BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 7 2012, pages 1001–1008
doi:10.1093/bioinformatics/bts081

Data and text mining

Advance Access publication February 13, 2012

ShapePheno: unsupervised extraction of shape phenotypes from biological image collections

Theofanis Karaletsos[1,*], Oliver Stegle[1,*], Christine Dreyer[2], John Winn[3] and Karsten M. Borgwardt[1,4]

[1] Machine Learning and Computational Biology Research Group, Max Planck Institute for Intelligent Systems and Max Planck Institute for Developmental Biology, [2]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany, [3]Machine Learning and Perception Group, Microsoft Research Ltd, Cambridge CB3 0FB, UK and [4]Zentrum für Bioinformatik, Eberhard Karls Universität, 72076 Tübingen, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Accurate large-scale phenotyping has recently gained considerable importance in biology. For example, in genome-wide association studies technological advances have rendered genotyping cheap, leaving phenotype acquisition as the major bottleneck. Automatic image analysis is one major strategy to phenotype individuals in large numbers. Current approaches for visual phenotyping focus predominantly on summarizing statistics and geometric measures, such as height and width of an individual, or color histograms and patterns. However, more subtle, but biologically informative phenotypes, such as the local deformation of the shape of an individual with respect to the population mean cannot be automatically extracted and quantified by current techniques.

Results: We propose a probabilistic machine learning model that allows for the extraction of deformation phenotypes from biological images, making them available as quantitative traits for downstream analysis. Our approach jointly models a collection of images using a learned common template that is mapped onto each image through a deformable smooth transformation. In a case study, we analyze the shape deformations of 388 guppy fish (*Poecilia reticulata*). We find that the flexible shape phenotypes our model extracts are complementary to basic geometric measures. Moreover, these quantitative traits assort the observations into distinct groups and can be mapped to polymorphic genetic loci of the sample set.

Availability: Code is available under: http://bioweb.me/GEBI

Contact: theofanis.karaletsos@tuebingen.mpg.de; oliver.stegle@tuebingen.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 22, 2011; revised on February 7, 2012; accepted on February 9, 2012

1 INTRODUCTION

With the advent of high-throughput genotyping techniques an unprecedented breadth of genotypic datasets can be generated, opening doors to large-scale association studies, promising sufficient power to understand the genetic underpinning of more subtle phenotypes that characterize the sample. As phenotyping often requires manual labor and expert knowledge, a major bottleneck now lies with the identification and quantification of informative traits. Currently, the quantification of phenotypic traits is predominantly done in a semi-manual fashion, rendering the task of analyzing large datasets expensive, time-consuming and error-prone. In order to address these shortcomings, the automated analysis of biological images has become a staple in modern biology.

High-throughput imaging techniques for various types of microscopy and other imaging modalities have become common in the experimental environment. Automated image analysis for bioimaging attempts to deal with the flood of data and subsumes a large variety of tasks and methods; for a comprehensive review, see Peng (2008) and Walter *et al.* (2010). Common tasks include the counting of cells in microscopy images and differential analysis of distinct cell types (Fuchs *et al.*, 2010; Pau *et al.*, 2010). Key challenges in bioimage informatics stem from the breadth and individuality of natural variation within these images and dealing with the inherent noise in biological imaging tasks. In order to deal with these factors, machine learning techniques have raised considerable attention and are used to tackle various complicated tasks in realistic settings (Ning *et al.*, 2005; Shamir *et al.*, 2010). For example, in the analysis of appearance phenotypes machine vision has been used to quantify the extent of existence of predefined visual features or detect interesting appearance features that characterize the data (Whibley *et al.*, 2006). Visual appearance features usually pertain to specific local properties of the depicted objects. However, more general visual phenotypes often are also biologically informative, such as the description of the shape of an object and the quantification of global (including size and height) as well as local (i.e. locally deformed parts of an image) shape variations. An example where such a method is useful is the characterization of the shapes of guppy fish, which so far can only be analyzed by labor-intensive manual geometric phenotype measurements on hundreds of fish, as performed in (Tripathi *et al.*, 2009).

Our goal is to automatically determine and quantify differences among observed shapes in biological images in order to interpret them as shape phenotypes and facilitate downstream analysis, for instance association tests of traits with putative causal factors in the genome. In this work, we propose an unsupervised machine learning method to quantify shape variations of a given object class depicted

*To whom correspondence should be addressed.

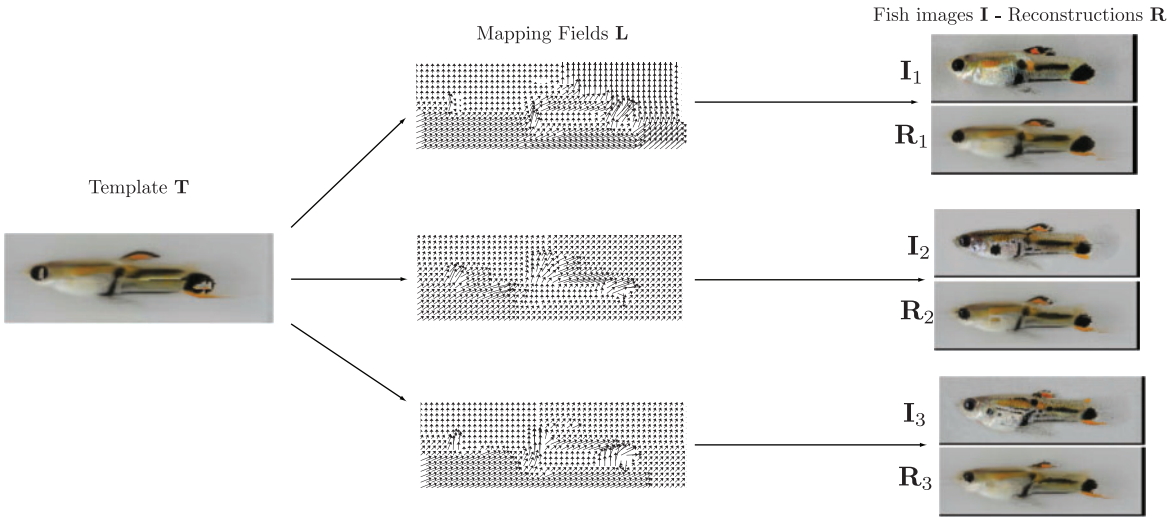

Fig. 1. A schematic overview of the generative process underlying the proposed approach, illustrated using three random images. From left to right: the learned common template $\boldsymbol{T}$, describing the mean-shape and appearance for the entire image collection. Image-specific mapping fields $\boldsymbol{L}$ capture the translation and deformation needed to map the template onto the observed images $\boldsymbol{I}$. Every pixel of each image has a corresponding vector that points onto the template pixel which it is drawn from. Images $\boldsymbol{R}$ denote the reconstruction of the corresponding raw image from the trained model.

in a set of images, one per individual or sample. We postulate the existence of an unobserved reference shape, called a template. We proceed with joint learning of this shared template and the image-specific shape deviation from this reference, allowing every image to be aligned to it. The resulting template iteratively converges to an idealized mean image from which the observed images are generated through *deformation fields* that explain the variation in shape of each image (Fig. 1). The converged model can also be run backwards, yielding a reconstruction of every image from the template and the mapping fields ($\boldsymbol{R}$ in Fig. 1).

The general task of aligning two or more images is also known as *registration*, where a correspondence between pixels of one image and pixels of another image is established. For example Saalfeld *et al.* (2010) perform a simpler form of registration, where images are aligned to a *known* template. In contrast to previous studies, our method does not require explicit knowledge of the template *a priori*; neither is supervision like setting of landmarks or outline selection/binarization on each image required. Instead, ShapePheno discovers and objectively quantifies deformation phenotypes on unannotated images in a fully unsupervised fashion while retaining interpretable features and results. Thus, our approach facilitates obtaining accurate non-trivial measurements on large datasets where human labor is costly and error-prone.

In Section 3, we present a case study of our method on guppy fish, *Poecilia reticulata*. The individuals in this dataset are subject to variation in appearance and shape. Interestingly, both appearance and shape variability have previously been shown to exhibit considerable genetic components (Tripathi *et al.*, 2009). In Section 3.3, we describe the basic approach of quantifying shape phenotypes using our model. We demonstrate that these quantitative traits are orthogonal to traditional geometric measures (Section 3.4) and show practical utility of these traits in the context of two fundamental types of downstream analyses. First, in Section 3.5, we show how learned shape features allow for grouping the observed images into plausible similarly shaped or deformed subgroups based on characteristic deformation patterns. Second, in Section 3.6, we show that quantitative shape phenotypes can be associated to variable genetic loci in the guppy genome. This analysis serves both as a step towards obtaining further knowledge as to the underlying biological processes that lead to variation of these phenotypes as well as a natural validation step, suggesting that the automatically determined phenotypes are biologically relevant.

2 METHODS

Extraction and quantification of shape features is carried out in two steps. In Section 2.1, we discuss how a graphical model based on Markov random fields (MRF) can be used to simultaneously learn the unknown template and recover smooth mapping fields, performing a flexible variant of deformable *registration*. We decompose the mapping fields into a sum of a technical translation component and shape-related deformation fields, both specific to each image. The overall setup of our probabilistic model largely follows the jigsaw model (Kannan *et al.*, 2007). Under this model, a set of N observed images $\boldsymbol{I}^i$, $i=1,\ldots,N$ is explained by a common latent template image $\boldsymbol{T}$. The training images are explained as function of the template through learned mapping vectors $\boldsymbol{L}^i$ between pixels in the template and observed pixels in each image i. The coordinate mapping accounts for an overall shift of the image with respect to the template, as well as local deformations, compressing or stretching specific parts of each image to match the common reference (Section 2.2). Subsequently, once the template and the deformation fields are learned, we extract quantitative traits from the information captured in the deformation fields. For this purpose, we employ linear dimensionality reduction (Section 2.3), yielding a compact set of features that explain the major axes of variation in the deformation fields for each image. For comparison, we also show how our model can be used to quantify length vectors within images, which can be directly related to established manual measurements of shape traits. Both types of features can be used for downstream analyses.

Summary overview of model parameters: In our model, we assume a set of N images $\boldsymbol{I}^i$ and corresponding mapping fields $\boldsymbol{L}^i$ of dimension $(d_{(x)} \times d_{(y)})$ for

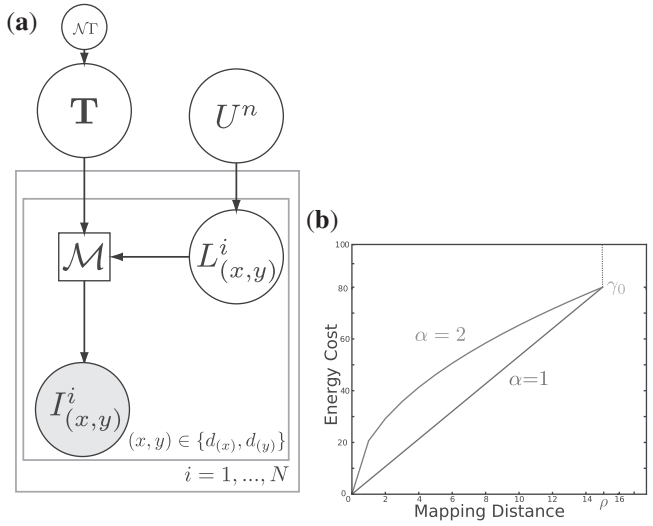

Fig. 2. Illustration of deformable registration model and deformation cost function. (**a**) Graphical model representation of the core ShapePheno model. Observed images $\boldsymbol{I}^1,\ldots,\boldsymbol{I}^N$ are modeled through common template image $\boldsymbol{T}$ and a coordinate mapping function $\mathcal{M}$. The mapping is parametrized by smooth deformation fields $\boldsymbol{L}^i_{(x,y)}$, denoting the coordinate offset between each image pixel and the template. The prior belief of smoothness of $\boldsymbol{L}$ is parameterized by an energy function for translation (E_{VT} and deformation E_{VD}) defined on pixel blocks $\boldsymbol{U}$. (**b**) Example choices of mapping energy function E_{VD} of order α, inducing the cost of non-neighboring mappings as a function of mapping distance. At maximum allowed deformation ρ, the energy cost diverges, restricting the effective range of deformations learned by the model.

$i=\{1,...,N\}$, where $\boldsymbol{L}^i$ is a matrix with entries $l^i_{(x,y)} \in \mathbb{Z}^2$ for each image i. The size of template image $\boldsymbol{T}$ is set to $(t_{(x)} \times t_{(y)})$. The model parameters μ_0, a, b and β define a prior on the template $\boldsymbol{T}$. The parameters determining the smoothness prior are given as: the prior on the translation component of the mapping field is parametrized by an energy constant γ_t; the deformable component of model is defined by as based cost at saturation, γ_0, the maximal allowed deformation distance ρ as well as α, parametrizing the order of the metric. $\boldsymbol{U}$ is describing the pixel block coupling.

2.1 Markov random fields for deformable registration

Each image pixel $I^i_{(x,y)}$ is related to the common template $\boldsymbol{T}$, linked by a transformation $\mathcal{M}^i$ that maps image coordinates (x,y) to the corresponding coordinates within $\boldsymbol{T}$:

$$\mathcal{M}^i_{(x,y)} = (x,y) - l^i_{(x,y)}. \tag{1}$$

The mapping is parametrized using a field of relative shifts $\boldsymbol{L}^i = \boldsymbol{L}^i_{(0,0)},\ldots,\boldsymbol{L}^i_{(d_x,d_y)}$, where contiguous (smooth) mappings of neighboring pixels corresponds to constant entries in $\boldsymbol{L}^i$.

For a given mapping field, $\boldsymbol{L}^i$, the generative model (Fig. 2a) of each image is then

$$\boldsymbol{I}^i = \boldsymbol{T}_{\mathcal{M}^i} + \boldsymbol{\psi}, \tag{2}$$

where $\mathcal{M}^i$ is parameterized by $\boldsymbol{L}^i$ and $\boldsymbol{\psi}$ denotes the reconstruction error of each pixel. Both, the template map $\boldsymbol{T}$ and the mapping fields $\boldsymbol{L}^i$ are unknown *a priori* and hence need to be learned from the image data alone. For a common template $\boldsymbol{T}$ with set dimensions $t_{(x)}, t_{(y)}$, a set of N observed images $\boldsymbol{I}^i$ and corresponding mapping fields $\boldsymbol{L}^i$, the joint probability under our model is

$$P(\boldsymbol{T}, \{\boldsymbol{I}^i, \boldsymbol{L}^i\}_{i=1}^N) = \underbrace{P(\boldsymbol{T})}_{\text{template prior}} \prod_{i=1}^{N} \underbrace{P(\boldsymbol{I}^i | \boldsymbol{T}, \boldsymbol{L}^i)}_{\text{likelihood}} \underbrace{P(\boldsymbol{L}^i)}_{\text{mapping prior}}. \tag{3}$$

The likelihood of the observation model corresponding to Equation (2) is a Gaussian mixture model independent for each image pixel (x,y):

$$P(\boldsymbol{I}^i | \boldsymbol{T}, \boldsymbol{L}^i) = \prod_{(x,y)} \left[(1-\pi) \mathcal{N}\left(I^i_{(x,y)} | \boldsymbol{\mu}_{\mathcal{M}^i_{(x,y)}}, \boldsymbol{\tau}^{-1}_{\mathcal{M}^i_{(x,y)}} \right) + \pi \text{ Uniform} \right]. \tag{4}$$

Here, $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ correspond to the means and the precisions at each position of the template image $\boldsymbol{T}$ and π is the mixture coefficient of the uniform background model to explain outliers that are not compatible with the template image. To ensure that the template is well-behaved, we choose a normal-gamma prior on the values of the template,

$$P(\boldsymbol{T}) = \prod_{(x,y)} \mathcal{N}\left(\mu_{(x,y)} | \mu_0, (\beta \tau_{(x,y)})^{-1} \right) \text{Gamma}(\tau_{(x,y)} | a, b). \tag{5}$$

Smoothness of the mapping fields $\boldsymbol{L}^i$ is encouraged through the choice of a Markov random field prior that couples neighboring mapping offsets in each image

$$P(\boldsymbol{L}^i) \propto \exp\left[- \sum_{(x,y),(x',y') \in \mathcal{E}_{(x,y)}} E_{\mathrm{V}}(l_{(x,y)}, l_{(x',y')}) \right]. \tag{6}$$

Here, $\mathcal{E}_{(x,y)}$ denotes the set of pixels in the direct neighborhood of (x,y) and the pairwise energy term E_{V} penalizes non-contiguous offsets of neighboring pixels. In ShapePheno, the mapping fields $\boldsymbol{L}^i$ are decomposed into a translation and deformation component with individual smoothness priors. The specific modeling choices will be discussed in Section 2.2.

Inference in the joint model implied by Equations (3–6) is feasible by means of iterative learning using expectation maximization. In this approach, we alternate between maximizing the joint probability (Equation (3)) with respect to mapping fields $\boldsymbol{L}^i$ and the unknown template $\boldsymbol{T}$, keeping the other variables fixed. The first task, determining the most probable template $\hat{\boldsymbol{T}}$ essentially boils down to parameter inference in a normal-gamma model, where closed form updates for mean and precisions of the template $(\boldsymbol{\mu}, \boldsymbol{\tau})$ are available (Bishop, 2006). Alternatively, for a fixed template $\boldsymbol{T}$, the most probable mappings $\hat{\boldsymbol{L}}^i = \text{argmax}_{\boldsymbol{L}^i} p(\boldsymbol{I}^i | \mathbf{T}, \boldsymbol{L}^i) p(\boldsymbol{L}^i)$ can be determined efficiently using graph cuts for each image (see Kannan *et al.* (2007) and Boykov *et al.* (2001) for details).

2.2 Design choices for MRF energy functions

Since the alignment of template and observed images requires a combination of *translation* and *deformation*, we choose $\boldsymbol{L}$ to be the sum of a translation and deformation field component $\boldsymbol{L} = \boldsymbol{L}_t + \boldsymbol{L}_d$. In this stacked two-layer Markov random field, $P(\boldsymbol{L}_t)$ accounts for the *global* shift of images to the template and $P(\boldsymbol{L}_d)$ specifies the prior probability of *local* deformations (see also Fig. 2.1). The joint prior probability of the mapping-field components can be expressed as

$$P(\boldsymbol{L}, \boldsymbol{L}_t, \boldsymbol{L}_d) = \delta(\boldsymbol{L}, (\boldsymbol{L}_d + \boldsymbol{L}_t)) P(\boldsymbol{L}_d) P(\boldsymbol{L}_t),$$

where $\delta()$ denotes the Dirac delta function. Accounting for both prior contributions, the effective energy term in Equation (6) becomes $E_{\mathrm{V}} = E_{\mathrm{VD}} + E_{\mathrm{VT}}$. Inference in the full model is done iteratively within the mapping updates, by first keeping $\boldsymbol{L}_d = 0$ fixed and updating $\boldsymbol{L}_t$. Next, $\boldsymbol{L}_t$ is kept at the learned value while updating $\boldsymbol{L}_d$. Both update steps can be done following the standard jigsaw inference (Section 2.1 and Kannan *et al.* (2007)). In the following, we will explain the modeling choices of each mapping prior separately.

2.2.1 Translation (rigid) model: Having defined a template that is of equal size as the images, the goal is to register images to it. In order to allow the shift field $\boldsymbol{L}_t$ to incorporate translation behavior, we employ a Pott's Model prior with energy function $E_{\mathrm{VT}}(l_p, l_q) = \gamma_t\, \delta(l_p, l_q)$. Here, the cost parameter γ_t is set to large value, such that all mappings $\boldsymbol{L}_i$ are forced to take on identical values, solely accounting for a constant overall image shift.

2.2.2 Deformable model: For the deformable prior $P(\boldsymbol{L}_d)$, we employ a non-rigid smoothness prior that encourages smooth deformation fields. In contrast to the rigid Pott's model, the energy costs is distant-dependent, favoring short-range deformations. More specifically, the energy function E_{VD} scales linearly with a particular choice of distance norm of order α:

$$E_{\mathrm{VD}}(l_p, l_q) = \begin{cases} \gamma_{\mathrm{def}} \|l_p - l_q\|_\alpha & \text{for } \|l_p - l_q\|_\alpha \le \rho \\ \infty & \text{otherwise.} \end{cases} \quad (7)$$

Here, ρ denotes the maximum permitted range of deformation, and α is a power (or order) where $\alpha \in \{1, 2n\}$ for $n \in \mathbb{N}_{>0}$ and scaling parameter $\gamma_{\mathrm{def}} = \frac{\gamma_0}{\max(1, \rho^{1/\alpha})}$.

Intuitively, one can imagine this function to apply elastic bands connecting pieces (in our case pixels) to their four neighbours with the E_{VD} part of the mapping costs being the equivalent of the elastic potential of the bands between all pieces. Figure 2b shows the energy function for two choices of α.

2.2.3 Robustifying deformable registration: We use various constraints on the registration to further improve the robustness of our method against noise and non-standardized images and reliably produce good results. We apply our deformation fields on pixel blocks, meaning that we constrain groups of pixels of block-size to obtain the same mapping via the prior $\boldsymbol{U}$ shown in Figure 2a. This leads to piecewise smooth deformation fields. Additionally, we constrain the parameter ρ of the deformation field itself. This makes large jumps prohibitively expensive and drives the model to use smoothly varying local deformation patches. A positive side effect of this constraint is a significant boost in computational efficiency and robustness against outlier-mistakes since the solution space is reduced. We also robustify alignments against appearance outliers with a mixture of densities used as the observation model in Equation (4), which allows for a background class.

2.3 Feature representation for deformation maps

The deformation fields $\boldsymbol{L}_d$ at pixel resolution, described in Section 2.2.2, capture the relevant information to explain local shape deformations of the samples in each image. Comparing deformation fields with non-equal objects at non-equal positions is hard, since we face the problem of correspondence. However, in our framework this problem can be elegantly circumvented using the common template all images are aligned to. To render individual deformation fields comparable between images, we first project these maps from observation space into the common reference coordinate system on the template. For this purpose, we apply the same mapping operator we used for image pixels now to the mappings themselves

$$\boldsymbol{D}^i_{\mathcal{M}^i} = \boldsymbol{L}^i_d,$$

resulting in deformation field representations in the object-centered template space. Thus, we obtain an easily interpretable shift-corrected field $\boldsymbol{D}^i$ of deformation for each image-mapping $\boldsymbol{L}^i_d$.

2.3.1 Low-rank representations of deformation fields: In order to extract meaningful features from the high-dimensional deformation fields, we reduce their dimensionality by representing individual deformation fields via a set of coefficients over a small number of bases Φ obtained via standard principal component analysis (PCA) (Fig. 3). Prior to running PCA, we can apply a binary mask to the template to select only relevant template regions for consideration as variance components. The resulting individual bases can be visualized and constitute local deformation factors.

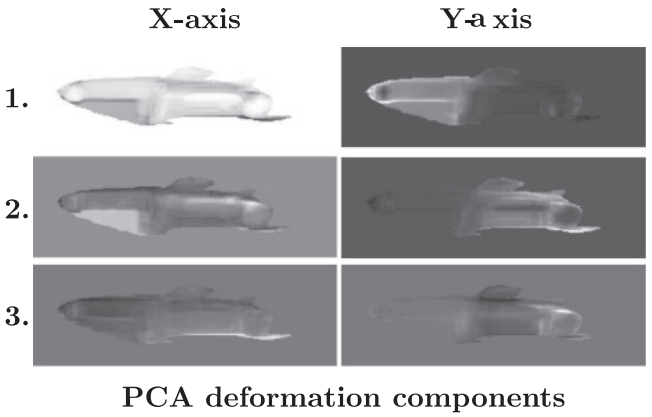

Fig. 3. Illustration of the first three PCA-bases of deformation fields in the X and Y deformation direction. Individual highlighted image parts such as black stripes correspond to image parts with most pronounced deformation. Here, these areas correspond to compression and stretching of fish tails and the main body (Section 3).

Building on the PCA basis functions, we use the corresponding coefficients for every image as a quantitative trait that characterizes the deformation field in a given sample. Formally, for a matrix of k linearized orthogonal bases Φ of dimensionality $p \times (d_{(x)} \cdot d_{(y)})$ (dimensions of an image), a coefficient matrix $\boldsymbol{W}$ of dimensionality $N \times k$ and a $N \times (d_{(x)} \cdot d_{(y)})$ observation matrix of linearized matrices $\boldsymbol{D}^i$ that contains the k-rank approximations to the observed, corrected and template-projected deformation fields $\boldsymbol{D}^i$ as described in Section 2.3, the low-dimensional projections per image have the form: $\boldsymbol{D}^i = \boldsymbol{w}^i \cdot \Phi$.

2.3.2 Geometric measurements as shape traits: We also use our method to measure distances in the images by exploiting the mapping fields to the common template. We measure geometric traits by selecting points of interest on the template and measuring their distance per image by projecting the annotated template-points into the reconstructed image through the mapping fields. Thus, a single annotation of the shared template yields an exhaustive annotation of all images explained by the model. These annotations can be used to measure geometric distances in the images, for example. Importantly, these geometric measurements are local measurements on an image in contrast to the holistic descriptor of shape we introduce in Section 2.3.1.

3 RESULTS

We applied ShapePheno to a dataset that shows the lateral aspect of male guppy fish, *P. reticulata*. The goal was to obtain local deformation patterns that are informative about typical distortions of the shape among the individuals, which also display considerable variation of appearance traits and size. We demonstrate further use of our method in two tasks: clustering of populations according to deformation patterns and association studies to link genotypes to deformation phenotypes.

3.1 Dataset

The 388 available individuals are second generation progeny (F2) of two parents representing geographically and genetically distant populations whose visual appearances also differ significantly. The male parent from Cumaná (Ve) (Alexander and Breden, 2004) has a slimmer posterior trunk and brighter orange ornaments as compared to the maternal population from the Quare river in East Trinidad.

This cross (157 Quare x Cumaná) has been subject of a genotyping project to establishing a comprehensive genetic map and to initiate conventional QTL (quantitative trait locus) mapping (Tripathi *et al.*, 2009). The raw images were rescaled such that the ratio of pixels to physical length is constant across the dataset. Due to rescaling, the image size was variable, ranging between 75×226 and 83×250 pixel size. To account for images taken at slightly varying distance, we chose to embed all images according to original size of the fish into empty images of the chosen format ($83\times$ 250). For this particular experiment, there were few outliers and thus setting the outlier ratio $\pi=0$ yielded good results. For each of the 388 individuals, the dataset included matching genotype information, covering a total of 1063 genome-wide single nucleotide polymorphisms (SNPs). After filtering, removing rare SNPs with a minimum rare allele frequency <5%, we obtained 814 polymorphisms that were considered for analysis.

3.2 Experimental settings

We chose the following parameters for the deformation model: $\gamma_0=40$, $\rho=10$ and $\alpha=1$, which resulted in robust registration in a series of test runs. We applied the deformation field to 2×2 pixel blocks in order to locally tie together image pixels to correct for appearance differences and to prevent excessive local deformation. The normal-gamma prior parameters (Equation (5)) were set such that the prior reflects the first and second empirical moments of the distribution of the raw image pixels (see also Kannan *et al.* (2007)). We ran a Python-based parallelized implementation of ShapePheno on an 8-core Intel Xeon machine where the full dataset could be run to convergence within 3 days. After convergence, we manually segmented the template fish from the background template to facilitate all downstream analyses (clustering and association mapping on foreground information only).

3.3 Shape phenotype determination

The converged ShapePheno model yielded a sharp template that resembles an average fish and mapping fields $\boldsymbol{L}^i$ for every image in the dataset (Fig. 1). The model perceives the shapes of the fish in individual images as locally stretched or smoothly distorted versions of the template and smoothly bypasses appearance differences that would counteract shape alignment. This suggests that the deformation fields $\boldsymbol{L}^i_d$ capture shape information corrupted by noise stemming from the difference in sizes of images and the background color similarity to the fish corpora. Next, we used linear dimension reduction (Section 2.3.1) to determine the corresponding deformation factors of the converged model; Figure 3 depicts the first three PCA-bases. These three main deformation features appear to divide the fish into the anterior and posterior part. Inflated anterior parts at the belly region as well as distorted posterior trunks are the main sources of shape variation. We also observed that local structure in the bases matches appearance features of the template that get distorted frequently. These findings reflect the set-up of the experiment in Tripathi *et al.* (2009) in agreement with our expectations, as the parents were originally chosen to exhibit these shape differences and the offspring shows strong variation at these features. Supplementary Figure S1 provides examples of inference results for extreme outliers within the data, here a singleton shape-mutant in our training set. Since the method is unsupervised, it requires shape mutants to be well-represented in the data in order to model their shape accurately.

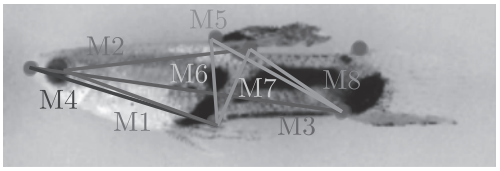

Fig. 4. Overview of the chosen points of interest on the fish and the template and the measured geometric phenotypes with corresponding names M1–M8. These measurements are chosen to capture shape differences between anatomically fixed points.

3.4 Quantification of geometric measurement accuracy

After the qualitative evaluation of the reconstructed shape template, we next characterized the accuracy of the shape representation captured by the model in a quantitative manner. For this purpose, we used the converged model to automatically measure geometric distances in images (Section 2.3.2). We comparatively evaluated eight geometric trait measurements (described in Figs 4 and 6), whose choice was motivated by primary analyses of the Guppy dataset (Tripathi *et al.*, 2009). Manual quantification was done on 50 individuals from our dataset chosen at random, measuring all 8 geometric distances in each raw image by 3 independent experts, as well as using the fully automated approach provided by the ShapePheno model. We assessed the correlation between manual and automated measurements, comparing the ShapePheno prediction to the mean of the manual quantification runs (Fig. 3.4). Encouragingly, all automated geometric measurements were in good agreement with the corresponding manual annotation. The correlation score for pairs of corresponding automated and manual measurements ranged between 0.65 (A2 versus M2) and 0.84 (A6 versus M6) with a mean correlation score of 0.76.

To better understand the magnitude of the variation between automated and manual measurements, we also considered the pairwise correlation between two of the three manual runs (Fig. 5b), yielding comparable results. Pairwise correlations here ranged from 0.81 (M7) up to 0.96 (M3) with a mean correlation score of 0.87.

This suggests that ShapePheno captures true variability in images and yields high levels of accuracy when used to quantify geometric measurements in place of a human expert. Detailed scatter plots, showing the correlations between manually and automatically determined traits are shown in Figure 6.

From either of the correlation analyses, it was also notable that geometric measurements correlated well with each other, reflecting the biological relatedness of growth-phenotypes that underlie the geometric measurements under consideration. In contrast to this observation, the correlation to the new PCA-deformation phenotypes described in Section 3.3 was weak, which shows that they capture orthogonal aspects of shape variation and hence are complementary to geometric measurements.

3.5 Clustering of populations based on deformation traits

We clustered populations of guppy fish according to their characteristic local deformation patterns, without any prior

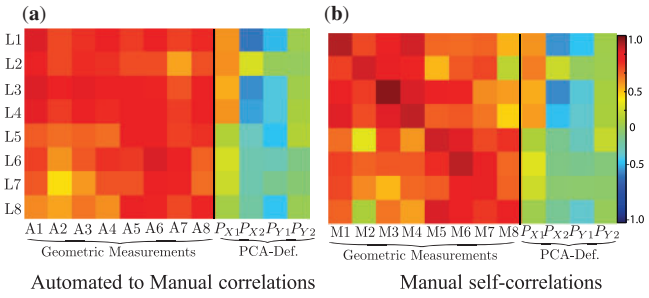

Fig. 5. Quantitative evaluation and comparison of the geometric traits as determined by ShapePheno (A1–A8), manually measured counterparts from a human expert (L1–L8) and novel PCA-deformation phenotypes ($P_{X1} - P_{Y2}$). (**a**) Correlation between automatic geometric measurements, manual measurements and PCA-deformation phenotypes. (**b**) Correlation between two manual quantification runs by an expert user (M1–M8 versus L1–L8). The results show that the reproducibility of manual expert labels is similar to automated measurements by ShapePheno, suggesting the model is able to achieve human-like results. Weak correlation between PCA-deformation phenotypes and geometric phenotypes shows that these new quantitative traits complement established measurements.

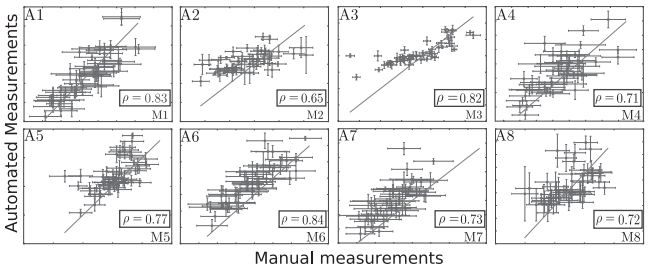

Fig. 6. Scatter plots with error bars, showing the relationship between manual measurements and automated geometric measurements as obtained by ShapePheno. Shown is data for each of the 8 length phenotypes, where each point corresponds to an instance of the 50 samples chosen for quantification. Error bars show ±1SD, estimated separately for each quantification approach. For manual quantification, error bars correspond to the empirical variation between three independent annotation runs. For automated quantification, uncertainty estimates stem from the variation of the 15 nearest neighbor assignments on the template to the selected point and measuring their SD. The green diagonals show the expected ideal correlation.

knowledge of their genetic constitution. Morphometric prototypes for the guppy have previously been determined from hand-annotated images and correlated to sex and environmental factors (Hendry *et al.*, 2006). We clustered deformation fields according to a linear kernel between low-rank projections of $\boldsymbol{D}^i$ (as described in Section 2.3.1) using affinity propagation (Frey and Dueck, 2007), a non-parametric clustering technique that uses deformation kernel values as inputs and yields a flexible number of clusters $|C|$. Reconstructing the mean low-rank vector field of each cluster given by its embedding in deformation space yields cluster-specific morphological deformation bases. Figure 7 provides a comparative overview of three characteristic clusters, indicating that independent factors can influence the shape of the anterior and posterior trunk of the examined guppy fish. Deformation bases correspond to low-rank projections of the cluster means, where only the characteristic local deformations per cluster are considered as shared elements of cluster members. The shape seen in an example image representing the median deformation field per cluster corresponds to our expectation given the profiles. Different clusters portray significant variability between their profiles, such as the regional focus and the extent of expected local deformation.

3.6 Association study of shape factors to genotype

Finally, we performed a genome-wide association study using the previously learned phenotypes and their measurements. The phenotypic measurements y are the per-image coefficients $\boldsymbol{w}^i$ of PCA-deformation bases Φ (Sections 2.3.1 and 3.3). We used a linear model that assesses how well a particular phenotypic value is modeled when genetic factors are taken into account, compared to when they are ignored. The relevant quantity is the log-odds (LOD) score,

$$\log_{10}\left\{\prod_j \frac{P(y_j|s_j,\theta)}{P(y_j|\theta_{\text{bck}})}\right\}$$

where s_j is a SNP measurement and y_j the phenotypic expression value for the j-th individual. The terms $\theta, \theta_{\text{bck}}$ are parameters for the genetic and background models, respectively. We thus obtain LOD score plots over a large genomic region to obtain an association plot. We used Storey's method (Storey and Tibshirani, 2003), a variant of Benjamini Hochberg, to assess genome-wide significance.

Although the available data has sparse genetic marker coverage, we still obtained statistically meaningful peaks as can be seen in Figure 8. Previous genetic QTL mapping in overlapping data has suggested markers of the proximal region of linkage group 12 (LG12) as relevant for size and body shape traits in male guppies, and in addition phenotypic sex has impact on these traits (Tripathi *et al.*, 2009). Among the significant hits in our mapping, Markers 398 (lod 7.7 on LG12) and 442 (lod 10.3, LG12) are found in the proximal region of LG12 while marker 229 (lod 11.9, LG12) is the most distal and closest to the putative male sex-determining locus. Depending on the trait analyzed, significant QTL were suggested within a region spanning ∼6 cM (∼7 Mb (Tripathi *et al.*, 2009)) in cross 157. Marker 442 was supported as a QTL for area of the posterior trunk for cross 158 (Tripathi, 2009). Additional loci were detected with good statistical support, in agreement with the observation that co-factors on various linkage groups contribute to complex traits.

4 DISCUSSION

We have proposed a generative probabilistic model that extracts deformation phenotypes by registering images to a latent, learned template in an unsupervised fashion. Our method presents a novel, clean framework for researchers to quantify and describe subtle local deformation patterns and use them for downstream analyses, like clustering or genetic association tests. We applied our method to a bioimaging task, where we discovered significant deformation patterns in images of guppy fish. We also showed that ShapePheno can be used for automated quantification of geometric measurements and showed good correspondence to manually labeled data.

More important than accurate geometric traits, ShapePheno yielded deformation fields that characterize the variability in shape and could be used to identify low-rank PCA factors

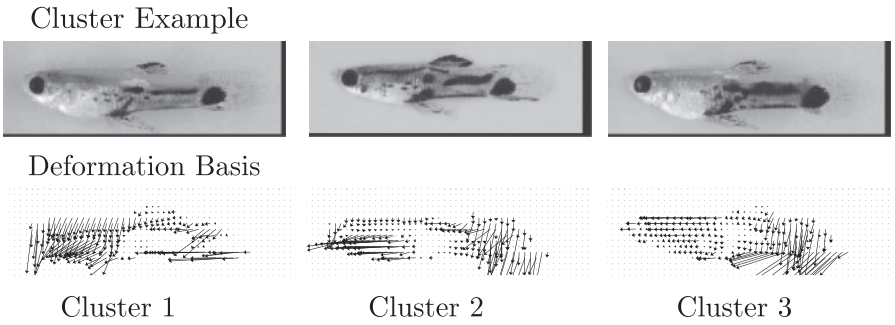

Fig. 7. Illustration of the clustering results obtained when using the PCA-deformation phenotypes from ShapePheno. From top to bottom the figure shows: exemplar of the cluster, illustrating the representative fish chosen by Affinity propagation to represent the given cluster. Deformation basis, showing the corresponding deformation of the exemplar projected onto the PCA-deformation fields used for clustering. The exemplars approximately correspond to visual categories of inflated anterior body, elongated anterior body and deformed tail, as present in the image collection.

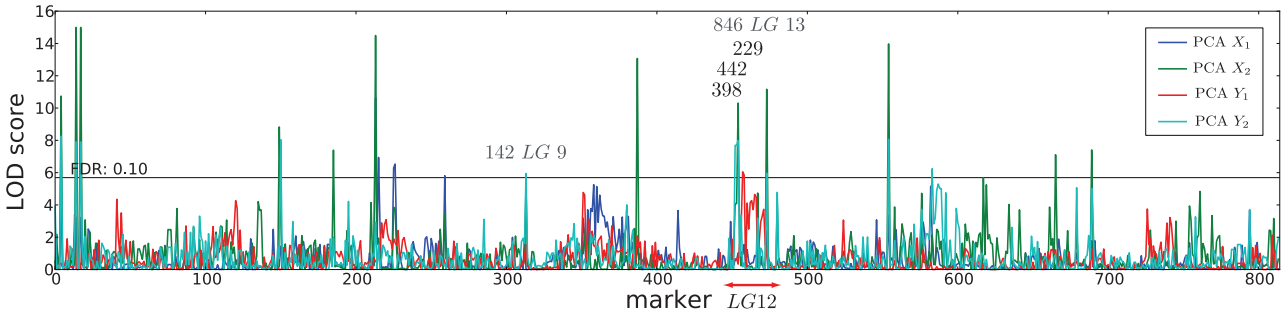

Fig. 8. Genome-wide association plot, showing the association strengths with the first two PCA-deformation phenotypes corresponding to the *X* and *Y* direction. The significance threshold of 10% false discovery rate (FDR) is shown as thin line in the diagram. SNPs are plotted in order of linkage groups, while significant hits on LG12 as described in Section 3.6 are highlighted. LG13 contains markers with function in sex determination yielding additional informative peaks for shape determination.

of shape variability. While simple distance measurements inter-correlate strongly, the deformation phenotypes we propose describe orthogonal shape factors and are thus novel holistic descriptors of shape. We showed practical utility of these PCA-deformation phenotypes in the context of clustering, grouping the data into clusters exhibiting characteristic deformation. We also performed a GWAs with the same traits, which yielded biologically sound results in agreement with previous results on geometric approximations of shape (see Tripathi *et al.* (2009) and unpublished observations of C.D.). We are convinced that comprehensive genomic analyses on larger datasets can be performed by using this method with a rigorous treatment of image acquisition, higher image resolution and higher marker density.

Unsupervised extraction and quantification of subtle morphological phenotypes, as done here, is the logical next step in automated image analysis. The relevance of these new types of methods is expected to rise quickly as dataset sizes increase, providing the necessary statistical power to identify and quantify complex phenotypic variation.

Funding: T.K. was supported by a Microsoft Research Cambridge stipend and O.S. was supported by a fellowship from the Volkswagen Foundation.

Conflict of Interest: none declared.

REFERENCES

Alexander,H.J. and Breden,F. (2004) Sexual isolation and extreme morphological divergence in the cuman guppy: a possible case of incipient speciation. *J. Evolution. Biol.*, **17**, 1238–1254.

Bishop,C. (2006) *Pattern Recognition and Machine Learning*. Vol. 4, Springer, New York.

Boykov,Y. *et al.* (2001) Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 1222–1239.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Fuchs,F. *et al.* (2010) Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol. Syst. Biol.*, **6**, 370.

Hendry,A.P. *et al.* (2006) Parallel evolution of the sexes? Effects of predation and habitat features on the size and shape of wild guppies. *J. Evolution. Biol.*, **19**, 741–754.

Kannan,A. *et al.* (2007) Clustering appearance and shape by learning jigsaws. In Schölkopf,B. *et al.* (eds) *Advances in Neural Information Processing Systems*. MIT Press, p. 2006.

Ning,F. *et al.* (2005) Toward automatic phenotyping of developing embryos from videos. *IEEE Trans. Image Process.*, **14**, 1360–1371.

Pau,G. *et al.* (2010) Ebimage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics*, **26**, 979–981.

Peng,H. (2008) Bioimage informatics: a new area of engineering biology. *Bioinformatics*, **24**, 1827–1836.

Saalfeld,S. *et al.* (2010) As-rigid-as-possible mosaicking and serial section registration of large system datasets. *Bioinformatics*, **26**, i57–i63.

Shamir,L. *et al.* (2010). Pattern recognition software and techniques for biological image analysis. *PLoS Comput. Biol.*, **6**, e1000974.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.*, **100**, 9440–9445.

Tripathi,N. *et al*. (2009) Genetic linkage map of the guppy, poecilia reticulata, and quantitative trait loci analysis of male size and colour variation. *P. Roy. Soc. B Bio.*, **276**, 2195–2208.

Tripathi,N. (2009) PhD Thesis, University of Tuebingen.

Walter,T. *et al*. (2010) Visualization of image data from cells to organisms. *Nat. Methods*, **7**(Suppl. 3), S26–S41.

Whibley,A.C. *et al*. (2006) Evolutionary paths underlying flower color variation in antirrhinum. *Science*, **313**, 963–966.