Auton Robot
DOI 10.1007/s10514-017-9644-y

Visual identification of biological motion for underwater human–robot interaction

Junaed Sattar[1] · Gregory Dudek[2]

Received: 16 February 2016 / Accepted: 1 June 2017

Abstract We present an algorithm for underwater robots to visually detect and track human motion. Our objective is to enable human–robot interaction by allowing a robot to follow behind a human moving in (up to) six degrees of freedom. In particular, we have developed a system to allow a robot to detect, track and follow a scuba diver by using frequency-domain detection of biological motion patterns. The motion of biological entities is characterized by combinations of periodic motions which are inherently distinctive. This is especially true of human swimmers. By using the frequency-space response of spatial signals over a number of video frames, we attempt to identify signatures pertaining to biological motion. This technique is applied to track scuba divers in underwater domains, typically with the robot swimming behind the diver. The algorithm is able to detect a range of motions, which includes motion directly away from or towards the camera. Once detected, the motion of the diver relative to the vehicle is then tracked using an Unscented Kalman Filter, an approach for non-linear estimation. The efficiency of our approach makes it attractive for real-time applications on-board our underwater vehicle, and in future applications we intend to track scuba divers in real-time with the robot. The paper presents an algorithmic overview of our approach, together with experimental evaluation based on underwater video footage.

Keywords Underwater robotics · Visual servoing · Visual tracking · Diver tracking · Human–robot collaboration

1 Introduction

Motion cues have been shown to be powerful indicators of human activity and have been used in the identification of their position, behavior and identity (e.g., see Sidenbladh et al. 2000, Nixon et al. 2005, Rashid 1980). In this work, we exploit motion signatures to facilitate visual servoing, as part of a larger human–robot interaction framework. From the perspective of visual control of an autonomous robot, the ability to distinguish between mobile and static objects in a scene is vital for safe and successful navigation. For the vision-based tracking of human targets, motion patterns are an important signature, since they can provide a distinctive cue to disambiguate between people and other non-biological objects, including moving objects, in the scene. We look at these features in this paper.

Our work exploits motion-based tracking as one input cue to facilitate human–robot interaction. An important sub-task for our robot, like many others, is for it to follow a human operator (as can be seen in Fig. 1a). We facilitate the detection and tracking of the human operator using the spatio-temporal signature of human motion [the psychological effect of which on human perception has been investigated by Rashid (1980)]. In practice, this detection and servo-control behavior is just one of a suite of vision-based interaction mechanisms. In the context of servo-control, we need to detect a human, estimate his image coordinates (and possible image velocity), and exploit this in a control loop. We use the periodicity inher-

Electronic supplementary material The online version of this article (doi:10.1007/s10514-017-9644-y) contains supplementary material, which is available to authorized users.

✉ Junaed Sattar
junaed@umn.edu

Gregory Dudek
dudek@cim.mcgill.ca

[1] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

[2] School of Computer Science, McGill University, Montréal, QC, Canada

Published online: 14 June 2017

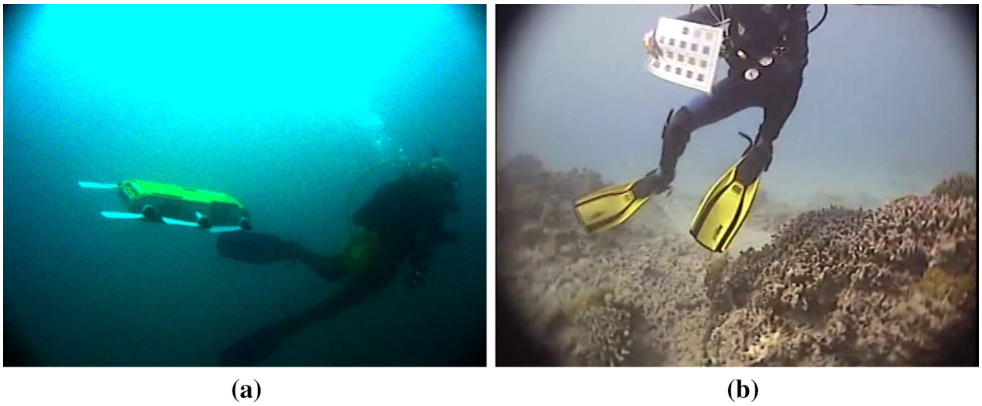


Fig. 1 External and robot-eye-view images of typical underwater scenes during target tracking by an autonomous underwater robot. **a** The Aqua underwater robot visually servoing off a target carried by a diver. **b** Typical visual scene encountered by an AUV while tracking scuba divers

ently present in biological motion (Bruce et al. 2006), and swimming in particular, to detect human scuba divers (Pendergast et al. 1996). Divers normally swim with a distinctive kicking gait which, like walking, is periodic, but also somewhat individuated. In many practical situations, the preferred applications of AUV technologies call for close interactions with humans. The underwater environment poses new challenges and pitfalls that invalidate assumptions required for many established algorithms in autonomous mobile robotics. In particular, the effect of distorted optics under water is a challenging problem to overcome for machine vision systems, thus for any autonomous system that is guided by visual sensing. Refraction, absorption and scattering of light are three major contributors for optical distortion in underwater domains (illustrated in Fig. 2), and can be encountered in most open-water (e.g., oceanic) environments. We do not track scuba divers through an air–water interface (e.g., from above the surface) in this work; however, changes in salinity, temperature variations or particulates in the water can cause refraction solely within the water medium. While not as significant as the other two phenomenon mentioned above, light refraction can cause significant visual distortion, e.g., by refracting different wavelengths of light (i.e., those belonging to different hues) by different amounts), and this in turn can affect the performance of visual perception algorithms, including visual tracking.

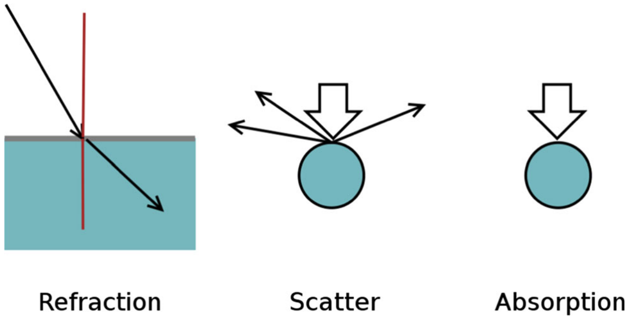


Fig. 2 Effects of light in light rays in underwater domains. An incident ray of light can encounter refraction at the air–water boundary, or at the boundary of water with different densities, arising from a variance of absorbent materials in the water. Both water molecules and suspended particles can scatter and absorb incident light rays, though the outcome of these phenomenon is dependent on the wavelength of light

While truly autonomous underwater navigation remains an important goal, having the ability to guide an underwater robot using sensory inputs also has important benefits; for example, to train the robot to perform a repetitive observation or inspection task, it might very well be convenient for a scuba diver to perform the task as the robot follows and learns the trajectory. For future missions, the robot can use the information collected by following the diver to carry out the inspection. This approach also has the added advantage of not requiring a second person tele-operating the robot, which simplifies the operational loop and reduces the associated overhead of robot deployment.

Our approach to track scuba divers in underwater video footage and real-time streaming video arises thus from the need for such semi-autonomous behaviors and visual human–robot interaction in arbitrary environments. The approach is computationally efficient for deployment onboard an autonomous underwater robot. Visual tracking is performed in the spatio-temporal domain in the image space; that is, spatial frequency variations are detected in the image space in different motion directions across successive frames. The frequencies associated with a diver's gaits (flipper motions) are identified and tracked. Coupled with a visual servoing mechanism, this feature enables an underwater vehicle to follow a diver without any external operator assistance, in environments similar to that shown in Fig. 1b.

The ability to track spatio-temporal intensity variations using the frequency domain is not only useful for tracking scuba divers, but also can be useful to detect motion of particular species of marine life (Bainbridge 1958) or

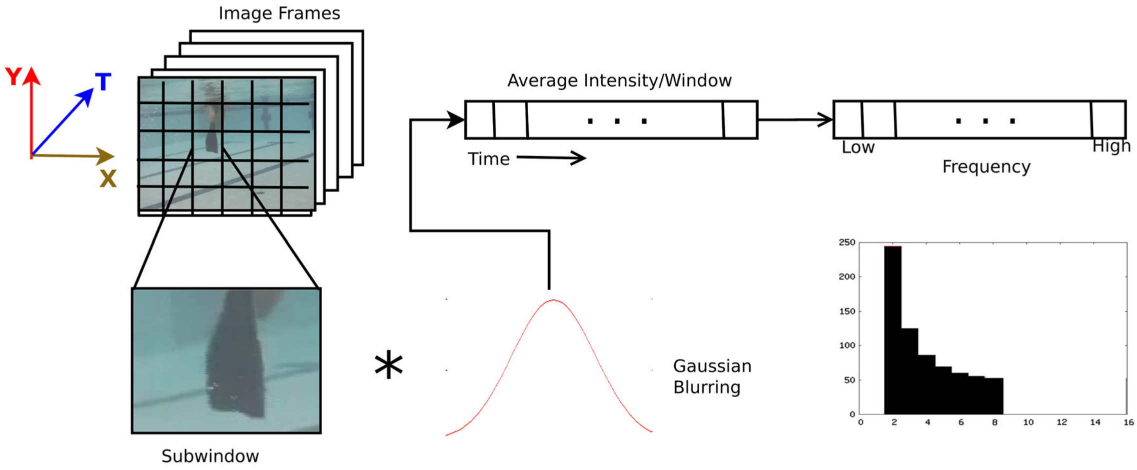


Fig. 3 Outline of the directional Fourier motion detection and tracking process. The Gaussian-filtered temporal image is split into non-overlapping sub-windows, and the average intensity of each sub-window is calculated for every timeframe. For the length of the filter, a one-dimensional intensity vector is formed, which is then passed through an FFT operator. The resulting amplitude plot can be seen, with the symmetric half removed

surface swimmers (Land 1990). It is also associated with terrestrial motion like walking or running, and our approach seems appropriate for certain terrestrial applications as well. It appears that most biological motion underwater as well as on land is associated with periodic motion, but we concentrate our attention to tracking human scuba divers and servoing off their position.

Our robot has been developed with marine ecosystem inspection as a key application area (Fig. 1a). Recent initiatives taken for protection of coral reefs call for long-term monitoring of such reefs and species that depend on reefs for habitat and food supply (Coral Reef Conservation Act 2003). We envision our vehicle to have the ability to follow scuba divers around such reefs and assist in monitoring and mapping of distributions of different species of coral. This application area is representative of the general class of deployments this technique can be applied to.

The paper is organized in the following sections: in Sect. 2 we look at related work in the domains of tracking, oriented filters and spatio-temporal pattern analysis in image sequences, Kalman filtering and underwater vision for autonomous vehicles. Our Fourier energy-based tracking algorithm is presented in Sect. 3. Experimental results of running the algorithm on video sequences are shown in Sect. 4.3.3. We draw conclusions and discuss some possible future directions of this work in Sect. 5.

2 Related work

The work presented in this paper combines previous research in different domains, and its novelty is in the use of frequency signatures in visual target recognition and tracking, combined with the Unscented Kalman Filter for tracking 6-DOF human motion. In this context, 6-DOF refers to the number of degrees of freedom of just the body center, as opposed to the full configuration space. In the following paragraphs we consider some of the extensive prior work on tracking of humans in video, underwater visual tracking and visual servoing in general.

A key aspect of our work is a filter-based characterization of the motion field in an image sequence. This has been a problem of longstanding relevance and activity, and were it not for the need for a real-time low-overhead solution, we would be using a full family of steerable filters, or a related filtering mechanism (Fleet and Jepson 1990; Freeman and Adelson 1991). In fact, since our system needs to be deployed in a hard real-time context on an embedded system, we have opted to use a sparse set of filters combined with a robust tracker. This depends, in part, on the fact that we can consistently detect the motion of our target human from a potentially complex motion field. Tracking humans using their motion on land, in two degrees of freedom, was examined by Niyogi and Adelson (1994). They look at the positions of head and ankles, respectively, and detect the presence of a human walking pattern by looking at a "braided pattern" at the ankles and a straight-line translational pattern at the position of the head. In their work, however, the person has to walk across the image plane roughly orthogonal to the viewing axis for the detection scheme to work.

There is evidence that people can be discriminated from other objects, as well as from one another, based on motion cues alone (although the precision of this discrimination may be limited). In the seminal work using "moving light displays", Rashid (1980) observed that humans are exquisitely sensitive to human-like motions using even very limited cues.

There has also been work, particularly in the context of biometric person identification, based on the automated analysis of human motion or walking gaits (Sidenbladh et al. 2000; Nixon et al. 2005; Sidenbladh and Black 2003). In a similar vein, several research groups have explored the detection of humans on land from either static visual cues or motion cues. Such methods typically assume an overhead, lateral or other view that allows various body parts to be detected, or facial features to be seen. Notably, many traditional methods have difficulty if the person is walking directly away from the camera. In contrast, the present paper proposes a technique that functions without requiring a view of the face, arms or hands (either of which may be obscured in the case of scuba divers). In addition, in our particular tracking scenario the diver can point directly away from the robot that is following him, as well as move in an arbitrary direction during the course of the tracking process.

While tracking underwater swimmers visually has not been explored in great depth in the past, some prior work has been done in the field of underwater visual tracking and visual servoing for autonomous underwater vehicles. Naturally, this is closely related to generic servo-control. In that context, on-line real-time performance is crucial. On-line tracking systems, in conjunction with a robust control scheme, provide underwater robots the ability to visually follow targets underwater (Sattar et al. 2005). Previous work on spatio-temporal detection and tracking of biological motion underwater has been shown to work well (Sattar and Dudek 2007), but only when the motion of the diver is directly towards or away from the camera. Our current work looks at motion in a variety of directions over the spatio-temporal domain, incorporates a variation of the Kalman filter and also estimates diver distance and is thus a significant improvement over that particular technique.

In terms of the tracking process itself, the Kalman filter is, of course, the preeminent classical methodology for real-time tracking. It depends, however, on a linear model of system dynamics. Many real systems, including our model of human swimmers, are non-linear and the linearization needed to implement a Kalman filter needs to be carefully managed to avoid poor performance or divergence. The Unscented Kalman Filter (Julier and Uhlmann 2004) we deploy was developed to facilitate non-linear control and tracking, and can be regarded as a compromise between Kalman Filtering and fully non-parametric condensation (Isard and Blake 1998).

3 Methodology

To track scuba divers in the video sequences, we exploit the periodicity and motion invariance properties that characterize biological motion. To fuse the responses of the multiple frequency detectors, we combine their output with an Unscented Kalman Filter (Julier and Uhlmann 2004). The core of our approach is to use periodic motion as the signature of biological propulsion and specifically for person-tracking, to detect the kicking gait of a person swimming underwater. While different divers have distinct kicking gaits, the periodicity of swimming (and walking) is universal. Our approach, thus, is to examine the amplitude spectrum of rectangular slices through the video sequence along the temporal axis. We do this by computing a windowed Fourier transform on the image to search for regions that have substantial bandpass energy at a suitable frequency. The flippers of a scuba diver normally oscillate at frequencies between 1 and 2 Hz (Wannier et al. 2001). Any region of the image that exhibits high energy responses in those frequencies is a potential location of a diver. The essence of our technique is therefore to convert a video sequence into a sampled frequency-domain representation in which we accomplish detection, and then use these responses for tracking (see Fig. 3). To do this, we need to sample the video sequence in both the spatial and temporal domain and compute local amplitude spectra. This could be accomplished via an explicit filtering mechanism such as steerable filters which might directly yield the required bandpass signals. Instead, we employ windowed Fourier transforms on the selected space-time region which are, in essence, 3-dimensional blocks of data from the video sequence (a 2-dimensional region of the image extended in time). In principle, one could directly employ color information at this stage as well, but due to the need to limit computational cost and the low mutual information content between color channels (especially underwater), we perform the frequency analysis on luminance signals only.

We look at the method of *Fourier Tracking* in Sect. 3.1. In Sect. 3.2, we describe the multi-directional version of the Fourier tracker and motion detection algorithm in the XYT domain. The application of the Unscented Kalman Filter for position tracking is discussed in Sect. 3.3. Section 3.4 looks at two parameters that affect the tracker, and lays out a set of experiments for quantitative assessment of tracker performance.

3.1 Fourier tracking

The core concept of the tracking algorithm presented here is to take a time-varying spatial signal (from the robot) and use the well-known discrete-time Fourier transform to convert the signal from the spatial to the frequency domain. Since the target of interest will typically occupy only a region of the image at any time, we naturally need to perform spatial and temporal windowing. The standard equations relating the spatial and frequency domain are as follows.

$$x[n] = \frac{1}{2\pi} \int_{2\pi} X(\mathrm{e}^{j\omega})\mathrm{e}^{j\omega} d\omega \tag{1}$$

$$X(\mathrm{e}^{j\omega}) = \sum_{n=-\infty}^{+\infty} x[n]\mathrm{e}^{-j\omega n} \tag{2}$$

where $x[n]$ is a discrete aperiodic function, and $X(\mathrm{e}^{j\omega})$ is periodic with length 2π and frequency ω. Equation 1 is referred to as the *synthesis* equation, and Eq. 2 is the *analysis* equation where $X(\mathrm{e}^{j\omega})$ is often called the *spectrum* of $x[n]$ (Oppenheim et al. 1996). The coefficients of the converted signal correspond to the amplitude and phase of complex exponentials of harmonically-related frequencies present in the spatial domain.

For our application, we do not consider phase information, but look only at the absolute amplitudes of the coefficients of the above-mentioned frequencies. The phase information might be useful in determining relative positions of the undulating flippers, for example. It might also be used to provide a discriminator between specific individuals. Moreover, by not differentiating between the individual flippers during tracking, we achieve a speed-up in the detection of high-energy responses, at the expense of sacrificing relative phase information.

Spatial sampling is accomplished using a Gaussian windowing function at regular intervals and in multiple directions over the image sequence. The Gaussian is appropriate since it is well known to simultaneously optimize localization in both space and frequency space (Duda et al. 2000). It is also a separable filter, making it computationally efficient. Note, as an aside, that a box filter for sampling can be simple and efficient, but these produce undesirable ringing in the frequency domain, which can lead to unstable tracking. The Gaussian filter has good frequency domain properties and it can be computed recursively making it exceedingly efficient.

One phenomenon that might have an effect of frequency detection is the robot's own motion, particularly in strongly perturbed waters. The motion of the Aqua robot is generated by beating flippers, but since the robot is motion stabilized with the aid of an inertial stability augmentation system (Giguere et al. 2013), the swimming motion of the robot is fairly stable, with minimal effect on the performance of the Fourier tracker. The range of data collected of swimming divers implicitly include different weather, water conditions, including effects from wave and surge effects. While we have not conducted exhaustive evaluations of the effects of the external forces exerted on the robot because of such phenomena, experimental results of the Fourier tracker conducted on footage from a variety of conditions all exhibit identical performance, indicating that the Fourier tracker compensates for robot motion to some degree implicitly.

3.2 Multi-directional motion detection

To detect motion in multiple directions, we use a predefined set of vectors, each of which is composed of a set of small, non-overlapping rectangular sub-windows in spatio-temporal space. The trajectories of each of these sub-windows are governed by a corresponding starting and ending point in the image. In any given time T, this rectangular window resides in a particular position along this trajectory and represents a Gaussian-weighted gray-scale intensity function of that particular region in the image. Over the entire trajectory, these windows generate a vector of intensity values along a certain direction in the image, producing a purely temporal signal for amplitude computation. We weight these *velocity vectors* with an exponential filter, such that intensity weights of a more recent location of the sub-window have a higher weight than another at that same location in the past. This weighting helps to maintain the causal nature of the frequency filter applied to this velocity vector. In the current work, we extract 17 such velocity vectors (as seen in Fig. 4) and apply the Fourier transform to them (17 is the optimum number of vectors we can process in quasi-real time in our robot hardware). The space formed by the velocity vectors is a conic in the XYT space, as depicted in Fig. 5. Each such signal provides an amplitude spectrum that can be matched to a profile of a typical human gait.

A statistical classifier trained on a large collection of human gait signals would be ideal for matching these amplitude spectra to human gaits [as exemplified by Begg and Kamruzzaman (2005)]. However, these human-associated signals appear to be easy to identify, and as such, an automated classifier is not currently used. Currently, we use two different approaches to select candidate spectra. In the first, we choose the particular direction that exhibits significantly higher energy amplitudes in the low-frequency bands, when compared to higher frequency bands. In the second approach, we pre-compute by hand an amplitude spectrum from video footage of a swimming diver, and use this amplitude spectrum as a true reference. To find possible matches, we use the Bhattcharyya (1943) measure to find similar amplitude spectra, and choose those as possible candidates.

3.3 Position tracking using an unscented Kalman filter

Each of the directional Fourier motion operators outputs an amplitude spectrum of different frequencies present in each associated direction. As described in Sect. 3.2, we look at the amplitudes of the low-frequency components of these directional operators, the ones that exhibit high responses are chosen as possible positions of the diver, and thus the position of the diver can be tracked across successive frames.

To further enhance the tracking performance, we run the output of the motion detection operators through an

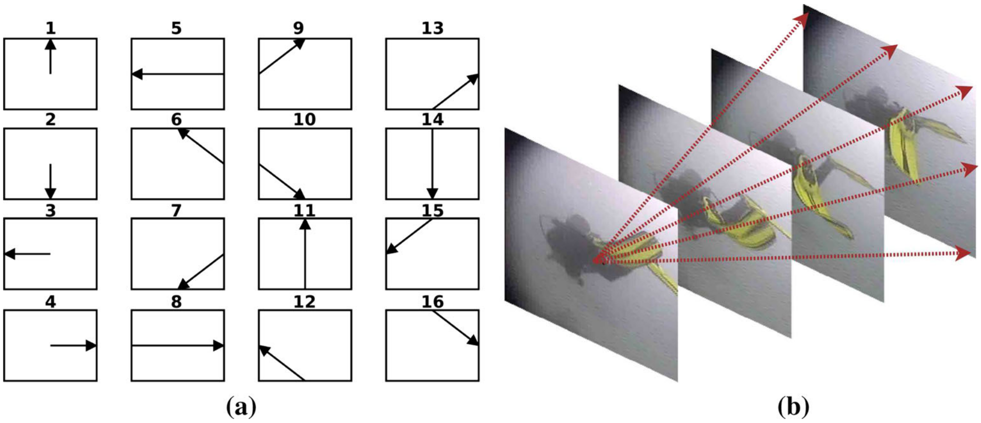


Fig. 4 Directions of motion for Fourier tracking, also depicted in 3D in a diver swimming sequence. **a** Motion directions covered by the various directional Fourier operators, depicted in a 2D spatial arrangement. **b** Image slices along the time axis showing 5 out of 17 possible track directions while tracking a scuba diver

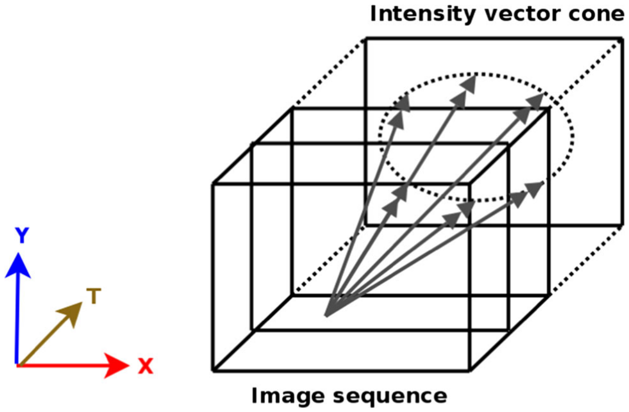


Fig. 5 Conic space covered by the directional Fourier operators

Unscented Kalman Filter (UKF), also referred to as the Sigma-Point Kalman Filter, or SPKF (although we use the term UKF throughout this paper). The UKF is a highly effective filter for state estimation problems, and is suitable for systems with a non-linear process model (Julier and Uhlmann 2004). Compared to the Extended Kalman Filter (EKF), the UKF captures the non-linearity in the process and observation models up to the second-order of the Taylor series expansion, whereas the EKF only captures up to the first-order expansion term. The track trajectory and the motion perturbation are highly non-linear, owing to the undulating propulsion resulting from flipper motion and underwater currents and surges. We chose the UKF as an appropriate filtering mechanism because of this inherent non-linearity, and also its computational efficiency.

According to the UKF model, an N-dimensional random variable $\mathbf{x}$ with mean $\hat{\mathbf{x}}$ and covariance P_{xx} is approximated by $2N + 1$ points known as the *sigma points*. The sigma points at iteration $k - 1$, denoted by $\chi^i_{k-1|k-1}$, are derived using the following set of equations:

$$\begin{aligned}
\chi^0_{k-1|k-1} &= \mathbf{x}^a_{k-1|k-1} \\
\chi^i_{k-1|k-1} &= \mathbf{x}^a_{k-1|k-1} + \left(\sqrt{(N+\lambda)(P)^a_{k-1|k-1}}\right)_i \\
& \quad i = 1 \ldots N \\
\chi^i_{k-1|k-1} &= \mathbf{x}^a_{k-1|k-1} + \left(\sqrt{(N+\lambda)(P)^a_{k-1|k-1}}\right)_{i-N} \\
& \quad i = N+1 \ldots 2N
\end{aligned}$$

where $(\sqrt{(N+\lambda)(P)^a_{k-1|k-1}})_i$ is the ith column of the matrix square-root of $((N+\lambda)(P)^a_{k-1|k-1})$, and λ is a predefined constant that dictates the spread of the sigma points.

For the diver's location, the estimated position $\mathbf{x}$ is a two-dimensional random variable, and thus the filter requires 5 sigma points. The sigma points are generated around the mean position estimate by projecting the mean along the X and Y axes, and are propagated through a non-linear motion model (i.e., the transition model) f, and the estimated mean (i.e., diver's estimated location), $\hat{\mathbf{x}}$, is calculated as a weighted average of the transformed points:

$$\chi^i_{k|k-1} = f\left(\chi^i_{k-1|k-1}\right) \quad i = 0 \ldots 2N \tag{3}$$

$$\hat{\mathbf{x}}_{k|k-1} = \sum_{i=0}^{2N} W^i \chi^i_{k|k-1} \tag{4}$$

where W^i are the constant weights for the state (*position*) estimator.

As an initial position estimate of the diver's location for the UKF, we choose the center point of the vector producing the highest low-frequency amplitude response. Ideally, the non-linear motion model for a scuba diver can be learned from training using video data, but for this application we use a hand-crafted model created from manually observing such footage. The non-linear motion model we employ predicts forward motion of the diver with a higher probability than vertical motion, which in turn is favored over lateral motion. For our application, a small number of iterations (approximately between 5 and 7) of the UKF is sufficient to assure convergence.

3.4 Parameter tuning

From the brief outline stated above, two distinct parameters are seen to be affecting the performance of the Fourier tracker—namely the size of the rectangular sub-windows, and the number of frames to look at in the time direction (i.e., the time duration in number of frames) for calculating the amplitude spectrum. The performance of the tracker, in terms of accuracy and speed, is directly dependent on these parameters. In this work, the component of the Fourier tracker which locates a diver in the image frame prior to tracking by the UKF is referred to as the *Fourier Detector*. To detect the intensity variations caused by the oscillation of the diver's flippers, we average the intensities at each frame (i.e., in each time-step) in these sub-windows. We refer to the size of the sub-windows as the *BoxSize* parameter, and refer to this parameter with the symbol κ. The duration over which these intensity variations are collected is the second parameter the tracker depends on. We call this the *RunLength* parameter, and denote it by the symbol λ. Figure 6 depicts these two parameters of the Fourier tracker over an image sequence. To assess the effect of these parameters on tracker performance, we perform a number of experiments, which are described in detail with our findings in the following sections.

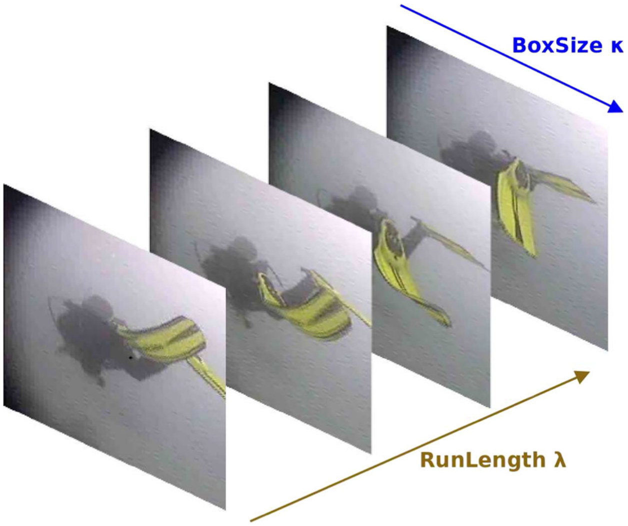


Fig. 6 A schematic showing the *RunLength* and *BoxSize* parameters for the Fourier tracker

4 Experimental evaluation

This section presents experimental results with the Fourier tracker. This algorithm has been experimentally validated on video footage recorded of divers swimming in open-water and closed-water environments. Both types of video sequences pose significant challenges due to the unconstrained motion of the robot and the diver, and the poor imaging conditions, particularly observed in the open-water footage due to suspended particles, water salinity and varying lighting conditions. The algorithm outputs a direction corresponding to the most dominant biological motion present in the sequence, and a location of the most likely position of the entity generating the motion response. Since the Fourier tracker looks backward in time every N frames to find the new direction and location of the diver, the output of the computed locations are only available after a "bootstrap phase" of N frames. We present the experimental setup below in Sect. 4.1 findings and the results in Sect. 4.2.

4.1 Experimental setup

We conduct experiments off-line on video sequences recorded from the cameras of an underwater robot. The video sequences contain footage of one or more divers swimming in different directions across the image frame, which make them suitable for validating our approach. We run our algorithm on a total of 2530 frames of a diver swimming in a pool, and 2680 frames of a diver swimming in the open-ocean, collected from open ocean field trials of the robot. In total, the frames amounted to over 10 min video footage of both environments. The *Xvid*-compressed video frames have dimensions of 768 × 576 pixels, the detector operated at a rate of approximately 10 frames/s, and the time window for the Fourier tracker for this experiment is 15 frames, corresponding to approximately 1.5 s of footage. Each rectangular sub-window is 40 × 30 pixels in size (one-fourth in each dimension). The sub-windows do not overlap each other on the trajectory along a given direction. Ground truth for evaluation was obtained by-hand, as discussed in Sect. 4.3.1.

For visually servoing off the responses from the frequency operators, we couple the motion tracker with a simple Proportional–Integral–Derivative (PID) controller (see Leigh 2004, for example), similar to the case with the Spatio-Temporal tracker. The PID controller accepts image space coordinates as input and provides as output motor commands for the robot such that the error between the desired position of the tracked diver and the current position is minimized. While essential for following any arbitrary target, the servoing technique is not an integral part of this motion detection algorithm, and thus runs independently of any specific visual tracking algorithm. Additionally, a wide-speed autopilot and stability augmentation system (Giguere et al. 2013) drives the robot in a stable and consistent manner towards waypoints received from any arbitrary controller, either autonomous or manual (e.g., tethered joystick controllers).

4.2 Results

Figure 7 shows a diver swimming along a diagonal direction away from the camera, as depicted by the green arrow. No part of the diver falls on the direction shown by the red arrow, and as such there is no component of motion present in that direction. Figure 8a, b shows the Fourier filter output for those two directions, respectively (the green bars correspond to the response along the green direction, and similarly for the red bars). The DC component from the FFT and the symmetric half of the FFT over the Nyquist frequency has been manually removed. The plots clearly show a much higher low-frequency response along the direction of the diver's motion, and almost negligible response in the low frequencies (as a matter of fact in all frequencies) in the direction containing no motion component (as seen from the amplitude values). Note that the lane markers on the bottom of the pool (that appear periodically in the image sequence) do not generate proper frequency responses to be categorized as biological motion in the direction along the red line.

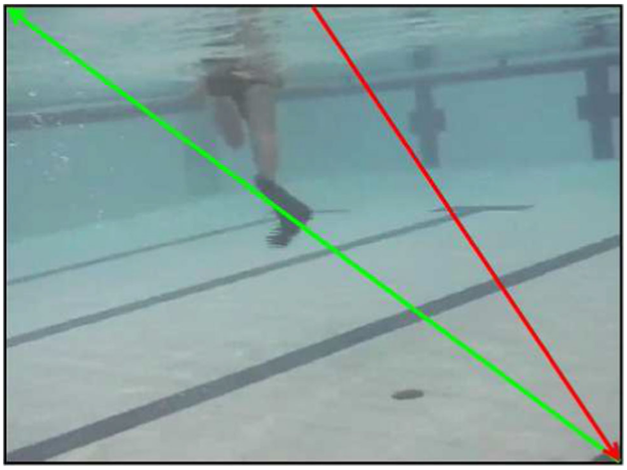


Fig. 7 Snapshot image showing direction of a swimmer motion (in *green*) and an arbitrary direction without a diver (in *red*) (Color figure online)

In Fig. 9, we demonstrate the performance of the detector in tracking multiple divers swimming in different directions. The sequence shows a diver swimming in a direction away from the robot, while another diver is swimming in front of her across the image frame in an orthogonal direction. The amplitude responses obtained from the Fourier operators along the directions of the motion for the fundamental frequency are listed in ascending order in Table 1. The first two rows correspond to the direction of motion of the diver going across the image, while the bottom three rows represent the diver swimming away from the robot. As expected, the diver closer and unobstructed to the camera produces the highest responses, but motion of the other diver also produces significant low-frequency responses. The other 12 directions exhibit negligible amplitude responses in the proper frequencies compared to the directions presented in the table. The FFT plots for motion in the bottom-to-top and left-to-right direction are seen in Fig. 10a, b respectively. Some additional Fourier detector responses are shown in Fig. 12. As before, the FFT plot has the DC component and the sym-

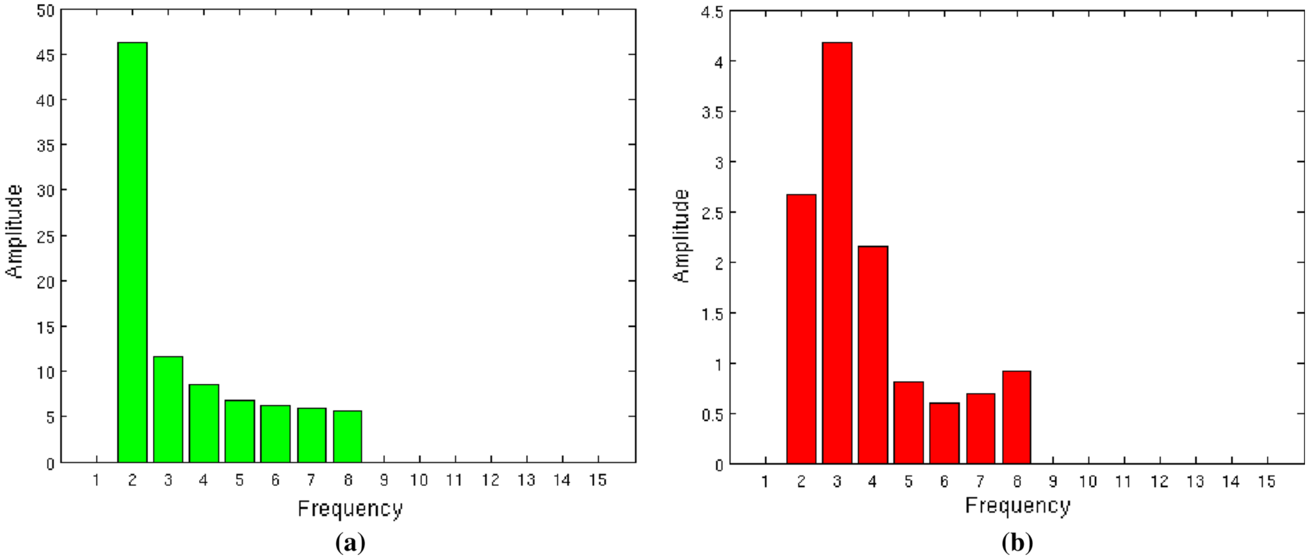


Fig. 8 Contrasting frequency responses for directions with and without diver motion in a given image sequence. **a** Frequency responses along the motion of the diver, depicted by the *green arrow* in Fig. 7. **b** Frequency responses along the direction depicted by the *red arrow*. Note the low amplitude values (Color figure online)

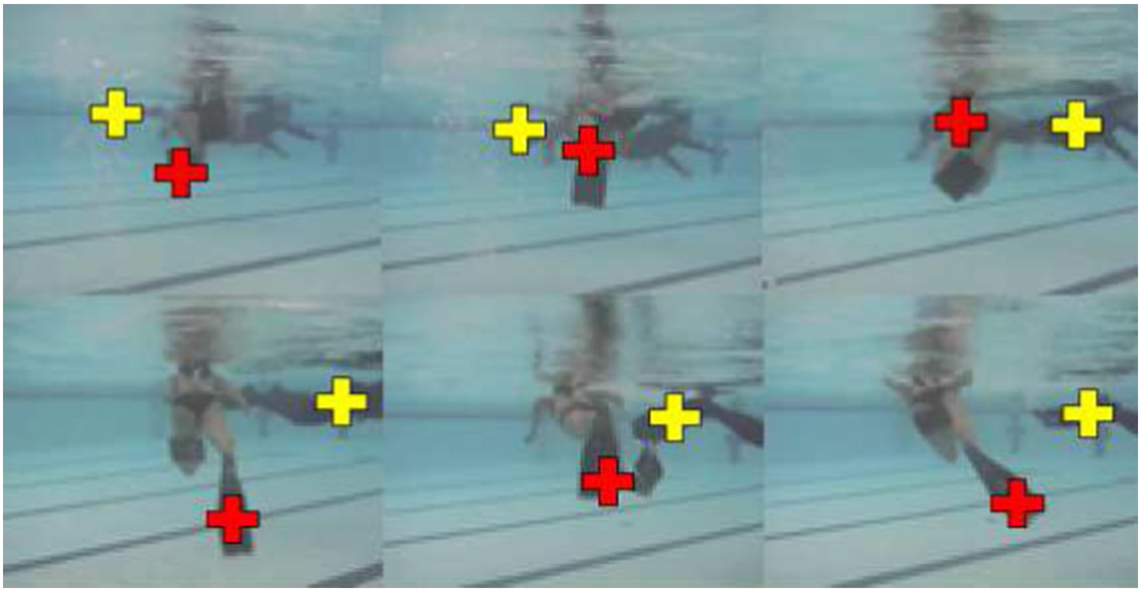


Fig. 9 An image sequence capturing two divers swimming in orthogonal directions

Table 1 Low-frequency amplitude responses for multiple motion directions

Direction	Lowest-frequency amplitude response
Left-to-right	205.03
Right-to-left	209.40
Top-to-bottom	242.26
Up-from-center	251.61
Bottom-to-top	281.22

metric half removed for presentation clarity. It is also worth noting that the diver swimming across the screen has almost minimal motion of his feet, and yet the Fourier tracker is able to detect his presence along the correct motion vector. Once obscured by the other diver, the Fourier tracker loses the frequency signature, but is able to reacquire it once he reappears.

An interesting side-effect of the Fourier tracker is the effect of the diver's distance from the robot (and hence the camera) on the low-frequency signal. The close proximity to the robot (i.e., camera) results in a lower variation

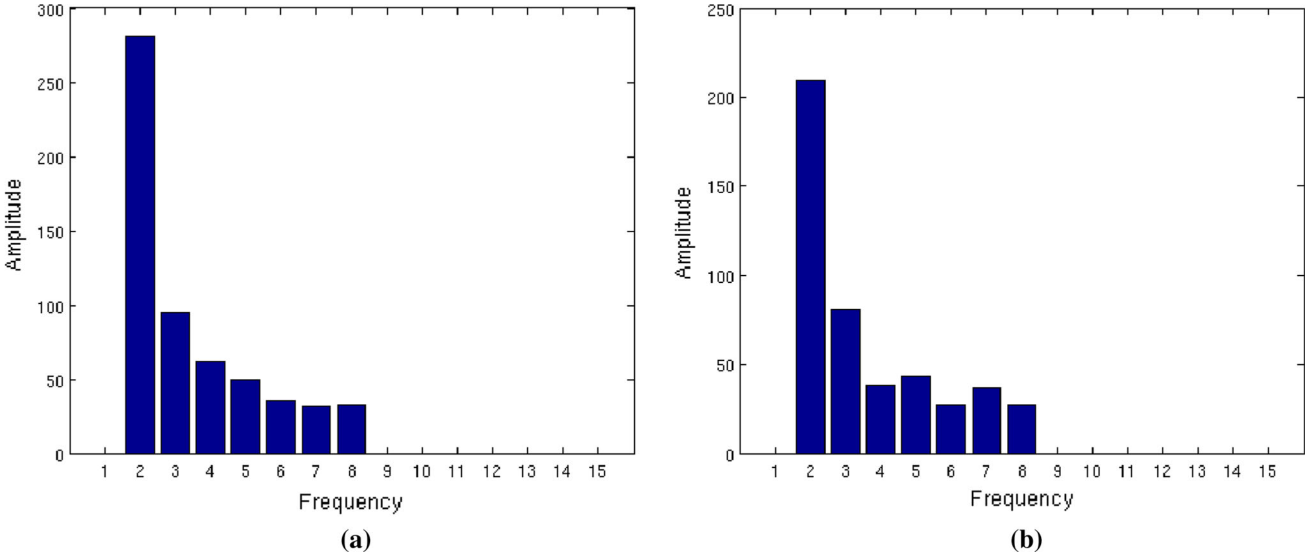


Fig. 10 Frequency responses for two different directions of diver motion in a single image sequence. **a** Frequency responses for the diver swimming away from the robot (*red cross*) in Fig. 9. **b** Frequency responses for the diver swimming across the robot (*yellow cross*) in Fig. 9 (Color figure online)

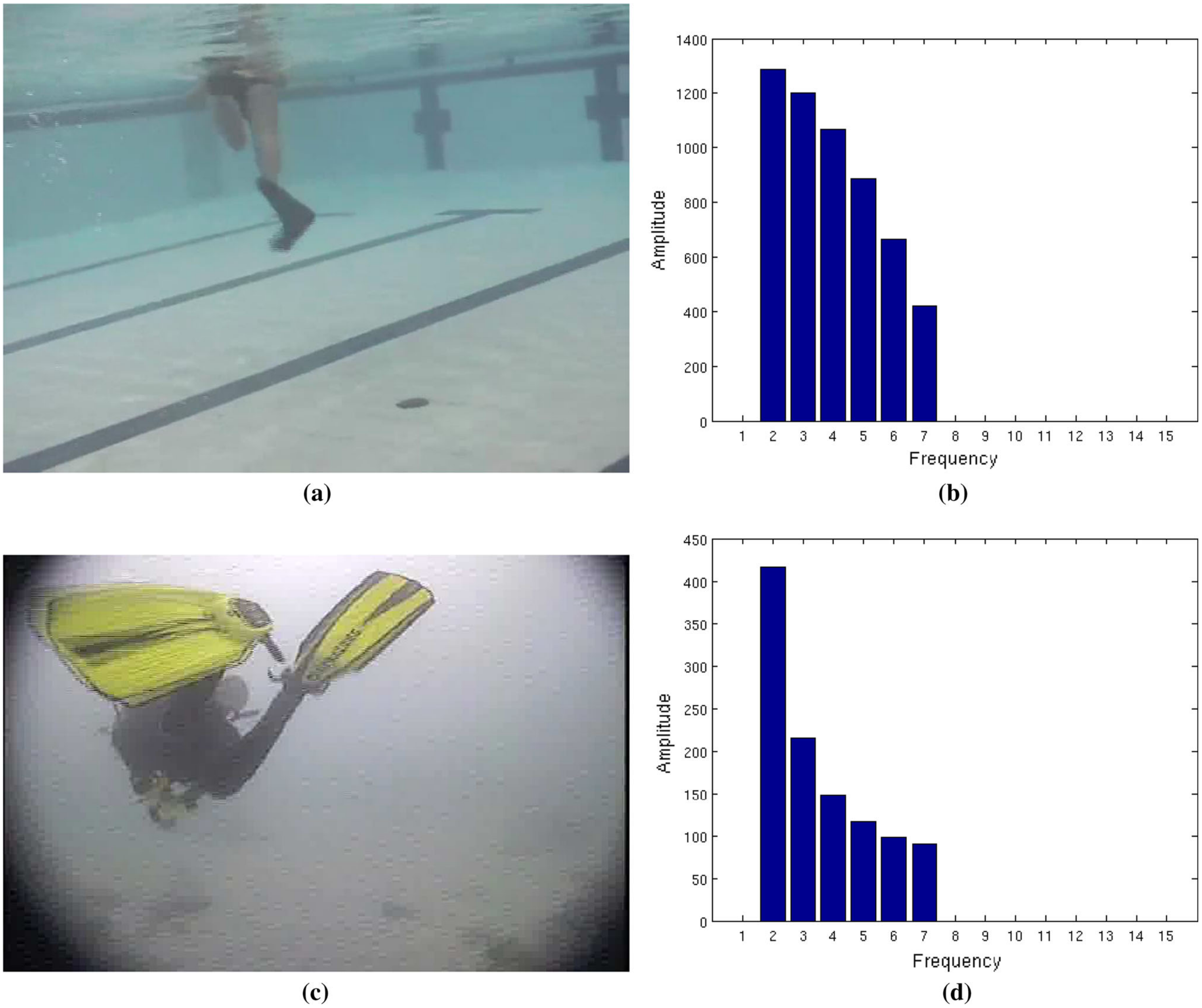


Fig. 11 Effect of diver's distance from camera on the amplitude spectra. Being farther away from the camera produces higher energy responses (Fig. 11b) in the low-frequency bands, compared to divers swimming closer (Fig. 11d)

of the intensity amplitude, and thus the resulting Fourier amplitude spectra exhibits lower energy in the low-frequency bands. Figure 11 shows two sequences of scuba divers swimming away from the robot, with the second diver closer to the camera. The amplitude responses have similar patterns, exhibiting high energy at the low-frequency regions. The spectrum on top, however, has more energy in the low-frequency bands than the one on the bottom, where the diver is closer to the camera (Fig. 12).

4.3 Performance evaluation

4.3.1 Datasets

To measure the effect of tracker accuracy and timing, we conducted a set of experiments by running the tracker on three different datasets (with available ground truth) of scuba divers swimming (Fig. 11a, c). These datasets were collected in a variety of environmental conditions—one dataset is of a diver swimming in a pool, the others are of scuba divers swimming in open-ocean environments. Each dataset contains approximately 3000 frames, which accounts for about a total of 40 min of footage. The datasets were collected from the robot's on-board cameras, running at a rate of 15 frames/s, and has dimensions of 720 × 480 pixels. While the footage is in color, we normalize the images to gray scale format, and only use the luminosity channel (average of the R, G and B channels) for Fourier tracking, discarding all color information.

4.3.2 Experiments

We created 15 configurations of the Fourier tracker by setting the *RunLength* (i.e., λ) and *BoxSize* (i.e., κ) arguments

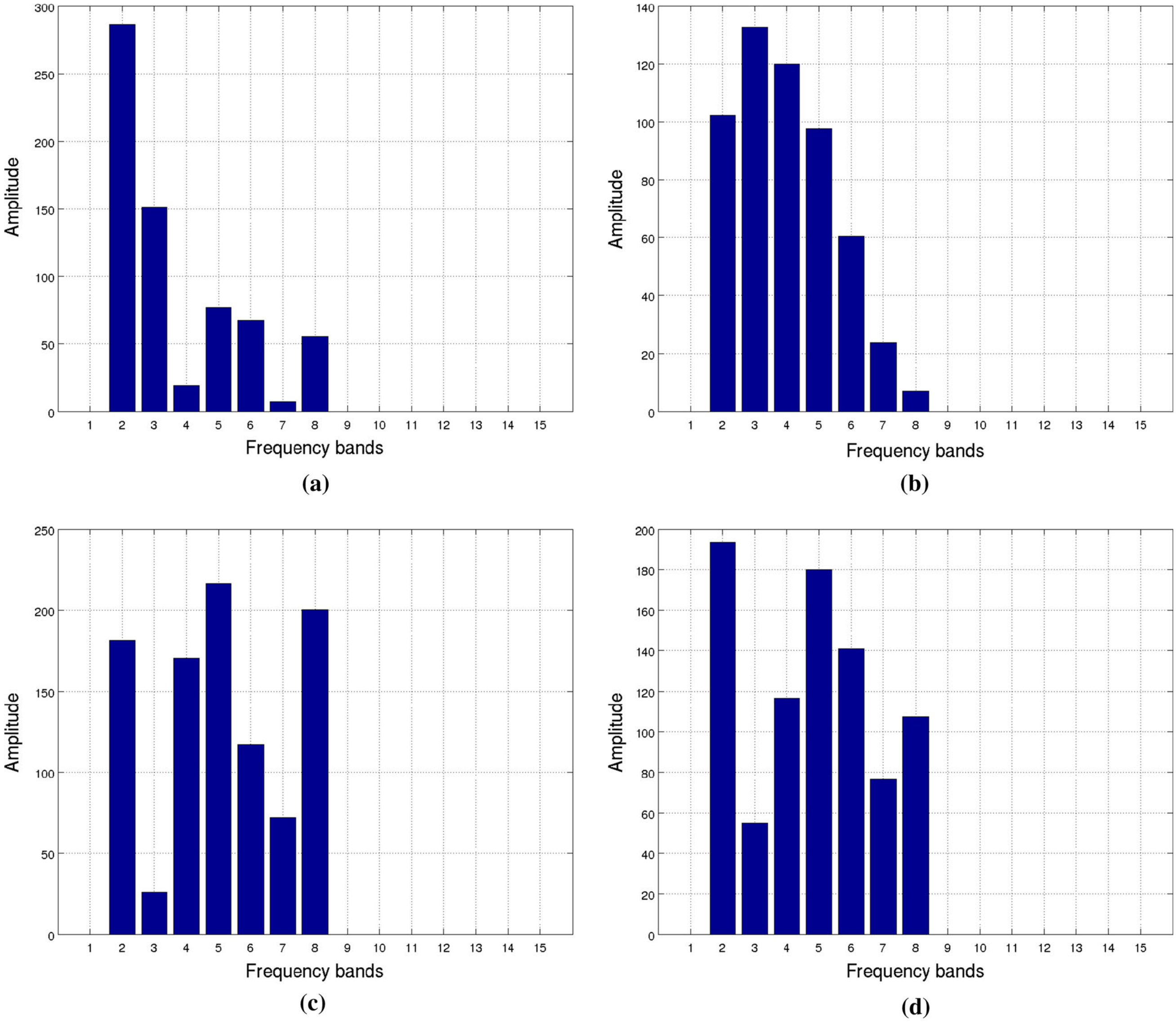


Fig. 12 Additional instantaneous amplitude responses at various times during diver tracking. Figure 12a, b are Fourier signatures of the diver's flippers, whereas Fig. 12c, d are examples of random (i.e., non-diver) locations

to different values, and ran all 15 configurations on the three datasets mentioned above. The λ parameter was set to $\frac{1}{2}$, same, and twice that of the camera frame-rate, corresponding to λs of 8, 15 and 30 frames/s. The κ parameter was set to maintain an aspect ratio of 5 : 4, at 5 different scales, with dimensions of 10×8, 20×16, 40×32, 80×64, and 160×128 pixels. We computed running time and accuracy for all configurations. In the experiments, we used the UKF component to track targets after the Fourier detector outputs the divers initial location. The accuracy measure includes output of the UKF, with the Fourier detector serving as measurement for the filter. Thus, while we vary the two parameters of the detector, the UKF parameters remain unchanged throughout the experiments.

To obtain a representative evaluation of the tracker's performance, we executed the test runs on-board the Aqua amphibious robot, thus ensuring that the execution occurs on the exact computing hardware available on the robot. The tracker is implemented in C++, and is part of a large suite of vision-based Human–Robot Interaction framework called *VisionSandbox* (Sattar and Dudek 2009a). The VisionSandbox framework is a collection of algorithms for visual tracking, diver following, learning-based object detection (Sattar and Dudek 2009b), gesture-based human–robot communications (Dudek et al. 2007) and risk assessment in human–robot dialogs (Sattar and Dudek 2011). Performance data was collected and stored in Matlab `.mat` format data files, for off-line analysis with Matlab.

4.3.3 Results

Figure 13a shows the time taken by the Fourier tracker per tracking sequence (i.e., over a length of λ frames). We observe from the experiments that the speed of the tracker increases linearly as the λ parameter is increased. As it can be seen, the time taken at each value of λ virtually doubles with the doubling of the λ parameter for up to the κ value of 80 × 64 pixels, and is almost five times higher when κ is set to 160 × 128 pixels. This is easily explained as the output of the tracker is not available until a vector of intensity values are available, and such vectors are only available after every λ frame. If the camera is operating at a frame-rate of C_f, then the time T_{iv} taken for each intensity vector to become available is

$$T_{iv} = \frac{RunLength}{C_f} \tag{5}$$

Thus, for a given camera (i.e., a given frame-rate), the time taken by the Fourier tracker is directly proportional to the value of the λ parameter.

The effect of variable κ parameters are not as linear, as shown in Fig. 13b. For the first three κ values of 10 × 8, 20 × 16 and 40 × 32 pixels, execution time does not increase proportionally. On the other hand, the time required by the 80 × 64 dimension κ is almost double of that taken by κ of dimension 40 × 32. For the sake of brevity, the plots are labeled with the width of the κ parameter only (i.e., "BoxSize 80" denotes a κ value of 80 × 64 pixels).

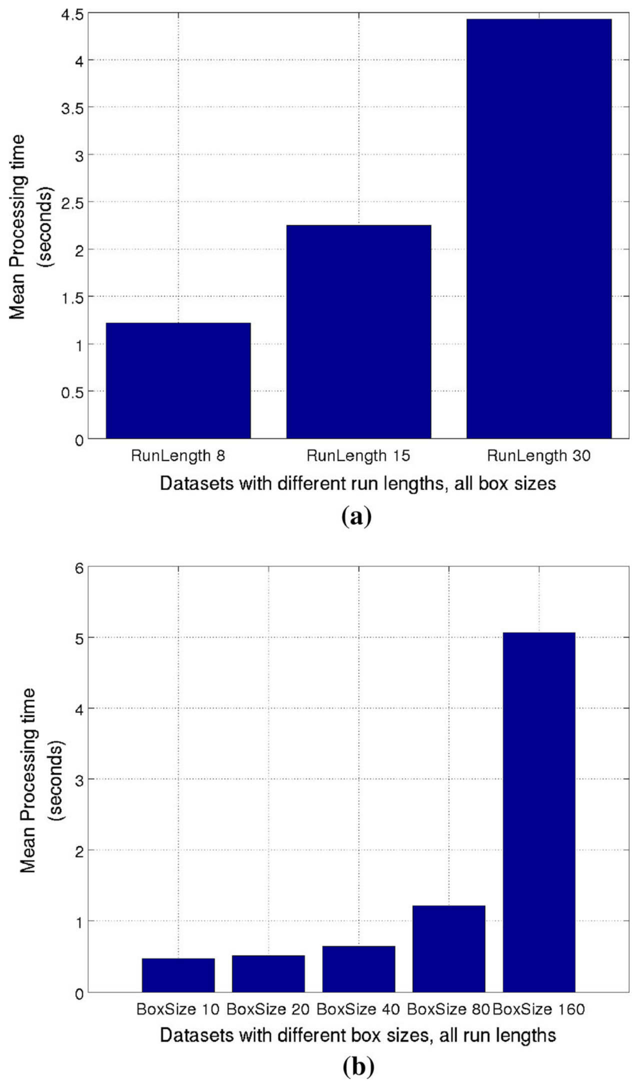


Fig. 13 Effect of the λ and κ parameters on tracker timing, with time taken shown per detection. **a** Timing with different RunLengths values, all BoxSizes. **b** Timing with different BoxSize values, all RunLengths

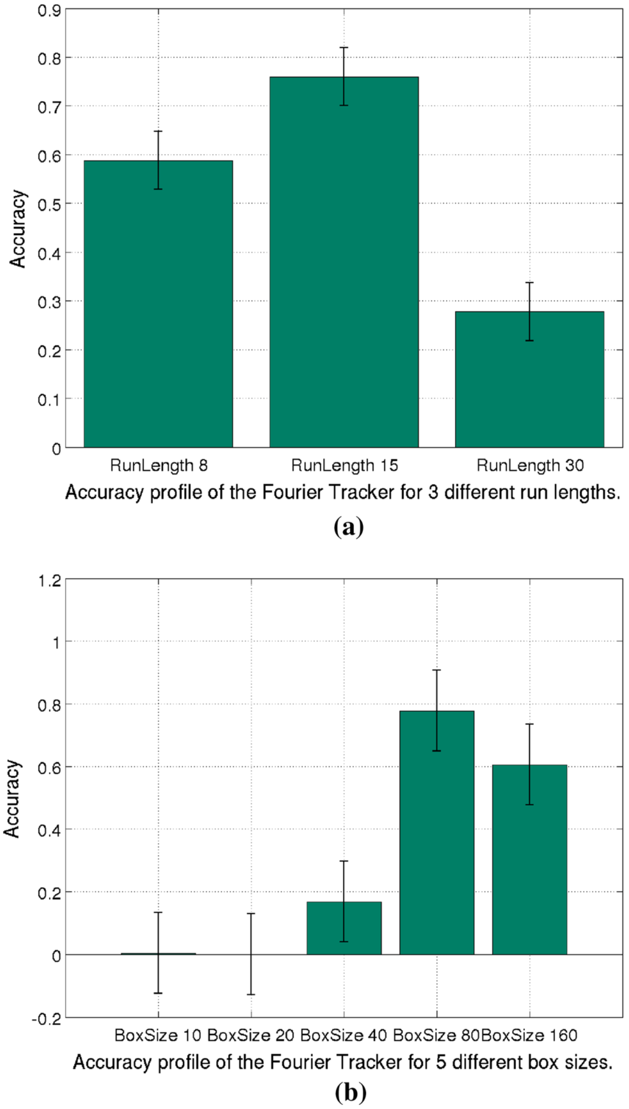


Fig. 14 Effect of the RunLength and BoxSize parameters on tracker accuracy. **a** Accuracy with different RunLengths values, all BoxSizes. **b** Accuracy with different BoxSize values, all RunLengths

The timing data suggests a larger value for κ and a longer λ would be detrimental for real-time performance that we seek. Smaller values for both would result in the Fourier tracker running at approximately 10 frames/s, and the Aqua vehicle would receive control inputs from the Fourier tracker at the same rate. For the aquatic environment, and considering the dynamics of the robot, a command rate of 10 Hz is more than sufficient to control the robot to follow targets based on visual input; given the fact that a scuba diver swims at a sustained speed of less than 0.5 m/s (Andersen 1969; Pendergast et al. 1996), a tracking rate of 1 Hz is sufficient for the robot to keep track of a diver using the Fourier tracker. From the timing data, it is thus evident that for all κ and λ values we tested, the Fourier tracker is able to run sufficiently fast.

Consequently, accuracy of the tracker for these different configurations become the significant statistic. We executed the Fourier tracker on the same dataset, annotated with ground truth locations for the diver's locations, and measured the number of frames the tracker was able to correctly locate the diver in. The results are summarized in Fig. 14. Figure 14a shows the accuracy rate of the Fourier tracker with different run lengths. A λ value of 15 yields the highest accuracy, while the significantly worst performer is the λ value of 30 frames. This suggests against using a longer duration for Fourier signature detection. On the other hand, as shown in Fig. 14b, the difference in accuracy across κ values is quite significant. A κ value of 80×64 pixels (shown as "BoxSize 80") yields very high accuracy across the entire dataset, and across all λ values. The κ values before and after 80×64 pixels show decreasing accuracy, indicating no utility in arbitrarily increasing κ. Moreover, from the timing data, it is evident that a larger κ contributes to a slow-running tracker, making larger values of κ undesirable.

5 Conclusions and future work

In this paper, we present a technique for robust detection and tracking of biological motion underwater, specifically to track human scuba divers. We consider the ability to visually detect biological motion an important feature for any mobile robot, and especially for underwater environments to interact with a human operator. In a larger scale of visual human–robot interaction, such a feature forms an essential component of the communication paradigm, using which an autonomous vehicle can effectively recognize and accompany its human controller. The algorithm presented here is conceptually simple and easy to implement. Significantly, this algorithm is optimized for real-time use on-board an underwater robot. While we apply a heuristic for modeling the motion of the scuba diver to feed into the UKF for position tracking, we strongly believe that with the proper training data, a more descriptive and accurate model can be learned. Incorporating such a model promises to increase the performance of the motion tracker.

While color information can be valuable as a tracking cue, we do not look at color in conjunction with this method. Hues are affected by the optics of the underwater medium, which changes object appearances drastically. Lighting variations, suspended particles and artifacts like silt and plankton scatter, absorb or refract light underwater, which directly affects the performance of otherwise-robust tracking algorithms (Sattar and Dudek 2006). To reduce these effects and still have useful color information for robustly tracking objects underwater, we have developed a machine learning approach based on the classic Boosting technique. In that work, we train our visual tracker with a bank of *spatio-chromatic* filters (Sattar and Dudek 2009b) that aim to capture the distribution of color on the target object, along with color variations caused by the above-mentioned phenomena. Using these filters and training for a particular diver's flipper, robust color information can be incorporated in the Fourier tracking mechanism, and be directly used as an input to the UKF. While this will increase the computational cost somewhat, and also introduce color dependency, we believe investigating the applicability of this machine learning approach in our Fourier tracker framework is a promising avenue for future research.

References

Andersen, B. G. (1969). Measurement of scuba diver performance in open ocean environment. In *American Society of Mechanical Engineers, Papers* (pp. 8–16).

Bainbridge, R. (1958). The speed of swimming of fish as related to size and to the frequency and amplitude of the tail beat. *Journal of Experimental Biology*, *35*(1), 109–133.

Begg, R., & Kamruzzaman, J. (2005). A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data. *Journal of Biomechanics*, *38*(3), 401–408.

Bhattcharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society*, *35*, 99–110.

Bruce, B. D., Stevens, J. D., & Malcolm, H. (2006). Movements and swimming behaviour of white sharks (Carcharodon carcharias) in Australian waters. *Marine Biology*, *150*(2), 161–172.

Coral Reef Conservation Act. (2003). Coral reef conservation act of 2000. P.L. 106–562, 16 U.S.C. 6401 et seq.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). Wiley. ISBN 0471056693.

Dudek, G., Sattar, J., & Xu, A. (2007). A visual language for robot control and programming: A human-interface study. In *Proceedings of the International Conference on Robotics and Automation ICRA, Rome, Italy* (pp. 2507–2513).

Fleet, D. J., & Jepson, A. D. (1990). Computation of component velocity from local phase information. *International Journal of Computer Vision*, *5*(1), 77–104.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(9), 891–906.

Giguere, P., Girdhar, Y., & Dudek, G. (2013). Wide-speed autopilot system for a swimming hexapod robot. In *2013 International Conference on Computer and Robot Vision* (pp. 9–15).

Isard, M., & Blake, A. (1998). Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, *29*(1), 5–28.

Julier, S. J., & Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE, 92*(3), 401–422.

Land, M. F. (1990). Optics of the eyes of marine animals. In P. J. Herring, A. K. Campbell, M. Whitfield, & L. Maddock (Eds.), *Light and life in the sea* (pp. 149–166). Cambridge: Cambridge University Press.

Leigh, J. R. (2004). *Control theory*. Institution of Electrical Engineers. ISBN 0863413390.

Nixon, M. S., Tan, T. N., & Chellappa, R. (2005). *Human identification based on gait*., The Kluwer international series on biometrics New York: Springer.

Niyogi, S. A., & Adelson, E. H. (1994). Analyzing and recognizing walking figures in XYT. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 469–474).

Oppenheim, A. V., Willsky, A. S., & Hamid Nawab, S. (1996). *Signals & systems* (2nd ed.). Upper Saddle River: Prentice-Hall Inc. ISBN 0-13-814757-4.

Pendergast, D. R., Tedesco, M., Nawrocki, D. M., & Fisher, N. M. (1996). Energetics of underwater swimming with SCUBA. *Medicine & Science in Sports & Exercise*, 28(5), 573. ISSN 0195-9131.

Rashid, R. F. (1980). Toward a system for the interpretation of moving light display. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*(6), 574–581.

Sattar, J., & Dudek, G. (2006). On the performance of color tracking algorithms for underwater robots under varying lighting and visibility. In *Proceedings of the IEEE International Conference on Robotics and Automation ICRA, Orlando, Florida* (pp. 3550–3555).

Sattar, J., & Dudek, G. (2007). Where is your dive buddy: Tracking humans underwater using spatio-temporal features. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Diego, California, USA* (pp. 3654–3659).

Sattar, J., & Dudek, G. (2009a). A vision-based control and interaction framework for a legged underwater robot. In *Proceedings of the Sixth Canadian Conference on Robot Vision (CRV), Kelowna, British Columbia* (pp. 329–336).

Sattar, J., & Dudek, G. (2009b). Robust servo-control for underwater robots using banks of visual filters. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA, Kobe, Japan* (pp. 3583–3588).

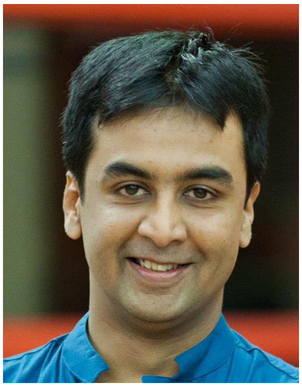
Sattar, J., & Dudek, G. (2011). Towards quantitative modeling of task confirmations in human–robot dialog. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA, Shanghai, China* (pp. 1957–1963).

Sattar, J., Giguère, P., Dudek, G., & Prahacs, C. (2005). A visual servoing system for an aquatic swimming robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Edmonton, Alberta, Canada* (pp. 1483–1488).

Sidenbladh, H., & Black, M. J. (2003). Learning the statistics of people in images and video. *International Journal of Computer Vision*, 54(1–3), 181–207. ISSN 0920-5691.

Sidenbladh, H., Black, M. J., & Fleet, D. J. (2000). Stochastic tracking of 3D human figures using 2D image motion. In *Proceedings of the European Conference on Computer Vision* (vol. 2, pp. 702–718).

Wannier, T., Bastiaanse, C., Colombo, G., & Dietz, V. (2001). Arm to leg coordination in humans during walking, creeping and swimming activities. *Experimental Brain Research*, *141*(3), 375–379. doi:10.1007/s002210100875. ISSN 0014-4819, PMID: 11715082.



Junaed Sattar is an Assistant Professor at the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities, and the founding director of the Minnesota Interactive Robotics and Vision Laboratory. Before coming to UoM, Junaed was an Assistant Professor at Clarkson University in Potsdam, NY, and was the founding director of the Robotics, Autonomy and Interactions Lab (RAIL). Junaed focuses his research on field robotics and human-in-the-loop autonomy for mobile robots. His research looks into field robotics, robot vision, human-robot interaction, applied machine learning for robotics, and assistive robotics. His research has been applied to variety of robotic platforms, including underwater, aerial, terrestrial, service and assistive robots. His work has been featured in a number of Canadian and International media outlets such as CTV National, Canal Savoir (in French) and the Discovery Channel. He has been a reviewer and program committee member in a number of international conferences, journals, and professional activities concerned with Robotics and Computer Vision.

Gregory Dudek research deals with sensing for robots, intelligent systems human-robot interaction and the development of underwater and amphibious robots. He is a professor at the School of Computer Science at McGill University, Director of the NSERC Canadian Field Robotics Network, and James McGill Chair. In 2010 he was awarded the Fessenden Professorship in Science Innovation and was also awarded the Canadian Image Processing and Pattern Recognition Award for Research Excellence and also for Service to the Research Community. Professor Dudek directs the McGill Mobile Robotics Laboratory and has authored over 200 research publications. This includes a book entitled "Computational Principles of Mobile Robotics" co-authored with Michael Jenkin and published by Cambridge University Press and now in its 2nd edition. In 2008 he co-founded the company Independent Robotics Inc. He has chaired, reviewed and been otherwise involved in numerous national and international conferences, journals, and professional activities concerned with Robotics, Machine Sensing and Computer Vision.