

# Anatomical Landmark Tracking by One-shot Learned Priors for Augmented Active Appearance Models

Oliver Mothes and Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Germany  
{oliver.mothes, joachim.denzler}@uni-jena.de

Keywords: Landmark Tracking, Active Appearance Models, Whitenened Histograms of Orientations

Abstract: For animal bipedal locomotion analysis, an immense amount of recorded image data has to be evaluated by biological experts. During this time-consuming evaluation single anatomical landmarks have to be annotated in each image. In this paper we reduce this effort by automating the annotation with a minimum level of user interaction. Recent approaches, based on Active Appearance Models, are improved by priors based on anatomical knowledge and an online tracking method, requiring only a single labeled frame. However, the limited search space of the online tracker can lead to a *template drift* in case of severe self-occlusions. In contrast, we propose a one-shot learned tracking-by-detection prior which overcomes the shortcomings of template drifts without increasing the number of training data. We evaluate our approach based on a variety of real-world X-ray locomotion datasets and show that our method outperforms recent state-of-the-art concepts for the task at hand.

## 1 INTRODUCTION

The profound investigation of animal locomotion plays an important role in many fields of research, e.g., zoology, biomechanics, and robotics. For those analyses an immense amount of data has to be recorded to be able to derive a model or to refine existing ones. In this context, it is necessary to evaluate the collected data in detail, which requires considerable expenses by biological experts in terms of manually annotating every single measure (Nyakatura et al., 2011; Andrada et al., 2013). Therefore, an automation of this task is highly preferable. In order to analyze the locomotor system *in vivo*, high-speed X-ray acquisition is applied. In an usual experimental setup, animals are placed on a treadmill which is enclosed by a C-arm X-ray acquisition system with two perpendicular detectors providing a top view (*dorsoventral* view) as well as a side view (*lateral* view) image of the entire locomotor system. To allow for a detailed biological evaluation, acquisition is performed at a high spatial and temporal resolution ( $1536 \times 1024$  pixels at 1000 FPS) on average for 1-2 seconds, resulting in up to 2000 frames. In order to avoid the time-consuming task of manual annotation of single images (Haase et al., 2013), an automation of this task at a minimum level of user interaction is of great interest.

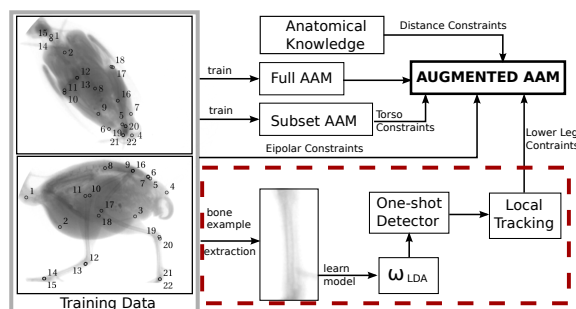


Figure 1: Based on few annotated biplanar recorded training images an Augmented AAM (Haase and Denzler, 2013) is trained, consisting of anatomical knowledge, a full multi-view AAM model, an AAM model of the torso landmark subset, epipolar constraints and a local tracking-by-detection prior introduced in this paper.

In (Haase and Denzler, 2013) Active Appearance Models (AAM) (Cootes et al., 2001) have been applied to several bipedal bird locomotion datasets. One crucial conclusion of this work is that AAMs need substantial constraints from various sources. With the support of additional anatomical knowledge, i.e. region segmentation, multi-view acquisition, and local landmark tracking, for the animals lower limb system, the resulting Augmented AAM (Haase and Denzler, 2013) provides robust results for the majority of the processed datasets. However, the applied online

tracking approach (Amthor et al., 2012) suffers from a potential *template drift* caused by severe and even full occlusion of the tracked objects.

Motivated by this shortcoming, we propose a one-shot learned tracking-by-detection approach which can handle these limitations by a global search. With one representative example of an annotated landmark subset a detector is learned. A two-staged graph-based tracking approach then provides motion trajectories through the whole sequence. Those trajectories are utilized as a prior in the Augmented Active Appearance Model framework together with priors from other sources as illustrated in Figure 1. In our experiments we show that this extension is able to improve previous results by up to 120 pixels in precision.

The remainder of the paper is structured as follows. In Section 2 we will give a brief overview of related work, followed by a short introduction to Active Appearance Models and its augmentation in Section 3. Afterwards, our one-shot learned tracking-by-detection approach will be introduced in Section 4. An evaluation of the detection and tracking results is presented in Section 5. Finally, Section 6 concludes the paper with a short discussion.

## 2 RELATED WORK AND MOTIVATION

For anatomical landmark tracking Haase et al. (Haase and Denzler, 2011) applied Active Appearance Models (Cootes et al., 2001) to X-ray locomotion scenarios. They showed that this generative model is well suited for the task at hand since training requires only a small amount of low contrast images. However, this approach has its weaknesses for a certain subset of landmarks—primarily landmarks of the lower limb system—undergo severe occlusions. They extended their approach in (Haase et al., 2011) to multi-view AAMs (Lelieveldt et al., 2003), which is more robust and accurate for torso landmark subsets compared to the single view approach. By concatenating corresponding landmarks of the second view the model became more general. The usage of additional constraints, especially for the distal limb landmarks, supporting the multi-view AAM, leads to a holistic model, referred to as Augmented AAM (Haase and Denzler, 2013). Anatomical knowledge, the multi-view information formulated as epipolar geometry, and a local tracking approach were used as priors for augmenting the standard AAM. Subtemplate Matching (STM) (Amthor et al., 2012) as a data-driven online tracking approach localizes landmarks of the distal limb segments. Based on the small number of

available training images, STM only needs one initial labeled frame for robust tracking, which renders the method highly preferable for the underlying task. However, online tracking fails in the case of severe occlusions and temporal disappearance of tracked objects. For example, subsequences with long-term occlusions of similar crossing objects considerably affects the tracking performance. As a consequence the template drift occurs, which results in a total loss of the structure to be tracked. Sequences with temporal disappearance of the object of interest produces a similar effect. An extension of STM is a pictorial structure approach (Amthor et al., 2014), where the distal limb system is formulated by a kinematic chain of single bones. Unfortunately, the extended method has the same weaknesses as STM. In contrast, our one-shot learned tracking-by-detection approach will tackle the template drift problem and handles strong texture shifts using a robust offline graph-based tracker even when the patch detection is missed in single frames.

Offline tracking algorithms are used to track objects in sequences (Andriluka et al., 2010; Li et al., 2015) and they are often formulated as a graph theoretical problem (Zhang et al., 2008; Berclaz et al., 2011; Jiang et al., 2013; Dehghan et al., 2015). First and foremost, reliable object detections serve as basis for all tracking approaches. Detection approaches are based on local image features like HOG (Dalal and Triggs, 2005; Felzenszwalb et al., 2010) or SIFT (Lowe, 2004) to detect objects in every single frame. In order to localize an object of interest a *Support Vector Machine* (SVM) is used for classifying positive and negative image patches in a sliding window manner (Felzenszwalb et al., 2010). However, SVM training is computationally expensive—especially when applying hard negative mining—and need a huge amount of training data.

Based on the fact that the amount of training data in our application scenario is limited, we use Whitened HOG features and an LDA model (Hariharan et al., 2012) for detecting landmark subsets, which only needs one single positive example for robust detector training.

More recently, Coarse-to-fine Convolutional Network Cascades (Sun et al., 2013; Zhou et al., 2013) are designed in a multi-level architecture for facial landmark detection. By fusing the outputs of each level of the multiple networks a robust and accurate estimation is possible. However, the Convolutional Neural Network frameworks have a complex structure and need a lot of data for training which is contrary to our pre-condition of a very limited number of training data.

The main contribution of this paper is a one-shot learned tracking-by-detection approach using a linear detector utilizing *Histogram of Oriented Gradients* (HOG) features and a classifier based on the *Linear Discriminant Analysis* in a sliding-window manner to detect the landmark subset of the lower limb system. The detection method provides two important advantages. On the one hand, many detector models for a sequence can be trained in a very short time and on the other hand the model training requires only one representative positive example which is important for the desired small annotation effort. Additionally, we use smart convolutions to speed up sliding window manner detections. Subsequently, a two-staged graph-based tracking algorithm is used to determine landmark subset trajectories through the whole sequence. In contrast to STM tracking, template drifts are reduced or even eliminated since landmark trajectories are optimized globally. The single landmark tracks of the lower limb system serve as important prior knowledge for the fitting task of an probabilistic Augmented AAM model, trained with only 10 annotated examples.

### 3 AUGMENTED ACTIVE APPEARANCE MODELS

Augmented AAMs (AAAMs) proposed in (Haase and Denzler, 2013) extend the fitting process of standard AAMs by providing additional prior knowledge.

An AAM is a parametric statistical generative model consisting of a shape component and a shape-free texture component. Training data consists of  $N$  images with corresponding landmark annotations for relevant anatomical structures. After aligning the  $N$  landmark shapes via *Procrustes Analysis* (Kendall, 1984), shape variation is parameterized by applying *Principle Component Analysis* PCA to the shape matrix  $\mathbf{S} = (\mathbf{s}_1 - \bar{\mathbf{s}}, \dots, \mathbf{s}_N - \bar{\mathbf{s}})$  where  $\bar{\mathbf{s}}$  represents the mean shape. The linear shape model  $\mathbf{s}$  with *shape parameters*  $\mathbf{b}_s$  and shape eigenvectors  $\mathbf{P}_s$  and the *mean shape*  $\bar{\mathbf{s}}$  is given by:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P}_s \mathbf{b}_s \quad . \quad (1)$$

Afterwards, each image texture  $\mathbf{I}_1, \dots, \mathbf{I}_N$  is warped into the mean shape  $\bar{\mathbf{s}}$ . To obtain the linear texture model  $\mathbf{g}$  the very same PCA-based procedure is applied to the shape-normalized image vectors

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \quad , \quad (2)$$

with the *texture parameters*  $\mathbf{b}_g$ , the *texture eigenvectors*  $\mathbf{P}_g$  and the *mean texture*  $\bar{\mathbf{g}}$ . By concatenating the variance-normalized shape and texture parameter vec-

tors and applying PCA again, we obtain the *combined parameters*  $\mathbf{b}_c$  (Cootes et al., 1998).

After training, the AAM model can be used to find a best fitting for the model parameter vector  $\hat{\mathbf{b}}_c$  to a given input image by minimizing the squared distance  $\delta \mathbf{g} = (\mathbf{g}_{image} - \mathbf{g}_{model})$  of the given image and the model appearance:

$$\hat{\mathbf{b}}_c = \underset{\mathbf{b}_c}{\operatorname{argmin}} \delta \mathbf{g}^\top \delta \mathbf{g} \quad , \quad (3)$$

where we make use of the linear relationship

$$\delta \mathbf{b}_c = \mathbf{A} \delta \mathbf{g} \quad . \quad (4)$$

However, the accuracy of standard AAMs is prone to occlusions and unseen cases due to the linear shape and texture modeling. To overcome these limitations prior knowledge from different sources can be incorporated. In addition to multi-view information (Haase et al., 2011), the authors of (Haase and Denzler, 2013) use various types of constraints to augment the standard AAM. Constraints like subset AAM, anatomical knowledge, epipolar geometry and a local tracking method (Amthor et al., 2012) for the animals lower limb system are used to increase the fitting performance and overcome the typical AAM weakness based on linear shape and texture modeling. We implement the idea of a combined approach by reformulating the AAM fitting as a maximum *a-posteriori* (MAP) framework as in (Haase and Denzler, 2013) with a *conditional independent* input image  $\mathbf{I}$  and every provided fitting constraints  $\boldsymbol{\pi}$ :

$$\begin{aligned} \hat{\mathbf{b}}_{c,MAP} &= \underset{\mathbf{b}_c}{\operatorname{argmax}} p(\mathbf{b}_c | \mathbf{I}, \boldsymbol{\pi}) \\ &= \underset{\mathbf{b}_c}{\operatorname{argmax}} p(\mathbf{I} | \mathbf{b}_c) \cdot p(\boldsymbol{\pi} | \mathbf{b}_c) \cdot p(\mathbf{b}_c) \quad . \end{aligned} \quad (5)$$

For input image data  $\mathbf{I}$  it is sufficient to use only a cropped version  $\mathbf{g}_{image}$ , defined by the AAM shape configuration. The likelihood can then be modeled as a Gaussian distribution  $\mathbf{g}_{image} | \mathbf{b}_c \sim \mathcal{N}(\mathbf{g}_{model}, \Sigma_{\mathbf{g}_{image} - \mathbf{g}_{model}})$  where  $\Sigma_{\mathbf{g}_{image} - \mathbf{g}_{model}}$  will be estimated in AAM training. The prior term  $p(\boldsymbol{\pi} | \mathbf{b}_c)$  performs the integration of all the constraints  $\boldsymbol{\pi}$  into the AAM fitting process, where  $\boldsymbol{\pi}$  represents the differences between the given target constraint values and the values based on the AAM parameters  $\mathbf{b}_c$ . Again a Gaussian distribution  $\boldsymbol{\pi} | \mathbf{b}_c \sim \mathcal{N}(0, \Sigma_\pi)$  will be assumed. The term  $p(\mathbf{b}_c)$  can be modeled as *maximum likelihood* estimation using a uniform distribution. For more information about the prior modeling of Augmented AAMs please refer (Haase and Denzler, 2013). A serious weakness of standard AAM is tracking landmarks of the lower limbs of the animal locomotor system. To overcome this drawback,

a local tracking constraint  $\pi_{local}$  with the results of an online tracking approach (Amthor et al., 2012), localizing those critical landmarks, is included in Augmented AAM framework.

## 4 ONE-SHOT LEARNED TRACKING APPROACH

For a reliable data-driven tracking of landmarks of lower limb landmarks, initially, a sophisticated detector is of great importance. As detection of single landmarks is more complicated, a detection of landmark subset patches is of advantage. The landmarks of single bones can be described as such a subset.

In the following sections we introduce a one-shot learned tracking-by-detection approach. In Section 4.1 the bone detection method will be discussed, while Section 4.2 focuses on bone tracking and landmark retrieval.

### 4.1 One-shot Learned Detector

To distinguish positive and negative examples, the combination of HOG features and SVM classification was the most commonly used approach in the past decade (Dalal and Triggs, 2005; Felzenszwalb et al., 2010). Unfortunately, SVM training and testing is computationally expensive, especially when applying hard negative mining with a huge amount of training data.

To overcome this limitation, Hariharan et al. introduces in (Hariharan et al., 2012) an object detection approach based on augmented HOG features (Felzenszwalb et al., 2010) and a classifier based on linear discriminant analysis (LDA).

Their model relies on the assumption that the distributions of object instances (positives) and background examples (negatives) follow both a Gaussian distribution. Thereby, the major computational effort is caused by the estimation of the background statistics (corresponding to the negative samples). Estimating the covariance matrix  $\Sigma_0$  and the mean vector  $\mu_0$  has to be done only ones.

For every positive class only the respective mean vector  $\mu_1$  has to be computed to obtain a discriminative linear separation of the two classes

$$\omega_{LDA} = \Sigma_0^{-1}(\mu_1 - \mu_0) \quad . \quad (6)$$

A sliding window-based method and template matching is used to compute similarity scores of a feature vector  $\mathbf{x}$  by a linear *Whitened Histograms of Orientations* (WHO) detector  $f(\mathbf{x}) = \langle \omega_{LDA}, \mathbf{x} \rangle$ . Dense sam-

pling of these features allows for matching the image templates.

To speed up the evaluation the authors of (Freytag et al., 2015) changed the order of computations and reformulated the sampling task as efficient convolutions. The entire set of window patches with  $D_C$  feature dimensions in the dense tiled grid of  $T \times T$  cells are evaluated by adding  $D$  convolutions of  $1 \times 1$  filters with corresponding feature planes. Consequently, derived from  $D = T \cdot T \cdot D_C$ , we obtain the feature extraction as efficient convolutions.

Unfortunately, the objects of interest in our application are rotated within a certain range. Accordingly, in the detection process, the input image needs to be rotated. As a result, the detection result at a specific location in the image contains a lot of multiple detections depending on the used angular resolution. Each detection contains position information, a detection angle and a detection score.

For the tracking algorithm the detection results are filtered to obtain object hypotheses. First, the normalized detection maps  $I_t$  of every frame  $t$  are smoothed by accumulating Gaussian filter kernels  $\mathbf{G}(x, y, \sigma)$  weighted by the corresponding detection score resulting in a smoothed detection map  $O_t$  with

$$O_t(i, j) = \sum_{x=-\frac{m}{2}}^{\frac{m}{2}} \sum_{y=-\frac{n}{2}}^{\frac{n}{2}} I_t(i+x, j+y) \mathbf{G}(x, y, \sigma) \quad , \quad (7)$$

where  $m = n$  describe the filter size of  $\mathbf{G}(x, y, \sigma)$ . In order to extract single detection hypotheses with new detection positions, *Non-Maximum Suppression* is applied to  $O_t$ .

Based on the assumption that the highest detection score yields the highest similarity with the model, the related detection score and rotation angle result from the rotation angle with the highest detection score within a local neighborhood defined by half the object size around the detection hypotheses is selected. Finally, the detection scores of all frames have to be normalized again.

### 4.2 Graph-Based Tracking

To associate the detection results of Section 4.1, a reliable tracking algorithm is necessary. The graph-based tracking approach based on (Jiang et al., 2013) uses the detection results of Section 4.1 and is divided into two steps. In the first step, the algorithm extracts segments of robust object trajectories by searching similar detected objects of subsequent frames.

Here, a Directed Acyclic Graph (DAG)  $\mathcal{G}$  is formulated where every detection hypothesis represents a node. We define detection hypotheses  $\mathbf{H} =$

$\{\mathbf{H}_0, \dots, \mathbf{H}_T\}$  with  $\mathbf{H}_t = \{\mathbf{h}_{t,0}, \dots, \mathbf{h}_{t,K_t}\}$  where  $\mathbf{h}_{t,i}$  represents the  $i^{\text{th}}$  detection hypothesis of frame  $t$ . Furthermore, we add a source  $\mathbf{h}_{source}$  and a sink node  $\mathbf{h}_{sink}$  to  $\mathcal{G} = (\mathbf{H}, \mathbf{E}, \mathbf{d})$ ,  $\mathbf{E} \subseteq \mathbf{H} \times \mathbf{H}$ , which are fully connected to all other nodes  $\mathbf{h}_{t,i}$ . The edge weights of the DAG depend on the number of detection feature weights  $d_p$ , as for example spatial  $d_s$ , temporal  $d_t$  and angular distances  $d_a$  but also detection scores or other detection results of adjacent detection hypotheses. However, the non-negative edge cost function  $d : \mathbf{H} \times \mathbf{H}$  with  $P$  detection features is calculated unlike (Jiang et al., 2013) as follows:

$$d(\mathbf{h}_{t,i}, \mathbf{h}_{t+\Delta t,j}) = \sum_{p=0}^P \alpha_p \cdot d_p(\mathbf{h}_{t,i}, \mathbf{h}_{t+\Delta t,j}) \quad (8)$$

*r.t.*  $\Delta t > 0$  .

The inner weight parameters  $\alpha_p$  of the single tracking priors with  $\sum_{p=0}^P \alpha_p = 1$  regularizes the influence of individual priors. Finding an optimal set of weights automatically is subject of future research. Before using a shortest path algorithm like *Dijkstra* or *Bellman-Ford*, thresholds,  $\theta_{min}$  and  $\theta_{max}$ , based on the used tracking priors have to be defined, such that:  $\theta_{min} \leq d_p \leq \theta_{max}$ . The thresholding sets constraints for the first stage (*tracklet extraction*) and set edges which do not match the pre-condition to infinity. Therefore, this constraints guarantee reliable path segments of related detection hypothesis and prevents mistakenly created shortest paths through the whole graph. In the DAG the edge weights of extracted paths are subsequently set to infinity to avoid multiple extraction of the same tracklets. The tracklet extraction process stops, if no further tracklet can be found, i.e., the tracklet length is smaller than 2.

Afterwards, the extracted tracklets are linked again to whole paths within a second DAG  $\mathcal{G}' = (\mathbf{H}', \mathbf{E}', d')$ ,  $\mathbf{E}' \subseteq \mathbf{H}' \times \mathbf{H}'$ , where  $\mathbf{H}' = \{\boldsymbol{\tau}_0, \dots, \boldsymbol{\tau}_{K'}\}$  are the estimated tracklet hypothesis and  $d' : \subseteq \mathbf{H}' \times \mathbf{H}'$  a non-negative cost function similar to equation 8.

## 5 EXPERIMENTS

In this section we evaluate the performance of the Augmented AAM framework extended by the introduced landmark detection and tracking techniques. We conduct our experiments on five avian locomotion datasets of several bird species with focus on sequences showing long-term object occlusion. The datasets were recorded by a high-speed X-ray acquisition system at 1000 Hz with a resolution of 1536 x 1024 pixels. Table 1 summarizes the analyzed datasets.

| Name | Species | Frames | Labeled Frames |
|------|---------|--------|----------------|
| Q1   | Quail   | 706    | 22             |
| Q2   | Quail   | 701    | 15             |
| T1   | Tinamou | 776    | 37             |
| J1   | Jackdaw | 1201   | 46             |
| J2   | Jackdaw | 1051   | 36             |

Table 1: An Overview of analyzed datasets

In Section 5.1 we compare different detection methods applied to two selected datasets with considerable self occlusions. Afterwards, results of the graph-based tracking algorithm are shown in Section 5.2 which uses the detections retrieved as described in Section 5.1. Finally, we use the tracking results in Section 5.3 as powerful priors for the Augmented AAM framework.

### 5.1 Comparison of the Detector Models

In general, detector models are learned from positive object examples. The number and quality of these examples is a crucial factor for their accuracy. However, in our application the number of annotated frames should be as much as necessary, but as few as possible. Another challenge for learning a reliable detector in our application is the visual appearance of the input images. X-ray acquisition systems provide grayscale images of low contrast. Accordingly, the detector has to overcome issues with respect to appearance and amount of positive training examples. Learned detector models using patches around landmarks are not representative enough concerning the high intra-class variability. Instead of using such landmark patches, the usage of subsets of landmarks is highly preferable to obtain a representative robust detector model based on examples with a low intra-class variability. Hence, the usage of corresponding landmarks of the lower limb bones (*proximal* and *distal* landmarks) are used in our application to define subsets and new patches for our detector. This is done by creating rotation normalized bounding boxes around the landmark subset. Figure 2 confirms our assumption that using landmark subset patches instead of patches around single landmarks for learning a detector model provides much better detection performance.

In our experiments we applied four different detector methods to the mentioned landmark subsets: HOG features trained with a linear SVM using all positive examples, a WHO model trained with all positive examples and as well as a HOG-SVM and a WHO

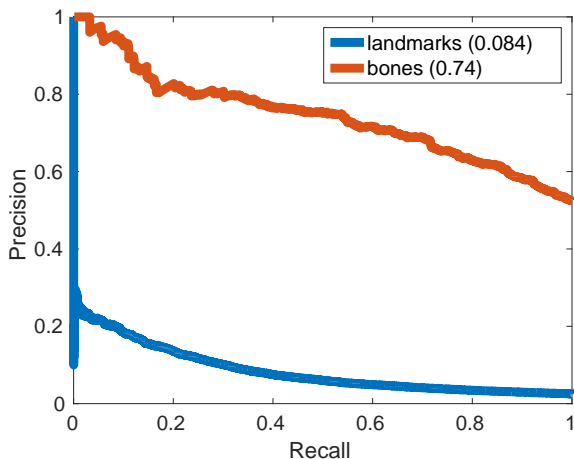


Figure 2: Detector models were trained for landmark subsets patches (bones) and the single landmark patches of the lower limb landmarks of all datasets of Table 1 .

model trained with only one-shot, in other words, one single example (one-shot learning).

**HOG.** The feature extraction is based on (Dalal and Triggs, 2005) with a gradient quantization to 9 orientation bins, a cell size of 8x8 pixels, a block size of 4x4 cells, and a block spacing stride of 8 pixels. A linear SVM model is trained using all annotated frames (positive bone examples) in the training data. Another SVM model is trained with only one representative example of the training data. To generate negative examples, window patches were clipped around the positive example patches. During detection the image is rotated between  $-90^\circ$  and  $+90^\circ$  with respect to biological constraints of the bone landmark subsets. For every rotation the detector obtains object hypotheses with information regarding position, detection score, and rotation angle using a sliding window technique. The chosen rotational resolution depends on the patch size. We used a degree step of 1.

**WHO.** For the WHO model  $\omega_{LDA}$  the background statistics has to be computed first. Therefore,  $N_0$  randomly unlabeled image patches were selected from the sequence and the mean  $\mu_0$  is estimated by computing the mean HOG feature  $\mu_0 = \frac{1}{N_0} \sum_{i=0}^{N_0} E[\mathbf{x}_i]$ . The covariance  $\Sigma_0$  is estimated using the *spatial autocorrelation function* (Hariharan et al., 2012). With the assumption of independent and identically Gaussian distributed positive and negative examples the model training can be performed using only one positive example. Since the LDA model  $\omega_{LDA}$  is not rotational invariant, the sliding window technique has to be performed for multiple image rotations, as well.

In a further experiment we compare the detection methods for all datasets of Table 1. We compared HOG-SVM models with WHO models. Both are

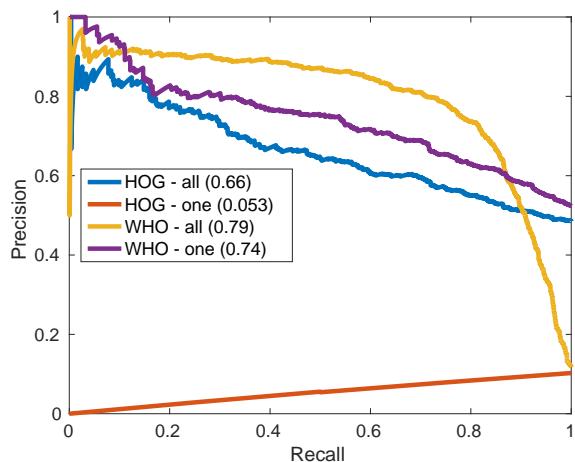


Figure 3: The Precision-Recall (PR) curve illustrates the performance comparison of the applied detection methods to all datasets of Table 1, respectively.

trained with all positive landmark and bone examples. Afterwards, models with only one representative example were trained. Since the detection method of Section 4.1 applied to landmark patches yields countless numbers of false positives and missed detections, we use for further experiments only those detection results based on the bone patches.

Figure 3 illustrates the comparison of the presented detection methods, which exhibit different detection performances. Both models using WHO features clearly outperform the results using HOG features, because the estimated background statistics  $\Sigma_0$  has a large influence on the linear separation. Regarding the whitening of the WHO features the performance of trained WHO models are nearly equivalent, regardless of the number of training samples. The poor performance of the model trained with one HOG feature example is caused by the weak representation of the class, as a linear separation is nearly impossible. Because of its very poor performance, the one-shot learned HOG-SVM model is ignored for the following experiments.

## 5.2 Online vs. Offline Tracking

In this section the graph-based tracking algorithm described in Section 4.2 is applied to the detection hypotheses of Section 5.1. Every detection provides a position information, a detection angle, a detection frame number, and a detection score. Based on this information the weights of the DAG  $\mathcal{G}$  are calculated as described in Section 4.2, where  $d_s$  represents the spatial distance,  $d_a$  the angular distance,  $d_t$  the temporal distance and  $d_c$  the inverted summarized detection scores between two nodes.

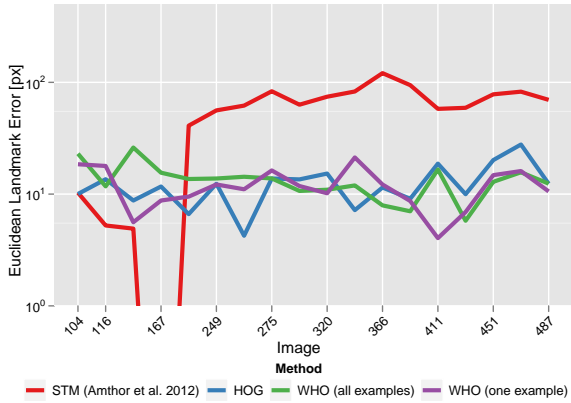


Figure 4: The detection result for the Q2-dataset from Section 5.1 were applied to the two-staged graph-based tracker and compared to the STM approach of (Amthor et al., 2012). The landmark error observed over time visualizes the drawback of the STM for cases of occlusion and temporal disappearance.

After normalizing all weights  $d_p$  we choose uniform inner weights  $\alpha_p$  for the edge cost function  $d$ . Additionally, in the second stage (*tracklet linking*), the mean velocity is calculated using position information of all detections covered by the respective tracklets. Based on the anatomical knowledge, especially the length of the extracted bone examples described in Section 5.1 and the detection angle information, it is possible to recover the *proximal* and *distal* landmark positions. Figure 4 illustrates the comparison of the STM baseline with our tracking results applied to the Q2-dataset. For all remaining datasets the Euclidean tracking error plots show similar results. The graph of the Euclidean tracking error clearly shows the template drift of the STM algorithm at time step 190 after the initial position at time step 167 where the pixel error was close to 0 pixels. All other graphs show robust trajectories of the introduced detection methods with an error of only 10 pixel on average. The trajectories based on the WHO one-shot detector model achieves the nearly same performance like the detectors trained with all positive examples while only one single training example was used.

### 5.3 A One-shot Learned Prior for AAAMs

As an extension of the Augmented AAM framework (Haase and Denzler, 2013), illustrated in Figure 1, we replaced the utilized local tracking prior  $\pi_{local}$  (Amthor et al., 2012) by our tracking-by-detection approach from Section 5.2 which is able to recover lost templates based on global optimization in contrast to

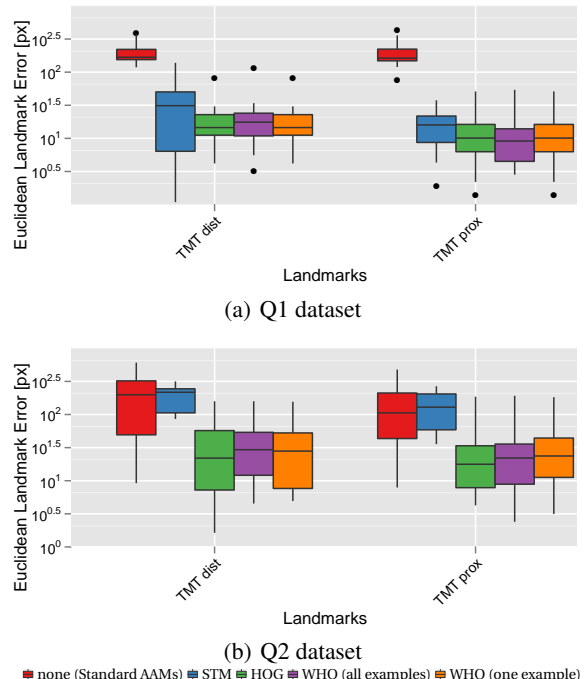


Figure 5: Prior for the Augmented AAM: four different local tracking approaches were analyzed with respect to Euclidean landmark errors of the *proximal* and *distal Tarsometatarsi* landmarks of the lower limb system. The Standard AAM and the Augmented AAM utilizing the STM approach as local prior serve as baseline for the comparison to our approach (see Section 5.2).

(Amthor et al., 2012). The comparison is conducted for all datasets of Table 1.

First, a multi-view AAM model of all landmarks and the torso landmark subset were trained with 10 annotated frames and is used as torso constraint. Based on both available views (*lateral* and *dorsoventral*) the epipolar geometry with the help of the Fundamental Matrix is estimated and is used as epipolar constraints. Anatomical knowledge, in terms of biological distance constraints were obtained via image segmentation as proposed in (Haase and Denzler, 2013). Together with one of the mentioned tracking approaches as lower leg constraints the AAAM is formulated as in Section 3.

In our experiments we compared the influence of the different tracking priors. Figure 5 illustrates the Euclidean landmark error of the local tracking approaches of Section 5.2 applied to the Augmented AAM framework of (Haase and Denzler, 2013). It can be clearly seen that the template drift problem of the Q2-dataset (see Figure 4) using the STM tracking approach substantially affects the performance of the entire AAAM framework. In contrast, our proposed tracking prior allows for highly accurate results of the

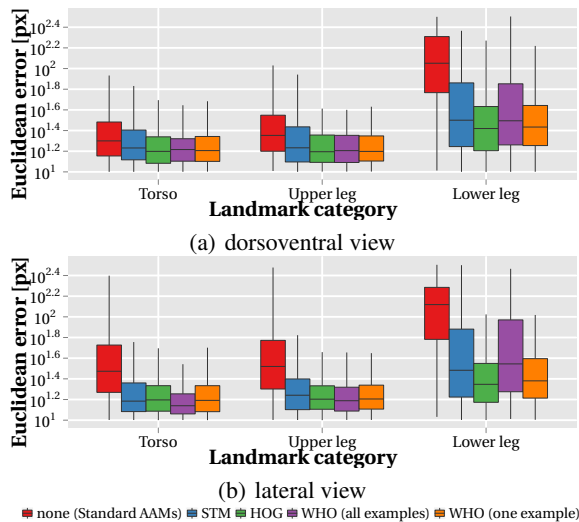


Figure 6: For the five datasets of Table 1 an averaged evaluation of the two different views illustrates the Euclidean landmark error of three landmark groups (torso, upper leg, lower leg).

combined AAAM approach without the loss of individual landmark subsets. In Figure 6 the average error over all sequences of Table 1 is shown. It can clearly be seen that our presented tracking-by-detection prior outperforms the online tracking approach even in the case of using only one single example for training.

## 6 CONCLUSIONS

In this paper we introduced a one-shot learned tracking-by-detection prior supporting an Augmented AAM framework for anatomical landmark retrieval in animal locomotion analysis. In particular, a linear detector was trained with only one representative positive example of a desired landmark subset in a very fast manner. A two-staged graph-based tracking algorithm generates whole trajectories of the detected hypotheses and recovers the single landmarks of the subset. Finally, the landmark tracking results were used as a prior for an AAM to support the model-driven baseline algorithm and solve the model fitting task for occluded and temporally disappeared landmarks. We compared our approach to another local tracking method using a frame-by-frame template matching strategy which is very accurate in sequences with partial self occlusion, but fails in case of long-term full occlusions. In our experiments we showed that this extension is able to improve previous results by up to 120 pixels in precision. To further improve the tracking precision of our proposed algorithm, a higher an-

gular as well as spatial resolution can be used, which, however, also increases detection runtime.

## ACKNOWLEDGMENTS

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

## REFERENCES

- Amthor, M., Haase, D., and Denzler, J. (2012). Fast and robust landmark tracking in x-ray locomotion sequences containing severe occlusions. In *International Workshop on Vision, Modelling, and Visualization (VMV)*. Eurographics Association.
- Amthor, M., Haase, D., and Denzler, J. (2014). Robust pictorial structures for x-ray animal skeleton tracking. In *International Conference on Computer Vision Theory and Applications (VISAPP)*. SCITEPRESS.
- Andrada, E., Nyakatura, J. A., Bergmann, F., and Blickhan, R. (2013). Adjustments of global and local hindlimb properties during terrestrial locomotion of the common quail (*coturnix coturnix*). *Journal of Experimental Biology*.
- Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE.
- Berclaz, J., Fleuret, F., Türetken, E., and Fua, P. (2011). Multiple object tracking using k-shortest paths optimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.
- Cootes, T., Edwards, G., and Taylor, C. (1998). Active appearance models. In *Computer Vision ECCV98*. Springer Berlin Heidelberg.
- Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE.
- Dehghan, A., Modiri Assari, S., and Shah, M. (2015). Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Freytag, A., Schadt, A., and Denzler, J. (2015). Interactive image retrieval for biodiversity research. In *German Conference on Pattern Recognition (GCPR)*. Springer.
- Haase, D., Andrada, E., Nyakatura, J. A., Kilbourne, B. M., and Denzler, J. (2013). Automated approximation of



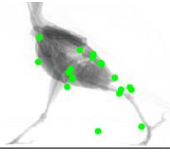
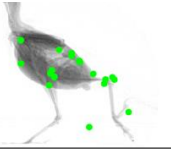
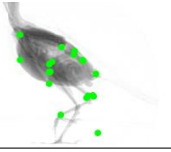
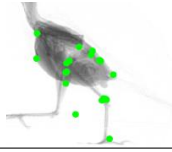
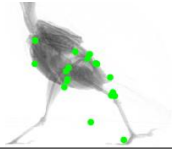
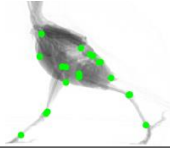
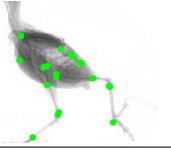
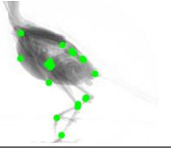
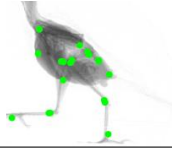
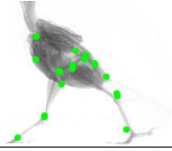
|       |   |   |   |  |   |
|-------|---|---|---|--|---|
| STM   |  |  |  |  |  |
| WHO   |  |  |  |  |  |
| frame | 300   | 350   | 400   | 450  | 500   |

Table 2: Qualitative results of selected frames of the J2-dataset illustrate the tracked landmarks by the Augmented AAM using the STM tracking prior (Amthor et al., 2012) and our one-shot learned tracking-by-detection prior.

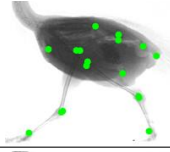
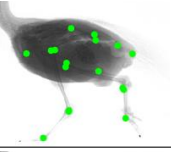
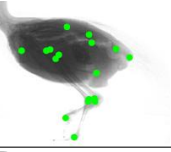
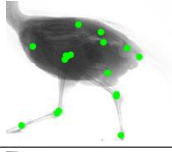
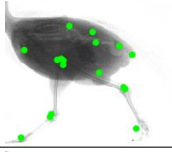
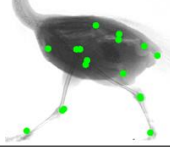
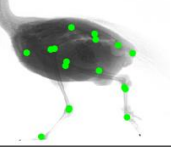
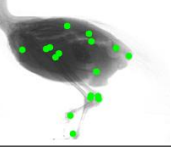
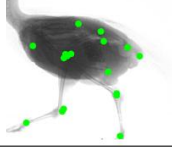
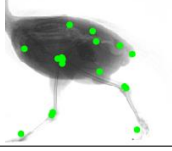
|       |   |   |   |  |   |
|-------|---|---|---|--|---|
| STM   |  |  |  |  |  |
| WHO   |  |  |  |  |  |
| frame | 225   | 265   | 285   | 305  | 385   |

Table 3: Augmented AAM using the STM tracking approach (Amthor et al., 2012) and our one-shot learned tracking-by-detection method as local tracking prior shows in comparison to each other nearly the same accurate results applied to the T1-dataset.

- center of mass position in x-ray sequences of animal locomotion. *Journal of Biomechanics*.
- Haase, D. and Denzler, J. (2011). Anatomical landmark tracking for the analysis of animal locomotion in x-ray videos using active appearance models. In *Scandinavian Conference on Image Analysis (SCIA)*. Springer.
- Haase, D. and Denzler, J. (2013). 2d and 3d analysis of animal locomotion from biplanar x-ray videos using augmented active appearance models. *EURASIP Journal on Image and Video Processing*.
- Haase, D., Nyakatura, J. A., and Denzler, J. (2011). 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)*. Springer.
- Hariharan, B., Malik, J., and Ramanan, D. (2012). Discriminative decorrelation for clustering and classification. In *Computer Vision—ECCV 2012*. Springer.
- Jiang, X., Haase, D., Körner, M., Bothe, W., and Denzler, J. (2013). Accurate 3d multi-marker tracking in x-ray cardiac sequences using a two-stage graph modeling approach. In *Computer Analysis of Images and Patterns*, pages 117–125. Springer.
- Kendall, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*.
- Lielieveldt, B., zmc, M., van der Geest, R., Reiber, J., and Sonka, M. (2003). Multi-view active appearance models for consistent segmentation of multiple standard views: application to long- and short-axis cardiac {MR} images. *International Congress Series*.
- Li, L., Nawaz, T., and Ferryman, J. (2015). Pets 2015: Datasets and challenge. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*. IEEE.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*.
- Nyakatura, J. A., Andrada, E., Blickhan, R., and Fischer, M. S. (2011). Avian bipedal locomotion. In *5th International Symposium on Adaptive Motion of Animals and Machines (AMAM)*. Elsevier.
- Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE.
- Zhang, L., Li, Y., and Nevatia, R. (2008). Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE.
- Zhou, E., Fan, H., Cao, Z., Jiang, Y., and Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*. IEEE.