate gaze estimation in natural environments. The method features eyetracking glasses for recording ground-truth and a smart inpainting GAN solution to provide realistic training samples. A new class of heavier offline algorithms uses Generative Adversarial Networks (GANs) like GazeDirector [63], head2head [34, 17] and implicitly deals with gaze detection by modifying gaze at will while also maintaining high fidelity video output.

3D Facial Capture: A recent monocular facial capture survey is offered by Zollhofer et.al. [74]. After a period where RGB-D cameras were frequently used for 3D facial capture [31], there is now a recent push with methods using GANs [18] or Diffusion Models [9], that exploit their ability for texture completion and extract a 3D morphable facial model along with texture, UV maps and BRDF maps [46]. Methods like EMOCA [15] employ very elaborate architectures with differentiable renderers and provisions like "Emotional Recognition" and "Emotional Consistency Losses". Other methods employ the same high quality model while also integrating a large number of frames from a video source to further improve visual fidelity [27]. Neural Radiance Fields have recently been coupled with morphable face models [5]. These methods perform very dense and high quality facial capture which is not at all possible in our case since our method just animates a sparse BVH skeleton which by design offers a more limited expressional output. Our method is much more lightweight in a manner similar to Kartynnik et.al. [33], however instead of regressing a facial tessellated mesh we employ purely rotational BVH configurations for the face. This is a much higher level representation that does not encode appearance at all, however it makes calculations (about e.g. how open are the eyes, or the mouth, the gaze, eyebrow tilt etc.) very accessible to potential applications since they can be directly accessed as Euler angle floating point numbers.

Facial Models: Several 2D and 3D facial models have been proposed, offering various degrees of detail and complexity. A Facial 3D model survey [19] offers a comprehensive list. Out of them two stand out having been used in many subsequent works, 3DMM [58] and the FLAME model [15]. FLAME combined with the SMPL model makes the SMPL-X [47] full body model which is commonly used in most Total Capture Methods in the literature that we will examine in the next section. 2D Facial models are also important and typically dictated by the popularity of facial databases and their ground truth annotation error. Very sparse facial models make annotation more accurate and inference during runtime faster while more complex facial models

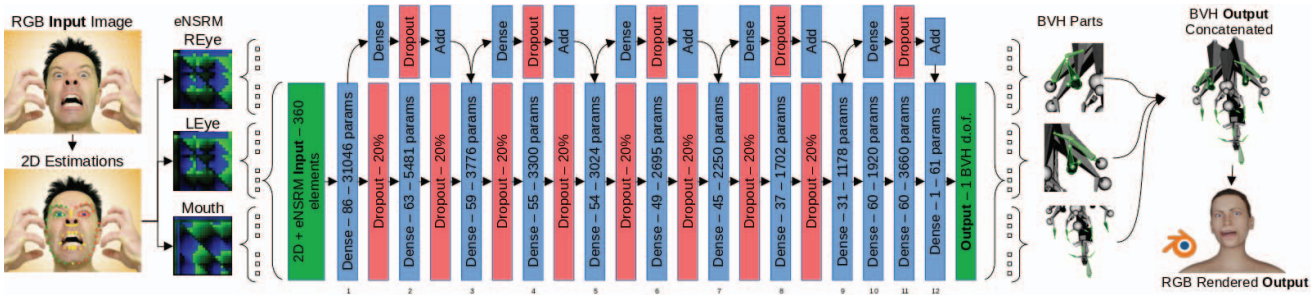

Figure 2. We propose three region separation of the human face with R/L eyebrows, eyelids and eyes and the nose/mouth. Upon receiving 2D facial joints we create three eNSRM [51] descriptors which are fed to an ensemble of encoders. The illustration highlights encoder architecture which is densely connected with skip connections every three layers and a relatively high 20% dropout until layer 8. Encoder output includes BVH inverse kinematics concatenate-able without any further processing and rendered using a skinned 3D model.

have the opposite effect. Influential 2D models include XM2VTS [42] and Multi-Pie/IBUG [28] model which is compatible with OpenPose [11] and thus very commonly used in the literature. For that reason we adopt IBUG and go into details regarding its keypoints in Section 3. Our 3D BVH output results can be used to re-animate a virtual skinned avatar of arbitrary appearance. We inherit the BVH armature of the baseline body pose estimation method [51] and use MakeHuman [37] making our work integrateable with open-source community developed tools.

Total Capture Methods: Our work combined with the baseline for body [50] and hands [51], yields a Total Capture Method. The first total capture methods appeared in 2019 [47, 66]. They were offline and regressed coarse 3D facial expressions without explicit 3D gaze estimation. These methods did not offer quantitative results on the facial capture task due to their main focus being the torso. The first real-time total capture method [72], used a combination of the SMPLH body model with a 3DMM [58] head model. In 2021 two more Total Capture methods appeared [70, 53] Notable recent methods include ZoomNAS [68] and a recent work during 2023 focusing on sign language [23].

3. Methodology

The MocapNET [51] baseline formulation that we used as the basis for the development of this work was initially conceived for body tracking purposes [49]. It thus required substantial methodology changes throughout its formulation in order to be successfully applied to our problem. We will attempt to provide a complete methodological description while also highlighting the most contrasting aspects. The overall design of the method is illustrated in Figure 2. Our framework receives an RGB image and performs 2D Joint Estimation resulting in 68 facial and 2 iris input 2D points. The appearance, shape and expression of a human face can vary wildly. Thus, these points are a very sparse 2D projection of a very high dimensional space. We encode these into eNSRM matrices [51] that represent the pairwise 2D joint relations. Depending on visibility, we feed the data to encoder ensembles that directly regress BVH [40] motion channel output (see Figure 2). Each trained encoder handles a specific degree of freedom (DoF) of the BVH frame. The first three encoder outputs are a positional component of the X, Y, Z location of the head. The rest are rotational components that describe the complete kinematic chain from the neck and upwards. Thin client applications at this point do not need to do any more inverse kinematics calculations and can consume this output directly. Just by concatenating the BVH channel outputs, for example a Raspberry Pi 4 gaze detector application, can perform forward kinematics and via a 3D ray/triangle intersection test, produce a "gaze attention" event. Each configuration angle for the skeleton is in relation to its parent and, thus, position/rotation invariant with respect to the global pose of the head. To produce "higher fidelity" output we can animate a skinned model as seen in parts 4, 5 of Figure 1 and in Figure 5. BVH is natively compatible with most 3D graphics engines and 3D editors like Blender [8] which we used for the illustrations of this paper. The total body capture system involves conditionally running our encoder ensemble along with the other MocapNETs [49, 50, 51, 48] depending on facial visibility and overwriting the observed facial BVH motion channels.

2D Joint Estimator: Our method is oriented towards in-the-wild, real-time operation using cheap off-the-shelf webcam grade RGB sources. We experimented using two different real-time RGB to 2D facial landmark estimators, OpenPose [11] and Mediapipe [6, 25]. The OpenPose [11] pose estimator produces 68 2D facial keypoints following the Multi-Pie/IBUG [28] configuration, appended by the two iris center locations. Mediapipe BlazeFace [6] produces a tessellated FaceMesh with 468 points and eye-gaze data needs to be extracted separately using MediaPipe Iris [25] output which also tracks the iris size and eyelids. We manually create 2D joint associations between the two models in order to work using the sparse Multi-Pie/IBUG [28] stan-

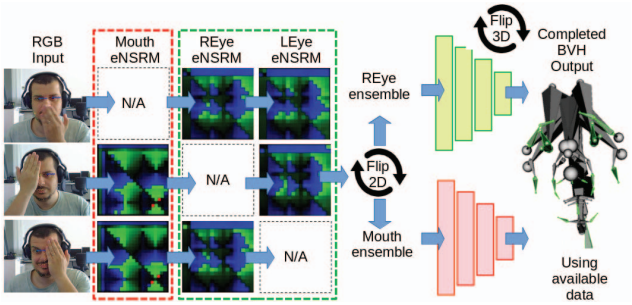

Figure 3. We visualize eNSRM [51] matrices using Blue color for negative values and Green color for positive values, scaling colors linearly according to each matrix element magnitude. The illustration showcases the constructed eNSRM matrices for three RGB inputs on the left showcasing substantial occlusions. Our divide and conquer approach gracefully handles such cases without affecting visible portions of the image. We use the same ensemble for L/R Eye by leveraging 2D input / 3D output symmetries.

dard so the two 2D joint sources become interchangeable[1].

Skinned Model: BVH skeletons are a very good fit for the baseline body pose estimation methods [49, 50, 51, 48] due to a 1:1 correspondence between BVH joints, the actual human skeleton and 2D joint estimations produced by e.g. OpenPose [11] using the BODY 25 [14] standard. Although in our case this holds true for eye gaze vectors which are common between the BVH armature and actual observed eyes, unfortunately all other facial landmarks reside on the skin and thus cannot be directly represented using a purely BVH representation. To remedy this problem we introduce a skinned model, rig it with our BVH armature, make sure that skin vertices are correctly weighted and assigned to skeleton bones, and that facial expressions can be sufficiently represented without rendering artifacts like candy-wrapping etc. We employ the Makehuman [37] creator utility which creates 3D skinned avatars that can be controlled using a wide range of controls (gender, race, weight, age) and internally maintains a 1:1 correspondence to a BVH skeleton armature. We also create a plugin [3] that uses the Blender [8] versatile Python SDK and the MakeHuman For Blender module [38] that allows seamless pose control and rendering while also being open-source and extendable. We map the parametric space over the facial controls and are thus able to control the neck and head pose, eye gaze, mouth movements, eyelids, eyebrows and nose sniffling. However, having a model and a way to create training facial renderings from BVH parameters requires one more step since our method relies on 2D Joints and not synthetic RGB renderings. To do so, we manually created 2D joint associations between the skinned MakeHuman models employed and our OpenPose/IBUG/Multi-Pie facial landmarks. The used facial model consists of 3 separate mesh geometries, one for body and face, one for eyebrows and one for the eyes. Having a 2D IBUG to 3D Makehuman model to BVH parameters triadic correspondance allows us to generate pixel perfect samples to train our neural network ensemble.

Model limits: The baseline body pose estimation method [49] used the BVH conversion [29] of the CMU MOCAP database [59], recorded using a VICON motion capture system, thus training on realistic and physically plausible motions. Its extension to hands [51] was based on the strict bio-mechanical limits of human hands. Unfortunately, we found no publicly available facial BVH motion capture data for this work, and due to the loose correspondence between skinned model and BVH skeleton we relied on manual probing of the model to discover its limits.

2D Input descriptor: The presented method falls into the 2-stage category. The RGB image is first processed through a convolutional network yielding discrete 2D points with a detection confidence for each one. Subsequently, 2D points are converted via our formulation to 3D/IK BVH output.

We receive 68 facial and 2 iris landmarks with normalized coordinates (a_x, a_y) in the range $[0, 1]$ for each 2D joint a. Each joint also carries a visibility parameter a_v which we threshold against a configurable lower limit marking and thus eliminating low confidence or occluded joints with $0, 0$. For each pair of visible 2D joints $a = (a_x, a_y)$, $b = (b_x, b_y)$ we can define a new point $c = (b_x, b_y - \mid b - a \mid)$ that translated the point b vertically by the length of vector ab. These three points a, b, c are used to encode the relation between points a and b as well as their relative rotation towards a fixed vertical axis as follows [51]:

$$eNSRM(a,b) = \begin{cases} \frac{2}{\pi} tan^{-1}\left(\frac{|\vec{ab} \times \vec{cb}|}{\vec{ab} \cdot \vec{cb}}\right) & a \neq b, \\ |\vec{aR}| & otherwise, \end{cases} \quad (1)$$

with $\cdot$ and $\times$ denoting inner and cross products. Each resulting $eNSRM(a, b)$ angle is invariant to 2D point cloud translation and scale. Scale is encoded in the diagonal with Euclidean distances from the R root joint, the relative position of joints using the rotation formed from triangle $\hat{abc}$, while preserving their relative orientation to the world coordinate system (with bc being parallel to the Y axis). In contrast to the baseline eNSRM [51] formulation we take into account that $tan^{-1}(\mathbb{R}) \longrightarrow (-\frac{\pi}{2}, \frac{\pi}{2})$ and multiply values with $\frac{2}{\pi}$ to normalize them in the range [-1,1]. A pictorial visualization of the matrix can be seen in Figure 3.

Sobol Sampling: Facial expressions differ much more subtly in relation to human hands and limbs. A person crossing

[1]For reproducibility we provide the association map between OpenPose/IBUG/Multi-Pie facial landmarks (1..68) to mediapipe's FaceMesh vertices, which is (34, 227, 137, 177, 215, 138, 170, 171, 152, 396, 395, 367, 435, 401, 366, 447, 264, 70, 53, 52, 65, 55, 285, 295, 282, 283, 300, 168, 197, 5, 4, 102, 79, 2, 19, 309, 331, 33, 160, 158, 133, 153, 144, 362, 385, 387, 263, 373, 380, 61, 40, 37, 0, 267, 270, 291, 321, 314, 17, 84, 91, 78, 38, 12, 268, 308, 316, 16, 86).

his arms and legs creates a vastly different 3D configuration and 2D projections compared to a person making a T-Pose and the same is true about the hands. Furthermore, being able to observe a person from any orientation $\pm 180^{\circ}$ produces a much bigger variety of training samples compared to faces. Thus, the face capture problem raises the unique requirement of having to have very fine tuned output on a relatively limited number of views. As already mentioned, the large magnitude of 2D changes and the presence of existing MOCAP data [49] give the original method the option to bypass this problem. For hands [51] the strict mechanical limits and randomization manage to achieve a similar effect. In the case of faces though we need a way to acquire 2D samples in the high dimensional 3D BVH parametric space that finely cover the multi-dimensional space in a homogeneous fashion. To this end we used Sobol sequences [57] that ensure a perfect quasi-random sampling distribution when the number of samples $N = 2^x$ was initially for x up to 51 and in recent implementations up to 1111 dimensions. Due to the exponential number of samples and limited available GPU NN training resources, picking a correct value for x is an important decision. Assuming a 3D armature of dimensionality D, in order for example to be able to sample 2 distinct configurations for each dimension, we would assume that we need $N = D^2$ samples. Thus, the "alloted" number of samples per dimension N_{dim}, assuming perfectly equidistant samples acquired using Sobol sequences [57] should be $N_{dim} = N^{(1/D)}$. Taking into consideration our ensemble facial partitions for $N = 2^{20} = 1M$ samples for the R/L eye which has 12 output dimensions we get $N_{dim} = 3.17$ unique samples per dimension which roughly means that partitioning their motion range in three non overlapping regions their min/max and mean areas should be covered. Mouth controls occupy 18 DoFs and thus for $N = 1M$, $N_{dim} = 2.16$ which gives less resolution, however still adequately covers our training space. We should keep in mind that using Sobol Sequences samples always uniformly cover the whole range of values for each dimension so the N_{dim} metric is not a hard-limit but rather used as a tool to conceptualize and quantify the effective number of unique samples per dimension.

Dimensionality, Occlusions and Symmetries: We described the basic building blocks that were used to facilitate the NN learning task. These components can be combined in many ways. A naïve approach would be to just create a single eNSRM descriptor describing the whole face in an attempt to tackle the full problem at once. However, trying this approach led to a number of unattractive properties.

First of all, as already mentioned, the number of samples needed to adequately cover a multi dimensional space scale exponentially to the number of dimensions. In more practical terms, required GPU VRAM to facilitate the combined problem was also a prohibitive factor towards this solution.

A second issue we encountered is joint imbalance in contrast to their importance. In a theoretical example where we would encode all 70 facial points of our IBUG+Iris representation using an eNSRM matrix we would have 4900 input elements to our neural network. Out of those, only two lines and two columns would encode the very important information about eye gaze using only 276 elements. Despite the very large importance of the eye gaze, roughly 95% of the eNSRM matrix would encode the rest of the face, thus making eye gaze much more challenging for the Neural Network to learn. The small relative motion of the eyes compared to the jaw/mouth, or even more pronounced motion of the whole head further exacerbated this issue.

A third very important cause for our decision to split the face into three areas is making our method occlusion tolerant. As seen in Figure 3, by splitting the problem into three sub-problems even in the complete absence of a whole quadrant of the face due to occlusions there is no impact on the visible regions. Having a single eNSRM for the whole face would cause a substantial part of it to go missing thus adversely affecting even the areas being perfectly visible.

A final attractive property that prompted us to break-up the problem into smaller ones is that the human face exhibits horizontal symmetry. We take advantage of this by only training an ensemble for the Right Eye and resolving the Left Eye by mirroring its X coordinates on the X axis and then multiplying 3D BVH outputs by -1 to mirror them back. This offers multiple benefits since the runtime requires only 2/3 of the memory it would otherwise require, there are no biases in handling either the left or right eye and we where able to conduct training with a $2\times$ speedup.

The mouth eNSRM descriptor uses the chin as its root joint, chin $\rightarrow$ nose as matrix alignment and adding 2 joints for inner and outer mouth, yields a 21×21 matrix. The right eye eNSRM descriptor uses nose as its root and is aligned horizontally in the virtual line that connects the edges of the eyes. This is done to make eNSRM value magnitudes bigger. The eNSRM is comprised of these elements plus the eyebrows, the eyelids, as well as the iris. We also include the chin and IBUG joint #1 to provide a larger frame of reference. The points mentioned until now sum up to 15 out of which the eye gaze is only encoded by a single row/column with 29 elements out of 225. Although this is better compared to a descriptor featuring the whole face, we proceed to create 3 more virtual points to better encode the position of the iris and make its features more prominent during training. The first virtual point is located halfway between the nose and iris, while we create two more virtual points by shifting the iris by (0.015,0.026) units and (-0.015,-0.026). This brings the REye eNSRM descriptor to a size of 18x18 out of which 4 lines/columns ($\approx 39\%$) encode the iris.

Model Training: We use Keras and Tensorflow 2 for training. Our ensemble is trained on a per encoder basis using

the ADAM optimizer, a learning rate of 0.00017, a batch size of 32 and a maximum of 24 epochs with early stopping when loss delta is less than 0.01. We use a mean squared error (MSE) loss function to incur heavier penalties on outliers. Training each encoder separately combined with sparse network input and output allows us to accommodate a very high number of training samples. Training follows the BVH hierarchical order with each encoder initialized with its parent in a form of transfer learning which seems to substantially help training. We use the SWISH [52] activation function, 32bit floating point precision, and a randomization seed set to 0 for reproducible experiments.

In contrast to the baseline [49, 50, 51] we normalize BVH outputs to $[-2, 2]$ to control our MSE loss maximum weight updates, since we both want to heavily suppress and penalize mistakes of the network, but at the same time not to perform very large weight updates that might destabilize network convergence. Due to our output normalization during training mean average error and other metrics ceasing to correspond to Euler angles after normalizing the BVH ground truth, in order to better quantify our training we also extract the R^2 coefficient of determination [62]. Using the configurations mentioned above we routinely achieve training $R^2 > 0.97$ for the encoders that deal with the REye and $R^2 > 0.90$ for the encoders that deal with the mouth. MocapNETs use a λ variable to scale the size of their hidden layers, this can be used to compress the network for further performance benefits, we use $\lambda = 1.8$ for the Mouth ensemble yielding $\approx 3M$ parameters and $\lambda = 0.6$ for the REye ensemble yielding $\approx 5M$ parameter ensembles with encoders with the shape shown in Figure 2. Due to different input dimensionalities using the same λ for the mouth makes each encoder $\approx 1M$ leading to a very large network.

NN Encoder/Ensemble Outline: After training, individual encoders are concatenated into two different ensembles for the Right Eye and Mouth. The Left Eye is handled by horizontal mirroring as already mentioned and seen in Figure 3. Our implementation thresholds 2D joints as occluded if their confidence is < 0.4. Ensembles with more than 15% missing input are not executed to reduce 3D output noise and unsettling grimaces by our model. In video streams we treat missing input with two policies. We remember the last confident configuration, or revert to a default facial expression where all BVH outputs for the region are set to 0. An optional assumption is eyes moving together allowing us to populate the occluded eye by mirroring the visible eye.

4. Experiments

To assess the performance and accuracy of our method we use a combination of known benchmarks featuring realistic samples and high quality ground truth. To quantitatively assess our method we use the combination of AFW [73], FRGC [43], LPFW [7], HELEN [71], IBUG [55] and XM2VTS [41] dataset 68 point annotations [10]. We use the Columbia Gaze Dataset [56] to assess 3D gaze and 3D head pose accuracy. For Qualitative results we use 300 Faces in the Wild (300W) dataset [54].

Method	**Gaze Error**
Deepwarp [24]	15.3°
He et al. [30]	8.7°
Xia et al. [65]	6.3°
Ours $[-30^\circ..15^\circ]$	14.8°
Ours $[-30^\circ..30^\circ]$	18.7°

Table 1. 3D gaze estimation accuracy comparison in degrees. As also seen in Figure 4, average accuracy is negatively impacted ($\approx +4^\circ$) including camera views diametrically opposed to the eye.

Dataset	**Median**	**Mean**	**St.Dev.**
AFW [73]	0.90%	1.44%	0.02
300W Outdoor [54]	0.95%	1.73%	0.02
300W Indoor [54]	1.46%	2.50%	0.03
IBUG [55]	1.57%	2.69%	0.03
XM2VTS [41]	2.01%	2.52%	0.02
FRGC [43]	2.26%	2.78%	0.09
HELEN [71]	3.42%	4.33%	0.03
LPFW [7]	3.36%	4.13%	0.03

Table 2. Quantitative results converting ground truth 2D data to BVH output with our method, rendering the skinned model on Figure 5, getting corresponding 2D joints out of the 3D model and comparing it to input after Procrustes analysis [26]. Results use Normalized Mean Error (NME) w.r.t. to input image resolution.

Method	**m.a.d.**	**Ex. time (sec)**
Martinez et al. [39]	0.0514	42.5
Čech et al. [12]	0.1047	4.05
Uřičář et al. [60]	0.0970	3.46
Deng et al. [16]	0.0226	1.97
Fan et al. [20]	0.0309	1.29
Ours	0.1623	0.03

Table 3. Comparison of the mean absolute deviation (m.a.d.) of 2D fitting results for 68 facial landmarks with the associated mean computational cost required. Our method uses 2D ground truth which is regressed to a BVH facial configuration and reprojected back to 2D points and compared to ground truth using procrustes analysis [26]. Percentages reflect normalized pixel distance w.r.t. to the d_{outter} metric defined in Figure 6 of paper [54].

Quantitative Experiments: The proposed ensemble of neural networks takes as input a list of 68 IBUG-compliant [55] 2D points, along with 2 iris locations and generates a 3D BVH frame that represents 3D facial configuration and gaze as relative rotations in a kinematic chain. While this design simplifies measuring gaze and head pose direction for benchmarking purposes, it also introduces challenges in studying the quality of 3D facial cap-

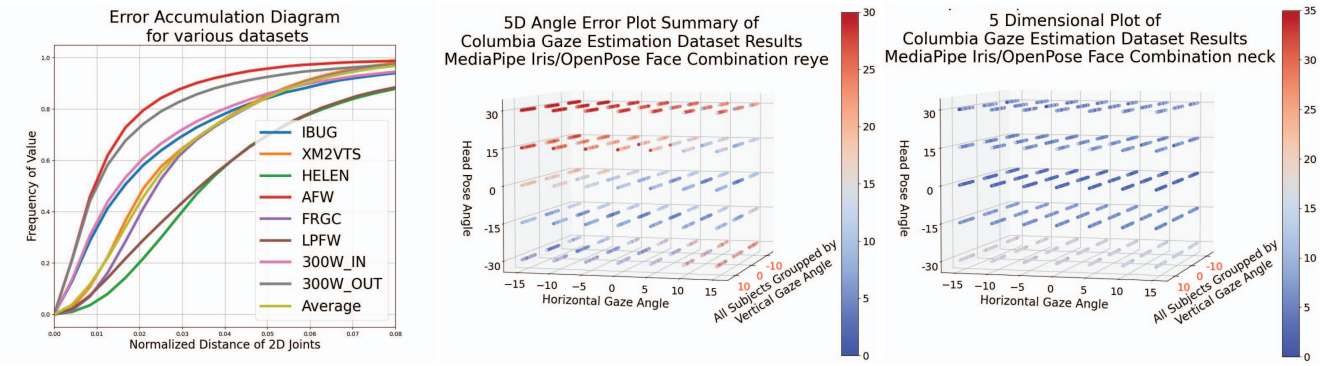

Figure 4. Left: Error accumulation graph for 68 joint 2D reprojections of our 3D output compared to ground truth using Procrustes analysis [26]. X-Axis uses normalized coordinates w.r.t the input image dimensions. Middle: Quantitative 3D eye gaze accuracy results measured against the Columbia Gaze Dataset [56]. The dataset contains RGB images from 56 subjects with gazes fixed at specific intervals. We regress and plot the angular error of the right eye using color, plotting all subjects adjacent one to the other. Each 3D line depicts results for all subjects. X, Y and Z axes depict horizontal/vertical gaze angle and head pose angle in relation to the camera. Right: Quantitative 3D neck/head pose accuracy seems uniformly good across all subjects with slightly elevated errors around the -30° limit.

ture. For head pose and gaze, we utilize the Columbia Gaze Dataset [56], which provides high-quality images (5184×3456 resolution) of 55 subjects. The subjects rest their heads on a calibrated height-adjustable metallic chinrest, with their eyes stabilized 70cm above the floor. The dataset captures a large range of horizontal and vertical gazes using a 5-camera system positioned 2m away from the subject. This yields 105 different gaze/pose combinations per subject. The ground truth for the dataset is given in terms of head pose and the relative angle of the person with respect to a wall located 2.5m away, featuring a grid with the various recorded focus points. Our method generates individual 3D vectors for each eye and the entire head. The ground truth vector can be easily converted to each individual eye by considering the inter-pupillary distance (IPD) measurement. The average IPD for human males is 0.064m and 0.0617m for females. We use an IPD of 0.06285 for our calculations slightly penalizing our measurements but making them gender neutral. We extract 2D landmarks from the RGB frames using OpenPose [11] + iris data supplied by Mediapipe Iris [25]. The results for gaze and head pose, shown in Figure 4 and Table 1, indicate satisfactory accuracy across subjects. However, accuracy degrades when the camera view is diametrically opposite to the eye. Since our method relies on 2D data, occlusions in the 2D space adversely affect our 3D BVH output, as expected. In contrast, head pose estimation appears to be more consistent across all subjects/views, which can be attributed to the availability of more 2D joints that enable better regression results.

Quantitatively assessing facial capture is more challenging since our method regresses all facial controls as 3D vectors rather than 2D or 3D points. Additionally, our method is not trained on any samples from the datasets, and employs a constant BVH kinematic skeleton for any input 2D face. While this guarantees consistent output adhering to strict limits and predictable facial configurations, it makes accuracy measurement more difficult. Direct output to ground truth comparisons are heavily influenced by the discrepancy between our constant model and each specific head.

Another challenge arises from image aspect ratios. Our proposed method assumes the commonly used 1920x1080 video resolution (aspect ratio 1.777) to accommodate a full view of the body for total capture and be compatible with off-the-shelf cameras. Head pose estimation methods on the other hand use rectangular patches centered around the head with an aspect ratio of 1. Since we receive 2D information in normalized coordinates, we need to pad 2D input to maintain consistency, but this shifts the 3D position of the output, requiring careful consideration in experiments.

To address these intricacies, we leverage multiple facial datasets, employ 2D comparison and use Procrustes analysis [26] to mitigate model discrepancies, aspect ratio differences, and positional offsets. Similar complexities might explain why most other total capture methods [47, 66, 53, 23] forego facial accuracy assessment. The two 3D total capture faces that don't , use MTC-Face [67]. Zhou et al. [72] reports landmark/photometric pixel error, while [70] tests on 750 frames of [67]. The body [49, 50] and hand [51] aspects of our baseline have been thoroughly investigated, so we focus on specialized head, gaze and face experiments establishing a comparative context within the broader landscape of standalone methodologies in the field.

We use OpenPose [11] and MediaPipe Iris [25] to acquire 2D landmarks from RGB images. We remove them as variability factors for accuracy measurements by directly using provided 2D ground truth as our input. After regressing the 3D BVH facial frame from ground truth 2D, we reproject 3D results back to 2D and study the error introduced by our NN. To remove positional, aspect ratio and model appearance error from the measurement, we perform

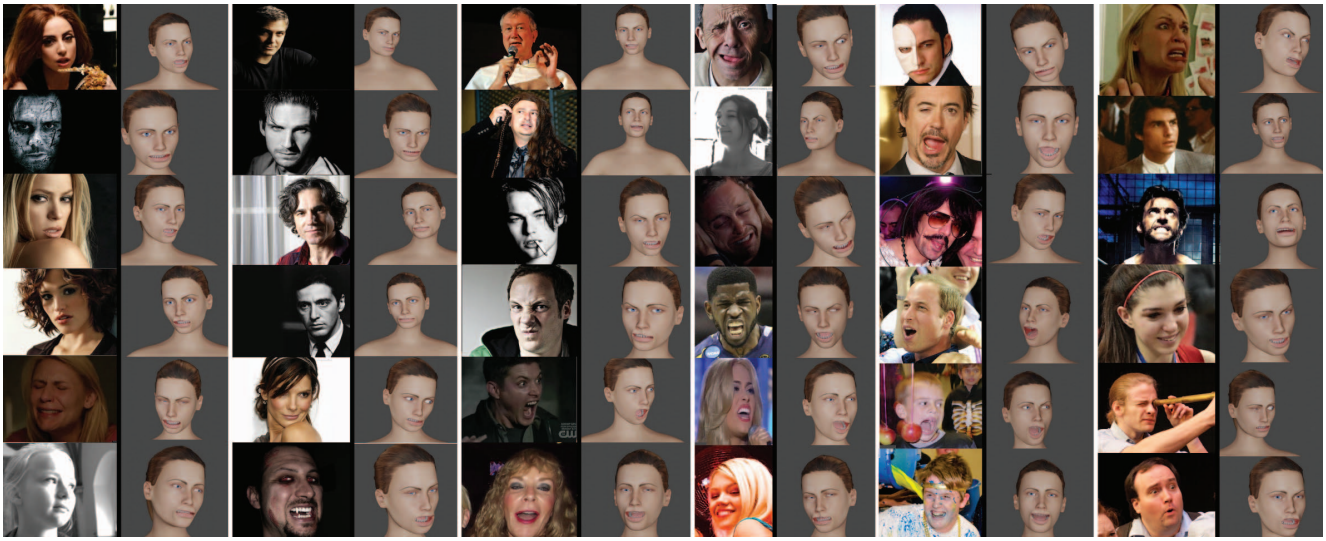

Figure 5. Qualitative illustrations with RGB input images (left) and retrieved renderings (right) from 300W [54] 2D landmarks and iris data extracted with MediaPipe Iris [25]. The BVH output acquired by our NN ensemble is rendered via MakeHuman [37] and Blender [8].

Procrustes analysis [26]. Results in Figure 4 and Table 2 show that overall errors are relatively small. Although introduced error is larger than some other methods (Table 3), our method's computational cost is orders of magnitude lower. Running 3 NN ensembles (REye/LEye/Mouth) in series, using TF-Lite, fp32 precision and no NN pruning or other optimizations on a Raspberry PI4 computer yields an interactive throughput of 4.43Hz. Running the same ensemble converted to Open Neural Network Exchange (ONNX) on a 10 year old Intel(R) Core(TM) i7-4790 CPU yields a real-time 30.87Hz using only CPU resources.

Qualitative Experiments: Figure 5 provides an illustration of our BVH results rendered in Blender using a MakeHuman generated model. Our method accurately translates head pose and gaze, but facial expressions have lower fidelity since the 3D model lacks the intricate wrinkled deformations seen in the RGB image. Nonetheless, expressions such as wonder, smiles, anger, fear, and talking are discernible, which was initially surprising given the limited number of bones used. Failure cases arise when extreme mouth deformations are not properly covered by the range of motion of the mouth bones. To understand why this happens, we can examine the middle of Figure 1 and imagine the mouth being made of soft fabric and being bent using five toothpicks. Although basic expressions like opening, closing, and smiling are possible, the range of motion remains limited due to the sparsity of the bones. Importantly, this limitation is not a drawback of the method or the BVH container itself. A BVH skeleton can have an arbitrary number of joints, so a more detailed 3D model, accompanied by additional BVH controls and, of course, 2D input points, could facilitate a wider range of expressions.

5. Summary

We proposed a novel approach for head pose, gaze estimation and 3D facial capture, by extending the MocapNET deep learning motion capture framework. The proposed method leverages RGB/2D joint estimators and converts 2D facial keypoints into a Bio-Vision Hierarchy (BVH) 3D skeleton in real-time via an ensemble of NN encoders. The approach is computationally very efficient and achieves good results on benchmarks, while architecturally excelling in challenging scenarios with significant facial occlusions. Its real-time capabilities and high level open-standard output make it a robust solution for applications requiring low-latency understanding of facial expressions and visual focus, with potential applications in human computer interaction and VR/AR. Furthermore, our research is a significant advancement for MocapNETs positioning them among a select few [47, 66, 72, 53, 68, 23] of state-of-the-art architectures that can perform holistic 3D total human pose capture.

Acknowledgements

The authors would like to gratefully acknowledge support for this research from the VMware University Research Fund (VMURF). This research was partially supported by BonsApps (EU H2020 Grant no. 101015848) AI Talent grant (Winner No. Bons_1OC_20) and the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for HFRI Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment", project I.C.Humans, number 91. We extend our appreciation to Dr. Iasonas Oikonomidis for his assistance with Sobol sequence sampling and Dr. Anastasios Roussos for his invaluable insights and suggestions.

References

[1] A. Qammaz. Mocapnet github repository. https://github.com/FORTH-ModelBasedTracker/MocapNET, 2019. [Online; accessed 13-June-2023].

[2] A. Qammaz. Mocapnet repository branch for "a unified approach for occlusion tolerant 3d facial pose capture and gaze estimation using mocapnets"'. https://github.com/FORTH-ModelBasedTracker/MocapNET/tree/mnet4, 2019. [Online; accessed 9-August-2023].

[3] A. Qammaz. Blender plugin for mocapnet/makehuman facial control using bvh armatures. https://github.com/FORTH-ModelBasedTracker/MocapNET/blob/mnet4/src/python/blender/blender_face.py, 2023. [Online; accessed 9-August-2023].

[4] Andrea F Abate, Carmen Bisogni, Aniello Castiglione, and Michele Nappi. Head pose estimation: An extensive survey on recent techniques and applications. *Pattern Recognition*, 127:108591, 2022.

[5] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20364–20373, June 2022.

[6] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

[7] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[8] Blender Community. *Blender - a 3D modeling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2023.

[9] Matyáš Boháček and Hany Farid. A geometric and photometric exploration of gan and diffusion synthesized faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 874–883, June 2023.

[10] C. Sagonas, E. Antonakos, G, Tzimiropoulos, S. Zafeiriou, M. Pantic. I-bug facial annotations for xm2vts, frgc ver.2, lfpw, helen, afw and 300w. https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/, 2023. [Online; accessed 13-June-2023].

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[12] Jan Čech, Vojtěch Franc, Michal Uřičář, and Jiří Matas. Multi-view facial landmark detection by using a 3d shape model. *Image and Vision Computing*, 47:60–70, 2016.

[13] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.

[14] CMU Perceptual Computing Lab. Openpose output format specifications, 2023. [Online; accessed 13-June-2023].

[15] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

[16] Jiankang Deng, Qingshan Liu, Jing Yang, and Dacheng Tao. M3 csr: Multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47:19–26, 2016.

[17] Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):31–43, 2021.

[18] Michail Christos Doukas, Evangelos Ververas, Viktoriia Sharmanska, and Stefanos Zafeiriou. Free-headgan: Neural talking head synthesis with explicit gaze control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[19] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

[20] Haoqiang Fan and Erjin Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016.

[21] Ali Pourramezan Fard, Hojjat Abdollahi, and Mohammad Mahoor. Asmnet: A lightweight deep neural network for face alignment and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1521–1530, 2021.

[22] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, pages 334–352, 2018.

[23] Maria-Paola Forte, Peter Kulits, Chun-Hao P Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J Kuchenbecker, and Michael J Black. Reconstructing signing avatars from video using linguistic priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12791–12801, 2023.

[24] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 311–326. Springer, 2016.

[25] Google. Mediapipe iris detection, 2023. [Online; accessed 13-June-2023].

[26] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[27] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022.

[28] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

[29] B. Hahne. *The Daz-friendly BVH release of CMU motion capture database*, 2010. Accessed: 2018-10-05.

[30] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6932–6941, 2019.

[31] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. Unconstrained realtime facial performance capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1675–1683, 2015.

[32] Eleni Ilkou and Maria Koutraki. Symbolic vs sub-symbolic ai methods: Friends or enemies? In *CIKM (Workshops)*, 2020.

[33] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.

[34] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 16–23. IEEE, 2020.

[35] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[36] Xiabing Liu, Wei Liang, Yumeng Wang, Shuyang Li, and Mingtao Pei. 3d head pose estimation with convolutional neural network trained on synthetic images. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1289–1293, 2016.

[37] MakeHuman Community. Makehuman. `http://www.makehumancommunity.org/`, 2023. [Online; accessed 13-June-2023].

[38] MakeHuman Community. Makehuman for blender 2 (mpfb2). `https://github.com/makehumancommunity/mpfb2`, 2023. [Online; accessed 13-June-2023].

[39] Brais Martinez and Michel F Valstar. L2, 1-based regression and prediction accumulation across views for robust facial landmark detection. *Image and Vision Computing*, 47:36–44, 2016.

[40] Maddock Meredith, Steve Maddock, et al. Motion capture file formats explained. *Department of Computer Science, University of Sheffield*, 211:241–244, 2001.

[41] Kieron Messer, Josef Kittler, Mohammad Sadeghi, Sebastien Marcel, Christine Marcel, Samy Bengio, Fabien Cardinaux, Conrad Sanderson, Jacek Czyz, Luc Vandendorpe, et al. Face verification competition on the xm2vts database. In *Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings 4*, pages 964–974. Springer, 2003.

[42] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, Gilbert Maitre, et al. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999.

[43] National Institute of Standards and Technology. Face recognition grand challenge (frgc), 2010. [Online; accessed 14-June-2023].

[44] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2192–2201, 2022.

[45] Pashalis Padeleris, Xenophon Zabulis, and Antonis A Argyros. Head pose estimation on depth data based on particle swarm optimization. In *IEEE Computer Vision and Pattern Recognition Workshops (CVPRW 2012)*, pages 42–49, Providence, Rhode Island, USA, June 2012. IEEE.

[46] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models, 2023.

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[48] Ammar Qammaz and Antonis Argyros. Compacting mocapnet-based 3d human pose estimation via dimensionality reduction. In *International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2023) (to appear)*, Corfu, Greece, July 2023. ACM.

[49] Ammar Qammaz and Antonis A Argyros. Mocapnet: Ensemble of snn encoders for 3d human pose estimation in rgb images. In *British Machine Vision Conference (BMVC 2019)*, Cardiff, UK, September 2019. BMVA.

[50] Ammar Qammaz and Antonis A. Argyros. Occlusion-tolerant and personalized 3d human pose estimation in rgb images. In *IEEE International Conference on Pattern Recognition (ICPR 2020)*, pages 6904–6911, January 2021.

[51] Ammar Qammaz and Antonis A Argyros. Towards holistic real-time human 3d pose estimation using mocapnets. In *British Machine Vision Conference (BMVC 2021)*, Virtual, UK, November 2021. BMVA.

[52] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[53] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.

[54] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[55] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013.

[56] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar. Gaze Locking: Passive Eye Contact Detection for Human?Object Interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 271–280, Oct 2013.

[57] Il'ya Meerovich Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802, 1967.

[58] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017.

[59] Carnegie Mellon University. Cmu graphics lab motion capture database. `http://mocap.cs.cmu.edu/`, 2003. Accessed: 2017-06-01.

[60] Michal Uřičář, Vojtěch Franc, Diego Thomas, Akihiro Sugimoto, and Václav Hlaváč. Multi-view facial landmark detector learned by the structured output svm. *Image and Vision Computing*, 47:45–59, 2016.

[61] Roberto Valle, Jose M. Buenaposada, and Luis Baumela. Multi-task head pose estimation in-the-wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2874–2881, aug 2021.

[62] Wikipedia contributors. Coefficient of determination — Wikipedia, the free encyclopedia, 2023. [Online; accessed 14-June-2023].

[63] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Gazedirector: Fully articulated eye gaze redirection in video. In *Computer Graphics Forum*, volume 37, pages 217–225. Wiley Online Library, 2018.

[64] Zhengyang Wu, Srivignesh Rajendran, Tarrence Van As, Vijay Badrinarayanan, and Andrew Rabinovich. Eyenet: A multi-task deep network for off-axis eye gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3683–3687. IEEE, 2019.

[65] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Wensen Feng. Controllable continuous gaze redirection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1782–1790, 2020.

[66] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019.

[67] Xiang, Donglai and Joo, Hanbyul and Sheikh, Yaser. Monocular total capture dataset. `http://domedb.perception.cs.cmu.edu/mtc.html`, 2019. [Online; accessed 13-June-2023].

[68] Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[69] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[70] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5560–5569, 2021.

[71] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 386–391, 2013.

[72] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4811–4822, 2021.

[73] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.

[74] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer graphics forum*, volume 37, pages 523–550. Wiley Online Library, 2018.