Multi-view Anatomical Animal Landmark Localization using Deep Feature Regression

Oliver Mothes
Computer Vision Group
Friedrich Schiller University Jena, Germany
Email: oliver.mothes@uni-jena.de

Joachim Denzler
Computer Vision Group
Friedrich Schiller University Jena, Germany
Email: joachim.denzler@uni-jena.de

***Abstract*—For animal locomotion analysis, biological experts have to evaluate an immense amount of recorded image data. The time-consuming annotation of important anatomical landmarks has to be done in every single recorded image to analyze different gaits or types of movement. In this paper, we introduce a method to reduce this effort by automating the annotation process with a minimum level of expert interaction. In contrast to recent approaches based on Augmented Active Appearance Models, our approach can deal with tracking single anatomical landmarks in non-cyclic locomotion sequences. Additionally, our approach is independent of anatomical knowledge. We evaluate our method on a variety of datasets and show that we achieve a performance comparable to that of biological experts.**

I. Introduction

Locomotion analysis is of great importance for zoologists, motion scientists, or in the domain of humanoid robotics to conduct profound investigations about different gaits and types of movement. In some cases, special video recording procedures are required for such studies. For a detailed understanding of bone movements, X-ray acquisition systems are used and applied to animals to see the inner bones of the locomotor system. A C-arm X-ray acquisition system with two perpendicular detectors providing a top view (*dorsoventral* view) and a side view (*lateral* view) image is used to provide an all-round-view around the entire locomotor system. In order to guarantee a detailed biological evaluation, a high spatial and temporal image resolution (1250 × 1250 pixels at 1000 FPS) is essential for acquisition, while a locomotion sequence has up to 2000 frames. The resulting immense amount of data incurs considerable expenses in terms of the evaluation for the biological experts, which have to manually annotate anatomical important points, so-called landmarks, in the individual recorded frames [15], [2], [14]. An automation of this task is of great interest to biologists to avoid the time-consuming manual annotation. In Figure 1, a recorded frame of the two different views with annotated corresponding landmarks is illustrated.

Motivated by the time-consuming task of manual annotation, we propose an automatic anatomical landmark localization approach for animal locomotion data by exploiting CNN layer activations as features for a landmark regression problem. In contrast to other recent approaches, we show in our experiments that we can deal with all available non-cyclic locomotion sequences, which was only evaluable for a small amount of sequences with the state-of-the-art approach. Opposed to cyclic locomotion sequences, where the animals are walking on a barrier-free treadmill, in non-cylic locomotion sequences the animals are running across a track with obstacles on it. In addition, our approach requires not more training data than existing state-of-the-art methods.

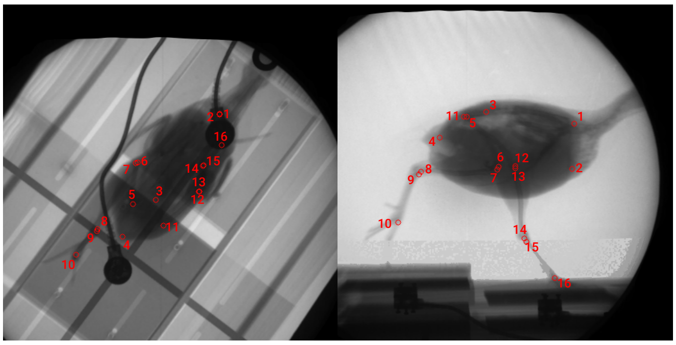

Fig. 1: The birds running on a track are recorded by a C-arm X-ray acquisition system providing a top view (left) and a side view (right). For a detailed understanding of the locomotion on ground, biological experts have to analyze anatomical landmarks of every single frame.

The remainder of the paper is structured as follows. A brief overview of related work is given in Section II. In Section III we introduce our automated landmark localization approach, followed by several experiments in Section IV. Finally, Section V concludes the paper with a short discussion.

II. Related Work

For motion analysis Haase and Denzler [8] apply *Active Appearance Models* [4] to X-ray animal locomotion datasets for tracking anatomical landmarks. Their results show that the generative model fits suitably using only a small amount of training data for AAM training. Unfortunately, standard AAMs show weaknesses with respect to certain landmark subsets, especially for landmark subsets of the lower limb system of the animals. An extension of their approach [11] to multi-view AAMs only shows a further improvement of the tracking performance of the animals torso landmark subsets compared to the single view AAM.

Motivated by the shortcomings, two more extensions [10], [14] lead to a holistic model with different constraints, especially for the lower limb landmarks. The various constraints support the multi-view AAM during the landmark fitting. However, the anatomical landmark tracking of the probabilistic Augmented AAM is limited to cyclic locomotion sequences of birds running completely inside the scene.

In contrast, we suggest a framework for anatomical landmark regression which can deal with X-ray video sequences for cyclic and non-cyclic locomotion even if not the whole bird is visible in the sequence. Hence, it is possible to track single landmarks as they enter the scene until they disappear. Furthermore, our regression framework does not depend on application-specific or anatomical knowledge, but it is possible to use context knowledge for improving the tracking results.

III. Method

This section describes the automatic landmark localization technique. While the first sub-section deals with the deep feature representation of the input image, the second subsection explains the landmark regression task using these powerful features.

A. Deep Feature Adjustment

In tasks where little training data is available, it is in most cases not possible to train a CNN for the desired task. Instead, a pre-trained CNN can be exploited for feature extraction from the early layers' activations [6], where afterwards these features can be used for training models using other classification or regression methods like Support Vector Machines [5] or Random Forests [3].

These activation features are called deep features or CNN features. In our landmark localization approach, we use an AlexNet architecture [13] trained on the ImageNet LSVRC-2012 dataset [16] for classifying 1000 different objects from everyday life to extract these features. To make the deep features more representative for our X-ray recorded image data, we use a domain adaption technique called *fine-tuning* [7] before. The parameters of several layers of the pre-trained AlexNet model are used as initial parameters of a new model with the same architecture. Afterwards, the CNN is trained for an auxiliary task for which more data of the same image type is available. Finally, the features are more suitable for images from the target image domain.

B. Multi-view Landmark Regression

After extracting the deep features of the L training images (both images of lateral and dorsoventral view) from one of the CNN layers, a linear model for regression can be trained jointly with the training landmark positions. The L training image sets are propagated sequentially through the fine-tuned pose classification CNN. Afterwards, the extracted deep features of both views are concatenated. As linear model we train an ϵ-SV regression [17]. The linear regression model uses the given training data $(x_1, y_1), ...(x_L, y_L) \subset \mathcal{X} \times \mathbb{R}^4$, where x_i denotes the concatenated deep features with $\mathcal{X} = \mathbb{R}^d$ and y_i a landmark position of the dorsoventral and lateral view. The goal of this regression task is to find a hyperplane $f(x) = \langle w, x\rangle + b$ with maximum deviation of ϵ from the target values y_i for all training data. Given the fact that the vector w is perpendicular to the hyperplane $f(x)$, we only need to minimize the norm of w, i.e. $||w||^2 = \langle w, w\rangle$. When working with real data in most cases it is not possible to find a decent solution for this convex optimization problem based on potential outliers. Using slack variables ξ_i and ξ_i^*, such infeasible conditions can be handled. We formulate the problem like [17]:

$$\underset{w,b,\xi_i,\xi_i^*}{\operatorname{argmin}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{L}(\xi_i + \xi_i^*)$$
$$s.t. \begin{cases} y_i - \langle w, x_i\rangle - b & \leq \epsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \tag{1}$$

where $C > 0$ is a constant.

For every landmark position pair of the dorsoventral and lateral view $y_i = (y_i^{d_1}, y_i^{d_2}, y_i^{l_1}, y_i^{l_2})$ we train a single regressor model. Hence, it is possible to track single landmarks through the whole scene without relying on other landmarks, in contrast to other methods [8], [11], [10].

IV. Experiments

In the following section, we evaluate the performance of our anatomical landmark localization approach using the introduced deep feature regression. For our landmark localization experiments, we used 39 bird locomotion sequences with non-cyclic movements of quails running across a track with obstacles on it. The birds have to overcome either a step up (a step of 2.5 Centimeter) or step down (holes of 1 Centimeter, 2.5 Centimeter and 5 Centimeter). While running, the quails were recorded by a high-speed X-ray acquisition system at 1000 Hz and a resolution of 1250 × 1250 pixels. Additionally, the acquisition system records from two different views, the dorsoventral view from above and the lateral view from the side. In Section IV-A we propose a feature adjustment technique to make the features for our landmark regression approach more representative. Afterwards, results of our multi-view landmark localization approach are shown in Section IV-B, followed by a 3D reconstruction [12] of the landmarks in Section IV-C.

A. Bipedal Locomotion Pose Quantization

X-ray recorded images are very different in appearance from natural images. We assume that extracted deep features of such low-contrast gray value images are not representative enough, due to the different image domain the CNN is trained on. Hence, we define an auxiliary task to guarantee a domain shift to our input data. For that, we fine-tune the pre-trained Alexnet CNN for bird pose classification. The deep features extracted from that fine-tuned model provide more representative features of our input data compared to deep features from the original AlexNet model. To define bird poses we use one element of *Active Appearance Models* (AAMs) [4]. The first parameter of the shape component of an AAM,

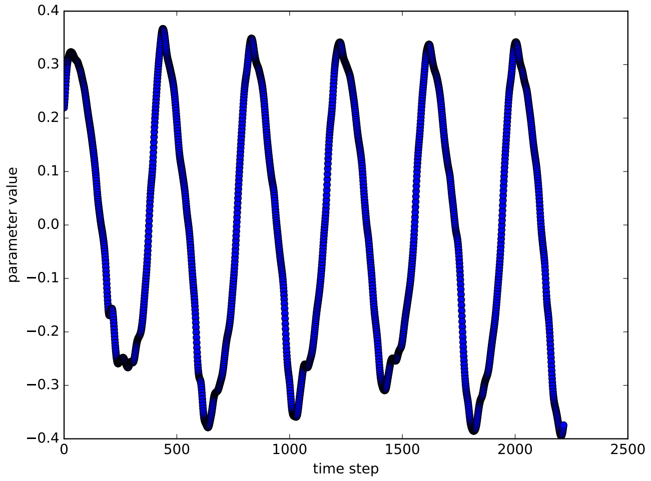

Fig. 2: With the first parameter of the shape component of an *Active Appearance Model* we can obtain the cyclicality of the cyclic locomotion datasets, which we can use to quantize bipedal locomotion poses.

computed by *Principle Component Analysis* (PCA), captures the largest landmark shape variance of the training data and influences the pose of the walking bird. To compute the shape models of single sequences, we use 2508 available frames of cyclic datasets with annotated landmarks [10] of both views. For every cyclic sequence we trained a multi-view shape model using 15 annotated frames. Afterwards, the first parameter of the shape component is extracted. In Figure 2, the locomotion cyclicality of a chosen cyclic dataset is shown. Based on the parameter value, frames are quantized into different numbers of pose classes. The CNN models are trained together with images of the dorsoventral view and lateral view. The results of the pose classification of different numbers of quantized pose classes (4,10,25,50,100 and 200 classes) show a decreasing accuracy with the increasing number of quantized pose classes. In the next sub-section we want to investigate how suitable the different trained models are for feature extraction.

B. Landmark Regression

We apply our landmark localization method of Section IV-B to 39 non-cyclic locomotion sequences containing 36348 frame sets (dorsoventral and lateral view) where 8088 frames are annotated. For training we use 15 labeled frames per sequence, like the state-of-the-art methods. The rest of the annotated frames of the sequence are for validation. As an initial investigation we evaluate the features of the CNNs trained to classify a different number of quantized bird locomotion poses. We trained with the different features of AlexNet's *conv5*-layer single landmark regressors. In Figure 3 we combined the localization errors of the results for the single CNN models for both views. It can be clearly seen that the number of pose classes has little influence on the quality of the features for regression. A closer look reveals that the model performs best with 10 pose classes which we use for further experiments. On average we obtain an Euclidean error of 10 pixels, which is comparable to the average error of human experts [9].

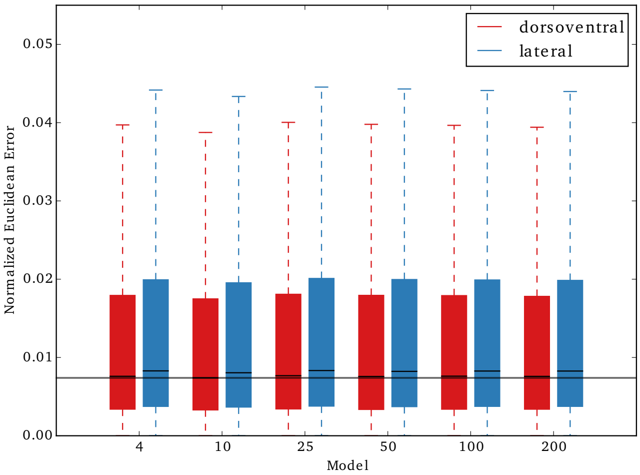

Fig. 3: The comparison of the CNN models trained to classify different numbers of quantized bird locomotion poses shows, that the regression quality among the models shows only little improvement of the model trained for 10 quantized pose classes. The horizontal line should indicate the smallest quantile.

In another experiment we investigate the influence of the different layers from which we extract the deep features. In Figure 4 the landmark regression performance of three selected layers (*conv4*, *conv5* and *fc6*) of the fine-tuned AlexNet can be seen. As can be seen, location information is lost in the fully-connected layer, while the convolutional layers retain it. It is also noticeable that the regression of the dorsoventral view has a better performance than the results of the lateral view. From the biological expert's point of view, however, this is the more difficult view when annotating.

As a final experiment we want to compare our landmark localization approach to the state-of-the-art method, the Augmented Active Appearance Models (AAAM) [10]. We found 6 non-cyclic locomotion datasets with which it was possible to train AAAM models for each sequence to compare the two different approaches. Both are trained with 15 annotated training frames together with both views. Figure 5 shows the view-combined results of the comparison between both approaches. It is clearly evident, that the deep feature regression outperforms the AAAM approach.

In further experiments we want to investigate influences of different sampling methods and feature dimension reduction to our results. Furthermore, we want analyze the lower bound of required training frames.

C. 3D Reconstruction

Recent analysis of animal locomotion uses 3D landmark information to make a better statement [1]. To exploit the multi-view sequence recordings, initially, images of a cubic calibration pattern for X-ray acquisition are also recorded. This enables a calibration of the recording system according to Hartley and Zisserman [12]. Concerning our four different

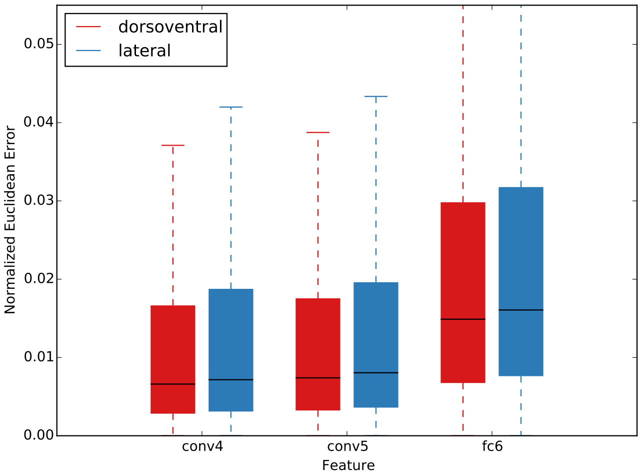

Fig. 4: Using activations of the convolutional layers (here *conv4* and *conv5*) as features of the CNN model shows that they are better suited for landmark localization than the activations of the fully-connected layers (here *fc6*), as they still contain position information.

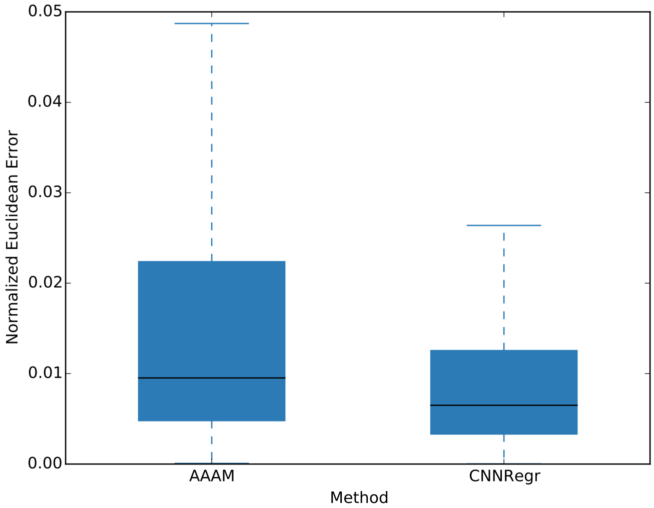

Fig. 5: We compared the Euclidean landmark error of our landmark localization approach (CNNRegr) with Augmented Active Appearance Models (AAAM) [10] using 6 non-cyclic animal locomotion datasets.

experiment setups based on the different obstacle types, the calibration is estimated four times. On average, we obtain a re-projection error (of the calibration pattern cube) of 0.25 Millimeters for the dorsoventral view and an error of 0.48 Millimeters in average for the lateral view. In further experiments we want analyze the reconstruction performance even if ground truth data is available.

V. Conclusion

In this paper, we introduced an anatomical landmark localization method based on regression models using deep features which are adjusted to the target image domain of low-contrast X-ray images. In our experiments we showed that we outperform state-of-the-art methods by evaluating on non-cyclic datasets and we achieve a comparable performance as biological experts when annotating manually. We also showed that the convolutional layers are better suited for landmark regression than the fully-connected layer behind them, because they have retained the location information. Finally, to exploit the multi-view sequence recordings, the localized 2D landmarks of the two different views can be reconstructed to 3D landmark points for a better evaluation by the biological experts.

Acknowledgment

The research was supported by grant DE 735/8-1 of the German Research Foundation (DFG).

References

[1] E. Andrada, D. Haase, Y. Sutedja, J. A. Nyakatura, B. M. Kilbourne, J. Denzler, M. S. Fischer, and R. Blickhan, “Mixed gaits in small avian terrestrial locomotion,” *Scientific Reports*, 2015.

[2] E. Andrada, J. A. Nyakatura, F. Bergmann, and R. Blickhan, “Adjustments of global and local hindlimb properties during terrestrial locomotion of the common quail (coturnix coturnix),” *Journal of Experimental Biology*, vol. 216, no. 20, pp. 3906–3916, 2013.

[3] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *PAMI*, vol. 23, no. 6, pp. 681–685, 2001.

[5] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *International conference on machine learning*, 2014, pp. 647–655.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[8] D. Haase and J. Denzler, “Anatomical landmark tracking for the analysis of animal locomotion in x-ray videos using active appearance models,” in *Scandinavian Conference on Image Analysis (SCIA)*, 2011, pp. 604–615.

[9] ——, “Comparative evaluation of human and active appearance model based tracking performance of anatomical landmarks in locomotion analysis,” in *Open German-Russian Workshop on Pattern Recognition and Image Understanding (OGRW)*, 2011, pp. 96–99.

[10] ——, “2d and 3d analysis of animal locomotion from biplanar x-ray videos using augmented active appearance models,” *EURASIP Journal on Image and Video Processing*, pp. 1–13, 2013.

[11] D. Haase, J. A. Nyakatura, and J. Denzler, “Multi-view active appearance models for the x-ray based analysis of avian bipedal locomotion,” in *Annual Symposium of the German Association for Pattern Recognition (DAGM)*, 2011, pp. 11–20.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] O. Mothes and J. Denzler, “Anatomical landmark tracking by one-shot learned priors for augmented active appearance models,” in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2017, pp. 246–254.

[15] J. A. Nyakatura, E. Andrada, R. Blickhan, and M. S. Fischer, “Avian bipedal locomotion,” in *5th International Symposium on Adaptive Motion of Animals and Machines (AMAM)*, 2011.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[17] V. N. Vapnik, “The nature of statistical learning theory,” 1995.