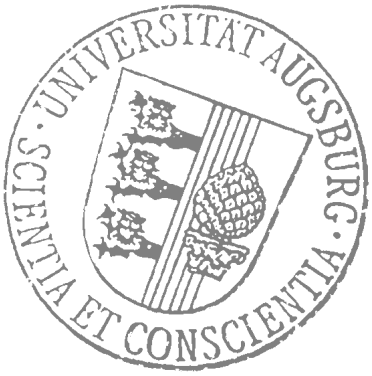
Monocular 3D Human Pose Estimation by Classification

Thomas Greif, Debabrata Sengupta, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Greif, Thomas, Debabrata Sengupta, and Rainer Lienhart. 2011. “Monocular 3D Human Pose Estimation by Classification.” Augsburg: Universität Augsburg.

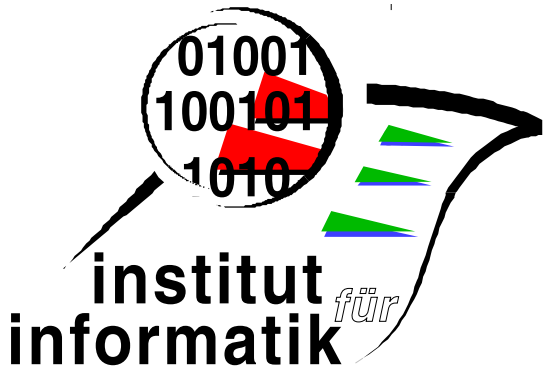
Universität Augsburg



Monocular 3D Human Pose Estimation by Classification

T. Greif, D. Sengupta, R. Lienhart

Report 2011-09 März 2011



Institut für Informatik
D-86135 Augsburg

MONOCULAR 3D HUMAN POSE ESTIMATION BY CLASSIFICATION

Thomas Greif, Rainer Lienhart

Multimedia Computing Lab
University of Augsburg, Germany
{greif,lienhart}@informatik.uni-augsburg.de

Debabrata Sengupta

IIT Guwahati
Guwahati - 781039, Assam, India
d.sengupta@iitg.ernet.in

ABSTRACT

We present a novel approach to 2D and 3D human pose estimation in monocular images by building on and improving recent advances in this field. We take the full body pose as a combination of a 3D pose and a viewpoint and in this way define classes that are then learned by a classifier. Compared to part based approaches, our approach does not suffer from self–occluded body parts since such occlusions are characteristic for certain classes and thus are captured during class definition. Moreover, we significantly relax the requirements posed on training data by the fact that we do neither require labeled viewpoints nor background subtracted images, and the carried out action does not need to be cyclic. By combining an efficient classifier with efficient image features, we present a generic and fast way to estimate human poses in images and achieve comparable results to state–of–the art approaches which we demonstrate on a public benchmark.

Index Terms— Pose estimation, human detection, random forests

1. INTRODUCTION

2D and 3D human pose estimation has a wide field of applicability, ranging from video indexing over security and safety to entertainment purposes. The recovery of 3D human poses in monocular image sequences, however, still poses a challenging problem to computer vision. Highly nonlinear human motions, cluttered backgrounds, partial or complete occlusions, and numerous different appearances already demand sophisticated 2D pose estimation approaches, and monocular 3D human pose estimation involves not only solving these problems, but also coping with the ambiguity between 2D and 3D poses.

Recent progress in computer vision and specifically in object detection, has helped to greatly facilitate the dealing with many of these difficulties. With the upcoming of powerful classifiers and robust features impressive results were achieved. Being inspired by state–of–the–art object detection algorithms [1, 2] that assume that each object is composed of smaller parts and thus allow a certain degree of deformation of these, which results in improved detection performance and invariance to viewpoint changes, part based models were adapted and extended [3, 4, 5, 6, 7] to pose estimation problems. In these approaches, classifiers for various body parts are trained and combined with suitable motion priors [8]. The disadvantage of such approaches, however, is that they usually cannot handle occluded body parts and that they perform best when action specific motion priors are used.

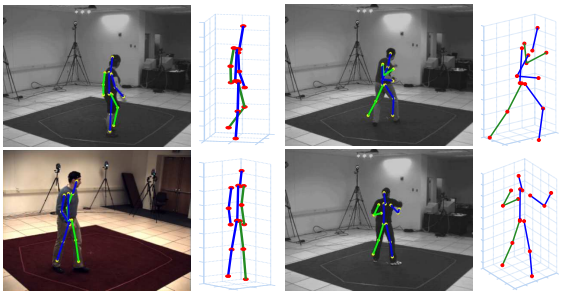


Fig. 1. Example results on walking and boxing sequences of the HumanEva [9] data set. Each subfigure shows the estimated 2D pose (overlay) as well as the corresponding 3D pose.

1.1. Overview of the proposed approach

We propose an efficient and generic approach to 2D and 3D human pose estimation in monocular images. For a given image without any additional information we want to detect all persons, and for each estimate the corresponding 2D and 3D body pose. The 2D pose is defined by the image coordinates of the person's prominent joints and the 3D pose by the 3D coordinates of the same joints with respect to a defined root joint. Figure 1 shows the estimated 2D and 3D poses for some images.

Using training images with annotated 2D and 3D joint coordinates (e.g. taken from a motion capture system as in the HumanEva [9] data set) for a certain action, we train an action specific classifier on features extracted from locations on the persons in these training images. Each training image with its 2D and 3D pose belongs to a discrete class. Every single class simultaneously defines a certain 3D pose within

978-1-61284-350-6/11/$26.00 ©2011 IEEE

the considered action and a viewpoint from which the person is observed. Pose estimation is thus treated as a classification problem and resolves to extracting image features in a region and predicting one of the learned classes. The estimated 2D and 3D pose are then the exemplars corresponding to the predicted class.

Contrary to training part specific classifiers, treating a complete body pose seen from a certain viewpoint as a single class enables us to already capture self–occlusion during the process of defining the classes, and no motion priors are required. While several approaches dealing with the problem of monocular 3D pose estimation include stages where the viewpoint is estimated, and therefore are restricted to data sets with those viewpoints annotated, we define our classes automatically from any training data supplying 2D and 3D pose information.

We demonstrate the performance of the presented approach on the walking and boxing sequences of the public HumanEva [9] benchmark.

1.2. Related work

In addition to the mentioned part based models used for 2D pose estimation, several approaches made significant improvements in the field of 3D human pose detection in monocular images and videos. Andriluka et al. [10] introduced a novel framework based on the combination of strong 2D image and pose evidence and a 2D–3D exemplar based pose lifting. The evidence is created using strong part detectors [3] and tracklets [11] that maximize the posterior over a short sequence of frames. They currently achieve the lowest published error on the HumanEva benchmark by solving several problems at once in their single framework. Moreover, they impressively show that they can decently track the 3D pose of multiple people in real world environments including partial and complete occlusions. Rogez et al. [12] are not using part based classifiers but define classes by a combination of the whole 3D body pose and the viewpoint from which the person is observed. Each class is thus a point on a 2D manifold which is used to model temporal relationships. They train a classifier based on hierarchical trees [13] and random forests [14] and use them to accurately estimate the poses in videos.

Our work shares some similarities with [12] in that we treat 3D human pose estimation as a pure classification problem and use a combination of histogram of oriented gradients [15] features and a random forest classifier. We, however, differ from and improve their approach in several ways. First, we significantly relax the requirements on training data because we (1) don't require labeled viewpoints, (2) don't have to perform a foreground/background separation on the training images to get proper gradient images, and (3) are not restricted to cyclic actions. This facilitates the evaluation of the approach on various public data sets and makes tedious preprocessing (e.g. background subtraction and viewpoint labeling) unnecessary. Second, we use improved and more efficient lower dimensional HOG features [2] that were recently proposed and help to improve the estimation performance as well as speed up the computations. Last, we show that we can achieve comparable results simply by (1) randomly sampling features on the person and (2) training a standard random forest rather than building a tree hierarchy by merging classes and choosing sample points according to the edge likelihood ratio within each class. This makes our approach extremely lightweight, efficient, and generic.

2. CLASS DEFINITION

In Rogez' original approach [12] each class is defined through a point on a 2D manifold; one dimension corresponds to the viewpoint, the other represents the 3D pose at a certain temporal progress of the considered action, e.g. walking or boxing. In order to create the initial class definitions, the viewpoint of each training image must be labeled, and the action has to be cyclic. The manifold, which constitutes a torus is then learned by temporally aligning the training data, and discretizing it accordingly.

A drawback of this approach to class definition is that, in general, pose estimation problems require a large number of training images to cover many aspects of human articulations and appearances, and annotating the viewpoint of many images (e.g. over 20,000 images when using the HumanEva data set with flipped images) by hand is quite tedious. Moreover, the demand for cyclic actions restricts one to data sets that include such cyclic actions. It would be advantageous if one could simply use any training data set for arbitrary actions, as long as the data set specifies the 2D and 3D joint coordinates. We therefore propose a simple unsupervised approach to generate the initial classes from a training set that does not require the viewpoint to be labeled nor the action to be cyclic.

We adhere to the class definition of Rogez, because the assumption to let a class represent a combination of viewpoint and 3D pose is reasonable: changing either results in a different projection of the pose in the image, which in turn yields different image features according to which we will classify a pose. According to this class definition we have to include pose as well as viewpoint information into our classes. This presents a problem since we don't want to rely on annotated viewpoints. If we, however, assume that the projection of a 3D pose onto the image plane given a fixed viewpoint is unique, i.e. each 2D pose is the result of a single 3D pose observed from a certain viewpoint, then a 2D pose in an image gives us information about both, the 3D

pose as well as the viewpoint. By partitioning 2D poses, we thus gain the correct class definitions.

Following the above considerations, we apply a standard k–means clustering to the 2D poses of the training set. For clustering, instead of using 2D joint coordinates in each image, it is more appropriate to use the angles between body parts if we assume a stick figure body representation [16]. We define the 2D pose in an image k by $\Phi_k = [\phi_1, \ldots, \phi_8]^T$. $\phi_{1 \leq j \leq 8}$ represent the absolute angles (in image coordinates) between a person's limbs. We found out that for clustering poses of walking and boxing persons, using only the arms and legs yields superior results over taking all angles into account. Applying k–means to the vectors supplied by the training set gives us our initial class labels. When using k–means, care has to be taken as we deal with circular quantities. This is important when computing the mean as well as the Euclidean distance of vectors. We therefore convert the angles to imaginary unit vectors and perform the cluster operations on them.

Figure 2 shows example images of eight different classes after clustering the HumanEva walking sequences. The shown classes roughly correspond to the same 3D pose and differ only in the viewpoint. Note that by using k–means we do not have to determine suitable values for the number of viewpoint and action clusters beforehand, which makes our approach more generic because — unlike using the 2D manifold — we do not need to restrict the training data to contain variations in pose as well as viewpoint. The clustering algorithm decides how to define classes according to the data at hand and the desired overall number of clusters.

The advantage of treating the complete body pose as one class over training multiple part classifiers is that self–occlusions arising from the carried out action can already be captured during class creation. The classifier can correctly classify such poses, because it implicitly adapted to such cases during training. For example, the left arm of a person walking from left to right is obviously not visible in some frames. This characteristic is learned while training our classifier. A part based classifier would fail to detect the left arm in such frames, and could only rely on additional temporal information or similar heuristics.

3. TRAINING THE POSE AND VIEWPOINT CLASSIFIER

Having defined the classes we now explain how our random forest [14] classifier is trained using features selected from prominent locations on the persons' bodies.

Prior to selecting feature locations and learning the classifier, we align all training images and annotations and center them in a normalized bounding box of fixed size. Feature locations will be specified relative to this box.

Fig. 2. Examples of images belonging to eight different classes. Each class shows roughly the same 3D pose observed from a different viewpoint.

We have chosen a random forest classifier over the popular SVM because the evaluation of a single location by a multi–class SVM is significantly slower than when using a random forest; a tree can, by a single decision, choose extremely fast at each node whether to propagate the example further or not.

3.1. Improved HOG features

For pose estimation we require features that are robust against changes in illumination as well as color. Gradient and edge based features like histograms of oriented gradients [15], SIFT [17] or shape context [18] have long since proven to fulfill these requirements and thus to be applicable for this task. Training the forest and testing an image at all positions and scales is, however, still very time consuming since it requires the computation of a very large number of features, independently of the classifier used. The chosen features should thus not only be robust but also be efficient and fast to compute.

We have chosen to rely on improved HOG features introduced by Felzenszwalb [2]. The original HOG descriptor [15] is a 36–dimensional feature vector, computed by normalizing the nine–dimensional gradient histogram of each cell by four different factors. Felzenszwalb found out that they can reduce the dimensionality to 13 dimensions, and augmenting the standard HOG descriptor by contrast sensitive features yields a powerful 31–dimensional feature. For a detailed explanation the reader is referred to [2].

3.2. Feature selection

We will train a standard random forest [14] classifier from features sampled from all training images. These features should capture the changes in pose depending on the class. Consequently it is important to choose the locations appropriately. Rogez et al. [12] take the edge likelihood of each class to determine such locations. They argue that locations with a high edge likelihood represent characteristic points for

(a) (b)

Fig. 3. Probabilities of locations being on a person's limbs for (a) walking and (b) boxing sequences (computed over the complete HumanEva [9] training set). Most of the features are sampled from orange and red locations.

a class, and thus should be picked with higher probability. While this is a reasonable assumption, background subtracted training images are crucial here, since otherwise background would introduce noise to the edge likelihoods leading to inaccurate locations. Obtaining background subtracted training images, however, is no easy task and always involves an additional preprocessing step with variable output.

Taking this into consideration, we choose a simpler approach that meets our requirements of making background subtraction unnecessary and simultaneously yielding equal results. We assume that changes are best captured at locations directly on the body and not on its boundary. Moreover, four HOG cells better capture the changes in the neighborhood of a location than does a single cell. Therefore, we compute for each location the percentage of it being on the body over all training images. From all locations above a certain threshold we randomly sample the desired number of locations and for each the four surrounding HOG cells will form the feature vector. For the walking and boxing sequences we consider only the limbs of a person, because the torso and head do not contribute as much to the motion. For other actions, where e.g. the torso moves a lot, other body parts should be considered as well. We use a threshold of 0.5 (i.e. the location has to be on a limb in 50 percent of all training images) and sample 100 locations in all of our experiments. Figure 3 shows the probabilities of each pixel being on a person's limbs for the two different actions.

3.3. Forest training

We extract improved HOG features at the specified locations of the aligned training images as described in the previous section. From these features and the corresponding class labels a random forest classifier is learned.

Following the notation of Breiman [14], a random forest consists of K random tree classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, \ldots, K\}$ where $\Theta_{1 \leq k \leq K}$ are independent identically distributed random vectors specifying which input variable to evaluate at each tree node. Each tree casts a unit vote for a class and the output of the random forest for input $\mathbf{x}$ is simply the majority vote among all trees.

We use random forests due to the fact that they are robust to outliers and noise and because they are extremely fast during classification, since negative examples can be rejected early in the tree. Moreover, during training no cross–validation is needed. By design each tree is trained on only roughly $2/3$ (63.2 percent [19]) of the training data, and the out–of–bag data is used to get an unbiased error estimate and improve the partitioning for each tree.

4. 3D HUMAN POSE ESTIMATION IN IMAGES

Being presented a new test image, the first task is to detect all persons and afterwards estimate their poses correctly. We think that the expensive multiscale search should be carried out by a declared and fast people detector rather than our pose classifier, so that extensive resources are only put into promising regions.

4.1. Localization

Classifying each possible location at each possible scale is computationally expensive, especially since each location has to be evaluated by the whole forest. While the classification of a single region is very fast, exploring the whole image is time consuming and prone to errors since using only peaks in the vote distribution as an indicator for the presence of a person is likely to produce false positives. We therefore rely on a state–of–the–art person detector [20, 2] trained on the INRIA person data set [15] which provides accurate localization of persons in images. This not only decouples the people detection stage from the pose estimation stage but also ensures that most resources are only spent on promising regions. Felzenszwalb's detector is very fast in detecting a person, and the advantage over using Dalal's HOG detector [15] lies in the re-usability of the improved HOG features: as the person detector is also utilizing them, the feature pyramid only has to be computed once for a single image.

4.2. Classification and pose estimation

Having localized all persons in an image, we let the forest classify each returned bounding box. We simply extract the precomputed HOG features at the determined locations (see section 3.2) to form our input vector $\mathbf{x}$ that we supply to each tree k which then predicts the class $c_k = h(\mathbf{x}, \Theta_k)$. As described in section 3.3 the predicted class of the whole forest is the majority vote over all trees. Moreover, we extract the vote distribution from the forest, which is simply a histogram over the trees' votes.

The estimated 2D and 3D poses of the predicted class are its exemplars. The exemplars are computed by taking the

mean relative joint coordinates of the class. Translating and scaling the 2D pose according to the location of the detection gives us our final estimated 2D pose in image coordinates. The absolute 3D pose cannot be estimated using only a single image without further information about extrinsic and intrinsic camera parameters.

Note that the people detector may return false positives which, although being quite rare, have to be dealt with. The vote distributions are a good enough indicator. Comparing the distributions of regions containing persons to regions that do not, one can observe that the latter case does not show peaks while in the first case clear peaks are apparent. Finding no peak in a vote distribution thus indicates a false positive which will be disregarded.

5. EXPERIMENTAL RESULTS

We evaluate our approach on different data sets to show its applicability. For comparison to other approaches, we use the public HumanEva benchmark and estimate the poses for walking sequences. Moreover, we demonstrate that our approach can be used for non–cyclic actions and apply it to the boxing sequences of the same benchmark.

5.1. Evaluation on data sets

We quantitatively compare our approach against two state–of–the–art approaches that currently achieve the lowest error on the HumanEvaII walking sequences. We use the complete training data of the HumanEvaI data set (three different subjects walking in circles) to train our model. We first align all images and poses so that the person is centered within a 150x240 pixel bounding box as described in section 3.3. The classes are automatically defined using the aligned poses and a random forest composed of 500 trees is trained. We use 192 classes, which — when inspected visually for the used training set — corresponds roughly to eight viewpoints and 24 action pose classes. Note that this partitioning is training set specific and we expect it to differ significantly for other training sets (e.g. ones that show only people from the sides). 2D and 3D poses of subject S2 are estimated during the first 350 frames of HumanEvaII. To compute the error in 2D, we use camera view C1 and C2 as suggested by the data set creators. Table 1 compares the errors on these sequences. Since [12] don't report the 3D error, ours can only be compared to [10]. The slightly higher 2D errors and standard deviations stem from the fact that we willingly forgo a temporal model and wrong classifications thus have great impact, whereas the worse error in 3D is due to the fact that we only estimate the rotation of the exemplar pose with respect to the camera and do not make use of other extrinsic parameters to get the correct scale. Nevertheless, we have shown that we can decently recover the 2D and 3D human body pose in monocular images very efficiently, without relying on any temporal information and without the need for background subtraction and viewpoint labels.

Fig. 4. Example results on images showing walking people of the INRIA Person data set [15].

To further demonstrate the wide applicability of our approach, we show that it can easily be applied to non–cyclic actions, and train the model on the boxing sequences of HumanEva. The errors on the whole validation set (over 8,000 frames) are 15.60 (2D) and 64 (3D). For the sake of completeness, we also report the errors on all walking sequences of the validation set (over 10,000 frames), which is 14.46 (2D) and 108 (3D). Moreover, we qualitatively evaluate the trained model by applying it to images showing only walking people of the INRIA data set where people appear in a variety of appearances in realistic environments. Some sample results are shown in figure 4. It can be seen that the poses of detected people are estimated quite accurately, and that our model — despite being trained on a different data set with static background — is able to generalize well enough to be used in unconstrained environments as well.

Cam	2D [12]	2D [10]	3D [10]	**2D**	**3D**
C1	12.98 (3.5)	10.49 (2.70)	107 (15)	14.09 (5.4)	142 (49)
C2	14.18 (4.38)	10.72 (2.44)	101 (19)	15.14 (6.3)	128 (33)

Table 1. Comparison of 2D (pixels) and 3D (millimeters) pose errors and standard deviations (in brackets) on HumanEva walking sequences for subject S2 and cameras C1 and C2.

5.2. Efficiency

The proposed approach achieves comparable low errors on the respective benchmarks, and it is concurrently very efficient. Rogez [12] report that their 100 tree forest takes 0.015s to classify a single bounding box. Evaluating all possible locations in a 640x480 px. image (like the ones provided by HumanEva) with their bounding box size results in $60 \cdot 48 = 2880$ locations that have to be evaluated. In other words, classifying the pose in a single scale already takes them 43.2 seconds.

With our implementation (using 500 trees in the forest), it takes us in average 2.3s (computed over the complete vali-

dation set of HumanEvaI walking sequences) to estimate the pose when evaluating all possible locations in all possible scales, with an octave interval of ten, and without any concurrency in the code. While still being far away from real time pose estimation, we considerably speed up the computations without losing accuracy.

6. CONCLUSIONS

We have introduced novel enhancements to 2D and 3D pose estimation in monocular images and videos. Automatic class definitions facilitate the usage of public data sets and make tedious manual labeling as well as expensive preprocessing steps with variable output obsolete. The actions that can be considered are not restricted to be cyclic and are only limited by the chosen data set. The presented approach significantly relaxes the requirements posed on the training data making it generically applicable, while still achieving comparable results on public benchmarks. Moreover, it is able to recover the poses with great efficiency as we have shown for walking and boxing sequences of the HumanEva data set.

7. REFERENCES

[1] P. Felzenszwalb, D. Mcallester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[3] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1014 –1021.

[4] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1 –8.

[5] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Pose search: Retrieving people using their pose," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1–8.

[6] Marcin Eichner and Vittorio Ferrari, "Better appearance models for pictorial structures," *BMVC-09*, 2009.

[7] B. Sapp, C. Jordan, and B. Taskar, "Adaptive pose priors for pictorial structures," in *Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Conference on*. IEEE, 2010, pp. 422–429.

[8] M. Vondrak, L. Sigal, and O.C. Jenkins, "Physical simulation for probabilistic motion tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1 –8.

[9] L. Sigal, A.O. Balan, and M.J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International journal of computer vision*, vol. 87, no. 1, pp. 4–27, 2010.

[10] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Conference on*. IEEE, 2010, pp. 623–630.

[11] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1 –8.

[12] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P.H.S. Torr, "Randomized trees for human pose detection," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1 –8.

[13] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural computation*, vol. 9, no. 7, pp. 1545–1588, 1997.

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

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Conference on*, 2005, vol. 1, pp. 886 –893 vol. 1.

[16] Thomas Greif and Rainer Lienhart, "A kinematic model for bayesian tracking of cyclic human motion," 2010, vol. 7543, p. 75430K, SPIE.

[17] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] S. Belongie, J. Malik, and J. Puzicha, "Shape context: A new descriptor for shape matching and object recognition," in *In NIPS*, 2000.

[19] Leo Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996, 10.1007/BF00058655.

[20] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively trained deformable part models, release 4," http://people.cs.uchicago.edu/ pff/latent-release4/.