Optimizing Human Pose Estimation for Automatic Analysis of Competitive Swimmers

Ir. Arne Vandendorpe
Ghent University
Ghent, Belgium

Prof. Dr. Ir. Tom Dhaene
Ghent University
Ghent, Belgium

Dr. Joachim van der Herten
Ghent University
Ghent, Belgium

***Abstract*—Advances in the field of human pose estimation have significantly improved performance across complex datasets. However, current solutions that were designed and trained to recognize the human body across a wide range of contexts, e.g. MS COCO, often do not reach their full potential in very specific and challenging environments. This impedes subsequent analysis of the results. Underwater footage of competitive swimmers is an example of this, due to frequent self-occlusion of body parts and the presence of noise in the water. This work aims to improve the performance of pose estimation in this context in order to enable an automatic analysis of kinematics. Therefore, we propose a framework that limits the search space for human pose estimation by using a set of anchor poses. More specifically, the problem is reduced to finding the best matching anchor pose and the optimal transformation thereof. To find this best match, we devise a method of assessing similarity between two poses and use the Viterbi algorithm to find the most likely sequence of anchor poses. Thereby, we effectively exploit the cyclic character of the swimming motion. This does not only improve pose estimation performance but also provides a method to reliably extract the stroke frequency, outperforming manual timings by a human observer.**

I. Introduction

With data analysis playing an ever-increasing role in sports, computer vision is often tasked with providing a cost-effective, flexible and scalable way of automated data collection. Compared to the use of wearable sensors, computer vision enables non-invasive collection of data as it involves minimal obstruction for the athletes. In this work, we further explore the applicability of computer vision for the analysis of swimming technique. In practice, analysis of swimming technique is often conducted by using video footage obtained through underwater windows or periscope systems. Therefore, this context provides a suitable use-case for an approach using computer vision.

The authors of [1] establish that, even for a human observer, competitive swimming is a particularly challenging subject for research and analysis due to the complexity of human motion and the aquatic environment. They also identify horizontal velocity as the most relevant metric for performance and break it down into its two constituent parts: stroke frequency and stroke length, the latter indicating the horizontal distance traveled during a full stroke cycle. This work proposes a framework that could act as a basis to extract both, based on human pose estimation with machine learning. We focus on extracting only the stroke frequency for one of the four competitive swimming strokes: the front crawl.

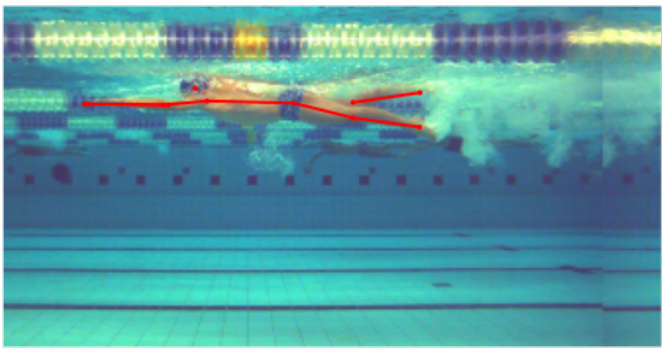


Fig. 1: Example of an underwater image with pose estimation result from the finetuned baseline model.

Human pose estimation is an active field where the state-of-the-art is rapidly improved upon and where implementations and pre-trained algorithms are often made publicly available. Examples include AlphaPose [2] and Mask-RCNN [3]. This greatly simplifies the adaptation of human pose estimation for real life use-cases such as sports analysis. Nevertheless, sports are usually associated with complex body positions, fast movements and noisy environments. Consequently, the performance of these solutions may be considerably worse in these contexts compared to their benchmarking datasets. This is true for competitive swimmers filmed underwater, as supported by [4]–[6].

A possible direction to improve existing human pose estimation networks involves retraining on relevant data of swimmers to improve pose retrieval. This would require creating a large labeled dataset of swimmers which would be costly and time-intensive. Instead, we collected a limited dataset to finetune the Mask-RCNN [3] model and benchmark its performance. Figure 1 shows an example of a prediction by this finetuned model, further referred to as the baseline model. In this work, we propose a framework to improve upon the baseline performance and enable the extraction of stroke frequency. The framework consists of three steps: baseline prediction, pose matching and extraction of the most likely sequence of pose matches. A visual overview is given by Figure 2.

The framework starts with the prediction by the baseline

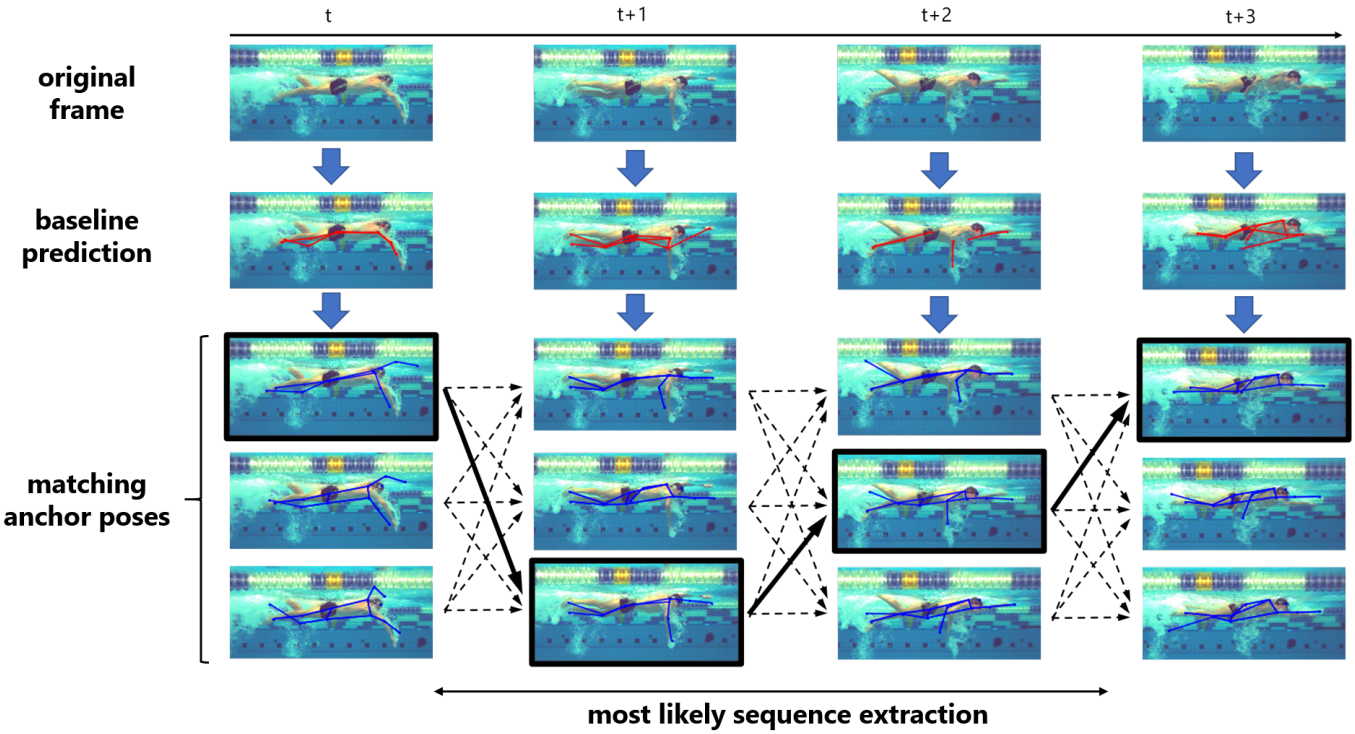


Fig. 2: The different steps of the framework are visualized: The first row displays the four original consecutive frames and the second row shows, in red, the predicted keypoints by the baseline model. Below each prediction, three transformed anchor poses with a high similarity score to the baseline prediction are shown. The most likely sequence of anchor poses as extracted by the Viterbi algorithm is indicated in bold.

model, illustrated by the second row of Figure 2, and exploits domain-specific knowledge in subsequent steps. We identify occlusion, i.e. a number of joints not being visible in the image, as the main source of baseline errors. More specifically, this often causes inversion errors, which indicate that a joint's location is predicted on the ground truth location of its symmetric counterpart. For example, when the swimmer's left wrist is occluded in the image, it is more likely to be falsely predicted on the location of the right wrist.

The second step of the framework involves pose matching and uses the baseline prediction as input. Furthermore, it receives as additional input a set of anchor poses, which is assumed to contain all poses that a swimmer could take during one full stroke cycle. A matching algorithm is then devised to assess the similarity between two poses. This enables replacing each baseline prediction with the most similar anchor pose, subject to translation or rotation. Hence, this step constrains the search space of the framework to anchor poses and transformations thereof. The bottom segment of Figure 2 shows how the three most similar anchor poses have been matched to each baseline prediction. Since the matching algorithm only considers the single-frame pose prediction, it is sensitive to baseline errors. As a result, it might select a wrong anchor pose as the best match. Moreover, a high number of inversion errors often leads to wrongfully selecting the anti-symmetric anchor pose.

For the final step, we construct a hidden Markov model to exploit the information provided by predictions in temporally neighboring frames. We assume that each frame can be assigned a hidden state that corresponds with exactly one anchor pose. Under this assumption, the Viterbi algorithm is used to find the most likely sequence of hidden states. By doing so, the effect of inversion errors on the selection of the best matching anchor pose can be eliminated. In Figure 2, the resulting sequence is indicated in bold for the example.

To conclude, we examine how the framework enables the extraction of stroke frequency. Furthermore, its accuracy is compared to methods that are currently used in practice.

This work is structured as follows: First, we briefly touch upon the related work in Section II. Thereafter, Section III goes into detail about the different steps of our proposed framework and also discusses the subsequent extraction of the stroke frequency. Finally, Section IV introduces our dataset and metrics and analyzes the effect of the proposed framework on pose estimation performance and the accuracy of stroke frequency extraction.

II. Related Work

A comprehensive survey of the evolution of human pose estimation is given by [7] and [8]. Early work in the domain puts an emphasis on the use of and research into predefined

human body models. A popular example of which was the pictorial structure model (PSM) [9], [10]. The paradigm shift from these classical approaches to the use of deep neural networks was marked by DeepPose [11]. This acted as a baseline which prompted further research in this direction [12]–[15]. Solutions for the extended task of multi-person pose estimation [16]–[19] were also proposed, as well as implementations that used the temporal context of videos for multi-frame pose estimation [20]–[22].

The authors of [23] identify MS COCO [8] as one of the largest collections of multi-instance person keypoint annotations with annual competitions and a wide impact on the community. They propose a taxonomy of typical multi-instance pose estimation errors that is also viable for single-pose estimation which will be used throughout this work. Furthermore, they also assess image complexity and its influence on errors. This allows for meaningful parallels to be drawn between the MS COCO dataset and our swimming dataset, such as the degree of occlusion.

A series of publications that are focused on using computer vision for the analysis of swimmers, give an indication of how the topic was influenced by the evolution of pose estimation. In [24], the classical approach was used to recognize predefined key poses, similar to our anchor poses, throughout the stroke cycle. The number of key poses ranged from two to eight and they were used to extract the stroke rate. Similarly, a mixture of poselets allowed to automatically select key poses in [25] by retaining only those poses that show a regular occurrence in time. Whereas these previous approaches were more concerned with directly determining the stroke frequency, [4] puts the problem of accurate pose estimation for swimmers first. However, the scope of the problem remains reduced to only handle four specified key poses. An extension to this solution is provided by [26] that lifts the restriction to key poses, accepts swimming stroke information as input and modifies the architecture for continuous pose estimation in videos. A particularly relevant approach is validated in [27]. It also devises a measure to compare the distance between two poses, based on which stroke cycles are detected and analyzed. Furthermore the concept of a set of anchor poses is also introduced, referred to as a reference pose clip, to assess cycle stability. The main concern of the latest research in this direction was the rectification of pose estimates by eliminating outliers in prediction and most importantly inversion errors, i.e. confusion between left and right joints. To this end, the velocity and acceleration of each joint are determined over a sequence of frames in [5]. The Viterbi algorithm [28] is then used to find the most likely pose sequences that do not introduce high velocities and accelerations due to inversion errors. The same idea provides the foundation of [6] where confusion between joints is more formally modeled using a kinematic partner graph. The pose rectification is then done using joint trajectory regression.

Finally, there are a number of papers that each offer a different approach to the matching and aligning of poses in general [29], [30] and also previous approaches of using human pose estimation in combination with Viterbi for action recognition [31].

III. Methodology

A. Human Pose

In this work, a pose $p \in \mathcal{P}$ is defined as a collection of 13 keypoints: a head and a left and right instance of shoulders, elbows, wrists, hips, knees and ankles. Each keypoint is represented by its location and a visibility flag v, which indicates whether the keypoint is occluded or not. We use notation $\hat{p}$ to denote a predicted pose.

$$p = \{k_j\}_{j \in \text{joints}} \quad \text{with} \quad k_j = (x_j, y_j, v_j) \tag{1}$$

Furthermore, the torso diameter d_{torso} of a pose is defined as the distance between the left shoulder and right hip.

B. Baseline

Our research uses Mask-RCNN [3], extended for human pose estimation and finetuned on our training dataset, as the baseline model. It is well-suited for three reasons:

- The authors note that minimal domain knowledge for human pose is exploited by the system because they expect domain knowledge to be added complementary to their approach.
- The model's performance is competitive and outperforms the winners of the MS COCO 2016 keypoint challenge.
- It has received high interest from the research community and is publicly available. Consequently, it has been used as a starting point for many domain-specific approaches.

Nevertheless, even after finetuning, a state-of-the-art model underperforms on images of swimmers. A high number of inversion errors [23], i.e. confusion between left and right instances of the same keypoint, are reported in [5] and [6]. Hence, rectifying these errors in subsequent steps by using domain-specific knowledge could yield a significant increase in performance.

C. Pose Matching

The pose matching step of our framework reduces the problem of human pose estimation for a given input frame to identifying the most similar pose in a provided set of anchor poses and subsequently transforming the selected pose onto the original frame. This is illustrated by Figure 3. Consequently, this discussion contains three parts. Firstly, the concept of an anchor set is discussed in more detail. Secondly, an algorithm to assess similarity between two poses is devised. To conclude, we explain how the final estimated pose is obtained from the selected anchor pose.

Anchor set: An anchor set, $A = \{a_i\}$, contains manually annotated poses that originated from a sequence of consecutive frames in a video. These frames are chosen to span exactly one stroke cycle of the swimmer in the video. The anchor set defines the search space for pose estimation by our framework. Hence, the anchor set should correspond to a

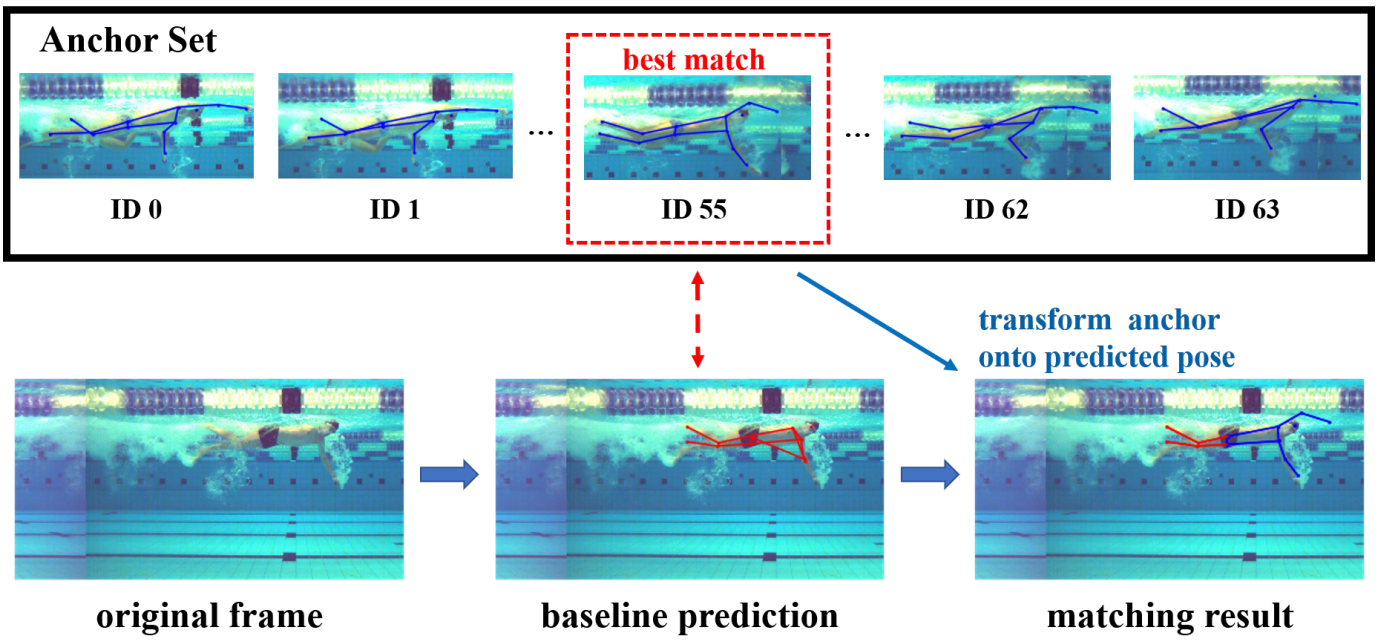


Fig. 3: Illustration of the use of anchor poses for pose estimation. The baseline prediction is used to find the most similar pose in the set of anchor poses. This pose is then transformed onto the orginal frame to provide a prediction for the upper body keypoints.

swimmer with similar body type and swimming technique as the swimmer that is to be analyzed by the framework. A brief discussion on how the choice of the anchor set impacts the similarity score can be found in [5], where a similar approach is used.

Pose Similarity: Before we commence the discussion on an algorithm to quantify similarity, it should be noted that the Mask-RCNN model associates with each keypoint detection a prediction score. To improve matching accuracy, only keypoints that are detected with a positive score according to the model are used. This is implicitly assumed in the remainder of this section.

Similarity between a pose and a target pose is in literature often determined with an approach based on Procrustes analysis [32]. This involves applying a transformation consisting of translation, scaling and rotation to the first pose such that the pairwise squared distance between corresponding joints from the transformed pose and the target pose is minimized. The distance between two poses of swimmers is evaluated like this in [27]. However, an additional constraint is imposed that partially detected poses should be discarded. Their design choice can be justified by noting that the transformation has too many degrees of freedom when only few keypoints are detected. Hence, it would then be possible to obtain high similarities for a partially detected pose to two very different target poses.

Our approach is similar but restrains the freedom of the transformation to accomodate partially detected poses: First, the pose is translated such that the translated pose's visible part of the hip is aligned with the hip of the target pose. The hip as reference for alignment is a suitable choice, since it is used as a stable point of reference throughout the stroke cycle in classic swimming stroke analysis [1]. Subsequently, the optimal rotation is obtained using the Kabsch algorithm [33] and applied to the translated pose to minimize the root-mean-square-deviation of keypoints from our transformed pose to the target pose.

The similarity score between a pose p and the target pose p_{target} with threshold t and p_{tf} the transformed pose of p, is determined by

$$sim\left(p, p_{\text{target}}\right) = \sum_{j \in joints} \frac{\max(0, (d_{\max} - d\left(p_{\text{tf},j}; p_{\text{target},j}\right))) \cdot w_j}{d_{\max}} \tag{2}$$

with

$$d_{\max} = t \cdot d_{\text{torso}} \tag{3}$$

An advantage of the Kabsch algorithm is that relative weights w_j can be assigned to the joints. In order to determine the relative importance of different joints in selecting the most similar anchor pose, we use an approach that resembles TF-IDF in information retrieval [34]: We cluster the ground truth poses in our dataset. Joints whose keypoint locations have a high standard deviation across all clusters, but a low per-cluster standard deviation can be identified as important joints for determining the anchor pose.

More specifically, we introduce buckets to represent the clusters. For this analysis an arbitrary subset of equidistant

anchor poses, $A' \subset A$, is temporarily selected from the complete anchor set. Each anchor pose, $a'_i \in A'$, is associated with one bucket $B_{a'_i}$. Every image in the dataset is assigned to the bucket of the most similar anchor pose based on its manually labeled, ground truth pose p_{gt}. Uniform weights were used for this in Equation (4).

$$p_{gt} \in B_{a'_i} \iff \arg\min_{a'_i \in A'}(sim(p_{\text{gt}}, a'_i)) \tag{4}$$

Subsequently, all poses in a bucket are aligned by the hip and for each joint the standard deviation of the location per bucket as well as across all buckets is calculated. Finally, the ratio between both standard deviation values indicates the relative weight of a joint in determining the similarity to an anchor pose. We only consider relative weights for the head, shoulders, elbows and wrists. This is because during front crawl the frequency of leg kicks can vary regardless of the stroke frequency of the arms [35] and because the hips have already been used to align the poses. Hence the hips, knees and ankles carry no weight in determining the most similar anchor pose.

Pose Estimation: The result of pose estimation after matching is obtained by transforming the selected anchor pose onto the original predicted pose. The same transformation that forms the basis of Equation (2) is used for this. Only the keypoints for ankles and knees from the baseline prediction are preserved as shown in Figure 3.

D. Most Likely Sequence

To further improve the performance we introduce the notion of time and construct a hidden Markov model. This enables finding the most likely sequence of anchor poses as hidden states by using the Viterbi algorithm.

We consider T consecutive frames $f_0, f_1, ..., f_{T-1} = f_{0:T-1}$ that will be the subject of analysis by the framework. An anchor set $A = \{a_i\}$ is assumed to be chosen such that each frame f_t has a hidden state S_t that corresponds to a unique anchor pose, i.e. $\forall t, \exists! a_i \in A : S_t = a_i$. Furthermore, the baseline model provides for each frame f_t a pose prediction $\hat{p_t}$. Additionally, the Viterbi algorithm requires to define prior probabilities $\pi(s)$ for the hidden states, emission probabilities for each hidden state and transition probabilities between hidden states.

Firstly, uniform prior probabilities are chosen for each state because no assumptions are made about the swimmers pose in the first frame f_0. This is shown in Equation (5).

$$\pi(s) = \frac{1}{|A|} \ , \ \forall s \in A \tag{5}$$

Secondly, the emission probability should represent how well an observed pose $\hat{p_t}$ explains a hidden state $S_t = a_i$. Hence, the similarity between both poses is used, as illustrated by Equation (6). This was defined in a similar fashion in [31].

$$Pr(\hat{p_t}|S_t = a_i) = sim(\hat{p_t}, a_i) \tag{6}$$

The final crucial component is the transition probability, under the Markov assumption, from state S_t to state S_{t+1}. Stroke frequencies ranging from 24 to 96 strokes per minute are discussed in [1]. To accommodate this wide range of stroke frequencies, transition probabilities are chosen such that transitions to anchor poses with up to distance 5 are possible as stated by Equation (7). A possibility would be to opt for equal probabilities for each of these transitions as in [31]. However, the assumption that the anchor set is chosen to be representative for the footage that is to be analyzed, suggests a similar stroke rate. Based on this, the transition probabilities presented in Table I were chosen.

$$Pr(S_{t+1} = a_j | S_t = a_i) > 0 \iff (j - i) \bmod |A| \leq 5 \tag{7}$$

TABLE I: Transmission Probabilities

Distance k	0	1	2	3	4	5	k $\geq$ 6
$Pr(S_{t+1} = a_{i+k} \| S_t = a_i)$	0.25	0.30	0.20	0.15	0.05	0.05	0.00

The Viterbi algorithm now allows to identify the most likely sequence of hidden states $s_{0:T-1} = s_0, s_1, .., s_{T-1}$ that is best explained by the sequence of observed pose predictions $\hat{p}_{0:T-1} = \hat{p}_0, \hat{p}_1, ..., \hat{p}_{T-1}$. Equation (8) shows the recursive property that is used by the Viterbi algorithm, with α a normalizing constant.

$$\begin{aligned} m_{0:T-1} &= \max_{s_{0:T-1}} Pr(s_{0:T-1}|\hat{p}_{0:T-1}) \\ &= \max_{s_{T-1}} \left(\alpha Pr(\hat{p}_{T-1}|s_{T-1}) \max_{s_{T-2}} ((Pr(s_{T-1}|s_{T-2}) m_{0:T-2}) \right) \end{aligned} \tag{8}$$

E. Stroke frequency extraction

Our approach so far is based on complementing a state-of-the-art human pose estimation algorithm with domain-specific knowledge. It is further improved by introducing a hidden Markov model to account for the correlation between consecutive frames. We now consider our approach for the extraction of swimming performance metrics. More specifically, stroke frequency is considered.

The Viterbi algorithm has selected for each frame f_t the most likely anchor pose a_i. Every new iteration through the anchor set marks the start of a new stroke cycle. Hence, the problem of stroke frequency extraction is reduced to recognizing crashes in the saw tooth pattern that is produced by plotting the anchor pose ID i in function of time t. A simple method is used as described in [36]. Given anchor set $A = \{a_i\}_{i=0}^{n-1}$ and considering $s_t \in A$ as the selected anchor for frame f_t, then the start of a new stroke cycle is predicted at time t if

$$s_{t-1} = a_{i_1} \wedge s_t = a_{i_2} \wedge i_1 - i_2 > \frac{n}{2} \tag{9}$$

IV. Results

A. Dataset and Metrics

Dataset: Our dataset consists of 600 images that were sparsely selected from videos of 15 different male and female swimmers. The footage was provided by the Flemish Swimming Federation and was captured at a resolution of 1408x672 at 50 frames per second. Only underwater cameras were used that do not record any body parts above the water surface. Each image contains a single swimmer and is manually annotated with a bounding box and 13 keypoints conform to our definition of a pose in Section III.

The dataset is divided in a train and test dataset of respectively 400 and 200 images. Half of the train dataset is used to finetune the baseline model and the other half is used to validate design choices and determine parameters specific to our approach. We also select 64 consecutive frames of a single male swimmer from the train dataset to use as the anchor set in this section. Even though the choice of anchor set has a significant influence on the performance, it was experimentally shown not to alter the reached conclusions.

Metric: The Percentage of Correct Keypoints (PCK) metric is used as in [37]. Contrary to the popular Object Keypoint Similarity (OKS), it allows to easily assess the detection of individual keypoints. PCK considers a keypoint detection $\hat{k}$ as correctly localized when it is within a maximum distance from its ground truth location. This maximum distance is chosen as a fraction t of the length of the torso diameter. This provides a intuitive method of measuring the performance for individual sets of keypoints across a range of thresholds.

Correctness of a predicted keypoint $\hat{k}$ with respect to the ground truth keypoint k at threshold t with d_{torso} the torso diameter is defined by

$$CK_t(k,\hat{k}) = \begin{cases} 1 & d(k,\hat{k}) < t \cdot d_{\text{torso}} \\ 0 & else \end{cases} \tag{10}$$

To measure PCK scores of groups of semantically similar joints, e.g. left and right wrist, we define $J \subset joints$. The PCK score for each image i in the dataset is

$$PCK_t(J) = \frac{\sum_i \sum_{j \in J} CK_t(p_j, \hat{p_j})}{\sum_i |J|} \tag{11}$$

An inversion error w.r.t a set J and threshold t can now be formally defined as

$$j, j' \in J : j \neq j' \wedge CK_t(p_j, \hat{p_j}) = 0 \wedge CK_t(p_j, \hat{p_{j'}}) = 1 \tag{12}$$

Degree of Occlusion: A first quantification of the complexity of our dataset is given by Figure 4. It shows the distribution of the number of visible keypoints in our dataset, which has been correlated by [23] to the performance of human pose estimation on the MS COCO dataset. It was shown that frequent occlusion reduced performance in terms of precision and recall. This high number of occlusions is the result of complex swimming movement, joints frequently located above the water surface and high levels of noise in the water during front crawl.

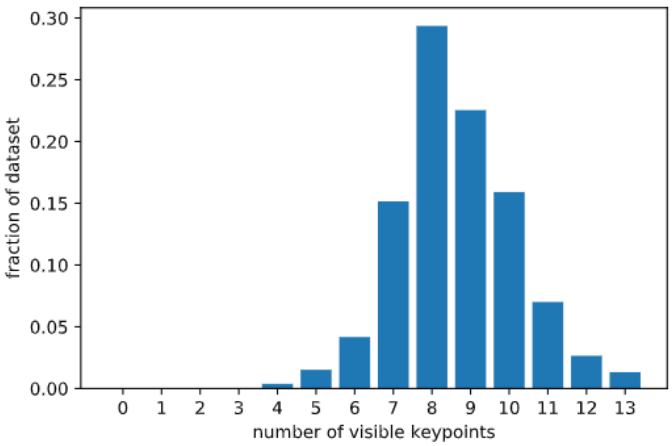


Fig. 4: The distribution of the number of visible keypoints per image in the front crawl dataset.

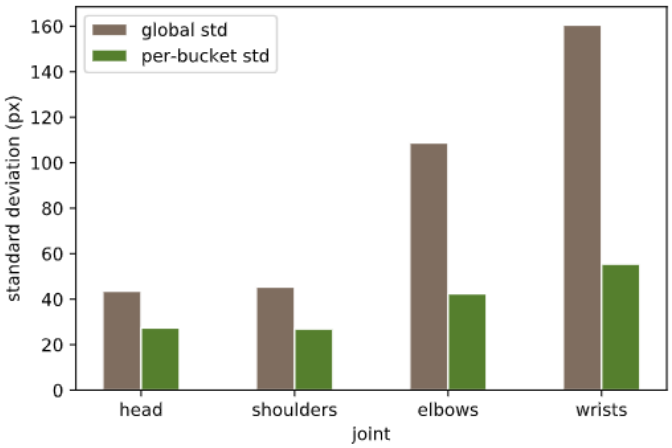


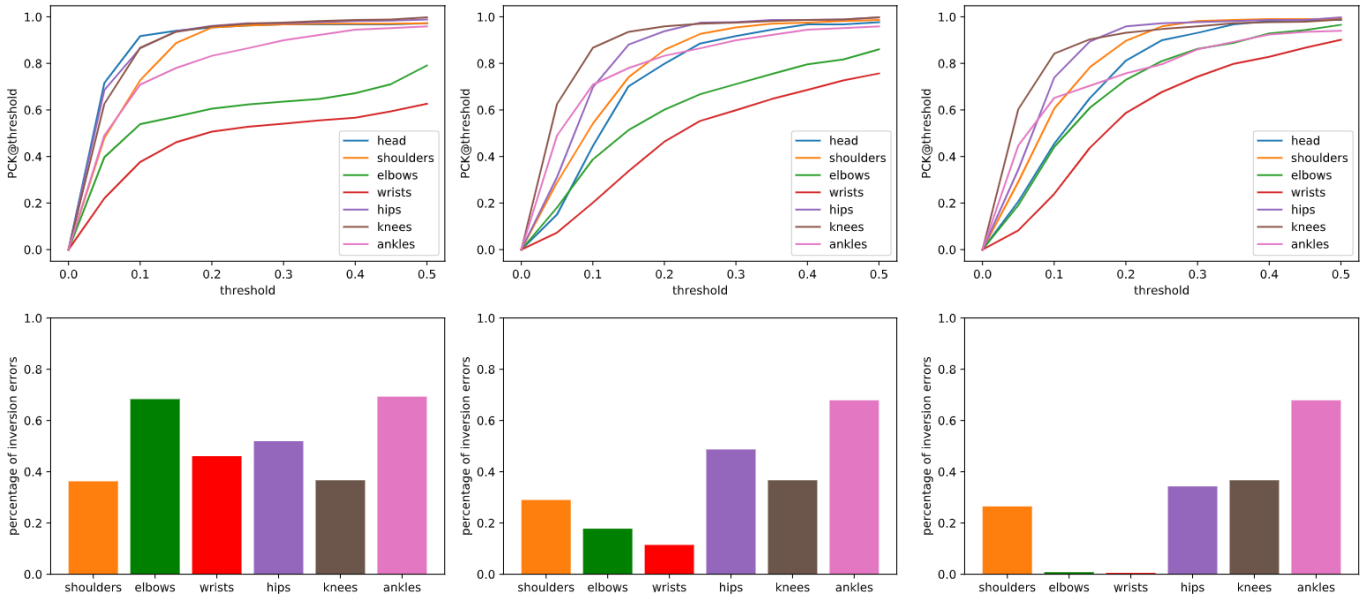
Fig. 5: Global and per-bucket standard deviation of keypoint location in pixels.

B. Joint Weights

Before pose matching can be used, the relative weights w_j for the head, shoulders, elbows and wrists in assessing pose similarity must be determined, as described in Section III. Therefore, we select a subset of 13 equidistant anchor poses, $A' \subset A$, from our anchor set and compute the average per-bucket standard deviation and the global standard deviation on our train dataset. Figure 5 illustrates the result. The weights corresponding to the ratio of both standard deviations are given by Table II. Elbows and wrists are clearly the most important joints when comparing two poses.

TABLE II: Relative weighting of joint groups for pose matching.

Joint Group	Head	Shoulders	Elbows	wrists
w_j	1.59	1.68	2.56	2.89



(a) baseline (b) matching (c) most likely sequence

Fig. 6: The influence of the different steps of the framework on the PCK score and percentage of inversion errors, at threshold=0.2, for the different joint groups.

C. Framework Performance

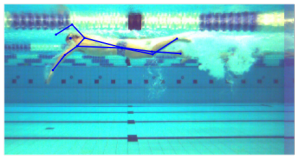
1) Baseline: Figure 6(a) shows the result of the baseline model on the test dataset. The most obvious observation is that elbows and wrists score significantly lower than the other joints. This can be explained by pointing out that elbows and wrists spend a significant portion of the stroke cycle above the water surface and are therefore not visible with the used camera setup. The occluded keypoints do however still contribute to the PCK metric, resulting in a lower score. The percentage of errors that can be contributed to inversion is also illustrated in Figure 6(a). These high values, especially for elbows and wrists, were also reported in [5] and [6] where a camera setup was used that also allowed for detecting joints above the water surface. These observations support our hypothesis that devising a framework, based on domain-specific knowledge, could yield a significant increase in performance by rectifying these inversion errors for elbows and wrists.

2) Pose Matching: The effect of the matching step on the PCK score is illustrated in Figure 6(b). Two observations can be made:

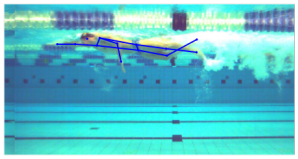
- In general, the score has decreased for all joint groups at low thresholds. This is due to the fact that the anchor poses do not perfectly match the analyzed swimmer's body type and swimming style. Hence, it cannot be expected to localize the keypoints with a very high accuracy. As the threshold increases, the scores converge to their values without matching for all joint groups except for elbows and wrists.
- For elbows and wrists, the score has increased for thresholds exceeding 0.2. This increase is driven by a reduction of the number of inversion errors.

The main obstacle for pose matching becomes apparent when observing the relative weights in Table II and the PCK scores for the baseline in Figure 6(a): Elbows and wrists are the most important joints in determining the best matching anchor pose, but also have the lowest reported PCK scores and the highest frequency of inversion errors. To visualize the impact of these seemingly contradictory statements, two of the buckets that were used to determine the weights are analyzed further. These buckets are chosen specifically to represent two distinct cases: one whose anchor pose has one arm completely occluded and one where both arms are visisble. For each image in either bucket the similarity score to each of the 64 anchor poses is calculated twice, once based on the manually annotated keypoints in the image and once based on the predicted keypoints by the baseline model. Figure 7 shows the resulting average similarity scores for these two buckets as well as the two corresponding anchor poses. In Figure 7(a), the swimmer's right arm is completely above the water surface which increases the chances of inversion errors. Figure 7(c) confirms this. There is only one peak for the similarity score based on the ground truth keypoints, correctly centered around anchor pose with ID 25. However, when using the predicted keypoints there is also a second peak approximately half a stroke cycle apart. This peak corresponds to the anti-symmetric pose. The other distinct case does not display this behavior. The anchor pose displayed in Figure 7(b) has both arms visible and consequently only has a single peak in Figure 7(d). Hence, the remaining inversion errors after pose matching can be

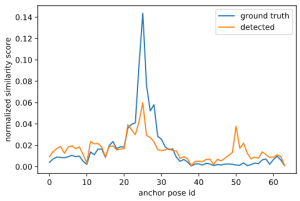
explained by wrongfully selecting an anti-symmetric anchor pose.



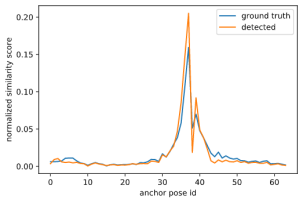
(a) anchor pose 25



(b) anchor pose 37



(c) average similarity score $B_{a_{25}}$



(d) average similarity score $B_{a_{37}}$

Fig. 7: a) Anchor pose with ID 25 and b) with ID 37. c) and d) average similarity score of each pose in the bucket to all anchor poses in the anchor set.

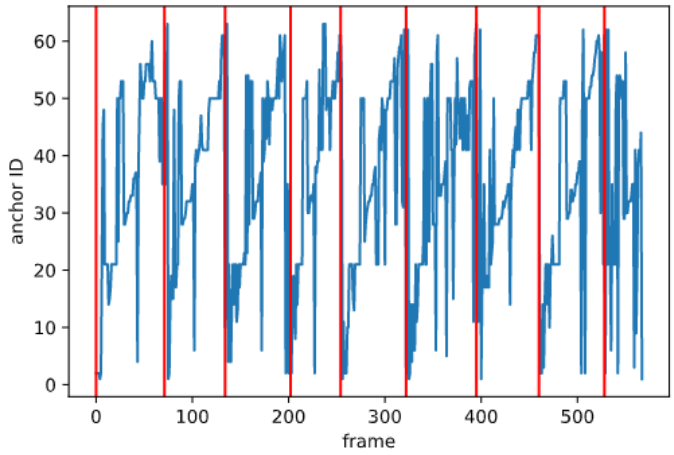
3) Most Likely Sequence: Figure 6(c) reveals how the final step of our framework affected the performance. The same trade-off as in Figure 6(b) is observed between a high PCK score at low thresholds and approximate localization of all keypoints, including those that are frequently occluded. It also shows yet another significant increase in score for the elbows and wrists at high thresholds and an elimination of nearly all of their inversion errors. The Viterbi algorithm has successfully removed the confusion between anti-symmetric poses.

D. Stroke Frequency Extraction

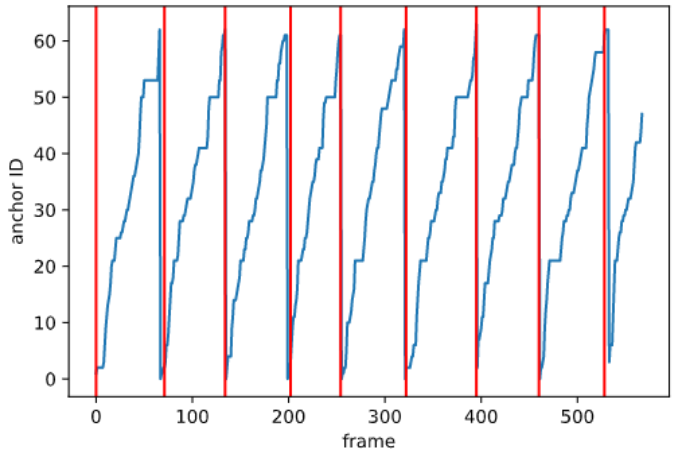
The analysis of how our framework affects the performance of pose estimation indicated that detection accuracy, i.e. at low thresholds, is sacrificed for consistency of detection to make an informed prediction even when a keypoint is occluded.

However, the true advantage of the approach is illustrated by Figure 8. It still uses our 64 anchor poses and considers a number of consecutive frames from a single video from the dataset. For each frame the ID of the matched anchor is shown after the matching step, Figure 8(a), as well as after the application of Viterbi, Figure 8(b). Additionally, frames containing the start of a new stroke cycle were manually annotated, and are indicated on the figure as well. It visually supports the hypothesis that the framework provides a method of reliably determining what part of the stroke cycle the swimmer is in. Furthermore, the importance of searching the most likely sequence is illustrated.

It is now possible to test the accuracy of our framework in determining the duration of stroke cycles and compare it to current manual methods. In practice, two methods of determining the stroke frequency are used. The most common method requires that a human observer uses a stopwatch to indicate the start time of every third stroke cycle. Each time



(a) sequence of best matches



(b) most likely sequence

Fig. 8: The sequence of anchor poses as predicted by the framework is illustrated in blue (a) after the matching step and (b) after selecting the most likely sequence. The vertical red lines indicate manual annotations of the start of a new stroke cycle.

interval is then divided by three to obtain the average stroke cycle duration during that interval. The authors of [38] studied the human response time to an anticipated visual trigger in swimming competitions with stopwatches and found an average error of 0.106s. This corresponds to an average error per stroke cycle of $\frac{0.106s}{3} \approx 0.035s$. Figure 9 illustrates how the average error in determining the stroke cycle duration when using manual timings compares to our framework's detection on the test dataset. Instead of considering only three successive stroke cycles, interval sizes ranging from 1 to 6 are used. Our framework outperforms manual timings collected by a human observer in real-time, across all interval sizes.

The second method involves a frame-by-frame analysis. Each frame that contains the start of a new stroke cycle is annotated. This method is not as frequently used due to its time-intensiveness. However, it is much more accurate as ex-

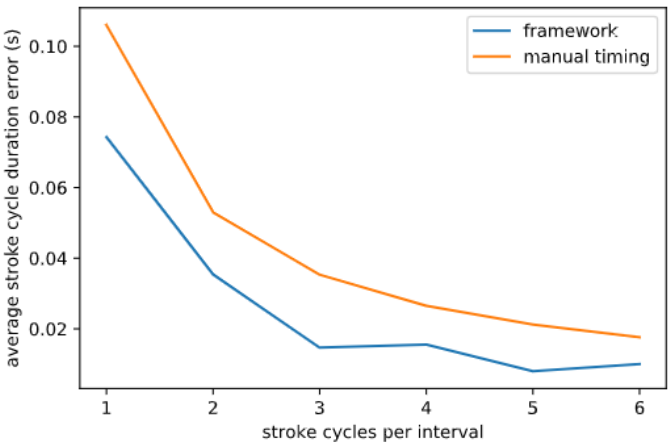


Fig. 9: The average error between the detected and ground truth duration of a stroke cycle when averaged over a different number of consecutive stroke cycles for both the framework and manual timings with a stopwatch.

perts claim that the start of each new cycle can unambiguously be identified. Hence, at a frame rate of 50 frames per second, the average error of the cycle duration is less than 0.02s. This level of accuracy is not achieved by our framework without the trade-off of having to consider multiple consecutive cycles.

V. Conclusion

In this work we examined the performance of a state-of-the-art human pose estimation model, Mask-RCNN, and identified its shortcomings on our dataset of swimmers to propose a framework that enables the extraction of swimming performance metrics by improving the human pose estimation result. We confirmed that the baseline model, consisting of Mask-RCNN for human pose estimation, underperforms on our dataset of swimmers, in particular for joints that are frequently occluded. A major component of this result can be attributed to inversion errors, i.e. confusion between left and right instances of joints. To mitigate this, our framework reduced the problem of human pose estimation to finding for each image the most similar pose in a set of anchor poses. The best matching anchor pose is then transformed onto the original prediction. This involved devising an approach to pose matching that accepts partially detected poses. However, due to the inversion errors, the matching algorithm still resulted in confusion between anti-symmetric poses. The Viterbi algorithm proved to be an effective option to deal with this confusion and correct the inversion errors. Consequently, the PCK metric showed that these steps resulted in a loss of the exact localization of detected keypoints, but increased the number of correct detections at larger distance thresholds. Furthermore, the main benefit of framework is that it enables the detection of stroke cycles. As a result, it was shown to enable stroke frequency extraction that is more accurate than extraction by an observer through real-time manual timings.

Further improvements on this work pertain to two specific directions. The first involves an extensive study on the impact of the choice of the anchor set on the performance of the framework. Additionally, multiple anchor sets corresponding to different swimming techniques could be provided. The Viterbi algorithm could then identify a most likely sequence for each anchor set and subsequently select the anchor set that best explained the observations. Hence, this could be used to automatically detect properties of the swimming technique such as coordination of index [39], which quantifies arm coordination during front crawl. A second direction for further research could extend the proposed framework for other swimming styles, e.g. breaststroke, and other swimming phases, e.g. the turn or dive. The latter could be achieved by using a hierarchical hidden Markov model [40] to recognize transitions between different phases if an anchor set is provided for each phase.

Acknowledgment

We would like to express our gratitude to the Flemish Swimming Federation (VZF) for providing the footage of the swimmers as well as guiding our research with their expertise.

References

[1] T. M. Barbosa, D. a. Marinho, M. J. Costa, and a. J. Silva., "Biomechanics of Competitive Swimming Strokes," *Biomechanics in Applications*, 2011.

[2] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, "RMPE: Regional Multi-person Pose Estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[4] D. Zecha, C. Eggert, and R. Lienhart, "Pose estimation for deriving kinematic parameters of competitive swimmers," in *IS and T International Symposium on Electronic Imaging Science and Technology*, 2017.

[5] D. Zecha, M. Einfalt, C. Eggert, and R. Lienhart, "Kinematic pose rectification for performance analysis and retrieval in sports," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[6] D. Zecha, M. Einfalt, and R. Lienhart, "Refining joint locations for human pose tracking in sports videos," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[7] H. B. Zhang, Q. Lei, B. N. Zhong, J. X. Du, and J. L. Peng, "A Survey on Human Pose Estimation," *Intelligent Automation and Soft Computing*, 2016.

[8] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.

[9] S. Zuffi, O. Freifeld, and M. J. Black, "From pictorial structures to deformable structures," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[10] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference, BMVC 2010 - Proceedings*, 2010.

[11] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[12] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using Convolutional Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 648–656, 2015.

[13] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 4724–4732, 2016.

[14] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912 LNCS, pp. 483–499, 2016.

[16] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Openpose," *arXiv*, 2018.

[17] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[18] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[19] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[20] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[21] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing Human Video Pose Estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[22] O. Sumer, T. Dencker, and B. Ommer, "Self-Supervised Learning of Pose Embeddings from Spatiotemporal Relations in Videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[23] M. R. Ronchi and P. Perona, "Benchmarking and Error Diagnosis in Multi-instance Pose Estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[24] D. Zecha, T. Greif, and R. Lienhart, "Swimmer detection and pose estimation for continuous stroke-rate determination," in *Multimedia on Mobile Devices 2012; and Multimedia Content Access: Algorithms and Systems VI*, 2012.

[25] D. Zecha and R. Lienhart, "Key-pose prediction in cyclic human motion," in *Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, 2015.

[26] M. Einfalt, D. Zecha, and R. Lienhart, "Activity-Conditioned Continuous Human Pose Estimation for Performance Analysis of Athletes Using the Example of Swimming," in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, 2018.

[27] R. Lienhart, M. Einfalt, and D. Zecha, "Mining automatically estimated poses from video recordings of top athletes," *International Journal of Computer Science in Sport*, 2018.

[28] G. D. Forney, "The Viterbi Algorithm," *Proceedings of the IEEE*, 1973.

[29] C. H. Chen and D. Ramanan, "3D human pose estimation = 2D pose estimation + matching," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[30] S. H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S. M. Hu, "Pose2Seg: Detection free human instance segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[31] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[32] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, 1975.

[33] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, 1976.

[34] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing and Management*, 2003.

[35] V. Gourgoulis, A. Boli, N. Aggeloussis, A. Toubekis, P. Antoniou, P. Kasimatis, N. Vezos, M. Michalopoulou, A. Kambas, and G. Mavromatis, "The effect of leg kick on sprint front crawl swimming," *Journal of Sports Sciences*, 2014.

[36] I. D. M. V. D. Brink, "A real-time detection algorithm for sawtooth crashes in nuclear fusion plasmas," no. 0502234, 2009.

[37] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[38] D. A. Faux and J. Godolphin, "Manual timing in physics experiments: Error and uncertainty," *American Journal of Physics*, 2019.

[39] D. Chollet, S. Chalies, and J. C. Chatard, "A new index of coordination for the crawl: Description and usefulness," *International Journal of Sports Medicine*, 2000.

[40] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: Analysis and applications," *Machine Learning*, 1998.