SWIM STROKE ANALYTIC: FRONT CRAWL PULLING POSE CLASSIFICATION

Hossein Fani[*], *Amin Mirlohi*[*], *Hawre Hosseini*[†], *Rainer Herpers*[‡]

[*]Faculty of Computer Science, University of New Brunswick, NB, Canada
[‡]Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany
[†]Department of Electrical & Computer Engineering, Ryerson University, ON, Canada

ABSTRACT

In this work, we automatically distinguish the efficient *high elbow* pose from *dropping* one in pulling phase of front crawl stroke in front view amateurly recorded videos. This task is challenging due to the aquatic environment and missing depth information. We predict the pull's efficiency through multi-class svm and random forest classifiers given arms key positions and angles as the feature set. We evaluate our approach over a labeled dataset of video frames taken from 25 members of masters' swim club at Ryerson University with different levels of expertise and physiological characteristics. Our results show the effectiveness of our approach with random forest classifier, yielding 67% accuracy.

Index Terms— Swim Stroke Analysis, Pose Estimation

1. INTRODUCTION

Majority of people, 65% to be precise, are visual learners who prefer to see what they are learning [1]. In sports, in particular, athletes attain their maximum potential not only by listening to the coaches but also by reviewing their own workouts in practice. In this respect and in order to improve understanding, training, and performance, computer vision algorithms find huge potential in automating video analysis of athletes and providing immediate feedback on poses' efficiency to the coaches as well as athletes themselves. For instance, action recognition provides swim coaches and swimmers with automatic performance assessment through statistics such as stroke rate based on visual perception of periodic body kinematic. In this work, we aim at analyzing swim stroke of a swimmer from her own swimming amateurly recorded by an off-the-shelf camera, for training purposes. This task is challenging due to the aquatic environment (e.g. water splashes and reflections) and missing information such as depth dimension in the low quality recorded videos.

We focus our study on front crawl swimming style and predict the efficiency of the arm stroke under water, called pulling, from the front view as seen in Figure 1. Pulling's objective is to generate moving drive forward through the water. Efficient pulling is initiated by maintaining the elbow close to the surface, so-called *high elbow* as shown in Figure 1 (left), followed by dropping the forearm below the elbow and a sweeping motion of the upper arm. This is in contrast to inefficient *dropping elbow*, as shown in Figure 1 (right), where the whole arm is down deep into the water. In this work, our goal is to predict the quality of pulling pose as of being high elbow versus dropping elbow given an underwater recording of front crawl strokes.

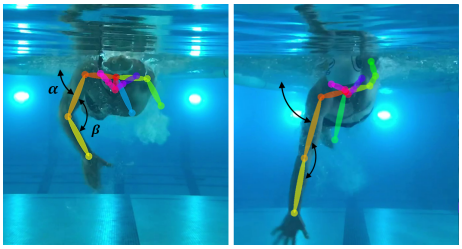

Fig. 1. High elbow (left) vs. dropping elbow (right).

Action recognition in sports, swimming in particular, has already been investigated in the literature [2, 3, 4, 5, 6]. Sha et al. [4] and Victor et al. [5] estimate the front crawl stroke rate. While the approaches are robust for the task of stroke rate estimation in natural videos, i.e., broadcast videos taken at races, they are not extensible to obtain performance indicators about the efficiency of body movements in swim strokes such as pulling pose. Zecha et al. [6], however, constrained their work to a lab environment, i.e., swimming glass channel, where the swimmer's whole body, under and above water, is visible from side view at all times such that almost all kinematic parameters are able to be quantitatively extracted. Our work is inspired by Zecha et al. but distinguishes itself in that we analyze the swimmer kinematic in a real aquatic setting, i.e., swimming pool, from her own recorded video with no prior knowledge about the environment. Needless to say that the swim glass channel is not accessible to all swimmers and the side view recording in a real swimming pool needs the camera to constantly follow the swimmer alongside the pool which not only adds complexity (active vision), but also requires sophisticated apparatus. In our work, however, we require front view recording from a fixed position which can easily be set up.

We detect the swimmer body parts using the state-of-the-art pose detection method OpenPose [7]. The choice of OpenPose is motivated for its functionality on image or video taken by webcam and IP cameras. This provides huge benefit in

comparison to the skeletal tracking capability of Microsoft Kinect or the likes which depend on depth information, i.e. three dimensional camera. We then measure the angles between upper arm and forearm, and upper arm and the water surface. Given the arm joints position and the respective angles, we train a classifier on our manually labeled dataset of swimmers with different levels of expertise and physiological properties[1]. Our results show that we are able to successfully infer that a swimmer's pose of elbow is either high or dropping in pulling phase of her front crawl stroke. The main contribution of this paper is not a new pose estimation algorithm, but rather an effective application of existing pose estimation method, i.e., OpenPose, in swim stroke analysis. Specifically,

1. We propose a novel application of computer vision in swimming performance feedback taking advantage of existing infrastructure;
2. Our framework performs on amateurly recorded video of swimming in a real aquatic environment of swimming pools, allowing performance assessments accessible to almost all end users; and,
3. We build a manually labeled dataset of video frames recorded from swimmers with different levels of expertise, i.e., beginner, novice, and professional, from a wide variety of physiological properties, i.e., gender, age and size, in order to train a model generalized enough to all types of swimmers.

The rest of the paper is organized as follows: we first present the related work in Section 2, then we continue with the problem definition and our proposed approach details in Section 3. Finally, our testbed is explained in Section 4, followed by a discussion in Section 5.

2. RELATED WORK

Closely related to our work but on other sports are the works by Vicente et al. [2] and Bidhendi et al. [3]. Vicente et al. [2] train a model on the athlete key poses, extracted from frames, to identify kick and punch in Taekwondo. Bidhendi et al. [3] do pose detection on overhead images of boxers to classify their punches to uppercut, straight, and hook. Like our proposed approach, these works rely on a pose detection API but with depth information in the frames as features. Vicente et al. extract pose features from frames by using RGB-D sensor (Microsoft Kinect) and Bidhendi et al. employ low resolution depth images. However, in our work we use OpenPose whose pose estimation does not require depth information.

[1]We publicly release the dataset as well as all of other artifacts such as codes, features, and the results at `goo.gl/mxz6tG`. We excludes the raw recording videos of swimmers due to the privacy.

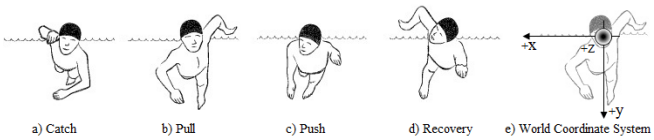

Fig. 2. Front crawl stroke phases [8] and respective world coordinate system.

3. SWIM STROKE ANALYSIS

3.1. Problem Definition

Stroke cycle of the front crawl swimming style consists of four phases in order: i) catch, ii) pull, iii) push, which happen under water, and iv) recovery which happens above the water [8] as shown from the front view in Figure 2. We focus on the pulling phase in our work for three reasons. Firstly, pulling generates almost the entire moving drive forward and its efficient pose dramatically improves swimming performance in front crawl [8]. Secondly, pulling pose in front crawl is shared among two other swimming styles, i.e., butterfly and breaststroke. So, our work is easily extensible to more swimming styles. Thirdly, pulling analysis needs simple-to-record video frames in our framework. It only requires front view frames captured by a camera which is fixed in just 10 centimeters under water (aligned with swimmer's head on the z-axis) for up to 30 seconds. A typical swimmer is able to do so by a water resistant mobile phone such as iPhone 7 or Samsung S7 and there is no need for highly developed waterproof camera and sophisticated underwater recording skills.

Now, given a set of amateurly recorded video frames of a swimmer who swims towards the camera on the optical axis, our task is to predict her pulling pose as of being either high or dropping elbow. Simply, this implies pulling classification problem given the frames as our observation. We formally define our swim stroke analysis as follows:

Definition 3.1. (Pulling Classification) *Let the camera and world coordinate systems be totally aligned with no translation but their z-axis be in opposite directions as shown in Figure 2(e). Given $\mathbb{V} = \{f_{1:N}\}$ the recorded video $\mathbb{V}$ consists of N frames, pulling classifier $c : \mathbb{V} \rightarrow \{-1, 0, 1\}$ is a function that maps a frame of a video $f \in \mathbb{V}$ to either 1; high elbow, or 0; dropping elbow, or -1; if the no recognizable pulling pose has been identified.*

Our swim stroke analysis seeks to learn the pulling classifier c given manually labeled frames from a set of $\mathbb{V}$.

3.2. Approach

To learn the pulling classifier c, we perform a supervised learning method on a set of frames $\{f_{1:N}\}$ and their respective pulling labels $\{y_{1:N}\}$ where $y \in \{1$ (high elbow), 0 (dropping elbow), -1 (no recognizable pulling pose)$\}$. Each frame f_i is represented by a d-dimensional feature vector

$\Phi(f_i) \in R^d; 1 \leq i \leq$ N, where the function Φ extracts features by analyzing the swimmer pose. Our proposed approach can be described in two steps: i) pose extraction, and ii) model training, each of which is explained in the following.

Pose Extraction. To learn the pulling classifier c, we map each frame to a feature space in which the arms' poses and angles in the frame are the features. Formally, given a frame $f \in \mathbb{V}$, we map it to a d-dimensional feature vector by the function $\Phi : \mathbb{V} \rightarrow R^d$. The feature vector is a concatenation of two types of arm's skeletal information as follows:

- Joints: the 2-dimensional position of the right and left shoulders, elbows, and wrists
- Joint angles: the angle between upper arms and the water surface, namely α, and the angle between upper arms and forearms, namely β

We show these features for a sample frame in Figure 1. We use OpenPose [7] to detect the joints. OpenPose[2] is a multi-person two dimensional (2D) pose detection. It is able to detect human body parts, including hand and facial keypoints (in total 130 key points) on single frame (image) in real-time. The choice of OpenPose is motivated for its functionality on image and video captured by webcam or IP camera which do not have extra depth information. OpenPose takes an image and identifies the locations of anatomical body points for each person which happen to be in the image using a multistage feedforward convolutional network. At each stage, it predicts a set of points from the body, the face, and the hands in the image plus a confidence score. We feed in OpenPose with each frame to detect only the joints, particularly the shoulder, elbow, and the wrist joints.

We calculate α and β based on $tan\phi = \frac{m_1 - m_2}{1+m_1 m_2}; m_1 \geq m_2$ where ϕ is the angle between two lines of slope m_1 and m_2. As we assume that the recording settings follows the Definition 3.1, we only need to calculate the slope of line linking shoulder and elbow for α since the x-axis slope is zero. We compute β based on the slopes of upper arm and forearm.

Model Training. To extract the arm's joints, we rely on the already trained model of OpenPose to detect the swimmers' arm joints under water. We learn the pulling classifier function c through the classification methods, namely random forest [9] and svm [10] over feature vectors. Random forest is an ensemble method which averages the predictions of several decision trees as base classifiers yielding an overall better model over a single decision tree. More, random forest due to its base classifier is inherently able to learn multi-class classifications such as our pulling classifier c. The number of trees is set to 100 and the criterion to measure the quality of a node split is Gini impurity for the information gain in our random forest. As an alternative baseline, we also use svm with one-versus-rest (ovr) multi-class strategy and linear kernel.

[2] github.com/CMU-Perceptual-Computing-Lab/openpose

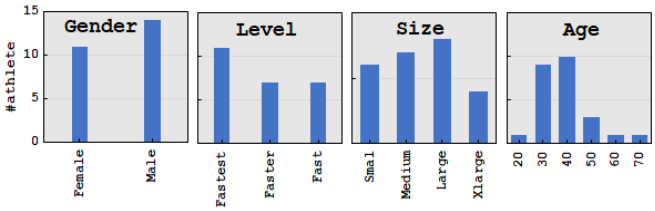

Fig. 3. Distribution of swimmers by level of expertise, gender, size, and age in our dataset.

Table 1. Distribution of different pulling poses in our dataset.

label	#	%
no pulling pose	1,274	48.39
dropping elbow	1,001	38.02
high elbow	358	13.60

4. EXPERIMENTATION

4.1. Setup

We used iPhone 7 and LG G6 both with 1,920×1,080×30fps to record the swimmers front crawl strokes under water. We fixed the camera 10 centimeters under water on the one end of the swimming pool wall and parallel to the water surface. In order to identify the body parts and poses, we extracted the video's frames and applied OpenPose library with its default settings, i.e., COCO model to identify 17 body parts with neural net resolution 656×368. We build and run OpenPose library on Intel Core i7-3770 quad-core processor with 16GB DDR3 of system memory and NVIDIA 1070 graphic card with 1,920 GPU cores and 8GB frame buffer. We used scikit-learn[3] to train and evaluate our baseline classifiers.

4.2. Data Acquisition

Our experiments include 25 swimmers of masters' swim club at Ryerson University[4]. The swimmers are from different levels of expertise. Fast (beginner), faster (novice), and fastest (professional). Also, they are sampled from a wide range of swimmers with different genders, ages, and body sizes as shown in Figure 3. As a result, the effect of physiological characteristics has been tried to be relieved. We record swimmers' front crawl pulling strokes under water from the front view as they approach, from one side of the pool, to the wall where our camera is installed (one lap swim of 25 yd = 22.86 m). We then extract frames and filter out those in which no arm joint has been detected by OpenPose. This way we filter out the starting frames where no trace of swimmer has been detected either by camera or OpenPose. This leads us from a dataset of 15,384 frames to 2,633 frames. All the frames were manually labeled by a swimming expert. Table 1 shows the distribution of different labels in our dataset. As shown, the distribution is skewed toward the 'no pulling pose' which could be due to the fact that either OpenPose falls short of

[3] scikit-learn.org

[4] www.rec.ryersonrams.ca

Table 2. The performance of our baselines.

model	feature set	precision	recall	f-measure	accuracy
random forest	joints+angles	**0.652**	**0.666**	**0.649**	**0.666**
	joints	0.643	0.656	0.64	0.656
	angles	0.546	0.557	0.547	0.557
svm	joints+angles	**0.482**	**0.498**	**0.442**	**0.498**
	joints	0.457	0.439	0.383	0.439
	angles	0.422	0.437	0.396	0.437

correctly detecting the swimmer's arm joints in a frame or the pose is in a phase other than pulling such as catch or push as shown in Figure 2, (a) and (c).

4.3. Evaluation and Results

We evaluate our classifiers, random forest and svm, on our dataset with three different feature subsets each of which includes only i) joints, ii) joint angles (α and β), and iii) both joints and angles. Due to label distribution imbalance, we did *stratified* 10-fold cross-validation and report the performance of each baseline by weighted average of all three classes for precision, recall, f1-measure, as well as accuracy in Table 2.

As evident in Table 2, all random forest baselines unanimously outperform the svm counterparts. We attribute this to the fact that random forest is an ensemble classifier whose prediction is based on the average over several independent decision trees. This is better than a single base classifier due to the lower variance. Among the random forest and svm baselines, we observe that joint features contribute more than angles to learning a better classifier in all metrics. For instance in random forest, while the angle features could not perform better than 56% in accuracy, joint features excels and reaches 66%. Nonetheless, joint plus the angle features leads to the best performance with minor 1% improvement in the random forest baseline.

5. DISCUSSION AND FUTURE DIRECTION

The goal of this study is to give feedback to the swimmer about efficiency of her pulling under water. We introduce the concept of swimmer *pulling map* which is able to give feedback on the swimmer's pulling efficiency over the course of different frames. For instance and for the swimmer in Figure 4, we show the respective pulling map in Figure 5. As seen in the frames in Figure 4, while the swimmer is doing dropping elbow on his left arm, he is doing high elbow on the right arm. Presumably, the swimmer is right handed and has some frailty in his left arm and needs to practice more on his left arm pulling. This is reflected in the pulling map in Figure 5 where the predicted pulling pose class of our best baseline, random forest with both joint and angle features, is shown. As seen, although there are misclassifications, the pulling map shows the predicted class of dropping elbow in left arm from frame# 100 to 106 and the predicted class of high elbow in right arm from frame# 120 to 125, respectively.

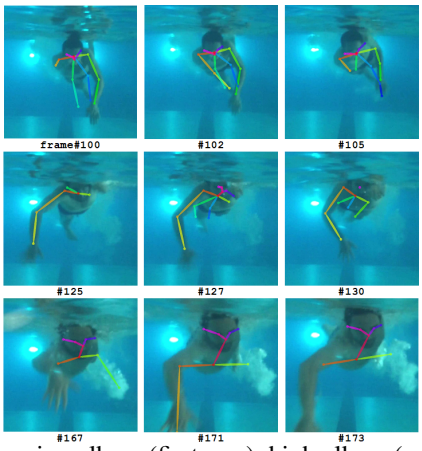

Fig. 4. Dropping elbow (first row), high elbow (second row), and no pulling pose (last row).

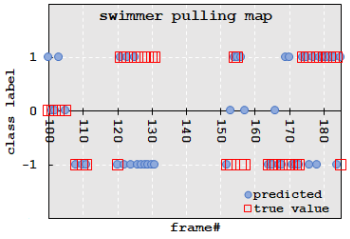

Fig. 5. Pulling map for a swimmer shown in Figure 4.

Two possible future directions to this work are: (1) the used pose detection library, OpenPose, is not perfect and shows false detection as in Figure 4 (last row). This is due to the fact that we use the pre-trained model of OpenPose in pose estimation which is not specifically trained for underwater environment. Moreover, we have not included any image prepossessing step on the video frames in our approach. These were intentional since we attempted to show the performance of our work with bare minimum configuration. An improvement to our work would be re-training the OpenPose's pose estimation model on the datasets of preprocessed frames where the positions of the joints are labeled as well already. (2) At a higher application level, we aim to extend our work to breaststroke and butterfly swim styles as the pulling pose analysis is very similar to front crawl.

6. ACKNOWLEDGEMENT

This work would not have been possible without the support of the swimmers at Ryerson University masters' swim club. We are especially indebted to Danica Vidotto, the head coach, who has been supportive of our project and worked actively to provide us with meetups to videotape the swimmers. We are also grateful to Computer Vision and Image Processing Laboratory at Ryerson University. We would especially like to thank Dr. Javad Alirezaie, the director, as he provided us with a gpu workstation to run our experiments. The last, not the least, we are immensely thankful to our colleague, Negar Arabzade at Ryerson University, who provided us with her iPhone 7 to do underwater recording.

7. REFERENCES

[1] William Bradford, "Reaching the visual learner: Teaching property through art," 09 2011.

[2] Claudio Marcio de Souza Vicente, Erickson R. Nascimento, Luiz Eduardo C. Emery, Cristiano Arruda G. Flor, Thales Vieira, and Leonardo B. Oliveira, "High performance moves recognition and sequence segmentation based on key poses filtering," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016, pp. 1–8.

[3] Soudeh Kasiri Bidhendi, Clinton Fookes, Stuart Morgan, David T. Martin, and Sridha Sridharan, "Combat sports analytics: Boxing punch classification using overhead depthimagery," in *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*, 2015, pp. 4545–4549.

[4] Long Sha, Patrick Lucey, Sridha Sridharan, Stuart Morgan, and Dave Pease, "Understanding and analyzing a large collection of archived swimming videos," in *IEEE Winter Conference on Applications of Computer Vision, 2014*, 2014.

[5] 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, CVPR Workshops, 2017*, 2017.

[6] Dan Zecha, Christian Eggert, and Rainer Lienhart, "Pose estimation for deriving kinematic parameters of competitive swimmers," vol. 2017, pp. 21–29, 01 2017.

[7] 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 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 1302–1310.

[8] D. Wilkie and K. Juba, *The Handbook of Swimming*, Pelham, 1996.

[9] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001.

[10] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995.