Utilizing Mask R-CNN for Waterline Detection in Canoe Sprint Video Analysis

Marie-Sophie von Braun
Laboratory for Biosignal Processing
Leipzig University of Applied Sciences
msvbraun@gmail.com

Patrick Frenzel
Laboratory for Biosignal Processing
Leipzig University of Applied Sciences
patrick.frenzel@htwk-leipzig.de

Christian Käding
Institute for Applied Training Science Leipzig
kaeding@iat.uni-leipzig.de

Mirco Fuchs
Laboratory for Biosignal Processing
Leipzig University of Applied Sciences
mirco.fuchs@htwk-leipzig.de

Abstract

Determining a waterline in images recorded in canoe sprint training is an important component for the kinematic parameter analysis to assess an athlete's performance. Here, we propose an approach for the automated waterline detection. First, we utilized a pre-trained Mask R-CNN by means of transfer learning for canoe segmentation. Second, we developed a multi-stage approach to estimate a waterline from the outline of the segments. It consists of two linear regression stages and the systematic selection of canoe parts. We then introduced a parameterization of the waterline as a basis for further evaluations. Next, we conducted a study among several experts to estimate the ground truth waterlines. This not only included an average waterline drawn from the individual experts annotations but, more importantly, a measure for the uncertainty between individual results. Finally, we assessed our method with respect to the question whether the predicted waterlines are in accordance with the experts annotations. Our method demonstrated a high performance and provides opportunities for new applications in the field of automated video analysis in canoe sprint.

1. Introduction

Recording and analysing video sequences is a common method for the quantification, logging and optimization of the technique of athletes performing canoe and kayak sprint [22, 23]. A particularly important form is the recording from the position of a motorboat that moves in parallel direction to the canoe [30, 34]. While moving at the same speed, the athlete is recorded from an approximately perpendicular perspective with respect to the movement direction. This ensures standardized recording conditions which

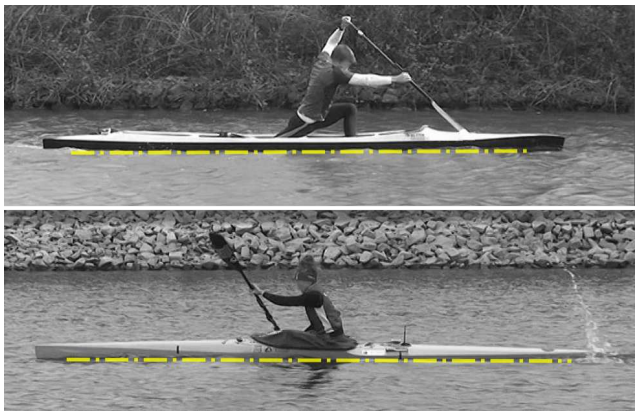

Figure 1. Illustration of the disciplines canoe sprint (top) and kayak sprint (bottom). The prediction of the waterline (dash-dotted yellow line) is crucial for the analysis of kinematic parameters to assess the technique of athletes.

are the basis to assess the performance of athletes and in particular of their technique. The actual analysis is then based on determining kinematic parameters and their comparison to well known target values.

The underlying recording conditions in the free environment are subject to variations (*e.g.* due to water movement, hand-held camera, inherent cyclic speed changes during paddling) that affect the comparability of the analyses performed on these recordings. Hence, an analysis is usually restricted to a narrow time range to minimize potential variations, *i.e.* often only a few or even a single paddle cycle. The technique is then analyzed based on several single images which are selected in accordance with a paddle phase model that defines the beginning, end and intermediate stage of a cycle [21, 15, 23]. The parameters determined in each of these images allow a comparison between multi-

ple athletes as well as between repeated training runs of the same athlete.

The analysis of these single images aims at measuring several distances and angles, *i.e*. the kinematic parameters, in the projected 2D image plane [11]. It comprises the identification of several landmarks as, in particular, key-points on the body of the athlete, a so called paddle line, that means a straight line through the longitudinal middle of the paddle shaft, and a waterline. As shown in Fig. 1, the latter approximates a boundary between the blurred water surface and the more well-defined hull of the canoe. The waterline is of particular interest for two reasons. First, it serves as a reference with respect to the kinematic parameters to be estimated. Second, its exact position and orientation is often difficult to specify in practice. Turbulences in water, waves and splashes, reflections, camera jitter, a varying camera position or even poor image quality give rise to ambiguity in this task.

The manual analysis procedure of these standardized video sequences is time consuming, requires expert knowledge and is also subject to individual errors. There is a general interest in an automatic procedure which provides a much faster analysis and which is also less prone to individual errors. The advances in image processing by means of deep neural networks in recent years pave the way for the development of a widely automated analysis of video recordings in canoe sprint. On the one hand, the rise of human 2D pose estimation algorithms [5, 6, 37, 17] and their task-specific optimization provides the opportunity to automatically determine key-points on an athlete's body in a given image [16]. The application of such algorithms for key-point detection has proven effective in a variety of applications such as, *e.g*., skeleton tracking of players in sports [3], swimming style classification [10] and stroke frequency detection [35], and pose mining in long jump [18]. There is reason to expect them to work in canoe sprint analysis as well. In fact, their potential use for video analysis in canoe sprint has recently been reported [12]. On the other hand, approaches for the pixel-wise segmentation of objects in an image such as Mask R-CNN [13] might serve as a basis for an automated detection of the canoe and the paddle which can subsequently be used to determine the waterline and the paddle line, respectively. However, the prospects of success are much less obvious compared to pose estimation.

This work presents an approach for an automated detection of the waterline. Our method is based on image segmentation using the Mask R-CNN network pre-trained on the COCO dataset [19] and a subsequent multi-stage procedure that includes two linear regression steps. As to the mentioned uncertainty of waterlines defined by several experts, we conducted an evaluation study and derived a gold standard to assess the predicted waterlines.

The contributions of this paper include (1) an adoption of Mask R-CNN for canoe segmentation and discrimination of the disciplines canoe sprint and kayak sprint (compare Fig. 1), (2) a procedure to estimate a waterline given the segmented shape of the canoe, (3) a gold standard to assess predicted waterlines with respect to human experts and (4) a performance analysis of the proposed waterline detection method.

2. Related Work

2.1. Mask R-CNN in Sports Applications

Since its superiority in the instance segmentation task of the COCO Challenge 2017, Mask R-CNN has been widely used for scene analyzing in sports videos. The applications range from ball detection [4] to jersey number recognition [20], and further to player tracking [28] and events identification [33, 27].

Challenges in these applications are amongst others the dynamics of the subject and the numerous occlusions of the tracked object that occur during the game. When analyzing canoe sprint videos though, there is an additional challenge: the peculiarity of the medium water, which render a robust and accurate detection of the waterline difficult.

2.2. Waterline Problem

Several works focus on estimating the waterline on images or videos of rivers or lakes in order to detect the sailing area of an autonomous boat. Wei and Zhang [36] present a waterline detection method based on texture analysis of river images with local binary patterns (LBPs) and gray level co-occurrence matrix (GLCM). Steccanella *et al*. [32] apply a supervised approach based on a Fully Convolutional Neural Network for obtaining a pixel-wise image segmentation.

These methods however rely on a detection of the water area and focus on its boundary with the horizon line. Hence, they cannot be used to estimate the separation line between the lower part of a canoe hull and the water surface. Our case requires a segmentation of the canoe hull within the water body. The use of Mask R-CNN for this case is the subject of several papers [31, 39, 25, 38, 24, 26]. However, these approaches are applied exclusively to satellite images. Due to the aerial perspective and their remote sensing character, these are not comparable to the video sequences of canoe sprints that are examined here.

3. Proposed Approach

3.1. Task definition

The goal of our work is to determine a waterline in an RGB image $I \in \mathbb{R}^{m \times n \times 3}$ drawn from video sequences (50 frames per second) in the disciplines canoe sprint and kayak sprint (see Fig. 1). Here, the spatial resolution is

$m = 1024 \times n = 576$ pixels. It is assumed that the images are recorded from an approximately perpendicular perspective with respect to the movement direction of the canoe. Moreover, only minor variations of the distance and the relative position between canoe and motorboat are expected for consecutive images selected from the short time window to be analyzed in the training run of an athlete.

The determination of the waterline is a regression problem. The goal is to approximate a straight line that separates the visible part of the canoe from the invisible part below the water surface. Waves, splashes and other disturbance that might occlude the canoe hull must be considered when approximating the line. The linear approximation of the waterline should be particularly accurate in the central part of the canoe right below the athlete, since this segment is notably important for the subsequent derivation of the kinematic parameters.

We propose a two-staged approach for waterline prediction. First, it is based on a pixel-wise segmentation of the canoe by means of a Mask R-CNN that we adjusted to this particular task using transfer learning. This is presented in Sec. 3.2. Second, it employs a multi-stage procedure to confine the pixels of the canoe segmentation to those close to the water surface which can finally be used to define a waterline. This is shown in Sec. 3.3.

3.2. Canoe Segmentation with Mask R-CNN

3.2.1 Method

The first stage of our approach is a pixel-wise instance segmentation of each canoe object contained in the image. The Mask R-CNN method proposed by He *et al*. [13] has evolved to a state of the art approach for pixel-wise instance segmentation. It is a two-stage framework built on top of a Faster R-CNN [29]: the first stage generates object proposals, while the second stage predicts the class of each object, refines its bounding box and generates a corresponding binary mask on pixel level. Both stages are connected to a backbone, in our case a ResNet101 [14] paired with a Feature Pyramid Network, that serves as a feature extractor. Hence, such a net is able to detect the set of objects Ω, with its elements $\omega_v = (M_v, c_v, p_v) \in \Omega, v \in \mathbb{N}$ defining the class c_v, its confidence value $p_v \in \mathbb{R}, 0 < p_v \leq 1$ and its binary mask $M_v \in \{0,1\}^{m \times n}$ of an instance. The latter provides a pixel-wise binary appearance of an object in an image.

The particular segmentation problem in canoe sprint video analysis is highly specific. That means that the algorithm is not required to identify any additional object in an image but only the canoe of an athlete. We exploited this fact and restricted the potential output objects to canoes used in canoe sprint and kayak sprint (*i.e*., $c_v \in [canoe, kayak]$), both of which are actually canoes but with a slightly different appearance. The segmented canoe as defined by its binary segmentation mask M_v can subsequently be used to determine the lower part of the outline of the canoe.

3.2.2 Dataset and Implementation Details

We adopted the Mask R-CNN implementation as provided by Matterport [2] which employs a ResNet101 architecture as a backbone [14]. It is implemented in TensorFlow [1] and Keras [7]. As described above, we restricted the output layer for the segmentation to only two types of objects, *i.e*. canoes for the disciplines canoe sprint and kayak sprint. We applied transfer learning to train our model. Therefore, we used the pre-trained weights that resulted from training the Matterport implementation on the COCO dataset [19]. We used the following training parameters: 400 iterations, 300 steps per iteration, SGD solver, learning rate 0.001, momentum 0.9, one image batch size and weight decay 0.0001.

We carried out image annotation to derive training and test sets as follows. Given were a total number of 66 video sequences from both disciplines from which 250 images were randomly selected. We used the VGG Image Annotator [9, 8] to define polygones that mimic the canoe hull in each image. Next, we used 210 images (58 from canoe sprint, 152 from kayak sprint) to define a training set and 40 for the validation set (11 canoe sprint, 29 kayak sprint). Moreover, we selected 30 of these images for the validation set in a way that ensured that they belong to video sequences from which no other image is used for training. In case of the other 10 images, the canoes they contain are already known to the model due to the training process.

During training, the image was either kept in its original form ($p = 0.5$) or processed using data augmentation as follows: flipping in horizontal direction ($p = 0.5$), rotation by 2 degree and cropping/padding in a range from -15 % to 15 % in both image dimensions. The former represents a zoom-in effect, the latter corresponds to zooming out.

3.3. Waterline Detection

The second stage of our approach comprises an iterative procedure to derive a waterline given the binary segmentation mask $M \in \{0,1\}^{m \times n}$ as provided by the segmentation approach presented in the previous section. The procedure is depicted in Fig. 2, step 1 shows the initial segmentation. First, all points $C \subset M$ that represent the contour of the canoe are determined (step 2). As a result, C contains at least two tuples for each image coordinate x_i where the canoe segment was found, *i.e*. (x_i, y_1) and (x_i, y_j), $j \in \mathbb{N}, j \geq 2$, all of which belong to the contour. It is obvious that the waterline is close to the bottom of the canoe and we therefore reject the tuples belonging to the upper part and keep the others. Since we defined $y = 0$ at the top of the image, the tuples with the larger coordinate, *i.e*.

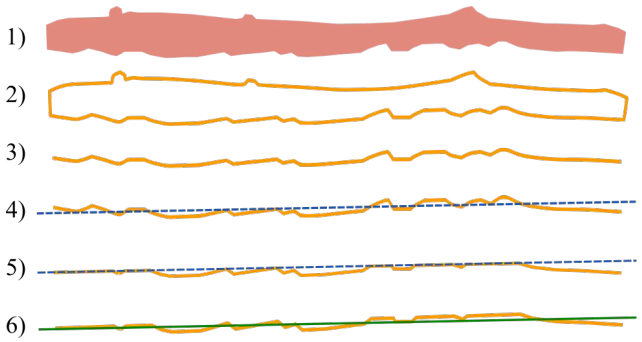

Figure 2. Illustration of the iterative procedure to predict a waterline on the basis of a canoe segmentation. Details for steps 1-6 are provided in the text. The dashed blue lines in steps 4 and 5 are the same. The solid green line in step 6 represents the predicted waterline. It corresponds to the waterlines illustrated in Fig. 1.

$max\,(y_1, ..., y_j)$, are used for further processing, leading to C'. Hence, $C' \subset C \subset M$ defines the set of points belonging to the bottom line of the canoe contour (step 3). Third, a linear regression $\mathcal{L}_1\,(C')$ is performed on the set of points contained in C' (step 4). All points above this regression line are subsequently removed to form $C'' \subset C'$ (step 5). This step mimics cropping of small waves and splashes. Finally, another linear regression step $\mathcal{L}_2\,(C'')$ is performed on the set of points in C'' (step 6). Its result defines the predicted waterline.

4. Evaluation

The evaluation of predicted waterlines requires reference data to assess its accuracy. However, quantitative reference data, for example, derived from passive optical markers does not exist. Moreover, it would only be hardly possible to collect this sort of data. Hence, the manual definition of waterlines by human experts is the only possibility to derive ground truth references. However, the task of defining a waterline in an image is subject to individual perception to some extent. As a result there often is no unique solution but rather different experts will provide several different waterline estimates for the same image. Hence, there is a narrow range within which waterlines defined by different experts can be expected. Since our goal is to provide a method that mimics an expert when solving the regression problem to define a waterline, it is necessary to answer the question whether the predicted waterline is in accordance with this narrow range of ambiguity. To put it more simple, the question should be answered whether a waterline prediction would be accepted by experts. We accounted for this and performed an evaluation as follows.

First, we assessed the quality of canoe segmentation as well as the inherently related classification of the particular disciplines (see Sec. 4.1). Second, we conducted a small study among several experts in the field of kinematic parameter analysis in canoe and kayak sprint (see Sec. 4.2). The study was the basis to define a ground truth reference as well as to quantify the uncertainty among experts. We subsequently assessed the accuracy of our predictions with respect to this gold standard.

4.1. Canoe Segmentation and Classification

The segmentation of the canoe is important for the subsequent waterline prediction. We assessed the segmentation quality of the adjusted and trained Mask R-CNN using the validation set according to a standard evaluation metric. We used the intersection-over-union (IoU) defined as

$$IoU = \frac{M' \cap M}{M' \cup M} \tag{1}$$

to measure the overlap between the predicted segmentation $M' \in \{0,1\}^{m \times n}$ and the ground truth mask $M \in \{0,1\}^{m \times n}$ of the canoe. The former was selected from the predicted objects Ω as the one with the highest confidence value p_v.

We also assessed the classification performance of the algorithm. To that end, we evaluated whether the algorithm predicts the disciplines canoe sprint and kayak sprint correctly. We therefore determined the true and false positives/negatives on the validation set separately for each discipline and used them to calculate the corresponding F1 scores.

4.2. Waterline Detection

The algorithm proposed in this work predicts the course of a waterline. Here, we present our evaluation procedure as well as the necessary preliminaries.

The parameterization of the waterline and an evaluation metric is presented in Sec. 4.2.1. An evaluation study that was carried out to derive individual annotations (*i.e.*, the determination of a waterline) for each image in the test set from several human experts is presented in Sec. 4.2.2. Finally, the actual derivation of the ground truth reference data and the ambiguity between different expert annotations that can finally be used for the performance analysis of the algorithm are presented in Sec. 4.2.3.

4.2.1 Parameterization and Evaluation Metric

The evaluation of a predicted waterline requires a suitable parameterization in order to perform comparisons with a reference. Each waterline is a linear function and is clearly defined by its slope and a bias, *i.e.* an interception with the coordinate axis at $x = 0$. Without the loss of generality, we used a different parameterization which is shown in Fig. 3. It allows a better interpretation and comparison of the evaluation results. Firstly, it consists of a height parameter h

which is defined as the y-coordinate of the waterline at the center postion of the image (in x-direction), thereby effectively representing the location of the line. Secondly, an angle α which defines the rotation of the waterline with respect to the horizontal line is another parameter.

Note that this parameterization implies a rough similarity between the waterlines in different images being evaluated. The consistence with respect to their position at and their extent along the x-direction is particularly important to derive an estimate for the uncertainty between experts (see below). Due to the reasonably controlled conditions during video recording in our particular application scenario, this convention can be considered as true.

Given a particular waterline, the deviation between a given ground truth height h_i and angle α_i and the predicted parameters h_i' and α_i' in the i-th image defines as

$$\epsilon_i^h = |h_i - h_i'| \tag{2}$$

$$\epsilon_i^\alpha = |\alpha_i - \alpha_i'| \tag{3}$$

The actual definition of a suitable ground truth for the parameters h_i and α_i for each image is subject of the evaluation study presented below.

4.2.2 Evaluation study

As mentioned before, a ground truth for the waterlines can only be defined on the basis of manual annotations. We conducted an evaluation study to derive multiple human annotations for each image in our test set for waterline evaluation. Therefore, we asked several experts from the field of kinematic parameter analysis in canoe sprint to determine a waterline. The implementation details were as follows.

The study was implemented on the basis of an interactive website. Given a test set (see below), we presented each image together with an initial guess of the waterline to each expert. The task of each expert was to carefully review the presented waterline and afterwards either modify its position and orientation by means of moving anchors at its ends or to accept the guessed line without any changes. Thereby, we controlled the initial guess as described below to prevent habituation to accept guessed lines without extensive review.

A total number of 130 images were selected from 66 videos to construct a test dataset T. 44 images were from canoe sprint and 86 from kayak sprint. We used these images to construct 4 groups within the test set. Group A: 90 images, the waterline as predicted by the algorithm without further modifications; Group B: 20 images drawn from group A, an additional offset of -3 pixels was added to the waterline; Group C: 10 images, a vertical shift of $+2$ pixels and a $-1.5\,^\circ$ rotation was added to the line as predicted by the algorithm; Group D: 10 images, similar processing

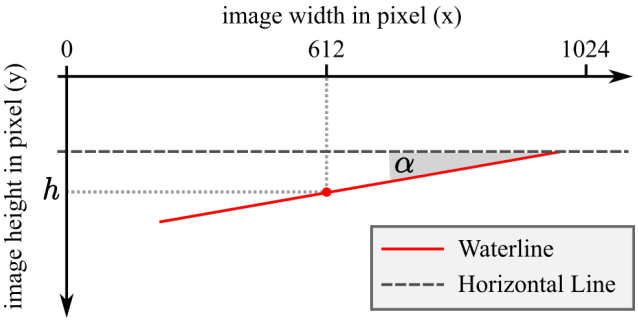

Figure 3. Waterline parameterization: h defines as the y-coordinate of the waterline at the center location in x-direction ($x = 612$); α is the angle between waterline and horizontal line.

as in group C, but with the rotation into the opposite direction, *i.e.* $+1.5\,^\circ$. The groups B-D were used to enforce a misalignment of the presented waterline so that the participants are expected to perform modifications. The size of the distortions were selected by means of explorative tests to achieve variations that are visible but not trivial. The images were presented in a random order and without disclosing the group.

4.2.3 Ground Truth and Ambiguity of Waterlines

The dataset T resulting from the evaluation study comprised multiple individual annotations for each image. We exploited this information to determine a ground truth waterline for each image as well as an estimate for the variation of the experts annotations as follows.

First, provided the individual parameters $h_{i,k}$ und $\alpha_{i,k}$, with $k \in N_i$ and N_i being the set of the individual experts, we calculated the mean value for each of the two parameters for each image $i, i \in T$, *i.e.*

$$h_i = 1\,/|N_i| \cdot \sum_{k \in N_i} h_{i,k} \tag{4}$$

$$\alpha_i = 1\,/|N_i| \cdot \sum_{k \in N_i} \alpha_{i,k} \tag{5}$$

$|N_i|$ denotes the number of experts who annotated the i-th image. Next, the deviation of each individual parameter $h_{i,k}$ and $\alpha_{i,k}$ to the corresponding mean values h_i and α_i were calculated by means of

$$\epsilon_{i,k}^h = h_{i,k} - h_i \tag{6}$$

$$\epsilon_{i,k}^\alpha = \alpha_{i,k} - \alpha_i \tag{7}$$

We employed these differences for further statistical analysis. We were particularly interested in the question whether

there is statistical evidence that the individual annotations provided by different experts are similar and can therefore be used to calculate an average annotation for each image as introduced before in Eqs. 4 and 5. We applied a Kruskal-Wallis test as a non-parametric method to compare the distributions of the individual differences to the ground truth estimates, separately for the height and rotation parameter. The null hypothesis is that the medians of these distributions are equal which would support the assumption that they originate from the same population. If the null hypothesis cannot be rejected in the light of the data, the parameters h_i and α_i as introduced above seem a plausible approximation for the ground truth in each image. Hence, they can be used as the reference for further performance analysis.

Given this ground truth for each image, we were still interested in the overall variation of the individual annotations. Based on the individual deviations according to Eqs. 4 and 5 for the entire dataset, we calculated the standard deviation for both waterline parameters, *i.e.*

$$\sigma_h = \sigma\left(\epsilon^h_{i,k}\right) \forall k \in N_i, i \in T, \text{and} \tag{8}$$

$$\sigma_\alpha = \sigma\left(\epsilon^\alpha_{i,k}\right) \forall k \in N_i, i \in T, \tag{9}$$

These estimates serve as a general measure for the uncertainty among all experts. Using these measures, we finally defined an acceptance range, that means an interval in the vicinity of the ground truth reference parameters h_i and α_i, within which a predicted waterline would be considered as a valid estimate. This interval was constructed as the u-fold σ_h and σ_α vicinity,

$$\Delta h = \pm u \cdot \sigma_h \tag{10}$$

$$\Delta \alpha = \pm u \cdot \sigma_\alpha \tag{11}$$

The parameter u was determined such that $95\,\%$ of all individual annotations are contained within this range.

5. Results

5.1. Canoe Segmentation and Classification

The segmentation quality on the test dataset measured in terms of the IoU was 0.82 on average with a standard deviation of 0.04 and minimal and maximal values of 0.72 and 0.88, respectively. This corresponds to a moderate and, more importantly, very consistent segmentation quality. The results clearly indicate a strong overlap between the true and the predicted masks of the canoe, which is particularly important for the subsequent waterline estimation.

The result of the classification performance analysis revealed a F1 score of 1.0 for both disciplines. It means that the Mask R-CNN was perfectly able to distinguish between canoe and kayak sprint. The classification performance itself is only less important for the subsequent estimation of the waterline. However, it might be utilized in an automated processing pipeline to discriminate different paths in the analysis procedure which depends on the particular discipline.

5.2. Ground Truth Waterline Parameters

A total number of 7 experts participated in our study, 6 of which processed all 130 images and one processed only 48. This led to a total of 828 annotated waterlines in the test dataset. The distributions of the individual deviations from the ground truth as determined according to Eqs. 10 and 11 are shown in Fig. 4. The visual comparison of the distributions reveals a general consent between experts. The rotation parameters are in stronger accordance to each other compared to the height parameters. This is also reflected in the upper and lower quartiles of these distributions, which are located within a range of only $\pm 0.22\,^\circ$ for the rotation and $\pm 1.96\,px$ for the height parameter. Note the minor different appearance of the results for participants 4, 5 and 7 for the height parameter and for subject 6 for the rotation parameter compared to the other participants. These distributions are shifted slightly downwards for the height parameter for the participants 4 and 5 and upwards for participant 7. Regarding the rotation, the distributions standard deviation for participant 6 is considerably larger compared to all others.

The Kruskal-Wallis test was performed separately for each parameter given the null hypothesis that the medians of the distributions are similar using a significance level of $p = 0.05$. We obtained the p-values $p = 0.65$ for the rotation and $p = 0.99$ for the height parameter. This indicates that the null hypothesis cannot be rejected in the light of the data. From this we concluded to estimate the ground truth reference for each waterline from the individual annotations provided by experts according to Eqs. 4 and 5 separately for each image.

Next, we determined the standard deviation of the expert annotations according to Eqs. 8 and 9 to $\sigma_h = 1.48\,px$ for the height parameter and $\sigma_\alpha = 0.20\,^\circ$ for the rotation parameter. Finally, we used Eqs. 10 and 11 to estimate $u = 2.5$ for the u-fold σ_h and σ_α vicinity around the ground truth parameters h_i and α_i, provided the assumption that $95\,\%$ of the individual expert annotations should be contained in this vicinity for both height and angle parameter. The resulting intervals are $\Delta h = \pm 3.70\,px$ and $\Delta \alpha = \pm 0.50\,^\circ$. The tolerance interval for the height parameter corresponds to less than $\pm 0.7\,\%$ of the spatial resolution (height dimension) of the images in the dataset.

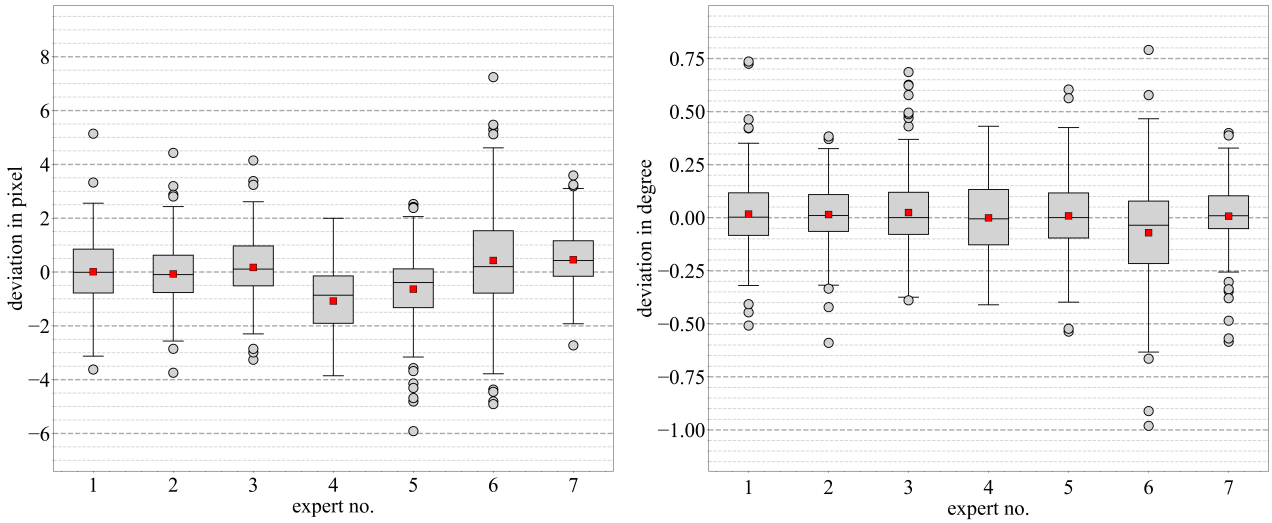

Figure 4. Distributions of the individual deviations from the estimated ground truth references shown for the height (left) and the rotation parameter (right). Upper / lower boundaries of boxes correspond to upper / lower quartiles. Whiskers denote the 1.5-fold of the inter quartile range; circles denote outliers; vertical bars / red squares inside the boxes denote median / mean values.

5.3. Accuracy of Predicted Waterlines

We assessed the accuracy of the predicted waterlines by calculating the absolute differences to the ground truth parameters by means of Eqs. 2 and 3. The results for each discipline and the combined results are shown in Fig. 5. It is obvious that the results obtained for canoe sprint appear to be slightly worse than for kayak sprint. Besides that it is shown that 50 % of the absolute differences for both height and rotation are less or equal than $1.26\,px$ and $0.19\,^{\circ}$, respectively, considering the distributions from the combined results. The largest values given the error metric are $5.57\,px$ for the height and $0.82\,^{\circ}$ for the rotation parameter.

Finally, we applied the u-fold σ_h and σ_α vicinity as determined in the previous section to assess whether a waterline can be considered as a valid expert estimate. It turns out that a total number of 85 % of all predicted waterlines are in accordance with this interval. If the parameters are considered separately, even 95 % of the results for the rotation and 89 % for the height parameter fall into these intervals.

6. Discussion and Conclusions

We introduced an approach for the automatic detection of the waterline in canoe sprint video analysis. Our general goal was to provide an estimate for the course of waterlines in an image that could also have been defined by a human expert or, in other words, that experts would accept this prediction as a valid estimate. We achieved this for 85 % of the images in the validation dataset. Our solution for this particular regression problem comprises the segmentation of canoes based on an adjusted Mask R-CNN approach and a subsequent multi-stage procedure to estimate a waterline. We demonstrated its performance on a real dataset for which the ground truth references were derived on the basis of human expert annotations. Our solution provided robust and accurate results while still leaving space for further improvements and optimizations. They might not only refer to the segmentation part and the waterline estimation procedure but also to the derivation of the ground truth references.

Defining a suitable reference that can be used for further performance evaluations is a particular problem of this kind of regression tasks, *i.e.* for which a ground truth cannot be

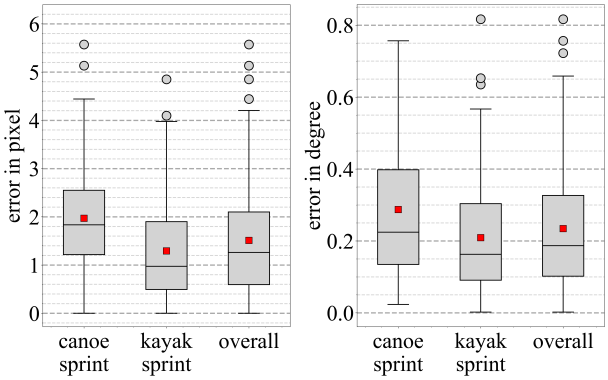

Figure 5. Error of predicted waterlines for height (left) and rotation parameter (right). Upper / lower boundaries of boxes correspond to upper / lower quartiles. Whiskers denote the 1.5-fold of the inter quartile range; circles denote outliers; vertical bars / red squares inside the boxes denote median / mean values. See text for details.

determined otherwise, *e.g*. by means of sensors. In fact, the actual problem here is that the definition of a waterline is an ambiguous problem (caused by waves, splashes, a.s.o.), since it is prone to errors due to individual perception. As a consequence the results obtained from several experts for the same test image are subject to small variations. Hence, the ground truth can only be defined on the basis of an average result from several experts. Moreover, it is necessary to determine the amount of the variation between individual annotations. In fact, only the latter provides an actual meaning for the predicted waterline during evaluation, namely that a prediction is in accordance with the consent of experts. The quantification of these variations needs to be done carefully.

Here, we derived a ground truth and an estimate of the variation by means of a study among a small group of experts. This is certainly a limitation of our procedure, since it led to only a rather small number of annotations per image and, therefore, might have reduced the validity of the ground truth references. However, we assume that the domain knowledge of experts to define the waterlines implies that their individual deviation from the average annotation is small at all, and so the mean value of their annotations can be considered an appropriate estimate. The calculation of the overall ambiguity among experts is less affected since it was derived from the distribution of all deviations in the dataset rather than from individual images. Nonetheless, our gold standard definition is only valid with respect to the experimental conditions of the video recordings, *e.g*. the perspective of the camera, the distance to the canoe and the spatial resolution of the images.

As mentioned before, 85 % of the predicted waterlines were in accordance with the gold standard we derived from the evaluation study and so 15 % were not. Importantly, our results show that the magnitudes of these outliers are still moderate. It is obvious that these proportions depend upon the definition of the u-fold σ_h and σ_α vicinities. Here, it was selected such that 95 % of individual annotations were part of this interval. Less strict assumptions would of course improve the success rate. Further data is needed to derive a more sophisticated selection for this value. Moreover, outlier samples are worth to be analyzed separately and in more detail in order to identify potential systematic errors and to achieve further optimizations.

We carried out a brief error analysis and found a clear pattern for images that were not in accordance with experts annotations. These waterlines were slightly shifted upwards in the frontal part of the canoes compared to the ground truth. There is reason to believe that this is caused by larger waves in the frontal area resulting from, *e.g*., the cyclic movement of the canoe in upward and downward direction which is inherent to these disciplines. As a result, the waves occlude visible parts of the canoes front which shifts the segmentations, their outlines and finally also the waterlines. A possible solution is to not only restrict the linear regression to more central areas of the canoe segmentation but also to improve the canoe segmentation itself.

The good quality of the segmentation performance of the adjusted Mask R-CNN is effectively reflected by a high average IoU value. The very small standard deviation underpins its general robustness, although the amount of data available for training and validation was fairly small. Increasing the amount of data might improve the performance significantly. Further limitations in the current dataset are unbalanced proportions of the samples with respect to the movement direction of canoes and to the actual discipline, *i.e*. canoe or kayak sprint. Moreover, the dataset does not contain any negativ samples, that means images without any canoe. However, this is only of minor relevance if it can be ensured that the algorithm is applied to application specific data.

The developed method provides an important component for future developments towards an automated derivation and analysis of kinematic parameters from video and image recordings in canoe sprint and kayak sprint. A straightforward extension is the application and optimization of algorithms for human pose detection, *e.g*. OpenPose [5], which can provide coordinates of key-points on limb and face of athletes. Assessing such key-point positions with respect to the waterline can be used to derive kinematic parameters comparable to those used in todays human analysis [12]. Another extention is the detection of the paddle, which is important for several reasons. First, it provides another reference for key-point positions and the subsequent kinematic parameter analysis. Second, it provides information on the relative time in a paddle cycle if evaluated in comparision to the waterline. The Mask R-CNN approach might be applied to the task of paddle segmentation as well.

Finally, the combination of these approaches paves the way for new applications in which not only several images but rather an entire video sequence can be analyzed. This provides new opportunities as, *e.g*., the utilization of temporal filters on the extracted parameters to achieve more robust predictions in single images but also to exploit the dynamics of kinematic parameters for the biomechanical analysis.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B Viégas, Oriol Vinyals,

Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *ArXiv*, abs/1603.04467, 2016. Version: 1.13.1.

[2] Waleed Abdulla. Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow. `github.com/matterport/Mask_RCNN`, 2017.

[3] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2487–2496, 2019.

[4] Matija Buric, Miran Pobar, and Marina Ivašić-Kos. Ball detection using Yolo and Mask R-CNN. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 319–323, 2018.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[6] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Bottom-up higher-resolution networks for multi-person pose estimation. *ArXiv*, abs/1908.10357, 2019.

[7] François Chollet et al. Keras. `keras.io`, 2015. Version: 2.2.4.

[8] Abhishek Dutta, Ankush Gupta, and Andrew Zisserman. VGG image annotator (VIA). `www.robots.ox.ac.uk/~vgg/software/via/`, 2016. Version: 2.0.5.

[9] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 2276–2279. ACM, 2019.

[10] Moritz Einfalt, Dan Zecha, and Rainer Lienhart. Activity-conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 446–455, 2018.

[11] Matthias Englert and Gerd Lehmann. Wissenschaftlicher Einsatz der Software utilius® easyInspect zur Analyse der individuellen Fahrtechnik im Kanusport. *Leistungssport*, 37(5):20–22, 2007.

[12] Mirco Fuchs, Piet Wagner, Gerold Bausch, and Patrick Frenzel. Kamerabasierte Erfassung von Skelettdaten und Vitalparametern. In *Proceedings 19. Frühjahrsschule Technologien im Leistungssport*, volume 13 of *Schriftenreihe für Angewandte Trainingswissenschaft*, pages 62–73. Meyer & Meyer Verlag, 2018. ISBN 978-3-8403-7628-3.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] Hartmut Herrmann. *Untersuchungen zum sporttechnischen Leitbild im Einer-Canadier des Kanu-Rennsports aus biomechanischer Sicht*. Dissertation A, Deutsche Hochschule für Körperkultur, 1979.

[16] Ying Huang, Bin Sun, Haipeng Kan, Jiankai Zhuang, and Zengchang Qin. FollowMeUp Sports: New benchmark for 2D human keypoint recognition. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 110–121. Springer, 2019.

[17] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. MultiPoseNet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, pages 437–453, 2018.

[18] Rainer Lienhart, Moritz Einfalt, and Dan Zecha. Mining automatically estimated poses from video recordings of top athletes. *International Journal of Computer Science in Sport*, 17(2):112–94, 2018.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[20] Hengyue Liu and Bir Bhanu. Pose-guided R-CNN for jersey number recognition in sports. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2457–2466, 2019.

[21] Ralph Mann and Jay T Kearney. Biomechanics of canoeing and kayaking. In *Scientific proceedings of the 1st international symposium on biomechanics in sports*, pages 145–151, 1983.

[22] Lisa K McDonnell, Patria A Hume, and Volker Nolte. An observational model for biomechanical assessment of sprint kayaking technique. *Sports biomechanics*, 11(4):507–523, 2012.

[23] Jacob S Michael, Richard Smith, and Kieron B Rooney. Determinants of kayak paddling performance. *Sports Biomechanics*, 8(2):167–179, 2009.

[24] Mingxian Nie, Jinjie Zhang, and Xuetao Zhang. Ship segmentation and orientation estimation using keypoints detection and voting mechanism in remote sensing images. In *Advances in Neural Networks - Proceedings (II) of 16th International Symposium on Neural Networks (ISNN)*, pages 402–413, 2019.

[25] Shanlan Nie, Zhiguo Jiang, Haopeng Zhang, Bowen Cai, and Yuan Yao. Inshore ship detection based on Mask R-CNN. In *2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 693–696, 2018.

[26] Xuan Nie, Mengyang Duan, Haoxuan Ding, Bingliang Hu, and Edward Wong. Attention Mask R-CNN for ship detection and segmentation from remote sensing images. *IEEE Access*, 8:9325–9334, 2020.

[27] Miran Pobar and Marina Ivašić-Kos. Mask R-CNN and optical flow based method for detection and marking of handball actions. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6, 2018.

[28] Miran Pobar and Marina Ivašić-Kos. Detection of the leading player in handball scenes using Mask R-CNN and STIPS. In Antanas Verikas, Dmitry P. Nikolaev, Petia Radeva, and

Jianhong Zhou, editors, *Eleventh International Conference on Machine Vision (ICMV 2018)*, volume 11041, pages 501–508. International Society for Optics and Photonics, SPIE, 2019.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 91–99, 2015.

[30] Michael G Robinson, Laurence E Holt, Thomas W Pelham, and Karen Furneaux. Accelerometry measurements of sprint kayaks: The coaches' new tool. *International Journal of Coaching Science*, 5(1):45–56, Jan 2011.

[31] David Schweitzer and Rajeev Agrawal. Multi-class object detection from aerial images using Mask R-CNN. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3470–3477, 2018.

[32] Lorenzo Steccanella, Domenico Bloisi, Jason Blum, and Alessando Farinelli. Deep learning waterline detection for low-cost autonomous boats. In *15th International Conference on Intelligent Autonomous Systems (IAS)*, pages 613–625, 2018.

[33] Lamia Kazi Tani, Abdelghani Ghomari, and Mohammed Kazi Tani. Events recognition for a semi-automatic annotation of soccer videos: a study based deep learning. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W16:135–141, 2019.

[34] Cheryl Sihui Tay and Pui Wah Kong. A video-based method to quantify stroke synchronisation in crew boat sprint kayaking. *Journal of Human Kinetics*, 65(1):45–56, 2018.

[35] Brandon Victor, Zhen He, Stuart Morgan, and Dino Miniutti. Continuous video to simple signals for swimming stroke detection with convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 122–131, 2017.

[36] Yangjie Wei and Yuwei Zhang. Effective waterline detection of unmanned surface vehicles based on optical images. *Sensors*, 16(10):1590, 2016.

[37] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, pages 472–487, 2018.

[38] Yanan You, Jingyi Cao, Yankang Zhang, Fang Liu, and Wenli Zhou. Nearshore ship detection on high-resolution remote sensing image via scene-Mask R-CNN. *IEEE Access*, 7:128431–128444, 2019.

[39] Yankang Zhang, Yanan You, Rui Wang, Fang Liu, and Jun Liu. Nearshore vessel detection based on scene-Mask R-CNN in remote sensing image. In *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 76–80, 2018.