COMPARATIVE EVALUATION OF HUMAN AND ACTIVE APPEARANCE MODEL BASED TRACKING PERFORMANCE OF ANATOMICAL LANDMARKS IN LOCOMOTION ANALYSIS[1]

Daniel Haase and Joachim Denzler

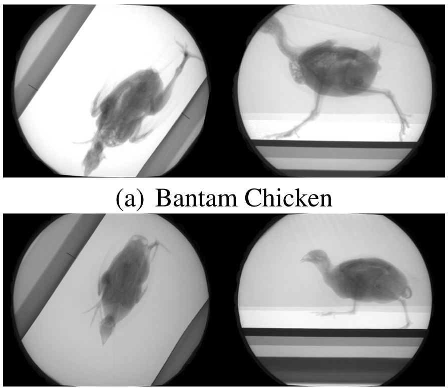
Chair for Computer Vision, Friedrich Schiller University of Jena, Germany
{daniel.haase, joachim.denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de/

The detailed understanding of animal locomotion is an important part of biology, motion science and robotics. To analyze the motion, high-speed x-ray sequences of walking animals are recorded. The biological evaluation is based on anatomical key points in the images, and the goal is to find these landmarks automatically. Unfortunately, low contrast and occlusions in the images drastically complicate this task. As recently shown, Active Appearance Models (AAMs) can be successfully applied to this problem. However, obtaining reliable quantitative results is a tedious task, as the human error is unknown. In this work, we present the results of a large scale study which allows us to quantify both the tracking performance of humans as well as AAMs. Furthermore, we show that the AAM-based approach provides results which are comparable to those of human experts.

Introduction and Related Work

The in-depth understanding of animal locomotion is an essential part of modern biology and has many applications in the fields of motion science and robotics. A key method of today's locomotion analysis is x-ray videography, which provides an insight into the course of motions at an extremely high level of precision [1]. For the practical realization, animals are placed on a treadmill and captured by a biplanar high-speed x-ray camera system with up to 2000 frames per second. In Fig. 1, exemplary images for two real world datasets are shown.

The biological evaluation of the recorded sequences is based on anatomical key points (*landmarks*) of the locomotor system, *e.g.* the knee and the hip joints. As a manual labeling of these points is extremely time-consuming due to the high frame rate, an automatic tracking method would be of great benefit. However, considering the nature of the data, several problems arise. Besides the low contrast, the greatest difficulties are x-ray occlusions in the images which make the use of local tracking methods impossible [2]. In [2, 3] it is shown that Active Appearance Models (AAMs) [4] can be used to overcome these specific problems encountered in this setting.


(a) Bantam Chicken

(b) Quail

Fig. 1. Example images of the two datasets used for the experiments in this paper.

All AAM-based studies presented so far, however, rely on datasets for which the ground truth landmark labeling was provided by only one single person [2, 3]. This situation makes it hard to distinguish between human errors in the ground truth data and automatic tracking errors. To be able to quantify both the performance of the AAM-based approach as well as the performance of human experts, a large scale study based on two datasets, each with 12 independent ground truth labelings, was carried out.

[1] This research was supported by grant DE 735/8-1 of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG)

The results of this survey are presented in this paper, together with a new practically important evaluation method.

Active Appearance Models (AAMs)

This section gives a brief overview of AAMs [4] in general and their application to landmark tracking. AAMs are statistical models which jointly describe the shape (landmarks) and the appearance (texture) of objects in digital images. Given a set of training images with annotated landmarks, an AAM can be trained and afterwards be fit to new images automatically.

Training Step. If $\boldsymbol{l}_n$ denotes the vectorized landmark coordinates of the n^{th} training image, the first step is to build a shape model by applying Principle Component Analysis (PCA) to the matrix $\boldsymbol{L} = (\boldsymbol{l}_1 - \boldsymbol{l}_\mu, \ldots, \boldsymbol{l}_N - \boldsymbol{l}_\mu)$, where $\boldsymbol{l}_\mu = {}^1\!/_N \sum_{n=1}^{N} \boldsymbol{l}_n$ is the *mean shape*. PCA yields the matrix $\boldsymbol{P_L}$ of *shape eigenvectors*, which can be used to describe an arbitrary shape $\boldsymbol{l}'$ via

$$\boldsymbol{l}' = \boldsymbol{l}_\mu + \boldsymbol{P_L}\boldsymbol{b}_{\boldsymbol{l}'}, \tag{1}$$

where $\boldsymbol{b}_{\boldsymbol{l}'}$ are the *shape parameters* of $\boldsymbol{l}'$.

In the next step, all training object textures are aligned into a common reference shape and vectorized. We denote the n^{th} texture by $\boldsymbol{g}_n$. Then, again PCA is applied on the matrix $\boldsymbol{G} = (\boldsymbol{g}_1 - \boldsymbol{g}_\mu, \ldots, \boldsymbol{g}_N - \boldsymbol{g}_\mu)$, where $\boldsymbol{g}_\mu$ is the *mean texture*. By using the resulting *texture eigenvectors* $\boldsymbol{P_G}$, any texture $\boldsymbol{g}'$ can be specified by its *texture parameters* $\boldsymbol{b}_{\boldsymbol{g}'}$ by means of

$$\boldsymbol{g}' = \boldsymbol{g}_\mu + \boldsymbol{P_G}\boldsymbol{b}_{\boldsymbol{g}'}. \tag{2}$$

To combine both the shape and the texture model, the according parameters are concatenated for each training example, and again a PCA is applied. Using the matrix $\boldsymbol{P_C}$ of *combined* or *appearance eigenvectors*, each model instance with the concatenated shape and texture parameters $\boldsymbol{c}'$ is described by

$$\boldsymbol{c}' = \boldsymbol{P_C}\boldsymbol{b}_{\boldsymbol{c}'}, \tag{3}$$

and $\boldsymbol{b}_{\boldsymbol{c}'}$ are the *appearance parameters* of each particular model instance.

Finally, only the most significant eigenvectors of $\boldsymbol{P_C}$ are used to obtain a compact parameterized model of the object shown in the training images.

Model Fitting. To fit an AAM to new images, a further training step is necessary. The known model parameters $\boldsymbol{b}_n$ of the training instances are varied by diverse values of $\Delta\boldsymbol{b}$. For each $\Delta\boldsymbol{b}$, the texture differences $\Delta\boldsymbol{g}$ between model and image are stored. Afterwards, a linear regression model $\Delta\boldsymbol{b} = \boldsymbol{R}\Delta\boldsymbol{g}$ is estimated. The matrix $\boldsymbol{R}$ can then be used to fit the model to unseen images, solely based on the current texture difference $\Delta\boldsymbol{g}$.

Application for Landmark Tracking. To make use of the biplanar camera setup, a combined model for both camera views is desirable. The combination of multiple views for AAMs is described in [5] and can be achieved in an easy manner.

Further adaptions of AAMs for the present scenario are described in [2, 3] and shall not be considered in detail here.

Experiments and Results

Datasets. The experiments presented in the following were conducted on two real world datasets of a bantam chicken in the one case, and a quail in the other case. Example images for both datasets are shown in Fig. 1. The datasets have an identical camera setup (top and side view) and image resolution of 1250×1250 pixels recorded at 1000 frames per second. Together with the recording times of 1024 and 1372 milliseconds, this results in image sequences with just as many frames.

In both cases, the ground truth landmark positions were provided independently by four biologists, three times each. Therefore, for any of the two datasets, a total of 12 independently labeled ground truth landmark sequences are available. The operating experiences of the experts ranged from very experienced (> 5 years) over experienced ($> 1{}^1\!/_2$ years) to novice. For the bantam dataset, every 10th frame was manually labeled, in the case of the quail dataset it was every 20th frame.

The anatomical landmarks covered by our experiments include the pelvis (two landmarks), the hip joints (two landmarks) and the knee joints (four landmarks), *i.e.* eight landmarks per camera view.

Evaluation Method. As general approach for the evaluation and comparison of the human and AAM-based tracking performance, we employed the following scheme separately for

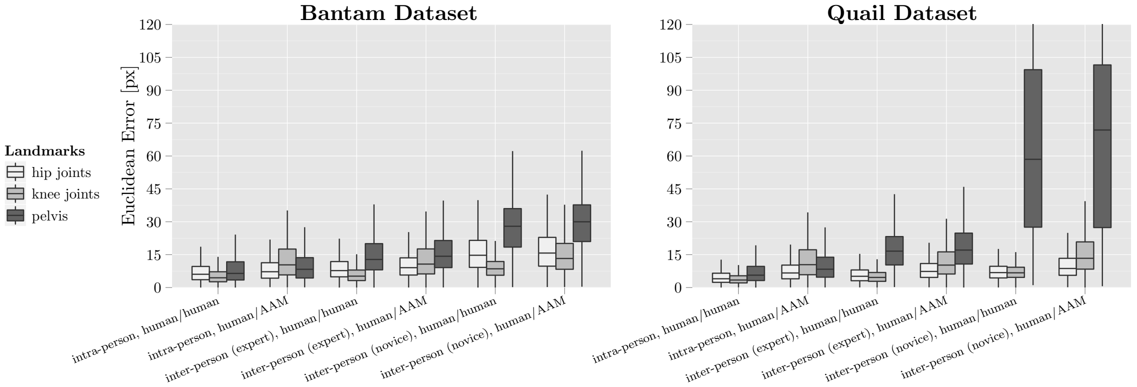


Fig. 2. Results of the Euclidean evaluation for the landmarks of the side view. For one person, an AAM which was trained on this person (intra-person, human/AAM) gives better results than a human novice (inter-person, human/human).

any of the two datasets. For all 12 ground truth landmark sequences, an individual multi-view AAM was trained on a subset of the labeled frames. The amount of training frames was 10 for the bantam and 14 for the quail dataset. Afterwards, the landmarks of interest were tracked for the entire locomotion sequence using these AAMs as described in [2, 3]. As a result, we obtained 12 artificially produced landmark sequences. We then performed a pairwise comparison for all 24 (12 human experts + 12 artificial AAMs) landmark sequences. For the comparisons of these sequences we selected two different error measures, as discussed in the following subsection.

To make the evaluation of the $\frac{24\cdot(24-1)}{2} = 276$ sequence pairs interpretable, they were grouped into the categories "*intra-person*" (different labeling sequences of one person), "*inter-person (expert)*" (labeling sequences of two distinct experts) and "*inter-person (novice)*" (labeling sequences of one expert and one novice). Together with the distinction between "*human vs. human*" and "*human vs. AAM*", this allows us not only to compare the tracking performances of AAMs with respect to humans, but also to quantify the error ranges of manual labeling to obtain a gold standard.

Error Measures. The most straightforward approach to quantify the differences between two given landmark sequences is by using the Euclidean distances between corresponding landmarks. In the literature, this error measure is known as *point to point error* [6] and is the only method used in this context so far [2, 3]. It has the advantages of being easily computable and feasible for both camera views.

However, the biological evaluation of the landmark positions is not based on their Euclidean coordinates, but rather on the *angles* between parts of the locomotor system. One example is the angle between femur and torso, which can be computed based on four landmarks of the side view, namely one knee, one hip joint and both pelvis landmarks. Thus, another error measure we used was the *angular error*, that is the difference between corresponding angles of certain parts of the locomotor system. Despite of its practical relevance, the drawback of the angular error is that it only incorporates a subset of landmarks and uses only landmarks of the side view.

As both error measures have advantages and disadvantages in this scenario, both of them were used for the evaluation of the results.

Euclidean Results. In Fig. 2, the results of the evaluation based on the Euclidean landmark distance are depicted for the side view of both datasets. To maintain a better overview, the eight landmarks were grouped into the three distinct anatomical classes "*pelvis*", "*hip joints*" and "*knee joints*". As can be seen, intra-person landmark sequences have the smallest errors. There is also a noticeable difference in tracking quality between experienced users and novices.

Obviously, the Euclidean errors are dependent on the landmark type. For human experts, the knee landmarks give the best results, while for the pelvis there may even be a *median error* of 60 pixels. A possible explanation for this result is that knees have a well observable local image structure, whereas the pelvis is located in a nearly homogeneous image region. As AAMs are global models, this is also the reason why AAMs perform better on pelvis and worse on knee landmarks compared to humans. In con-

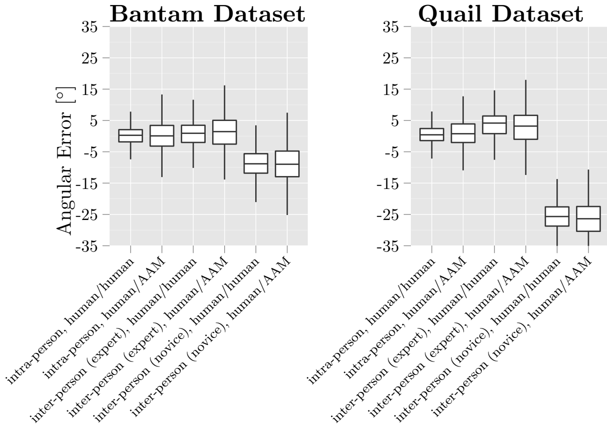


Fig. 3. Results of the angular evaluation for the angle between left femur and torso. An AAM trained on one person (intra-person, human/AAM) gives more accurate results than another experienced biologist does compared to that person (inter-person, human/human).

clusion, the Euclidean based evaluation shows that for one person, an AAM which was trained on this person gives better results than a human novice.

Angular Results. The results of the angular error based evaluation are shown in Fig. 3 for the angle between the left femur and the torso. It can be clearly seen that, again, the human intra-person class has the best performance, as there is virtually no bias and a maximum angular error of only $\pm 8^\circ$. When humans are compared to other human experts, the error range increases to about $\pm 12^\circ$, and—at least for the quail dataset—a bias of $+4^\circ$ is introduced. Another striking observation is that there is a huge bias of -9° (bantam dataset) and -26° (quail dataset) between human experts and novices. The variance, however, is comparable to the inter-person results of two human experts. This effect is caused by the fact that the novice *consistently* assigned an incorrect position for certain landmarks, in this case for the pelvis.

A general observation is that the angular results of AAMs show a larger error variance compared to their corresponding human counterparts. However, a major advantage of AAMs is that they practically have no bias in their tracking results, as opposed to human experts. The explanation for this property is that AAMs only use the given training data, whereas humans may have different levels of experience and anatomical knowledge. Another interesting effect can be seen in the results of the quail dataset. Here, the intra-person performance of the AAMs is superior to the inter-person performance of human experts. Consequently, this implies that an AAM trained on one person gives more accurate results than another experienced biologist does compared to that person.

Conclusions and Further Work

We have presented and analyzed the results of a large scale study to quantify both the human as well as the AAM-based tracking performance in x-ray locomotion analysis. The evaluations were based on two datasets, each having 12 independent ground truth labelings available. We used two different error measures and showed that AAMs provide results which are comparable to those of human experts.

Future work should take into account that—compared to humans—AAMs have deficits in using *local* gray value information. Thus, a possible extension should cover a local refinement step after standard AAM fitting.

Acknowledgements

The authors would like to thank the following people from the Institute of Systematic Zoology at the Friedrich Schiller University of Jena: John Nyakatura for providing the datasets and landmark labelings, as well as Irina Mischewski, Katharina Müller-Langer and Gabriele Unterhitzenberger for providing landmark labelings.

References

[1] M. Fischer and K. Lilje, *Dogs in Motion*. VDH, 2011.

[2] D. Haase and J. Denzler, “Anatomical Landmark Tracking for the Analysis of Animal Locomotion in X-ray Videos Using Active Appearance Models,” in *Proceedings of the 17th Scandinavian Conference on Image Analysis (SCIA 2011)*, ser. LNCS, A. Heyden and F. Kahl, Eds., no. 6688. Springer, 2011, pp. 604–615.

[3] D. Haase, J. A. Nyakatura, and J. Denzler, “Multiview Active Appearance Models for the X-Ray Based Analysis of Avian Bipedal Locomotion,” in *Proceedings of the 33rd DAGM Symposium (DAGM 2011)*, ser. LNCS, R. Mester and M. Felsberg, Eds., no. 6835. Springer, 2011, pp. 11–20.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active Appearance Models,” *IEEE T. Pattern Anal.*, vol. 23, no. 6, pp. 681–685, 2001.

[5] E. Oost, G. Koning, M. Sonka, P. V. Oemrawsingh, J. H. C. Reiber, and B. P. F. Lelieveldt, “Automated Contour Detection in X-Ray Left Ventricular Angiograms using Multiview Active Appearance Models and Dynamic Programming,” *IEEE T. Med. Imaging*, vol. 25, no. 9, pp. 1158–1171, 2006.

[6] M. B. Stegmann, “Active Appearance Models: Theory, Extensions and Cases,” Master’s thesis, Technical University of Denmark, DTU, 2000.