A Kendall Shape Space Approach to 3D Shape Estimation from 2D Landmarks

Martha Paskin[1], Daniel Baum[1], Mason N. Dean[2], and
Christoph von Tycowicz[1,3]

[1] Zuse Institute Berlin, Germany {paskin,baum,vontycowicz}@zib.de
[2] City University of Hong Kong, China mndean@cityu.edu.hk
[3] Freie Universität Berlin, Germany

Abstract. 3D shapes provide substantially more information than 2D images. However, the acquisition of 3D shapes is sometimes very difficult or even impossible in comparison with acquiring 2D images, making it necessary to derive the 3D shape from 2D images. Although this is, in general, a mathematically ill-posed problem, it might be solved by constraining the problem formulation using prior information. Here, we present a new approach based on Kendall's shape space to reconstruct 3D shapes from single monocular 2D images. The work is motivated by an application to study the feeding behavior of the basking shark, an endangered species whose massive size and mobility render 3D shape data nearly impossible to obtain, hampering understanding of their feeding behaviors and ecology. 2D images of these animals in feeding position, however, are readily available. We compare our approach with state-of-the-art shape-based approaches, both on human stick models and on shark head skeletons. Using a small set of training shapes, we show that the Kendall shape space approach is substantially more robust than previous methods and results in plausible shapes. This is essential for the motivating application in which specimens are rare and therefore only few training shapes are available.

Keywords: 3D shape estimation, 2D-to-3D, shape space approach, Kendall's shape space, sparse data.

1 Introduction

The reconstruction of a 3D shape from 2D images is an important problem of computer vision that has been the subject of research for decades [34,6]. It is a widespread challenge with sample applications including motion recognition and tracking, autonomous driving, as well as robot-assisted surgery and navigation. Generally, estimating the 3D shape of an object from single monocular 2D images is an ill-posed problem in the sense of Hadamard. To tackle this problem, several approaches have been proposed, the choice of which greatly depends on the type and amount of data, and the application.

One common approach is to use a series of 2D images [15,7] or multiple cameras [19,16,24]. This is used, for example, in photogrammetry. In order to calculate the depth, corresponding points in 2D images need to be identified [29,7,1].

Another approach is based on physical modeling of the object to be estimated [32,2]. Here, the object is described as a kinematic tree consisting of segments connected by joints. Each joint contains some degrees of freedom (DOF), indicating the directions it is allowed to move. The object is modeled by the DOF of all joints. Length constraints can also be added, such as limb lengths constraints in a human pose. This helps reduce the depth ambiguity by constraining the relative position of the joints.

In lieu of additional viewpoints or physical models, another line of work regularizes the estimation problem by employing prior knowledge learned from examples of the respective object class. Some works in this category are based on the concept of shape spaces. Here, the 3D shape of an object is estimated by interpolating through a set of pre-defined 3D shapes, called training shapes, that represent possible deformations of the object. Given a 2D projection of the object, the aim is to find the ideal camera parameters and the shape coefficients describing the interpolation. Most commonly, *active shape model* (ASM) techniques are used [23,14,30,35,38,37]. ASM has been used intensively for the purpose of recognizing and locating non-rigid objects in the presence of noise and occlusion [13,4]. It aims at providing a robust approach, which accommodates shape variability by arguing that when the objects deform in ways characteristic to the class of objects they represent, the method should be able to recognize them. Here, the *shape* of a class of objects is described by a set of labeled *landmark* points, each representing a particular part of the object or its boundary. Given a set of pre-defined *basis shapes* of an object, the assumption is that any shape of the object can be described by linearly interpolating between basis shapes. The legitimacy of the model is determined by the set of training shapes, which needs to include the variability in the shape of the object.

Apart from these more classical approaches, recent years have mostly been devoted to the development of deep learning-based methods that exploit domain-specific data repositories with abundant samples (e.g., *ShapeNet, Human3.6M*). In 2017, a kinematic skeleton fitting method using a *convolutional neural network* (CNN) was proposed that regresses 2D and 3D joint positions jointly in real-time [21]. More recently, *Procrustean regression networks* (PRN) were proposed [26]. Using a novel loss function, they solve the problem of automatically determining the rotation. Another recent deep learning-based method, *KShapeNet* [11], makes use of Kendall's shape space by modeling shape sequences first as trajectories on Kendall's non-linear shape space, which are then mapped to the linear tangent space. The authors report [11] that the non-linear motion can be captured better using their approach than using previous geometric deep-learning methods.

In general, the training of deep neural network models requires a considerable amount of data. While the generalization of learned models to new classes with few available examples has become a highly popular research topic (see e.g. [22]), such base models are rarely available in many applications. There is therefore a strong impetus to develop so-called shallow learning approaches that can be applied even if few-shot learning is not an option.
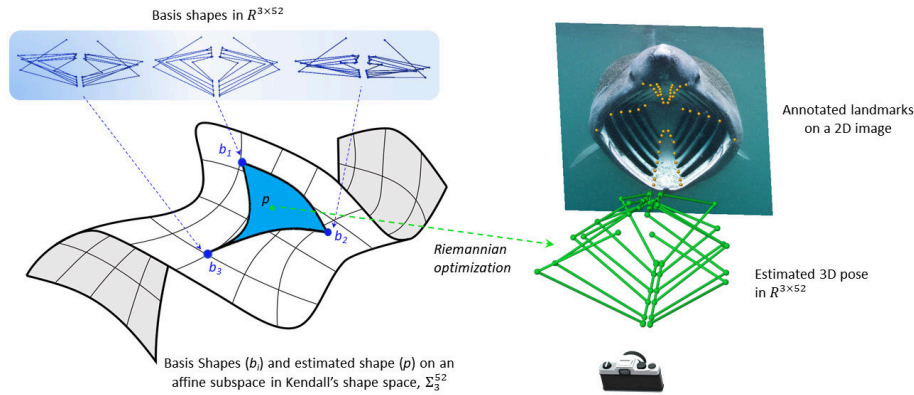
Basis shapes in $R^{3\times52}$

Annotated landmarks
on a 2D image

b_1

p

b_2

b_3

Riemannian
optimization

Estimated 3D pose
in $R^{3\times52}$

Basis Shapes (b_i) and estimated shape (p) on an
affine subspace in Kendall's shape space, Σ_3^{52}

Fig. 1: Methodology diagram for the proposed method.

Our work contributes to this area by deriving a novel shape space-based approach. We demonstrate the effectiveness of the proposed scheme in an application to the study of marine animal behavior. In this application, we are faced with the retrieval of the 3D shape of a non-rigid (biological) object the head skeleton of a basking shark from single monocular 2D images. Basking sharks (*Cetorhinus maximus*) are massive animals and thought to be one of the most efficient filter-feeding fish in terms of the throughput of water filtered through their gill rakers [31]. Details about the underlying morphology of the basking shark filter apparatus and particularly how it is configured when the animal feeds have not been studied due to various challenges in acquiring real-world data. In this study, we address this issue, using several computed-tomography (CT) scans to create skeletal representations of basking shark head structure. From these, and known aspects of the biology (e.g. shark cranial anatomy, feeding behaviors), we semi-automatically created plausible configurations of the skeleton which we use as training shapes to describe the shape space. The ultimate goal is, given a photograph of a basking shark in feeding position, to estimate a plausible 3D shape of the skeletal representation from the training data and annotations of the skeletal joints in the form of 2D landmarks.

Since the number of training shapes is very limited in our application, deep learning-based methods cannot be applied. When using only a single 2D image, photogrammetry-based methods also cannot be applied. And for a physical modeling-based approach, information about the movement of the skeletal elements that is supported by the joints would be necessary. This information is not available. We therefore resort to shape space-based methods. In particular, we develop a new *Kendall shape space* (KSS) approach for the reconstruction of 3D shape from 2D landmarks. Our overall workflow is sketched in Fig. 1. In order to evaluate our approach, we compare it qualitatively and quantitatively with state-of-the-art ASM approaches [37] on both human poses and basking shark skeletons. Due to the limited number of training shapes for the basking

sharks, we are mainly interested in the performance of the approaches w.r.t. a small number of training shapes.

The results of our experiments suggest that our method is substantially more robust than previous shape space-based methods. This is particularly true for real world data both of human beings and shark data. It is also computationally very efficient and provides plausible interpolations without auxiliary regularization since the estimated shape will always lie on a subspace in Kendall's shape space spanned by the training data.

The source code of the proposed method together with the derived basking shark shapes have been made publicly available [27].

2   Methods

In this work, we approach the problem of estimating the shape of 3D objects from single images. To this end, we employ a landmark-based representation, viz. a set of ordered 3D points $(x_i)_{i=1,\ldots,k}$ that correspond to homologous points (e.g. joint centers in a skeleton) common to all subjects belonging to the object class under study. By assuming the weak perspective camera model, the projection or forward map is given by the linear system

$$W = \Pi R X + t \mathbf{1}_k^T, \tag{1}$$

where $X \in \mathbb{R}^{3 \times k}$ and $W \in \mathbb{R}^{2 \times k}$ denote the matrices of stacked landmarks and their 2D projections, respectively. The rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^2$ encode the camera coordinate system while the projection is carried out by the matrix

$$\Pi = \begin{pmatrix} \alpha \ 0 \ 0 \\ 0 \ \alpha \ 0 \end{pmatrix},$$

where the scalar α depends only on the focal length and distance to the object. The inverse problem of finding the 3D configuration corresponding to given projections is commonly formulated in terms of a least-squares problem. However, the underdeterminedness renders such a problem ill-posed requiring inclusion of prior knowledge in order to obtain feasible solutions. An established strategy based on the active shape model [9] is to restrict the search space to linear combinations of basis shapes $B_1, \ldots, B_n \in \mathbb{R}^{3 \times k}$, i.e.

$$X = \sum_{j=1}^n c_j B_j, \tag{2}$$

with weights c_j satisfying the constraint $\sum_j c_j = 1$. The quality of this approach, however, depends on the validity of the assumption that instances of the object class are well approximated by a hyper-planar manifold. In general, the shape of an object is considered to be invariant under rotation, translation and scaling and, hence, takes values in a nonlinear space with quotient structure. In particular, for landmark-based representations, this concept gives rise to the well-known Kendall shape space [17]. The non-Euclidean nature of shape space not only violates the linearity assumption but furthermore prevents direct application of

algebraic expressions such as Eq. (2) due to the lack of a global vector space structure. Nevertheless, the framework of (computational) differential geometry provides us with a rich toolset to generalize the ASM approach for use within the reconstruction problem.

2.1   Geometric model

Before we formulate a manifold version of the least-squares estimation, we first recall essential notions and properties from Kendall's shape space. We refer the reader to the comprehensive work [18] for an in-depth introduction.

Let $X \in \mathbb{R}^{m \times k}$ be the matrix holding k landmarks $x_i \in \mathbb{R}^m$ as columns. We remove the effect of translations and scaling by subtracting the row-wise mean from each landmark and then dividing X by its *Frobenius norm* (denoted $\| \cdot \|$). The set of all such configurations constitutes the *pre-shape space*

$$\mathcal{S}_m^k = \{X \in \mathbb{R}^{m \times k} : \textstyle\sum_{i=1}^k x_i = 0 \,,\, \|x\| = 1\}.$$

Now, the left action of $SO(m)$ on $\mathcal{S}_m^k$ given by $(R, X) \mapsto RX$ defines an equivalence relation given by $X \sim Y$ if and only if $Y = RX$ for some $R \in SO(m)$. Kendall's shape space is defined as the quotient $\Sigma_m^k = \mathcal{S}_m^k / \sim$. Now, denoting the canonical projection of $\sim$ by π and the spherical distance by $d_{\mathcal{S}}(X, Y) = \arccos \langle X, Y \rangle$, the induced distance between any two shapes $\pi(X)$ and $\pi(Y)$ is given by

$$d_{\Sigma}(X, Y) := \min_{R \in SO(m)} d_{\mathcal{S}}(X, RY) = \arccos \textstyle\sum_{i=1}^n \lambda_i, \qquad (3)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m-1} \geq |\lambda_m|$ are the pseudo-singular values of YX^T [25]. The points X and Y are said to be *well-positioned* (denoted by $X \overset{w}{\sim} Y$), if and only if, YX^T is symmetric and $d_{\mathcal{S}}(X, Y) = d_{\Sigma}(X, Y)$. Note that for $m \geq 3$, the shape space Σ_m^k contains singularities making it a strata of manifolds with varying dimensions.

In order to generalize the ASM approach, we first note that Eq. (2) effectively parametrizes the affine subspace spanned by the basis shapes. A common approach to obtain equivalent notions in Riemannian manifolds is to define *geodesic subspaces* that are spanned by generalized straight lines (geodesics) emanating from a reference point and with directions restricted to a linear subspace of the tangent space [10]. However, as geodesic subspaces rely on tangent vectors, they can only be defined within regular strata of the shape space. Therefore, we opt for an alternative construction by Pennec [28] called *barycentric subspaces*, which are defined as the locus of points that are weighted means of the basis shapes. Specifically, let $b_1 = \pi(B_1), \ldots, b_n = \pi(B_n) \in \Sigma_m^k$ be a set of n distinct basis shapes, then the *Fréchet barycentric subspace* is given by

$$FBS_{(b_j)_j} = \left\{ \arg\min_{p \in \Sigma_m^k} \sum_{j=1}^n c_j \, d_{\Sigma}^2(p, b_j) : \sum_{j=1}^n c_j \neq 0 \right\}.$$

With a notion of affine submanifold at hand, we next derive a reprojection error that is defined on Kendall's shape space. Consistently to the 3D case, we will treat the 2D projections as points in shape space Σ_2^k and employ the geodesic distance as objective for the reconstruction problem. To this end, we re-formulate the projection (1) as a map from 3D KSS to 2D KSS by

$$p \in \Sigma_3^k \mapsto \pi \circ \nu(R_{[1,2]}\pi^\dagger(p)) \in \Sigma_2^k, \tag{4}$$

where ν performs normalization (i.e. projects onto the sphere) and the subscript $[1,2]$ refers to the first two rows of $R \in SO(3)$. To assure class invariance under $\sim$, we introduce the map $\pi^\dagger$, which yields a unique representative for a shape by well-positioning to a fixed reference pre-shape $X \in \mathcal{S}_3^k$, viz.

$$\pi^\dagger : p \in \Sigma_m^k \mapsto P \in \mathcal{S}_m^k \quad s.t. \ \pi(P) = p, P \overset{w}{\sim} X. \tag{5}$$

Note that due to the quotient structure, the re-formulation no longer depends on the scaling and translation camera parameters α and t found in the original formulation (1).

We are now in the position to obtain a geometry-aware counterpart of the estimation problem in terms of the projection (4) and the Fréchet barycentric subspace resulting in the following optimization problem

$$\min_{p,R} d_\Sigma^2\left(W, \nu(R_{[1,2]}\pi^\dagger(p))\right) \quad s.t. \ p \in FBS_{(b_j)_j}, R \in SO(3). \tag{6}$$

2.2 Algorithmic treatment

In order to solve the reconstruction problem (6), we employ an alternating optimization strategy taking advantage of the product structure of the search space, viz. $FBS_{(B_j)_j} \times SO(3)$. Each outer iteration thus consists of two steps: First, rotation R is minimized while considering the current shape estimate p fixed. To this end, we perform Riemannian optimization on the *Stiefel* manifold

$$V_2(\mathbb{R}^3) := \{X \in \mathbb{R}^{3\times 2} : X^T X = I_2\} \cong SO(3)$$

utilizing the `Pymanopt` [5] library. Second, we fix the rotation R and solve the resulting sub-problem w.r.t. shape p parametrized by barycentric weights. For Riemannian computations on Kendall's space space, we employ `Morphomatics` [3]. As both steps are guaranteed to weakly decrease the reconstruction objective, we obtain a simple, yet convergent optimization scheme. In particular, both sub-problems are solved employing steepest descent strategies, where gradients are computed via automatic differentiation with `JAX` [12].

Due to the non-Euclidean structure of the shape space, there are no closed-form expressions for determining the (weighted) Fréchet mean of a given collection of sample points. Here, we adopt an efficient, inductive estimator to derive a parameterization of the Fréchet barycentric subspace that exhibits only a linear computational cost in the number of basis shapes. In particular, let

$b_1, \ldots, b_n \in \Sigma_3^k$ and $w_1, \ldots, w_n \in \mathbb{R}$ such that $\sum_{j=1}^n w_j \neq 0$, then the parameterization $(w_j)_j \mapsto \mu_n \in \Sigma_3^k$ is given by the recursion

$$\mu_1 = b_1 \qquad \mu_j = \gamma_{\mu_{j-1}}^{b_j} \left(\frac{w_j}{\sum_{l=1}^j w_l} \right), \qquad (7)$$

where $\gamma_{\mu_{j-1}}^{b_j}$ denotes the shortest geodesic from μ_{j-1} to b_j. Note that under certain conditions (cf. [8]) the recursion can be shown to converge to the actual weighted Fréchet mean. In any case, we expect the Fréchet barycentric subspace to be contained in the image of the above parameterization.

3   Experiments

We perform qualitative and quantitative evaluation of the proposed Kendall shape space (*KSS*) approach in experiments concerning the estimation of human poses as well as head skeletons of basking sharks. Throughout the experiments, we additionally provide results for the vanilla ASM approach as well as a recent convex relaxation thereof presented by Zhou et al. [37]. We will refer to the former and latter as *ASM non-convex* and *ASM convex*, respectively.

3.1   Human Pose Estimation

The design of this experiment closely follows the setup proposed in [37] in order to provide comparability. In particular, we employ the MoCap dataset[4] for defining training and test shapes. Sequences from subject 86 are used as training data and sequences from subjects 13, 14 and 15 are used as testing data. All sequences are composed of a variety of poses, including running, climbing, jumping, and dancing, and can hence be used to define a comprehensive set of basis shapes. While 41 landmarks are used to encode human poses in the MoCap database, we use a reduced representation comprising a subset of 15 landmarks mimicking the construction in [37]. To reduce the effects of rotation and translation as much as possible for the Euclidean approaches, we perform generalized Procrustes analysis of the data during pre-processing.

As this dataset features abundant training data, we apply a strategy to learn an appropriate set of a restricted number of basis shapes. An immediate choice to determine such a set would be to perform barycentric subspace analysis [28], thus finding an optimal subspace w.r.t. unexplained variance. However, as this could introduce a positive bias for our approach when comparing to the Euclidean ones, we opt for the established k-means clustering [20]. Specifically, we set the basis shapes as the arithmetic means of each determined cluster thereby providing a homogeneous sampling according to the Euclidean metric.

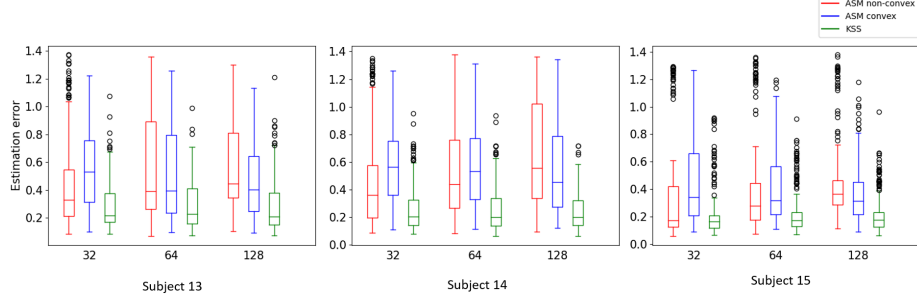[4] Carnegie Mellon University - Graphics Lab - motion capture library http://mocap.cs.cmu.edu/

Fig. 2: Boxplots of shape estimation errors of human pose test shapes using the three approaches.

Quantitative Testing Similar to the creation of basis shapes, 200 shapes were chosen for each of the test subjects by performing k-means clustering on the pose sequences. The shapes were then projected along the z-direction to create the 2D landmarks. We then estimated the 3D shapes from the 2D landmarks for all approaches with different numbers of basis shapes (32, 64, and 128) and computed for each reconstruction the estimation error as the Procrustes distance (3) between the true and estimated shapes. Fig. 2 shows boxplots of estimation errors obtained for the derived test shapes for the ASM non-convex, ASM convex, and our approach (listed as KSS), respectively. Table 1 gives an overview of the error distribution. Statistical Friedman tests were conducted for the different combinations of test subjects and number of basis shapes to examine the effect that the reconstruction methods had on estimation errors. Results showed that the methods lead to statistically significant differences in estimation errors ($p < 0.001$). The Kendall shape space approach out-performs the other approaches by estimating the unknown 3D shapes with consistently lower errors for all sets of basis shapes and test subjects. The accuracy of the approach is directly proportional to the number of basis shapes used.

Table 1: Mean(variance) estimation errors for human pose test shapes.

|	#Basis shapes	32	64	128
	ASM NC	0.462(0.121)	0.548(0.142)	0.571(0.104)
Subject 13	ASM Conv	0.550(0.071)	0.513(0.100)	0.461(0.059)
	KSS	**0.295(0.031)**	**0.295(0.031)**	**0.288(0.038)**
	ASM NC	0.475(0.129)	0.544(0.127)	0.644(0.128)
Subject 14	ASM Conv	0.583(0.066)	0.566(0.077)	0.533(0.087)
	KSS	**0.267(0.029)**	**0.258(0.028)**	**0.242(0.019)**
	ASM NC	0.358(0.136)	0.406(0.122)	0.457(0.089)
Subject 15	ASM Conv	0.450(0.092)	0.416(0.069)	0.366(0.043)
	KSS	**0.221(0.034)**	**0.231(0.027)**	**0.221(0.021)**
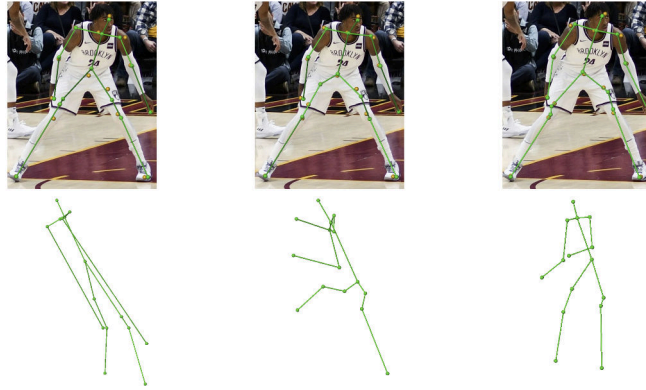
Fig. 3: Comparison of the three approaches for 3D shape estimation applied to an image of a basket ball player. Top row: Fitting of the estimated shapes (green) to the landmarks (yellow). Bottom row: side views of the estimated shapes. Left: ASM non-convex approach. Middle: ASM convex approach. Right: KSS approach (ours).

Qualitative Testing Next, we investigate the performance of the approaches in estimating the 3D pose of people in real-world images with manually labelled landmarks. Fig. 3 shows the estimated 3D poses of a basket ball player from different viewpoints using all three approaches. The ASM approaches perform very well in estimating a 3D shape that fits to the 2D landmarks, but the estimated shapes are implausible, which can be seen by viewing them from the side. Moreover, they are vastly different from the basis shapes. The proposed approach, although compromising an exact fit to the 2D landmarks, is able to estimate the 3D pose with greater plausibility. This may be attributed to more consistent interpolation between the basis shapes adhering to the Kendall shape space constraints, which lead to reduced distortion of the basis shapes.

A similar scenario is shown in Fig. 4. Here, however, even though the KSS results in a plausible shape, it does not fit to the shape we would expect to obtain. This is most likely attributed to the fact that the ballet pose cannot be estimated from the given basis shapes.

3.2 Estimating Head Skeletons of Basking Sharks

In this section, we give the results for the application that motivated the new approach for shape estimation, that is, the estimation of the shape of the head skeletons of basking sharks in feeding position. As specimens are hard to come by and the species is heavily protected under conservation laws in many countries, the ability to extract the 3D structure of the head region non-invasively, using real-world underwater images is a great contribution to the field of biology. We use computed tomography (CT) scans of the heads of three basking sharks (*Cetorhinus maximus*) from two museum collections (BMNH: British Museum of Natural History; ZMUC: Zoological Museum, University of Copen-
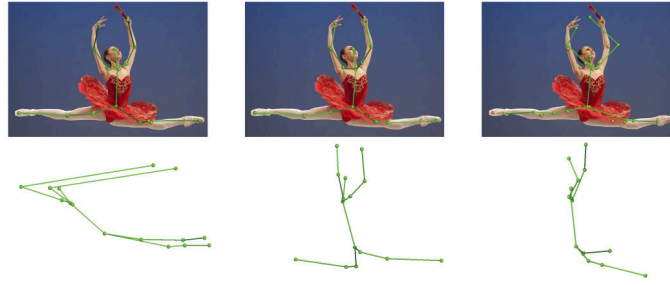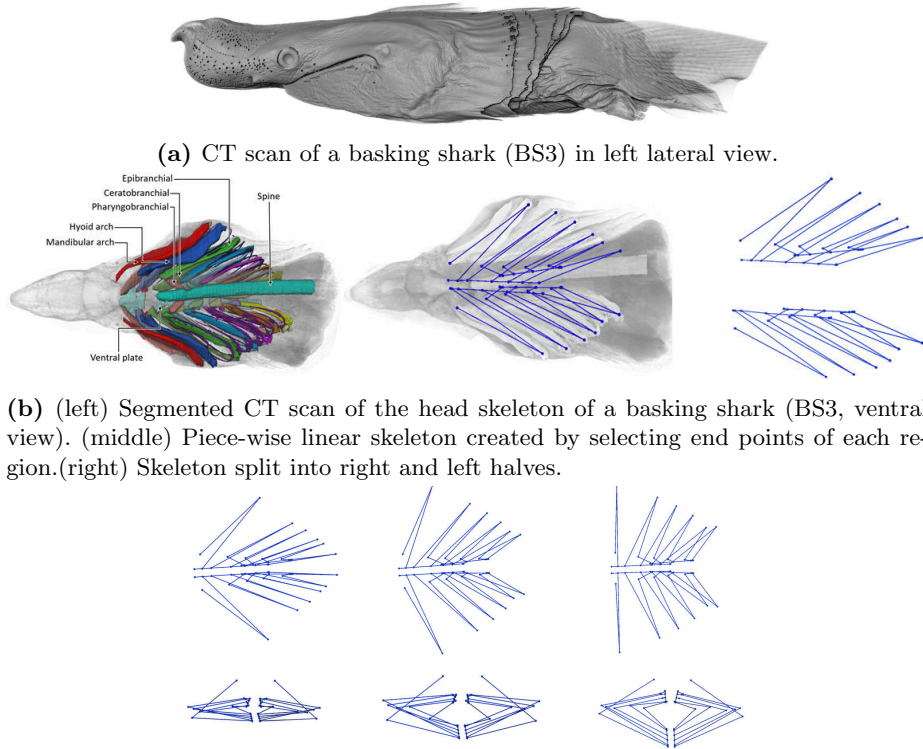
Fig. 4: Comparison of the three approaches for 3D shape estimation applied to an image of a ballet dancer. Top row: Fitting of the estimated shapes (green) to the landmarks (yellow). Bottom row: side views of the estimated shapes. Left: ASM non-convex approach. Middle: ASM convex approach. Right: KSS approach (ours).

hagen); sample and scan information are as follows: BS1: BMNH 1978.6.22.1, size: $512 \times 512 \times 1389$; BS2: BMNH 2004.4.15.30, size: $512 \times 512 \times 1654$; BS3: ZMUC (no accession number), size: $512 \times 512 \times 1862$. BS1 and BS2 were scanned at the Royal Brompton Hospital, BS3 at the Aarhus University Hospital. CT data were used to derive basis shapes and test the methods on real-world images of the species with manually labelled landmarks. BS3 is shown in Fig. 5a.

Basking sharks feed by holding their mouths agape and ballooning the pharynx (throat region), allowing water to flow out the gill slits on the sides of the head [31,33,36]. The pharynx is supported by jointed branchial arches, comprising rod-like skeletal elements linked together to form a mobile skeletal basket around the throat, which can expand and collapse. The kinematics of the pharyngeal skeleton are key to the filtering mechanism, but cannot be visualized from outside the body. From the CT scans of basking sharks, the relevant elements of the head skeleton (the jaw, hyoid and branchial arches, skull, and spine) are isolated using manual and automatic segmentation techniques. These isolated elements are then used to create a piece-wise linear skeleton of the sharks' pharyngeal region by labelling the start and end points of each skeletal element (e.g. the rod-like ceratobranchial cartilages; Fig. 5b) and connecting each point pair via a linear segment. Due to preservation conditions (e.g. specimens fixed in non-physiological positions), sharks in the CT scans showed non-physiological deformations (e.g. the branchial basket collapsed asymmetrically under the shark's weight), structures needed correction before creating the set of basis shapes. Due to the bilateral symmetry of basking sharks, their right and left halves are essentially identical: skeletons were split into right and left halves which could later be mirrored to create full skeletons. A set of optimization problems which aim to conserve the lengths of segments were used to move the spine and the ventral plate (a portion of the branchial basket) of the half skeletons into an anatomically plausible position [36]. The mouth of the skeleton was opened using the same method. The corrected skeletons were then mirrored w.r.t. the sagittal plane. Fig. 5c visualizes this process. The anatomical arrangements of the final

(a) CT scan of a basking shark (BS3) in left lateral view.

(b) (left) Segmented CT scan of the head skeleton of a basking shark (BS3, ventral view). (middle) Piece-wise linear skeleton created by selecting end points of each region.(right) Skeleton split into right and left halves.

(c) Basking shark skeleton shapes created by mirroring optimized half skeletons. (top) Top/dorsal-view showing the expansion of gill arches. (bottom) Front/anterior-view showing varying degrees of openness of the mouth.

Fig. 5: From CT scan to skeletal representation of shark head. (a) Volume rendering of CT scan. (b) Segmentation, skeletonization, and separation of left and right parts of the skeleton. (c) Opening of skeleton to create open-mouth configurations of the head.

skeletons were approved by experts in the field (M. Dean & F. Mollen, personal communication).

Analysis of reconstruction accuracy Similar to human pose estimation, quantitative and qualitative analyses were performed for all approaches. In the quantitative experiments, we used the head skeleton shapes (29 in total) of one of the basking sharks to estimate the shapes (28/29) of the other two. Two types of projections, bilaterally symmetric and asymmetric, are used to estimate the 3D shape. The former is the case when viewing the shark anteriorly (from the front) and the latter when viewing from the side (e.g. where one side of the throat is more visible than the other). Both scenarios are equally important when extracting the 3D shape from images. Fig. 6 shows the errors in estimation using all the approaches and Table 2 provides some error statistics. Again, statistically signif-
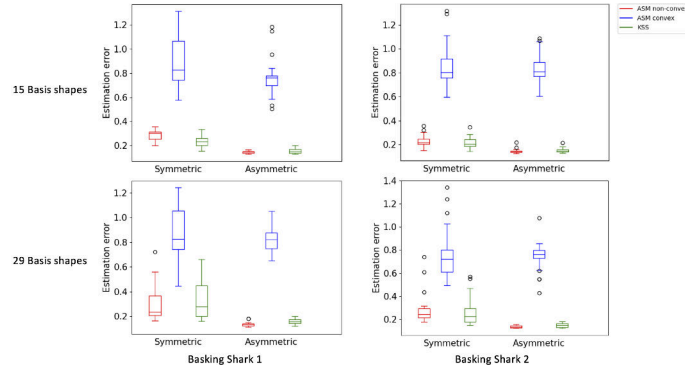
Fig. 6: Boxplots of shape estimation errors of basking shark test shapes for the three approaches using the basis shapes from BS3.

icant differences have been confirmed via Friedmann tests ($p < 0.001$) for each combination of subject and viewpoints. Contrary to the human pose estimation, the ASM non-convex and our approach provide similar estimation errors for the basking shark experiment with the former being slightly superior for 5 out of the 8 cases. These results indicate that the basking shark data might be located in a region of relatively low curvature in shape space that is well approximated by a hyper-plane.

Qualitative validation on underwater imagery We investigated the performances of the approaches in estimating the 3D structure of basking shark heads in underwater images with manually placed landmarks. Shapes from all three sharks (86 in total) were used as basis shapes. Figs. 7 and 8 show the results obtained using the two ASM approaches and the novel KSS approach. Similarly to the results for the estimation of real-world human poses, the shape estimations obtained from the two ASM approaches fit the 2D landmarks very well but fail to produce convincing results when looking from a different view.

---

[5] Source: https://www.amustard.com/library/albums/1/AMU_12_300611_0014.jpg

[6] Source: https://www.manxbaskingsharkwatch.org/shark-files/biology-behaviour/feeding

Table 2: Mean(variance) estimation errors for basking shark test shapes for symmetric and asymmetric projections using (15/29) basis shapes from Basking Shark 3.

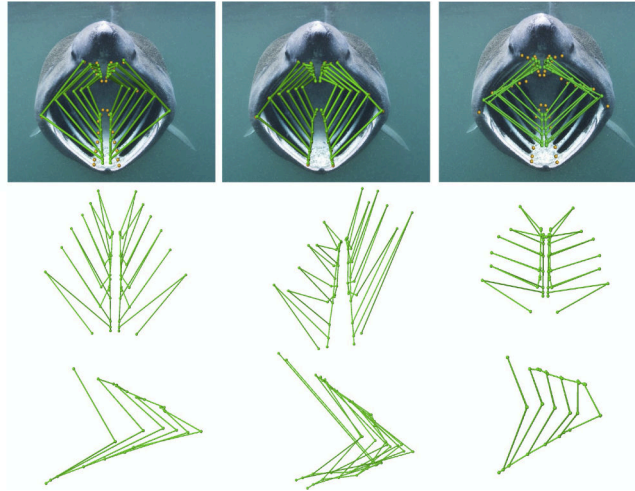|	#Basis Shapes	15 (sym)	15 (asym)	29 (sym)	29 (asym)
	ASM NC	0.231(0.002)	**0.146(3e-04)**	**0.306(0.020)**	**0.136(2e-04)**
BS1	ASM Conv	0.854(0.030)	0.836(0.013)	0.857(0.036)	0.820(0.011)
	KSS	**0.215(0.001)**	0.151(3e-04)	0.336(0.025)	0.156(0.0003)
	ASM NC	0.286(0.001)	**0.145(1e-04)**	0.281(0.014)	**0.137(9.7e-05)**
BS2	ASM Conv	0.906(0.041)	0.757 (0.021)	0.755(0.041)	0.747(0.013)
	KSS	**0.236(0.001)**	0.153(3e-04)	**0.256(0.012)**	0.150(3.5e-04)

Fig. 7: Comparison of the three approaches for 3D shape estimation applied to an under-water image[5] of a basking shark. Top row: Fitting of the estimated shapes (green) to the landmarks (yellow). Middle row: top views of the estimated shapes. Bottom row: side views. Left: ASM non-convex approach. Middle: ASM convex approach. Right: KSS approach (ours).

This is particularly striking for the ASM convex approach but is also noticeable for the ASM non-convex method—the estimated shape of which is unnaturally elongated. In contrast, applying the KSS approach results in 3D shapes that are very plausible and show a widely opened mouth.

4   Conclusion

In this work, we have proposed a novel, non-Euclidean approach for the estimation of 3D shapes from single monocular 2D images. In particular, we generalized the established active shape model approach to Kendall's shape space combining concepts from geometric statistics and computational differential geometry. The resulting geometric model provides efficient, yet highly consistent estimation of 3D shapes due to a natural encoding of constraints.

    We have demonstrated the performance of our approach in experiments on the reconstruction of human poses as well as head skeletons of basking sharks. The advantages of the proposed scheme are particularly apparent in qualitative comparison to previous approaches that rely on assumptions of linearity. While former approaches attained lower reprojection errors, the corresponding 3D shapes show defects like non-physiological distortions indicating a need for further domain-specific regularization. For example, the convex relaxation of the ASM proposed in [37] developed strong asymmetries although basis shapes are symmetric. In contrast, our geometric method yielded highly physiological re-
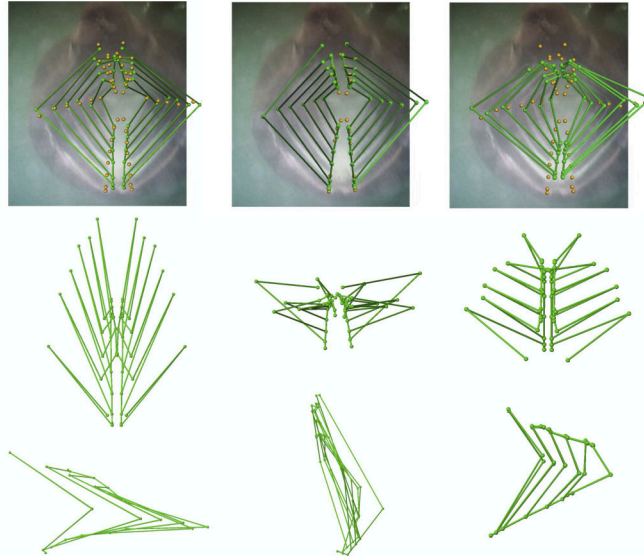
Fig. 8: Comparison of the three approaches for 3D shape estimation applied to an under-water image[6] of a basking shark. Top row: Fitting of the estimated shapes (green) to the landmarks (yellow). Middle row: top views of the estimated shapes. Bottom row: side views. Left: ASM non-convex approach. Middle: ASM convex approach. Right: KSS approach (ours).

sults even in challenging situations where the target shape differed greatly from the training distribution.

In future work, we plan to extend our approach to more advanced camera models to be able to account for strong perspective effects. Furthermore, while we use the 2D Kendall shape space distance as reprojection error, weighted or ℓ^1-like extensions could be employed to incorporate uncertainty in 2D landmark detection or to increase robustness against outliers, respectively.

References

1. Agudo, A., Agapito, L., Calvo, B., Montiel, J.M.: Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 1558–1565 (2014)
2. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 1446–1455 (2015)
3. Ambellan, F., Hanik, M., von Tycowicz, C.: Morphomatics: Geometric morphometrics in non-Euclidean shape spaces (2021). https://doi.org/10.12752/8544
4. Ambellan, F., Lamecker, H., von Tycowicz, C., Zachow, S.: Statistical shape models: understanding and mastering variation in anatomy. In: Biomedical Visualisation, pp. 67–84. Springer (2019)
5. Boumal, N., Mishra, B., Absil, P.A., Sepulchre, R.: Manopt, a matlab toolbox for optimization on manifolds. The Journal of Machine Learning Research **15**(1), 1455–1459 (2014)
6. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: Proc IEEE Comput. Vis. Pattern Recognit. vol. 2, pp. 690–696. IEEE (2000)
7. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. ACM Transactions on Graphics (TOG) **32**(4), 1–10 (2013)
8. Chakraborty, R., Bouza, J., Manton, J., Vemuri, B.C.: Manifoldnet: A deep neural network for manifold-valued data with applications. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
9. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Comput Vis Image Underst **61**(1), 38–59 (1995)
10. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans Med Imaging **23**(8), 995–1005 (2004)
11. Friji, R., Drira, H., Chaieb, F., Kurtek, S., Kchok, H.: Kshapenet: Riemannian network on Kendall shape space for skeleton based action recognition. arXiv preprint arXiv:2011.12004 (2020)
12. Frostig, R., Johnson, M.J., Leary, C.: Compiling machine learning programs via high-level tracing. Systems for Machine Learning **4**(9) (2018)
13. Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: A review. Med Image Anal **13**(4), 543–563 (2009)
14. Hejrati, M., Ramanan, D.: Analyzing 3d objects in cluttered images. Advances in neural information processing systems **25** (2012)
15. Howe, N.R.: Silhouette lookup for automatic pose tracking. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop. pp. 15–22. IEEE (2004)
16. Jiang, H., Liu, H., Tan, P., Zhang, G., Bao, H.: 3d reconstruction of dynamic scenes with multiple handheld cameras. In: European Conference on Computer Vision. pp. 601–615. Springer (2012)
17. Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London mathematical society **16**(2), 81–121 (1984)
18. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and shape theory. John Wiley & Sons (2009)
19. Larsen, E.S., Mordohai, P., Pollefeys, M., Fuchs, H.: Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)

20. Lloyd, S.: Least squares quantization in pcm. IEEE Trans Inf Theory **28**(2), 129–137 (1982)
21. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: VNect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4), 1–14 (2017)
22. Michalkiewicz, M., Parisot, S., Tsogkas, S., Baktashmotlagh, M., Eriksson, A., Belilovsky, E.: Few-shot single-view 3-d object reconstruction with compositional priors. In: European Conference on Computer Vision. pp. 614–630. Springer (2020)
23. Mori, G., Malik, J.: Recovering 3d human body configurations using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7), 1052–1062 (2006)
24. Mustafa, A., Kim, H., Guillemaut, J.Y., Hilton, A.: General dynamic scene reconstruction from multiple view video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 900–908 (2015)
25. Nava-Yazdani, E., Hege, H.C., Sullivan, T.J., von Tycowicz, C.: Geodesic analysis in Kendalls shape space with epidemiological applications. Journal of Mathematical Imaging and Vision **62**(4), 549–559 (2020)
26. Park, S., Lee, M., Kwak, N.: Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations. In: European Conference on Computer Vision. pp. 1–18. Springer (2020)
27. Paskin, M., Dean, M., Baum, D., von Tycowicz, C.: A Kendall shape space approach to 3D shape estimation from 2D landmarks – source code and data (2022). https://doi.org/10.12752/8730
28. Pennec, X.: Barycentric subspace analysis on manifolds. The Annals of Statistics **46**(6A), 2711–2746 (2018)
29. Plänkers, R., Fua, P.: Tracking and modeling people in video sequences. Comput Vis Image Underst **81**(3), 285–302 (2001)
30. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3d human pose from 2d image landmarks. In: European conference on computer vision. pp. 573–586. Springer (2012)
31. Sanderson, S.L., Roberts, E., Lineburg, J., Brooks, H.: Fish mouths as engineering structures for vortical cross-step filtration. Nat. Commun. **7**(1), 1–9 (2016)
32. Sanzari, M., Ntouskos, V., Pirri, F.: Bayesian image based 3d pose estimation. In: European conference on computer vision. pp. 566–582. Springer (2016)
33. Sims, D.W.: Sieving a living: a review of the biology, ecology and conservation status of the plankton-feeding basking shark *Cetorhinus maximus*. Advances in marine biology **54**, 171–220 (2008)
34. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. Int. J. Comput. Vis. **9**(2), 137–154 (1992)
35. Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W.: Robust estimation of 3d human poses from a single image. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 2361–2368 (2014)
36. Wegner, N.C.: Elasmobranch gill structure. In: Fish Physiology, vol. 34, pp. 101–151. Elsevier (2015)
37. Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3d shape estimation from 2d landmarks: A convex relaxation approach. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 4447–4455 (2015)
38. Zia, M.Z., Stark, M., Schiele, B., Schindler, K.: Detailed 3d representations for object recognition and modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2608–2623 (2013)