

DEVICE AND METHOD FOR ESTIMATING THE SHAPE AND POSE OF A BODY

TECHNICAL FIELD

[0001] Various aspects of this disclosure relate to devices and methods for estimating the shape and pose of a body.

BACKGROUND

[0002] Monocular 3D human shape and pose estimation is a fundamental task in computer vision, which aims to recover the unclothed human body as well as its 3D pose from a single input image and can be used in many applications including visual tracking, virtual/augmented reality, motion generation, image manipulation, and neural radiance fields.

[0003] Different from multi-view 3D reconstruction where the solution is well regularized by epipolar constraints, this task is in particular challenging due to the inherent depth ambiguity of a single 2D image, which usually requires strong knowledge and informative patterns learned from large amounts of data to generate reasonable results.

[0004] Accordingly, efficient methods for estimating the shape and pose of a body from a single image are desirable.

SUMMARY

[0005] Various embodiments concern a method for estimating the shape and pose of a body comprising obtaining an image of the body, extracting image features from the image, and transforming the image features to a set of parameter values describing the shape and pose of the body by means of a transformer comprising a cascade of transformer units, wherein each transformer unit operates on a representation of a respective version of the set of parameter values and comprises a cross-attention layer modifying the representation of the respective version of the set of parameter values by the image features weighted by an attention of the respective version of the set of parameter values to the image features in sequence with a self-attention layer modifying the representation of the respective version of the set of parameter values by components of the representation of the respective version of the set of parameter values weighted by an attention between the components of the representation of the respective version of the set of parameter values to themselves.

[0006] According to one embodiment, the self-attention layer comprises at least one self-attention unit performing respective attention operation with query, key and value, for which each of the query, the key and the value are set to, up to a modification by positional encoding (which is optional), the representation of the respective version of the set of parameter values.

[0007] According to one embodiment, the cross-attention layer comprises at least one cross-attention unit performing respective attention operation with query, key and value, for which the query is set to, up to a modification by positional encoding (which is optional), the representation of the respective version of the set of parameter values and the key and the value are set to, up to a modification by positional encoding, the image features or a subset of the image features.

[0008] According to one embodiment, the image features comprise image features for each of a plurality of scales and wherein the cross-attention layer comprises a cross-attention unit for each layer modifying the representation of the respective version of the set of parameter values by an attention from the representation to image features of the scale.

[0009] According to one embodiment, the cross-attention layer outputs a modified version of the representation of the set of parameter values which it generates by averaging results of the modifying performed by the cross-attention units for the scales.

[0010] According to one embodiment, the set of parameters comprises values of parameters of a parametric body model.

[0011] According to one embodiment, the method further comprises generating a mesh representation of the body from the values of the parameters of the parametric body model.

[0012] According to one embodiment, the set of parameter values comprises values for a set of parameters including rotation parameters for each of a plurality of parts of the body.

[0013] According to one embodiment, the cross-attention layer comprises a local attention unit modifying the representation of the respective version of the set of parameter values by, for each of the rotation parameters, the image features of a patch showing the body part weighted by an attention of the respective version of the set of parameter values to the image features of the patch showing the body part.

[0014] According to one embodiment, the cross-attention layer outputs a modified version of the representation of the set of parameter values which it generates by summing a result of the modifying performed by the local attention unit and a result of the modifying of the representation of the respective version of the set of parameter values by the image features

weighted by an attention of the respective version of the set of parameter values to the image features.

[0015] According to one embodiment, the set of parameter values comprises values for a set of shape parameters.

[0016] According to one embodiment, the image is obtained by means of a camera and the set of parameter values comprises values for a set of camera parameters.

[0017] According to one embodiment, the method comprises extracting the image features by means of a convolutional neural network.

[0018] According to one embodiment, the method comprises training the transformer by supervised learning.

[0019] According to one embodiment, the transformer is trained using a ground truth for at least one of 2D vertices of a model of the body and joint configurations of the body.

[0020] According to one embodiment, the transformer is trained using a ground truth for rotation parameters for each of a plurality of parts of the body.

[0021] According to one embodiment, the cross-attention layer disregards attention of the representation of the respective version of the set of parameter values to itself and disregards attention of the respective version of the set of parameter values to the image features.

[0022] According to one embodiment, the self-attention layer disregards attention disregards attention of the respective version of the set of parameter values to the image features and disregards attention of the image features to the respective version of the set of parameter values.

[0023] According to one embodiment, the method comprises generating an initial version of the set of parameter values from the extracted image features and generating a sequence of successively updated versions of the set of parameter values by generating each updated version of the set of parameter values by updating a previous version of the set of parameter values in the sequence according to a result of modifying a representation of the previous version of the set of parameter values by one or more of the transformer units and mapping a result of the modifying of the representation of the previous version of the set of parameter values to a corresponding version of the set of parameter values (e.g. transforming a result of the modifying of the representation of the previous version of the set of parameter values to an update of the set of parameter values and adding the update to the previous version of the set of parameter values).

[0024] Modifying by multiple transformer units means (since they form a cascade) modifying the representation by one transformer unit, then modifying the result by the next etc.

[0025] According to one embodiment, the mapping comprises an operation by a linear layer (and, if the result of the linear layer is an update, an addition).

[0026] According to one embodiment, a data processing device configured to perform the method of any one of the embodiments described above is provided.

[0027] According to one embodiment, a computer program element is provided comprising program instructions, which, when executed by one or more processors, cause the one or more processors to perform the method according to any one of the embodiments described above.

[0028] According to one embodiment, a computer-readable medium is provided comprising program instructions, which, when executed by one or more processors, cause the one or more processors to perform the method according to any one of the embodiments described above.

[0029] It should be noted that embodiments described in context of the method for estimating the shape and pose of a body are analogously valid for the data processing device and vice versa.

BRIEF DESCRIPTION OF THE DRAWINGS

[0030] The invention will be better understood with reference to the detailed description when considered in conjunction with the non-limiting examples and the accompanying drawings, in which:

- FIG. 1 shows a human shape and pose estimation arrangement according to an embodiment.
- FIGs 2A to 2C illustrate different transformer designs for shape and pose estimation.
- FIG. 3 illustrates a machine learning model according to an embodiment.
- FIG. 4 shows a transformer block according to an embodiment.
- FIG. 5 shows a transformer unit according to an embodiment.
- FIG. 6 shows a flow diagram illustrating a method for estimating the shape and pose of a body.

DETAILED DESCRIPTION

[0031] The following detailed description refers to the accompanying drawings that show, by way of illustration, specific details and embodiments in which the disclosure may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the disclosure. Other embodiments may be utilized and structural, and logical changes may be made without departing from the scope of the disclosure. The various embodiments are not necessarily mutually exclusive, as some embodiments can be combined with one or more other embodiments to form new embodiments.

[0032] Embodiments described in the context of one of the devices or methods are analogously valid for the other devices or methods. Similarly, embodiments described in the context of a device are analogously valid for a vehicle or a method, and vice-versa.

[0033] Features that are described in the context of an embodiment may correspondingly be applicable to the same or similar features in the other embodiments. Features that are described in the context of an embodiment may correspondingly be applicable to the other embodiments, even if not explicitly described in these other embodiments. Furthermore, additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment may correspondingly be applicable to the same or similar feature in the other embodiments.

[0034] In the context of various embodiments, the articles “a”, “an” and “the” as used with regard to a feature or element include a reference to one or more of the features or elements.

[0035] As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0036] In the following, embodiments will be described in detail.

[0037] FIG. 1 shows a human shape and pose estimation arrangement 100 according to an embodiment.

[0038] A camera 101 takes an image 102 (e.g. an RGB image) of a human person 103. The image 102 is fed to a data processing device 104 (typically a computer). The data processing device 104 comprises a processor (e.g. CPU) 105, a system memory 106 and may comprise further data processing components like a hard-disk memory, various input/output (I/O) devices etc.

[0039] According to various embodiments, the data processing device 104 operates as a shape and pose estimator. For this, a machine learning model 117 is stored in the memory (in

the form of a part of a computer program) and the processor 105 performs a shape and pose estimation method using the machine learning model 117.

[0040] According to various embodiments, the data processing device 104 performs monocular 3D shape and pose estimation (of a human, but other objects, in particular animals but also robots or other mechanical movable devices, are also possible), i.e. estimates the shape and pose of the human person 103 from a single 2D image 102. This is a difficult task due to the inherent depth ambiguity of a single 2D image.

[0041] For this task a transformer (deep learning model) may be used for the machine learning model 117. In particular, a transformer may be used to estimate a 3D human mesh as it has powerful capabilities in representation learning and pattern recognition. Such a transformer may follow a design of a ViT (Vision transformer) as it is illustrated by FIG. 2A.

[0042] FIG. 2A illustrates a transformer design where the target embeddings are first concatenated with the input features and then processed by a full attention layer that models all pairwise dependencies including target-target, target-feature, feature-target, and feature-feature by corresponding attention, i.e. attention of representation to itself (i.e. of components of the representation to themselves), attention of components of the representation to the image features, attention of the image features to the representation and attention of the image features to themselves, respectively.

[0043] While a transformer of this design can achieve impressive results for 3D human shape and pose estimation, it leads to quadratic computation and memory complexity with respect to the image feature length F , i.e., $O((N + F)^2)$ as illustrated in FIG. 2A where N is the length of the target embedding. Thus, a transformer with this design can typically only employ low-resolution image features where F is fairly small, i.e. this design prevents the transformer from using higher-resolution features for more accurate 3D human recovery.

[0044] In view of the above, according to various embodiments, a machine learning model 117 is used for shape and pose estimation which has a transformer architecture which allows better exploitation of high-resolution image features for high-quality 3D body shape and pose reconstruction. Specifically, the machine learning model 117 includes the following two design characteristics (which are, according to various embodiments, used in combination but which may also be used independently from each other).

[0045] First, different from the ViT style transformer where the image features are learned by attention operations, the machine learning model 117 uses a feature extractor like a convolutional neural network (CNN) to extract image features from the image 102 and it uses attention operations only to aggregate the features to improve the target embeddings. Thus, it is less important to model feature-feature and feature-target correlation (as in design of FIG. 2A). Accordingly, according to various embodiments, the full attention operation is decoupled into a target-feature attention and a target-target attention that are cascaded to avoid the quadratic complexity.

[0046] FIG. 2B illustrates a transformer design with a decoupled attention according to various embodiments which leads to a computation and memory complexity of $O((N + F)N)$ which is linear with respect to the feature length F .

[0047] Second, while the decoupling strategy effectively improves the efficiency of the transformer compared to the ViT style, straightforwardly using it for 3D human reconstruction still leads to a considerably heavy cost of computation and memory, if a vertex-based target (i.e. human body) representation for the transformer is used where the embedding length N is often quite large, equalling the number of vertices on a body mesh. To address this issue, according to various embodiments, a representation based on a parametric body model (e.g. SMPL (Skinned Multi-Person Linear Model)) is used, where a target embedding only needs to be learned for a small number of body parts that account for 3D body part rotations, plus one query for human body shape. A parametric body model may in this context be understood as a model of a body which defines a shape and pose of the body by, in particular, rotations of body parts (e.g. limbs) with respect to each other.

[0048] FIG. 2B illustrates a transformer design with decoupled attention (according to FIG. 2B) and further using a SMPL representation which further lessens the computation and memory burden compared to the vertex representation with $Q \ll N$.

[0049] So, in combination, the above two design characteristics lead to a more concise attention learning process (compared to ViT style design) and in particular allows the machine learning model 117 to use high-resolution (image) features (i.e. a larger F) for high-quality 3D human mesh reconstruction. This means that by improving the attention formulation and the target representation transformer architecture is provided that can exploit higher-resolution features for accurate monocular 3D human shape and pose estimation.

[0050] According to various embodiments, an omni-scale transformer is used which effectively aggregates multi-resolution features as well as local image information.

[0051] FIG. 3 illustrates the processing pipeline of a machine learning model according to an embodiment.

[0052] The machine learning model for example corresponds to the machine learning model 117. It comprises an image feature extractor 301 (e.g. a CNN) for extracting image features 305 (of length F) from an input image 304 (e.g. corresponding to image 106), a transformer 302 which receives the image features 304 as input and outputs a shape and pose parameter values 306 (including parameter values describing the shape and pose of the body 103) and an SMPL conversion 303 which generates a human body mesh 307 from the shape and pose parameter values 306 as the output of the shape and pose estimation.

[0053] SMPL is a flexible and expressive human body model for 3D human shape and pose modelling. It is parameterized by a compact shape vector $\beta \in \mathbb{R}^{10}$ and a set of rotation parameters $\{\theta_i \in \mathbb{R}^3, i \in 1, 2, \dots, J\}$.

[0054] The body shape is represented by a low-dimensional space learned with Principal Component Analysis from a training set of 3D human scans, where β is the vector of coefficients of the principal basis vectors. The body pose is defined by a skeleton rig with $J = 24$ joints including the body root, and θ_i is the rotation of the i-th body part in the axis-angle form whose skew symmetric matrix lies in the Lie algebra space $\mathfrak{so}(3)$.

[0055] From $\{\theta_1, \theta_2, \dots, \theta_J\}$ and β the data processing device 104 can calculate a 3D body mesh

$$V = f_{\text{SMPL}}(\{\theta_1, \theta_2, \dots, \theta_J\}, \beta) \quad (1)$$

where $V \in \mathbb{R}^{N \times 3}$ is the vector of vertices of a triangulated surface with, according to one embodiment, $N = 6890$ and where f_{SMPL} is realized by the SMPL conversion 303.

[0056] From the body mesh, the data processing device 104 can estimate the 3D locations of the body joints X using a pre-trained linear mapping matrix $\mathcal{M} \in \mathbb{R}^{J \times N}$:

$$X \in \mathbb{R}^{J \times 3} = \mathcal{M}V. \quad (2)$$

[0057] With the 3D human joints, the data processing device 104 can further obtain the 2D joints using weak perspective projection. Denoting the camera parameters as $\mathcal{C} \in \mathbb{R}^3$ that include the scaling factor and the 2D principal-point translation in the projection process, the 2D keypoints can be formulated as:

$$\mathcal{J} \in \mathbb{R}^{J \times 2} = \Pi_{\mathcal{C}}(X) \quad (3)$$

where $\Pi_{\mathcal{C}}$ is the projection function.

[0058] The transformer comprises a sequence (or, in other words, a cascade) of transformer blocks 308 which each perform multiple attention operations, i.e. has multiple attention heads. An attention operation can be seen as the central component of a transformer and can be formulated as:

$$\tilde{\mathcal{Q}} = \text{Softmax} \left(\frac{(\mathcal{Q} \cdot W_q)(\mathcal{K} \cdot W_k)^\top}{\sqrt{d}} \right) (\mathcal{V} \cdot W_v) \quad (4)$$

where $\mathcal{Q} \in \mathbb{R}^{Q \times d}$ and $\mathcal{K}, \mathcal{V} \in \mathbb{R}^{K \times d}$ are the input of the attention operation, representing query, key, and value, respectively. $\tilde{\mathcal{Q}} \in \mathbb{R}^{Q \times d}$ is the output of the attention operation. Q is the length of the query $\mathcal{Q}$ and K is the length of $\mathcal{K}$ and $\mathcal{V}$.

[0059] The key and value are generally the same, while $\mathcal{Q}$ and $\mathcal{K}$ could be the same or different (correspondingly, the attention given in equation (4) is called self-attention or cross-attention). d denotes the size of the channel dimension. $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ represent learnable linear projection weights. It should be noted that while single-head attention is illustrated here for clarity, multi-head attention may be used by applying different linear projection weights to different attention heads.

[0060] Essentially, equation (4) can be seen to model dependencies between element pairs from $\mathcal{Q}$ and $\mathcal{K}$ with dot product and obtains an enhanced query embedding (i.e. an updated embedding (also referred to as “representation”) of the parameter set describing pose and shape) $\tilde{\mathcal{Q}}$ by aggregating information from $\mathcal{V}$, where the aggregation weights are decided according to the dependencies between $\mathcal{Q}$ and $\mathcal{K}$.

[0061] Recovering the 3D body mesh 307 by regressing the vertex locations V leads to a redundant target representation $[q_{v_1}, q_{v_2}, \dots, q_{v_N}]$ where $q_{v_i} \in \mathbb{R}^d$ is the embedding of the i -th vertex v_i . The length of this target embedding is often quite large ($N = 6890$ by default for the SMPL mesh), resulting in heavy computation and memory cost in attention operations even after mesh downsampling as illustrated in FIG. 2A. To address this issue, according to various embodiments, instead of a vertex-based target representation, the target representation of the transformer 302 is an SMPL-based representation and the transformer 306 thus estimates values of a (compared to the 3D mesh) smaller set of parameters $P = \{\theta_1, \theta_2, \dots, \theta_J, \beta, \mathcal{C}\}$ that describe the 3D human shape and pose as well as the camera as the shape and pose parameter values 306. Specifically, the SMPL-based target embedding (also referred to as (target) representation (of the parameter value set, e.g. SMPL parameter set, describing the shape of pose of the body) can be written as $T = [q_{\theta_1}, q_{\theta_2}, \dots, q_{\theta_J}, q_{\beta}, q_{\mathcal{C}}]$ which separately correspond to the variables in P . The length of this representation is much shorter than the vertex-based embedding, thereby facilitating the efficient design of the machine learning model (as illustrated in Fig. 2C. In addition, the target representation allows restricting the solution to lie in the SMPL body space, which naturally ensures smooth meshes, whereas the vertex representation lacks proper smoothing regularization and may lead to a spiky body surface.

[0062] The CNN 301 generates, from the input image 304 the image features 305 in the form of a multi-scale feature map $\Phi = \{\phi^s \in \mathbb{R}^{F^s \times d}, s = 1, 2, \dots, S\}$ where S is the number of scales, and F^s is the number of pixels in the feature map ϕ^s .

[0063] The lowest-scale feature ϕ^S is sent into a global average pooling layer 309 followed by a three-layer MLP 310 to obtain an initial estimation of the 3D human parameter P_0 which is progressively refined by the transformer blocks 308 (of number B). Specifically, the refinement process can be written as:

$$T_b = f_{\text{TB}}^{(b)}(T_{b-1}, P_{b-1}, \Phi), \quad (5)$$

$$P_b = f_{\text{linear}}(T_b) + P_{b-1}, \quad b = 1, 2, \dots, B, \quad (6)$$

where T_b and P_b are respectively the target embedding and the estimated parameters from the b -th transformer block $f_{\text{TB}}^{(b)}$. Here, f_{linear} is a linear layer 311 that transforms the embedding into the desired residual output. To bootstrap the refinement process, the initial target embedding T_0 is generated by combining P_0 and the image feature ϕ^S : $T_0 = f_{\text{GP}}(\phi^S) + f_{\text{linear}}(P_0)$ where f_{GP} is the global pooling function.

[0064] FIG. 4 shows a transformer block 400 according to an embodiment (which may be used for each of the transformer blocks 308).

[0065] The transformer block mainly consists of U cascaded transformer units 401, which improve the respective target embedding T with image features Φ according to equation (4).

[0066] A way to design such a transformer unit 401 is to set each of $\mathcal{Q}, \mathcal{K}, \mathcal{V}$ in equation (4) as a concatenation of T and Φ . However, as illustrated in FIG. 2A, this way models the full dependencies between target-target, target-feature, feature-target, and feature-feature, which leads to quadratic complexity with respect to the feature length, hindering the utilization of high-resolution image features for high-quality 3D human mesh recovery.

[0067] Therefore, according to various embodiments, a decoupled attention is used by combining a cross-attention layer and a self-attention layer as illustrated in FIG. 2B. The cross-attention uses T as the query and Φ as the key and value, which transfers information from the image to the target embedding. The self-attention uses T for all of $\mathcal{Q}, \mathcal{K}, \mathcal{V}$, modelling the interactions among embeddings of different body parts q_{θ_i} and body shape q_{β} , and camera $q_{\mathcal{C}}$. It should be noted that as the features Φ are extracted with a strong feature extractor 301, e.g. a deep CNN, it is less important to model feature-feature dependencies that are usually used for improving the image feature representation. This decoupling design achieves linear-complexity attention regarding to the feature length, allowing our transformer to use high-resolution features, e.g., ϕ^1 , for accurate 3D human reconstruction.

[0068] As mentioned above, according to various embodiments, the features 305 are multi-scale features. A straightforward strategy for the incorporation of multi-scale-features is to combine all-scale features of Φ into a single feature vector and then use this combined feature

as $\mathcal{K}$ and $\mathcal{V}$ in equation (4). However, this strategy uses the same projection weights for different scale features and thereby models target-feature dependencies in the same subspace, which ignores the characteristics of different scales and is less flexible to fully exploit the abundant multi-scale information.

[0069] To address this issue, according to various embodiments, an Omni-scale transformer unit (OTU) is used.

[0070] FIG. 5 shows an Omni-Scale transformer unit (OTU) 500 according to an embodiment (that may be used as the transformer units 401). Specifically, it shows the u-th transformer unit 401 of the b-th transformer block 308.

[0071] The OTU employs different projection weights for different scales by having a cross-attention layer having a cross-attention module 501 for each scale s (i.e. having a component or cross-attention (sub-)layer for each scale s ; these components are also referred to as cross-attention units). Specifically, the embeddings for the body shape β and the camera $\mathcal{C}$ can be obtained by $\hat{q}_\beta = h(q_\beta, \Phi)$ and $\hat{q}_\mathcal{C} = h(q_\mathcal{C}, \Phi)$ where h averages the attention outputs of different scales:

$$h(q, \Phi) = \frac{1}{S} \sum_{s=1}^S \text{Softmax} \left(\frac{(qW_q^s) \cdot (\phi^s W_k^s)^\top}{\sqrt{d}} \right) (\phi^s W_v^s). \quad (7)$$

Here, W_q^s, W_k^s, W_v^s are the linear weights of the query, key, and value for scale s , respectively.

[0072] In addition, for estimating the rotation of each body part θ_i , the OTU 500 includes a local attention module 502 which can be seen to be mainly based on the observation that the local articulation status around human joints strongly implies the relative rotation between neighbouring body parts and thereby it can be beneficial to adequately focus on local image features in the attention operation. Therefore, according to various embodiments, the embedding for θ_i is calculated by combining the multi-scale global attention in equation (7) and the local attention:

$$\hat{q}_{\theta_i} = h(q_{\theta_i}, \Phi) + \text{Softmax} \left(\frac{(q_{\theta_i} W_q) \cdot (\phi_{\mathcal{N}(\mathcal{J}_i)}^1 W_k)^\top}{\sqrt{d}} + \beta \right) (\phi_{\mathcal{N}(\mathcal{J}_i)}^1 W_v), \quad (8)$$

where $\mathcal{N}(\mathcal{J}_i)$ denotes an image patch (i.e. a patch of image features) around the i -th joint $\mathcal{J}_i$ with size $r \times r$ and $\phi_{\mathcal{N}(\mathcal{J}_i)}^1 \in \mathbb{R}^{r^2 \times d}$ represents the local features sampled at $\mathcal{N}(\mathcal{J}_i)$ from ϕ^1 . The 2D joint $\mathcal{J}_i$ can be obtained by Eq. 3 with the current parameter estimation P . It should be noted that use the highest-resolution feature ϕ_1 is used for the local attention here, as $\mathcal{N}(\mathcal{J}_i)$ can cover a large portion of smaller feature maps such that local attention on lower-resolution features becomes similar to global attention.

[0073] Finally, the output of the 500 OTU is generated by a self-attention layer 503 (comprising at least one self-attention unit) to model the interactions among $[\hat{q}_{\theta_1}, \dots, \hat{q}_{\theta_J}, \hat{q}_\beta, \hat{q}_c]$. It should be noted that the proposed OTU 500 realizes the design illustrated by in FIG. 2B of decoupling the full attention into cascaded cross-attention and self-attention. In the example of FIG. 5, this happens in an omni-scale manner, unifying multi-scale, global, and local image information.

[0074] As the attention operation itself in equation (4) is position agnostic, positional encoding may be used to inject the position information of the input. According to various embodiments, a learnable positional encoding is used for both the target embedding and the image feature. These are realized by a T-PE block 402 and F-PE block 403 (see FIG. 4).

[0075] In particular, the positional encoding of image features can be written as

$\{\delta_s \in \mathbb{R}^{F_s \times d}, s = 1, 2, \dots, S\}$ which are directly added to Φ before the transformer units 401. The positional encoding may learn δ_s for all scales. However, this not only results in excessive parameters, but also ignores the relationship between the locations across different scales, e.g., similar spatial locations at different scales should have similar positional embeddings. To address this issue, according to various embodiments, the positional embedding is only learned for the highest scale, i.e., δ_1 , and the embeddings for the other scales

are produced by average pooling of δ_1 : $\delta_s = f_{\text{pool}}^{(2^{s-1})}(\delta_1)$, where $f_{\text{pool}}^{(2^{s-1})}$ is the average pooling layer with a stride and window size of 2^{s-1} .

[0076] For training the machine-learning model, the following loss function may be used:

$$\ell_{\text{basic}} = w_V \cdot \|V^B - V^{\text{GT}}\|_1 + w_X \cdot \|X^B - X^{\text{GT}}\|_2^2 + w_{\mathcal{J}} \cdot \|\mathcal{J}^B - \mathcal{J}^{\text{GT}}\|_2^2, \quad (9)$$

where $V^B, X^B, \mathcal{J}^B$ are the predicted vertices, 3D joints, and 2D joints that are computed with equation (3) using the final output 306, i.e. P_B , of the transformer 302. “GT” indicates the corresponding ground truth. $w_V, w_X, w_{\mathcal{J}}$ are hyper-parameters for balancing different terms.

[0077] While equation (9) works reasonably well for vertex regression based methods, it may be insufficient for a regression output 306 based on 3D rotations of body parts and cannot be well trained without proper rotation regularization. To address this issue, according to various embodiments, a rotation loss is further introduced:

$$\ell_{\text{rotation}} = w_R \cdot \frac{1}{J} \sum_{i=1}^J \|R(\theta_i^B) - R(\theta_i^{\text{GT}})\|_1, \quad (10)$$

where the predicted rotation θ_i^B from P_B is encouraged to be close to the ground truth θ_i^{GT} . w_R is also a hyper-parameter. Since the axis-angle representation is not unique with respect to rotation, e.g., increasing the angle by 2π gives the same rotation, the axis-angle may be transformed into a rotation matrix using the Rodrigues’ formula R , which computes the exponential mapping from Lie algebra $\mathfrak{so}(3)$ to Lie group $\text{SO}(3)$ (i.e. the 3D rotation group).

[0078] So, according to various embodiments, the machine learning model is trained with a training loss comprising a combination of (e.g. a sum of) ℓ_{basic} and ℓ_{rotation} . It should be noted that the CNN 301 and the transformer 302 may be trained together with this loss or the CNN 301 may be pre-trained and the transformer 302 may then be trained using this loss.

[0079] In summary, according to various embodiments, a method for estimating the shape and pose of a body is provided as illustrated in 6.

[0080] FIG. 6 shows a flow diagram 600 illustrating a method for estimating the shape and pose of a body.

[0081] In 601, an image of the body is obtained.

[0082] In 602, image features are extracted from the image.

[0083] In 603, the image features are transformed to a set of parameter values describing the shape and pose of the body by means of a transformer comprising a cascade of transformer units, wherein each transformer unit operates on a representation of a respective version of the set of parameter values and comprises a cross-attention layer modifying (e.g. replacing) the representation of the respective version of the set of parameter values by the image features weighted by an attention of the respective version of the set of parameter values to the image features in sequence with a self-attention layer modifying (e.g. replacing) the representation of the respective version of the set of parameter values by components of the representation of the respective version of the set of parameter values weighted by an attention between the components of the representation of the respective version of the set of parameter values to themselves (i.e. in particular each of the components to one or more of the other components).

[0084] According to various embodiments, in other words, shape and pose estimation is performed according to cross-attention from target representation to image features decoupled from self-attention from target representation to target representation (as illustrated by FIG. 2B and FIG. 2C) rather than a “full” attention unit which includes attention from image features to target representation and attention from image features to image features (as illustrated in FIG. 2A).

[0085] Modifying in sequence by means that one of the attention layers modifies the representation and the other attention layer further modifies the representation (depending on the order in which they are arranged in the sequence). The further modification may depend on the result of the earlier modification (i.e. the attention of the further modification is then based on the representation resulting from the earlier modification).

[0086] The estimated shape and pose may for example be used in for visual tracking, virtual/augmented reality, motion generation, image manipulation, and neural radiance fields.

[0087] The method of FIG. 6 is for example carried out by the data processing device 104 of the arrangement of FIG. 1.

[0088] The methods described herein may be performed and the various processing or computation units and the devices and computing entities described herein may be

implemented by one or more circuits. In an embodiment, a "circuit" may be understood as any kind of a logic implementing entity, which may be hardware, software, firmware, or any combination thereof. Thus, in an embodiment, a "circuit" may be a hard-wired logic circuit or a programmable logic circuit such as a programmable processor, e.g. a microprocessor. A "circuit" may also be software being implemented or executed by a processor, e.g. any kind of computer program, e.g. a computer program using a virtual machine code. Any other kind of implementation of the respective functions which are described herein may also be understood as a "circuit" in accordance with an alternative embodiment.

[0089] While the disclosure has been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the invention as defined by the appended claims. The scope of the invention is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.

CLAIMS

1. A method for estimating the shape and pose of a body comprising:
Obtaining an image of the body;
Extracting image features from the image;
Transforming the image features to a set of parameter values describing the shape and pose of the body by means of a transformer comprising a cascade of transformer units,
wherein each transformer unit operates on a representation of a respective version of the set of parameter values and comprises
a cross-attention layer modifying the representation of the respective version of the set of parameter values by the image features weighted by an attention of the respective version of the set of parameter values to the image features;
in sequence with
a self-attention layer modifying the representation of the respective version of the set of parameter values by components of the representation of the respective version of the set of parameter values weighted by an attention between the components of the representation of the respective version of the set of parameter values to themselves.
2. The method of claim 1, wherein the self-attention layer comprises at least one self-attention unit performing respective attention operation with query, key and value, for which each of the query, the key and the value are set to, up to a modification by positional encoding, the representation of the respective version of the set of parameter values.
3. The method of claim 1 or 2, wherein the cross-attention layer comprises at least one cross-attention unit performing respective attention operation with query, key and value, for which the query is set to, up to a modification by positional encoding, the representation of the respective version of the set of parameter values and the key and the value are set to, up to a modification by positional encoding, the image features or a subset of the image features.

4. The method of any one of claims 1 to 3, wherein the image features comprise image features for each of a plurality of scales and wherein the cross-attention layer comprises a cross-attention unit for each layer modifying the representation of the respective version of the set of parameter values by an attention from the representation to image features of the scale.
5. The method of claim 4, wherein the cross-attention layer outputs a modified version of the representation of the set of parameter values which it generates by averaging results of the modifying performed by the cross-attention units for the scales.
6. The method of any one of claims 1 to 5, wherein the set of parameters comprises values of parameters of a parametric body model.
7. The method of any one of claims 1 to 6, wherein the method further comprises generating a mesh representation of the body from the values of the parameters of the parametric body model.
8. The method of any one of claims 1 to 7, wherein the set of parameter values comprises values for a set of parameters including rotation parameters for each of a plurality of parts of the body.
9. The method of any one of claims 1 to 8, wherein the cross-attention layer comprises a local attention unit modifying the representation of the respective version of the set of parameter values by, for each of the rotation parameters, the image features of a patch showing the body part weighted by an attention of the respective version of the set of parameter values to the image features of the patch showing the body part.
10. The method of claim 9, wherein the cross-attention layer outputs a modified version of the representation of the set of parameter values which it generates by summing a result of the modifying performed by the local attention unit and a result of the

modifying of the representation of the respective version of the set of parameter values by the image features weighted by an attention of the respective version of the set of parameter values to the image features.

11. The method of any one of claims 1 to 10, wherein the set of parameter values comprises values for a set of shape parameters.
12. The method of any one of claims 1 to 11, wherein the image is obtained by means of a camera and the set of parameter values comprises values for a set of camera parameters.
13. The method of any one of claims 1 to 12, comprising extracting the image features by means of a convolutional neural network.
14. The method of any one of claims 1 to 13, comprising training the transformer by supervised learning.
15. The method of claim 14, wherein the transformer is trained using a ground truth for at least one of 2D vertices of a model of the body and joint configurations of the body.
16. The method of claim 14 or 15, wherein the transformer is trained using a ground truth for rotation parameters for each of a plurality of parts of the body.
17. The method of any one of claims 1 to 16, wherein the cross-attention layer disregards attention of the representation of the respective version of the set of parameter values to itself and disregards attention of the respective version of the set of parameter values to the image features.
18. The method of any one of claims 1 to 17, wherein the self-attention layer disregards attention disregards attention of the respective version of the set of parameter values to the image features and disregards attention of the image features to the respective version of the set of parameter values.

19. The method of any one of claims 1 to 18, comprising generating an initial version of the set of parameter values from the extracted image features and generating a sequence of successively updated versions of the set of parameter values by generating each updated version of the set of parameter values by updating a previous version of the set of parameter values in the sequence according to a result of modifying a representation of the previous version of the set of parameter values by one or more of the transformer units and mapping a result of the modifying of the representation of the previous version of the set of parameter values to a corresponding version of the set of parameter values.
20. The method of claim 19, wherein the mapping is an operation by a linear layer.
21. A data processing device configured to perform the method of any one of claims 1 to 20.
22. A computer program element comprising program instructions, which, when executed by one or more processors, cause the one or more processors to perform the method of any one of claims 1 to 20.
23. A computer-readable medium comprising program instructions, which, when executed by one or more processors, cause the one or more processors to perform the method of any one of claims 1 to 20.