Single-Image Insect Pose Estimation by Graph Based Geometric Models and Random Forests

Minmin Shen(✉), Le Duan, and Oliver Deussen

INCIDE Center, University of Konstanz, Konstanz, Germany
minmin.shen@uni-konstanz.de

Abstract. We propose a new method for detailed insect pose estimation, which aims to detect landmarks as the tips of an insect's antennae and mouthparts from a single image. In this paper, we formulate this problem as inferring a mapping from the appearance of an insect to its corresponding pose. We present a unified framework that jointly learns a mapping from the local appearance (image patch) and the global anatomical structure (silhouette) of an insect to its corresponding pose. Our main contribution is that we propose a data driven approach to learn the geometric prior for modeling various insect appearance. Combined with the discriminative power of Random Forests (RF) model, our method achieves high precision of landmark localization. This approach is evaluated using three challenging datasets of insects which we make publicly available. Experiments show that it achieves improvement over the traditional RF regression method, and comparably precision to human annotators.

Keywords: Insect pose estimation · Landmark detection · Random forest

1 Introduction

Automated image based tracking and pose estimation receives increasing interests of both biology and computer science community, as its developments enable remotely quantify and understand individual behavior previously impossible [1]. Therefore, automatic insect tracking techniques have been an research topic in biological image analysis [2–4]. The movements of harnessed insects' bodyparts, such as antennae or mouthparts, provide information for behavioral study. Motivated by latest behavioral studies in biology [5], we aim to localize the landmark as the tips of bodyparts (e.g. a bee's antennae or tongue shown in Fig. 1) to provide detailed pose information.

In contrast to most existing works that aim at estimating the center of mass (position), detecting the detailed body posture and position of appendages (pose) is more challenging. Most existing tracking/pose estimation algorithms are not applicable for our task, due to a number of specific challenges and constraints:

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part I, LNCS 9913, pp. 217–230, 2016.
DOI: 10.1007/978-3-319-46604-0_16

- **Occlusion.** Insect body parts are highly clustered due to their small sizes, thus self occlusions are prevalent in insect body parts (Fig. 1b). As a result, it is difficult to estimate the pose.
- **Unstructured appearance.** Insect body parts have dark appearance, similar shape and no texture. Moreover, our image data is a set of 2D videos, and does not contain depth information. Their unstructured appearance makes them difficult to differentiate.
- **Complex motion.** A varying number of body parts are observed in consecutive video frames (e.g. the bee tongue does not appear in Fig. 1c), thus we have incoherent motion paths and the trajectories have long tracking gaps. Motion cues furthermore provide only little information to predict the current pose.

In the videos to be analyzed, the insect will be fed with sugar water by a stick and it may respond by extending its tongue. It is required to infer the presence of the tongue before localizing it. Similarly, an antenna may be absent when it moves above the head (Fig. 1d) or suffers from heavy motion blur (Fig. 1e). As pointed out in [6], the state-of-the-art tracking algorithms do not perform well when applied to our task. In [6], a track linking approach was used for estimating landmarks in merge conditions, i.e. the tips of two body parts are bounded within the same bounding box (BB). In contrast to our fully automated method, the approach in [6] is a multi-frame pose estimation framework, and it requires additional human intervention to rectify probable erroneous hypotheses at some frames. In this paper, we formulate this problem as inferring a mapping from the appearance of an insect to its corresponding pose. We focus on single-image pose estimation, because this strategy could further improve the performance of a multi-frame framework as per-frame initialization and recovery.

Some related pose estimation techniques have been proposed, such as model based methods [4,7] and Random Forest (RF) regression methods [8]. There are some recently emerged studies on pose estimation of humans [9], heads [10], and hands [11], as well as medical image analysis for localization of landmarks [12,13]. The success of these RF regression methods comes from two factors. One is the discriminative power of RF model. The other factor is the strategy of localizing landmarks by estimating its relative displacements with regards to other image patches, making it suitable for highly structured objects but not for our case. For example, these relative displacements do not change dramatically in medical images in 2D grids. Similarly, relative positions between joints of body parts are relatively stable for humans, heads or hands in 3D coordinates, so that these methods work well with depth images [9–11]. Our task is more challenging, due to the large variance of relative position between landmarks in 2D grids. Besides, it is difficult to infer the configuration (pose) with varying number of landmarks based on local appearance.

To address these aforementioned issues, we present a unified framework that incorporates the geometric model as the prior, and utilizes the RF model to estimate the possible positions of body part in pixel precision. Under this framework, the maximum a posteriori (MAP) estimation is found as the landmarks positions.

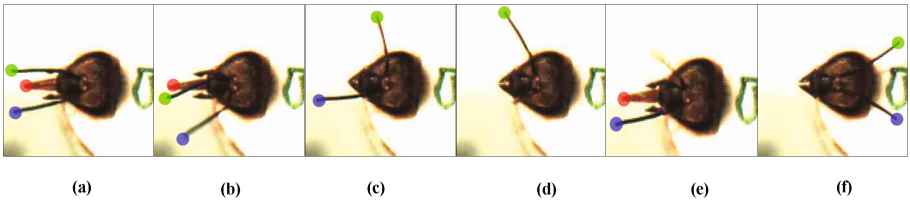

Fig. 1. Example frames of various poses and the outputs by our method. The tips of three body parts are marked in different colors: a blue circle represents for the left antenna, red for the tongue and green for the right antenna. (a) all body parts are present; (b) the tongue is partially occluded by the right antenna; (c)–(e) some body parts are absent; (f) in some rare cases, the antennae move backwards. (Color figure online)

The main contributions of our work are:

- To the best of our knowledge, the proposed framework is the first method for detailed insect pose estimation from a single frame.
- This is an exemplar of random forest based pose estimator with data-driven regularization. Given the landmark candidate positions predicted by Random Forests, we propose a data-driven approach to adaptively weight and select precise landmark positions, incorporating probable global structure of the anatomy to be estimated.
- We benchmark our approach on a set of large challenging datasets and make it publicly available for future studies.

2 Related Work

Animal tracking. Recently, a number of tracking techniques emerged in biology for tracking various types of animals [1]. Particle filtering is used in some insect tracking algorithms to maintain the identity of objects throughout a whole video sequence [2,14,15]. However, as pointed out in [16], particle filtering is often only effective for short tracking gaps and the search space becomes significantly larger for long gaps. Similarly, data association techniques that have been applied in [17] are also not able to tackle the tracking gaps. To develop a more efficient algorithm, some studies incorporate higher level attributes that characterize specific insect motion into a learning diagram. In [18], overlapping larvae are separated by assigning object labels to each pixel, given user-annotated examples of encounters of two larvae as boundary conditions. For modeling occluded spatiotemporal regions, dozens of examples of encounters of two larvae need to be selected. A behavior model is proposed in [14] by firstly abstracting local motions and by modeling the behavior as a dynamical model on such local motions. However, the Markov model used in [14] for behavior limits its applications to some latest behavioral studies which require multi-target tracking, because the number of

parameters to specify the transition and observation models is exponential in the number of moving objects.

Pose estimation. While some animal tracking algorithms also provide some information about pose (e.g. orientation of the animal) based on contour models [4], few works have been done for directly addressing the pose estimation problem. Given a rough initial guess of the pose parameters, a cascaded pose regression is proposed to progressively refine the pose estimation until it converges [7]. Our task is more challenging in that we have to address the self-occlusion problem. Moreover, instead of an iterative solution, we present a one-pass algorithm. Different from [6] and [7], our algorithm does not require an initial guess of the target position.

Random Forest based pose estimator/landmark detector. Random Forests (RF) [19] describe an ensemble of decision trees trained independently on a randomized selection of features. Recent works on Hough Forests [8] have shown that objects can be effectively located by training RF regressors to predict the position of a point relative to the sampled region, then running the regressors over a region and accumulating votes for a likely position. In these studies, the final estimation is found as the center of the densest vote mass by mean shift [20]. If the objects are occluded or missing on the test image, the candidates of the target point returned by the trained RF regressor will contain outliers. Therefore, it is crucial to remove such outliers to guarantee a reasonable estimation under the condition of occlusion. In [10], the authors assume the pose parameters stored in a single leaf will follow a Gaussian distribution, and discard the leaves with high variances by simple thresholding when performing mean shift. Similarly, the hand pose estimator in [11] is refined by a mean shift based method to recover the poorly detected joints when they are occluded or missing. Besides the occlusion problems our input images have also less informative visual features for disambiguating objects with similar appearance. As pointed out in [12], landmark detectors based on classification may produce highly interchangeable responses due to very similar local appearance patterns of different anatomical body (sub)parts. Further disambiguation is required that incorporates the global structure of objects. Existing works on medical image analysis for localization of landmarks exploit global landmark relation represented by either repetitive anatomical patterns [12] or shape models [13] for regularization. In our task, however, such stable global landmark relation is not applicable. In this paper, we propose a method to learn the global landmark relation from training data based on a graph based geometric model.

3 Problem Statement

In this paper, we aim to estimate the pose $\mathbf{X} = \{\mathbf{x}^n | 1 \leq n \leq N, \forall \mathbf{x}^n \in \Re^5\}$ from a single image I, where the state of part n is defined as $\mathbf{x}^n = \{\mathbf{y}^n, \theta^n, s^n, t^n\}$. $\mathbf{y}^n$ is the position of the landmark in our defined coordinate system as shown in Fig. 2, θ^n is the angle, and s^n is the scale. Specifically, $t^n = \{0, 1\}$ is the

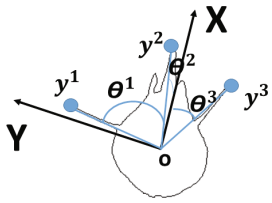

Fig. 2. The coordinate is defined by setting the centroid of head as the origin **o**, the line from the origin towards the mouth center as the x-axis. The head centroid and the mouth center are assumed to be known.

state indicating the presence of part n. The images are acquired from the top view of observed individual insects. Six example frames of an insect images with the outputs of our method are shown in Fig. 1 to visualize various insect poses. Taking a bee for example, we aim to estimate the tips of its three body parts: left antenna (blue circle), tongue (red circle) and right antenna (green circle). Figure 1b illustrates that these body parts are highly clustered, and self-occlusion usually occurs. Body parts may be present or absent. The antennae usually move forwards while occasionally backwards.

4 Combined Landmark Position Proposals

To map the image I to the corresponding pose $\mathbf{X}$ is difficult because the mapping from visual input to poses is highly complicated, our framework imposes constraints to the solution space based on two cues: local appearance and global structure. On one hand, we use RF model for predicting the class label c ($c = \{1$: right antenna, 2: left antenna, 3: tongue$\}$ and 0 as background.) at pixel precision based on patch appearance. On the other hand, we propose a method based on Pictorial Structure (PS) model [21,22] to learn the global structure. We do *not* make any specific assumption about the anatomical model of an insect's head (which was done in [6]), instead we use a common assumption that holds for generic objects: for the same type of insects, they have similar appearances when they are in similar poses. The global structure of an insect is represented by its silhouette, which is a d-dimensional datapoint $\mathbf{f} \in \Re^d$. Based on this assumption, it is expected that these datapoints will lie on or near a low dimensional manifold, in which the neighborhood of each datapoint is preserved. The likelihood of global structure of the unknown pose is learned by the nearest neighbor (NN) method, and used as a constraint to regularize the mapping estimated by the RF model.

Given the image I, the posterior probability $p(\mathbf{X}|I)$ is computed as

$$p(\mathbf{X}|I) \propto p(I|\mathbf{X})p(\mathbf{X}) \tag{1}$$

where $p(I|\mathbf{X})$ is the likelihood of the image evidence given a particular pose, and the $p(\mathbf{X})$ corresponds to a tree prior according to the PS model. Both terms

are learned from training data. Specifically we propose a method to adaptively construct the graph for the insect geometric model.

4.1 Geometric Model

Typical PS models assume that, an object can be decomposed into parts connected with pairwise constraints that define the prior probability of part configurations. As we aim to estimate the tips of body parts appended to the insect's head, we build a complete graph based on the pairwise relations between tips as well as each tip $\mathbf{x}^n$ and the centroid of the head $\mathbf{x}^0$. Then the PS model $G = (V, E)$ is learned from the training data by computing the minimum spanning tree of a graph. The resultant E is the set of pairs of each tip and the centroid, and the centroid is the root. Based on the assumption that part likelihoods are conditionally independent [21], Eq. (1) is factorized as

$$p(\mathbf{X}|I) \propto p(\mathbf{x}^0) \prod_{n=0}^{N} p(\mathbf{r}^n|\mathbf{x}^n) \prod_{(i,j)\in E} p(\mathbf{x}^i, \mathbf{x}^j|\Phi_{ij}) \tag{2}$$

where $p(\mathbf{r}^n|\mathbf{x}^n)$ is the likelihood of position based on local appearance $\mathbf{r}^n$. The joint probability $p(\mathbf{x}^i, \mathbf{x}^j|\Phi_{ij})$ indicates the spatial relations between $\mathbf{x}^i$ and $\mathbf{x}^j$ with parameters Φ_{ij}.

Although we model the global structure as the form of Eq. (2), it is different from typical PS models that assume all parts are present. As we focus on pose estimation rather than detecting the insect head, $\mathbf{x}^0$ is assumed to be known. To represent the global structure of a pose, we extract a feature vector $\mathbf{f} \in \Re^d$ for each image combining five types of silhouette features: Edge Histogram Descriptor [23], the Geometrical Feature [24], the Shape Signature Histogram [25], the Fourier Descriptor [26] and the Hu moments [27]. The dimensionality of $\mathbf{f}$ is $d = 88$, thus no dimensionality reduction is required. For a new visual input represented by silhouette features $\mathbf{f}$ of image I, we find its K neighbors in the visual feature space and construct the neighborhood $\mathbf{K}$ in the pose space.

As the nodes of landmarks of G are all leafs, and the likelihood of individual landmarks are conditionally independent, the node representing $\mathbf{x}^n$ will be removed if the frequency of the presence of $\mathbf{x}^n \in \mathbf{K}$ is lower than a threshold value τ. The new graph $G^* = (V^*, E^*)$ remains a tree structure. Based on G^*, Eq. (2) is changed to

$$p(\mathbf{X}|I) \propto \prod_{n\in E^*} p(\mathbf{r}^n|\mathbf{x}^n) p(\mathbf{x}^n, \mathbf{x}^0|\Phi_{n0}) \tag{3}$$

We assume that Φ_{n0} takes the form of unimodal Gaussian distribution over $\mathbf{y}^n, s^n$, and the von Mises distribution over θ^n. Thus, the model parameters $(\boldsymbol{\mu}_{\mathbf{y}}^n, \Sigma_{\mathbf{y}}^n)$, (μ_s^n, σ_s^n) and $(\mu_\theta^n, \sigma_\theta^n)$ are learned from the training samples in $\mathbf{K}$. The joint probability is computed as

$$p(\mathbf{x}^n, \mathbf{x}^0|\Phi_{n0}) = \mathcal{N}(\mathbf{y}^n|\boldsymbol{\mu}_{\mathbf{y}}^n, \Sigma_{\mathbf{y}}^n)\mathcal{N}(s^n|\mu_s^n, \sigma_s^n)\mathcal{M}(\theta^n|\mu_\theta^n, \sigma_\theta^n) \tag{4}$$

4.2 Random Forest Based Classifier

We use RF model to compute the likelihood of landmark position based on local appearance evidence $p(\mathbf{r}^n|\mathbf{x}^n)$. Taking bee images for example, each datapoint in the training dataset D is an image patch ($I(\mathbf{y}_i)$, where $\mathbf{y}_i$ is the image coordinate) sampled in the following way: We randomly sample the patches with centroids located along the contour (see the light blue contours in Fig. 3) as examples of corresponding class c and the patches with centers inside or outside the contour as examples of background. A class labels c is assigned to every datapoint. As shown in Fig. 3a, a class label $c \in [0, N]$ of each patch (colored square) is the index of its closest tip (colored circle) along the contour: {1: right antenna, 2: left antenna, 3: tongue} and 0 as background. A similar sampling strategy is applied for ant images, as shown in Fig. 3c. $N = 3$ for a bee and $N = 2$ for an ant. We classify left or right antenna to balance the distribution of classes, since the samples of class $c = 3$ are much fewer than the class $c = 1$ or $c = 2$.

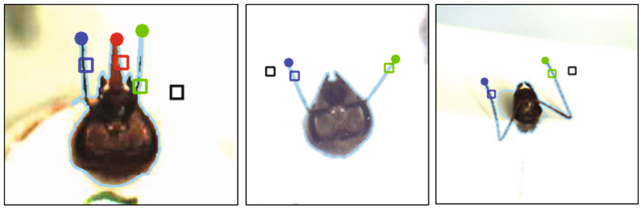

Fig. 3. The image sizes of Dataset A, B and C are (a) 275 × 235, (b) 350 × 320 and (c) 415 × 420 in pixels, respectively. The patch size (denoted by a square) is 16 × 16 pixels. The images have been scaled for better visualization.

To train the forest R, we randomly extract patches from D and use the information gain criterion to select the split function. The split function of a node is represented by a simple two-pixel test as in [9]. The forest R is constructed, with each leaf L_j created when the maximum depth is reached or a minimum number of patches are left. Each leaf stores the patches from D that end here. Each patch of a given test image I passes down a tree and ends in a leaf $L_j(\mathbf{y}_i)$, which gives the class probabilities $p(c|I(\mathbf{y}_i))$. A class label c is assigned to each pixel with the highest $p(c|I(\mathbf{y}_i))$, and $p(\mathbf{r}^c|\mathbf{x}^c)$ is set to be 1. Specifically, we set $p(\mathbf{r}^c|\mathbf{x}^c) = 1, \forall c = 1, 2$, since the RF model may not correctly differentiate the left or the right antenna. We will address this problem by solving Eq. (3).

4.3 Final Landmark Localization

According to Eq. (3), the posterior probability of configuration $p(\mathbf{X}|I)$ is computed as the product of the posterior probability of the landmarks $n \in E^*$. We construct a response image by computing the posterior probability for each of

these landmarks at an image coordinate. To localize a landmark, a simple strategy by selecting the pixel with highest probability may fail due to the outlying pixels. Instead, we use mean shift [28] with a flat kernel to find the modes of probability mass for each part, and assume that it lies in the connected component that contains the landmark. Finally, the landmark is simply estimated as the furthest pixel within the connected component.

5 Experiments

As pointed out in [6], the state-of-the-art tracking algorithms do not perform well when applied to our task. In this experiment, as our method combines the strength of the traditional regression forest [8] and Pictorial Structure model, it is compared with the methods directly applying these two concepts.

5.1 Datasets and Evaluation Metric

In this experiment, our method was evaluated on three challenging datasets of individual insects (i.e. bees and ants) during a behavioral experiment, among which two datasets of individual bees are recorded in different light conditions or other experimental settings. For example, Dataset A contains images from a video recording a bee in different trials of experiments, while Dataset B is of various bees in different trials. The image data comes from our biological partner:

- Dataset A (bee): 5633 training images, 2788 testing images
- Dataset B (bee): 3625 training images, 9003 testing images
- Dataset C (ant): 215 training images, 238 testing images

The spatial resolution is 39 pixels per μm for bee images and 22 pixels per μm for ant images. More details about the three datasets are shown in Fig. 3.

As our method directly estimates the position of landmarks, it does not produce bounding box (BB) hypotheses. Some popular pose estimation metrics, such as average precision of keypoints (AFK) [29], require ground truth BBs for evaluation, thus not suitable for our method. Results of our method are compared with the two aforementioned methods as well as ground truth landmark positions, which are manually annotated by a human. We compute the rate of false positives (FP) and false negatives (FN) of inferring the presence of each landmark to validate the adequacy of our geometric model, e.g. a FP of the tongue indicates that the tongue is inferred to be present while it is absent actually. The accuracy of localization is measured by the average Euclidean distance in pixels between the results and the groundtruth.

5.2 Implementation Details

For learning the geometric model, we found $K = 100$ nearest neighbors to construct $\mathbf{K}$ in Sect. 4.1. We construct 10 trees for the RF and each tree has the

depth of 5. Each tree converges until the maximum tree depth is reached or the amount of remaining patches is less than 50. The bandwidth of meanshift kernel is set as 0.05 for all images.

The complexity of the algorithm is measured by the processing time. For constructing a random forest, we use the code of Hough Forest [8]. Using a Matlab implementation of our method, testing takes around 4 seconds per frame on an Intel i7 machine with 8 GB RAM. It takes about 1 s for RF classification, and 3 s for computing Eq. (3) and final landmark localization.

5.3 Results and Discussion

For quantitative evaluation, Table 1 shows the average position error (pixels) in three datasets. The average position errors of each part are merely 10.2, 5.3 and 18.0 pixels, respectively. They are rather small compared to the size of an insect head (shown as Fig. 3). The position error on Dataset C is larger than the others because ant images have more severe motion blur (e.g. the right antenna in Fig. 5f), and the exact positions of landmarks are ambiguous.

Table 1. Quantitative evaluation on three datasets.

Datasets	Position error (pixels)	FN (%)	FP (%)
A	10.2	5	0
B	5.3	3	0.2
C	18.0	0	0

To validate the advantage of our method over RF and the typical PS model, we compare the three methods on Dataset A, which is the most challenging one due to the complex background. The typical PS model in Eq. (2) assumes that all landmarks are present, and the spatial relations between parts are learned from all the training samples. As shown in Table 2, this method produces a high FP rate when inferring the presence of the tongue. With the learned geometric model in Eq. (3), our method achieves a significant improvement over both the PS model and the RF regressor in terms of both the localization precision and the ability to disambiguate different objects.

Table 2. Comparison of three methods on dataset A.

Methods	L. antenna			Tongue			R. antenna		
	pos. error	FN(%)	FP(%)	pos. error	FN(%)	FP(%)	pos. error	FN(%)	FP(%)
Ours	10.2	4	0	14.9	11	0	8.4	3	0
PS	17.6	2	0	18.4	4	13	16.7	4	0
RF	24.5	2	0	26.3	26	44	25.3	8	0

Figure 4 shows the localization errors of each landmark from Dataset A for a more detailed discussion. We show nine image sequences where either the RF or the PS method produces large localization errors, while our method achieves very low position errors in most of the frames.

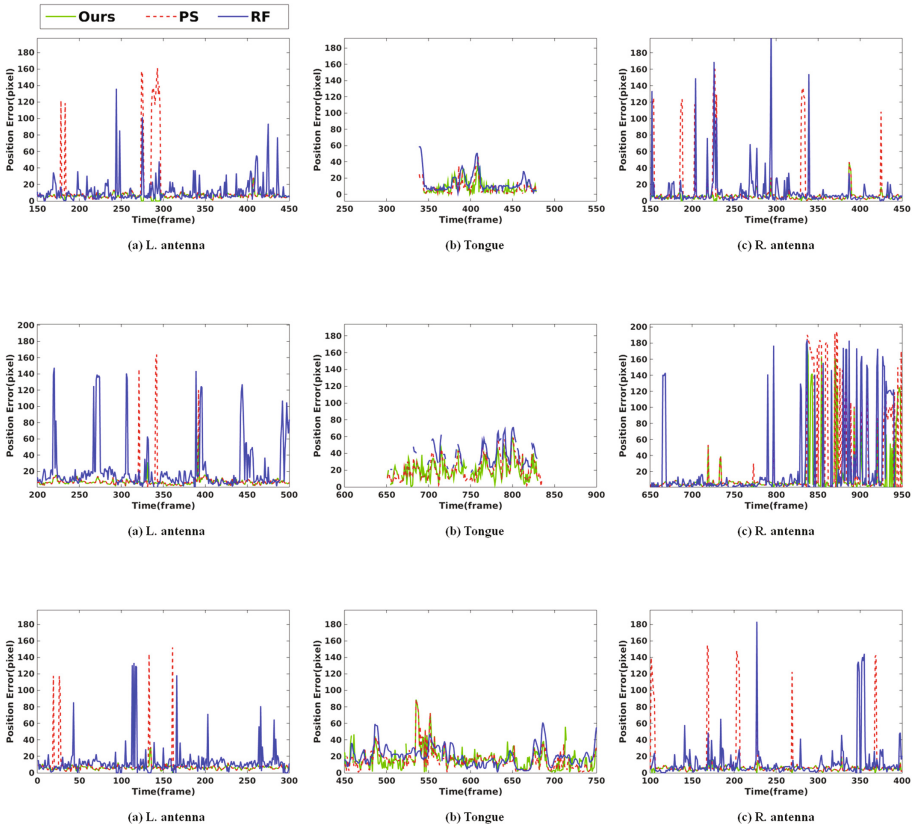

Fig. 4. Quantitative results by our method (green), RF regressor (blue) and PS model (red). (Colour figure online)

We visualize more results in Figs. 5 and 6 to discuss the advantage of our method in more details. As shown in Fig. 5a and e, our method successfully localizes the two antennae even in complex background, while the RF regressor fails to distinguish the left antenna from the noise of background and thus produces high position errors. Besides, the RF regressor may incorrectly recognize an antenna as the tongue (as shown in Fig. 5c–d), indicating a high FP rate of the tongue in Table 2. In contrast, our method is able to disambiguate the tongue and the right antenna even when they are very close to each other (see Fig. 5a).

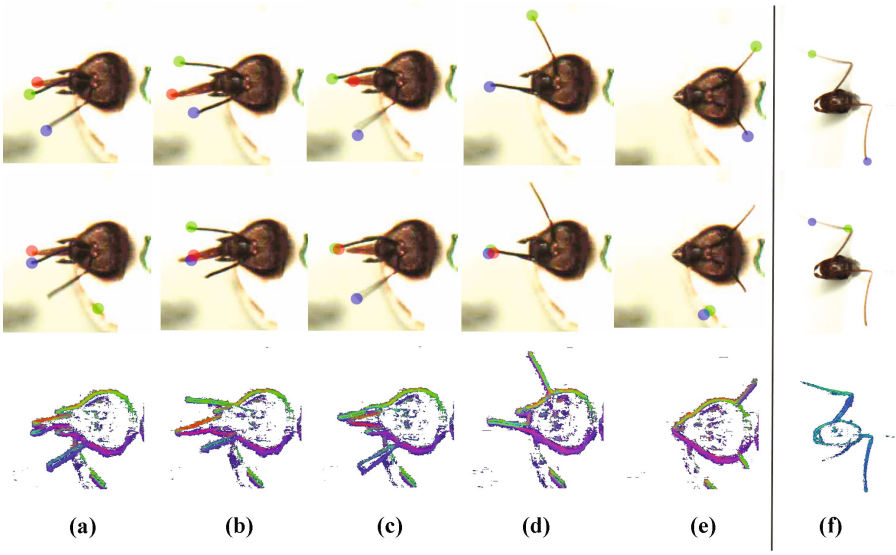

Fig. 5. Qualitative results by our method (the first row), RF regressor (the second row) and RF classification (the bottom row).

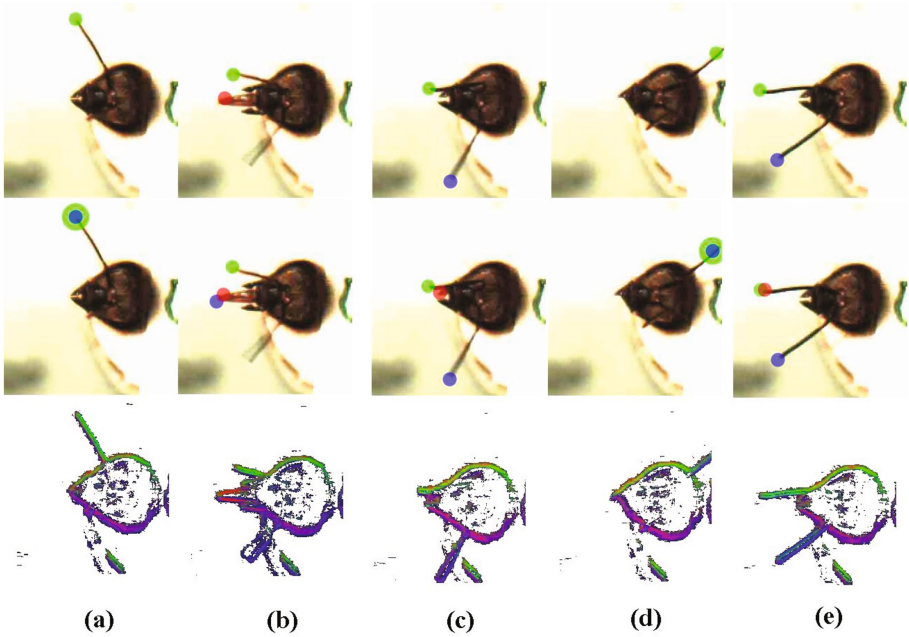

Fig. 6. Qualitative results by our method (the first row), PS model (the second row) and RF classification (the bottom row).

Figure 6 shows the advantage over typical PS models in inferring the number of landmarks present. Without inferring the presence of a landmark, it will be localized at some position. A naïve approach for rejecting the potentially incorrect position is that a landmark falling inside the region of insect head will be rejected, e.g. the tongue has been rejected in Fig. 6a and d. But it cannot deal with more general cases as shown in Fig. 6b and e. Moreover, the spatial relations between parts learned from all training data provides little information in our case, since the possible positions of antennae in all training images are nearly uniformly distributed. As shown in Fig. 6a and d, the left antenna is incorrectly located in the right antenna. In contrast, our method is capable of inferring the absence of a landmark by the learned geometric model.

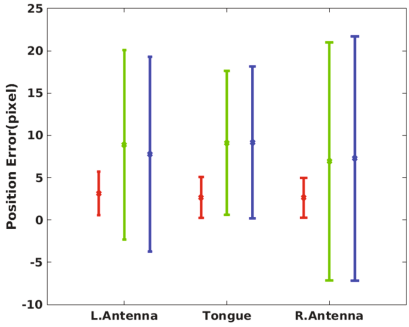

Fig. 7. Comparison between human annotators and our method: the means and standard deviations of the Euclidean distance between the landmark positions of human annotator A vs. the results of our method (blue), human annotator B vs. our method (green) and between human raters only (red). (Colour figure online)

As for the running time of the three methods, our method is the fastest. The PS model takes 8 s for computing Eq. (2), while ours only takes 3 s for Eq. (3).

To validate that the localization precision of our method are comparable to human annotators, we also compare the pixel error between two annotators and our method. The means and standard deviations of the Euclidean distance between the landmark positions of human annotator A vs. the results of our method (blue), human annotator B vs. our method (green) and between human raters only (red), are illustrated in Fig. 7. The results show that the estimation of our method is comparably accurate as human annotators in most cases.

6 Conclussion and Future Work

In summary, we presented a new algorithm exploiting local appearance and global geometric structure of an insect to infer its pose from a single image. Our method is a data-driven approach to incorporate geometric constraints. The model parameters are learned from the training data. Our method addresses

the issue of interchangeable estimations by solely using RF model for landmark detection, and presents nice interplay between RF and PS model. The performance of our method has been validated on three large challenging datasets of different types of insects, which achieves comparable position accuracy to that of human annotators.

Future work includes incorporating our method into a multi-frame pose estimation framework. Given the high accuracy of pose estimated based on single frames by our method, merging it into an interactive tracking framework such as [6] could result in a new approach for insect behavioral analysis.

References

1. Dell, A.I., Bender, J.A., Branson, K., Couzin, I.D., de Polavieja, G.G., Noldus, L.P., Pérez-Escudero, A., Perona, P., Straw, A.D., Wikelski, M., et al.: Automated image-based tracking and its application in ecology. Trends Ecol. Evol. **29**(7), 417–428 (2014)
2. Branson, K., Belongie, S.: Tracking multiple mouse contours (without too many samples). In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 1039–1046. IEEE (2005)
3. Khan, Z., Balch, T., Dellaert, F.: MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 1960–1972 (2006)
4. Branson, K., Robie, A.A., Bender, J., Perona, P., Dickinson, M.H.: High-throughput ethomics in large groups of drosophila. Nat. Method **6**(6), 451–457 (2009)
5. Huston, S.J., Stopfer, M., Cassenaer, S., Aldworth, Z.N., Laurent, G.: Neural encoding of odors during active sampling and in turbulent plumes. Neuron **88**(2), 403–418 (2015)
6. Shen, M., Li, C., Huang, W., Szyszka, P., Shirahama, K., Grzegorzek, M., Merhof, D., Duessen, O.: Interactive tracking of insect posture. Pattern Recogn. **48**(11), 3560–3571 (2015)
7. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085. IEEE (2010)
8. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough forests for object detection, tracking, and action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11), 2188–2202 (2011)
9. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
10. Fanelli, G., Gall, J., Van Gool, L.: Real time head pose estimation with random regression forests. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 617–624. IEEE (2011)
11. Tang, D., Yu, T.H., Kim, T.K.: Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In: IEEE International Conference on Computer Vision (ICCV), pp. 3224–3231. IEEE (2013)
12. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. Med. Image Anal. **17**(8), 1304–1314 (2013)

13. Chen, C., Xie, W., Franke, J., Grutzner, P., Nolte, L.P., Zheng, G.: Automatic x-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. Med. Image Anal. **18**(3), 487–499 (2014)
14. Veeraraghavan, A., Chellappa, R., Srinivasan, M.: Shape and behavior encoded tracking of bee dances. IEEE Trans. Pattern Anal. Mach. Intell. **3**, 463–476 (2008)
15. Landgraf, T., Rojas, R.: Tracking honey bee dances from sparse optical flow fields. FB Mathematik und Informatik FU, pp. 1–37 (2007)
16. Perera, A., Srinivas, C., Hoogs, A., Brooksby, G., Hu, W.: Multi-object tracking through simultaneous long occlusions and split-merge conditions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 666–673 (2006)
17. Balch, T., Khan, Z., Veloso, M.: Automatically tracking and analyzing the behavior of live insect colonies. In: Proceedings of the Fifth International Conference on Autonomous Agents, pp. 521–528. ACM (2001)
18. Fiaschi, L., Diego, F., Gregor, K., Schiegg, M., Koethe, U., Zlatic, M., Hamprecht, F., et al.: Tracking indistinguishable translucent objects over time using weakly supervised structured learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2736–2743. IEEE (2014)
19. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
20. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. **17**(8), 790–799 (1995)
21. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: people detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1014–1021. IEEE (2009)
22. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
23. Frigui, H., Gader, P.: Detection and discrimination of land mines in ground-penetrating radar based on edge histogram descriptors and a possibilistic K-Nearest neighbor classifier. Fuzzy Syst. **17**(1), 185–199 (2011)
24. Li, S.Z.: Shape matching based on invariants. In: Omidvar, O. (ed.) Shape Analysis, Progress in Neural Networks, pp. 203–228. Ablex, Norwood (1999)
25. Zhang, D., Liu, G.: Review of shape representation and description techniques. Pattern Recogn. **37**(1), 1–19 (2004)
26. Zhang, D., Lu, G.: A comparative study of curvature scale space and fourier descriptors for shape-based image retrieval. J. Visual Commun. Image Represent. **14**(1), 39–57 (2003)
27. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Trans. Inf. Theor. **8**(2), 179–187 (1962)
28. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
29. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2878–2890 (2013)