

Where is your dive buddy: tracking humans underwater using spatio-temporal features

Junaed Sattar and Gregory Dudek
Centre for Intelligent Machines,
McGill University, 3480 University Street,
Montreal, Québec, Canada H3A 2A7.
{junaed, dudek}@cim.mcgill.ca

Abstract— We present an algorithm for underwater robots to track mobile targets, and specifically human divers, by detecting periodic motion. Periodic motion is typically associated with propulsion underwater and specifically with the kicking of human swimmers. By computing local amplitude spectra in a video sequence, we find the location of a diver in the robot’s field of view. We use the Fourier transform to extract the responses of varying intensities in the image space over time to detect characteristic low frequency oscillations to identify an undulating flipper motion associated with typical gaits. In case of detecting multiple locations that exhibit large low-frequency energy responses, we combine the gait detector with other methods to eliminate false detections. We present results of our algorithm on open-ocean video footage of swimming divers, and also discuss possible extensions and enhancements of the proposed approach for tracking other objects that exhibit low-frequency oscillatory motion.

I. INTRODUCTION

In this paper propose a technique to allow an underwater robot to detect specific classes of biological motion using visual sensing. We are specifically interested in tracking humans, which has many applications including servo-control. Development of underwater autonomous vehicles (UAV) has made rapid progress in recent times. Equipped with a variety of sensors, these vehicles are becoming an essential part of sea exploration missions, both in deep- and shallow-water environments. In many practical situations the preferred applications of UAV technologies call for close interactions with humans. The underwater environment poses new challenges and pitfalls that invalidates preassumptions required for many established algorithms in autonomous mobile robotics. 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 repeatative 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 executions, the robot can utilize 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.

Keeping such semi-autonomous behaviors in mind, we present a novel application of tracking scuba divers in underwater video footage and real-time streaming video for

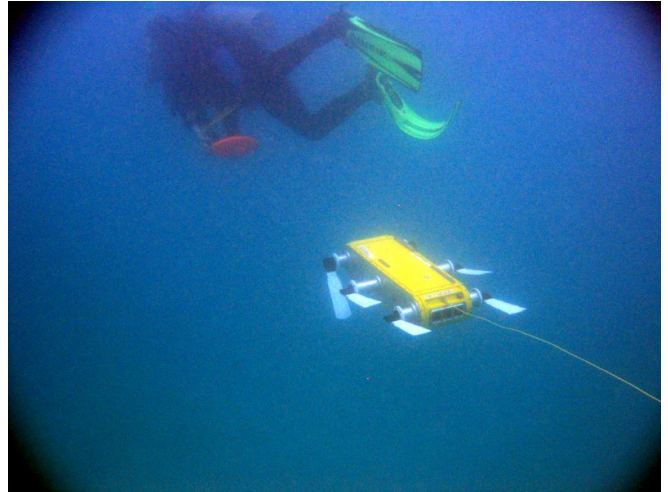


Fig. 1. The Aqua robot following a diver

on-board deployment in an autonomous underwater robot. Visual tracking is performed in spatio-temporal domain in the image space; that is, spatial frequency variations are detected in the image space across successive frames. The frequencies associated with a divers gaits (flippers motions) are identified and tracked in successive frames. Coupled with a visual servoing mechanism, this feature enables an underwater vehicle to follow a diver without any external operator assistance.

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 or surface swimmers. It appears that most biological motion underwater is associated with periodic motion, but in this paper we confine our attention to tracking human scuba divers and servoing off their position. Our platform, the Aqua amphibious legged robot [1], is being developed with marine ecosystem inspection as a key application area. 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. From the perspective of an autonomous vehicle, this can be classified as a Site Acquisition and Scene Reinspection (SASR) task. The robot can visually follow a scuba diver over a reef as he swims around the site, and for future

reinspections, make use of this information.

The paper is organized in the following sections: in Sec. II we look at related work in the domains of tracking, oriented filters and spatio-temporal pattern analysis in image sequences, as well as underwater vision for autonomous vehicles. The Fourier energy-based tracking algorithm is presented in Sec. III. Experimental results of running the algorithm on video sequences are shown in Sec. IV. We draw conclusions and discuss some possible future directions of this work in Sec. V.

II. RELATED WORK

The work presented in this paper combines previous work done in different domains, and its novelty is in the use of frequency domain information in visual target recognition and tracking. 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.

Our work is based on estimating amplitude spectra in the temporal domain of live video. If the computational resources were available we might seek to estimate a full ensemble of spatio-temporal orientations using a technique such as the well-established steerable filter [2].

Niyogi and Adelson have utilized spatio-temporal patterns in tracking human beings on land [3]. 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.

Several researchers have looked into the task of tracking a person by identifying walking gaits. Recent advancements in the field of Biometrics also have shown promise in identifying humans from gait characteristics [4]. It appears that different people have characteristic gaits and it may be possible to identify a person using the coordinated relationship between their head, hands, shoulder, knees, and feet.

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 is typically points directly away from the robot that is following them.

While tracking underwater swimmers visually have not been explored in the past, some prior work has been done in the field of underwater visual tracking and visual servoing for AUVs. Naturally, this is closely related to generic servo-control. The family of algorithms developed are both of the offline and online variety. The online tracking systems, in

conjunction with a robust control scheme, provide underwater robots the ability to visually follow targets underwater [5].

III. METHODOLOGY

The core of our approach is to use periodic motion as the signature of biological propulsion and, specifically for person-tracking, to use it 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 local amplitude spectra of the image in the frequency domain. We do this by computing a windowed Fourier transform on the image to search for regions that have substantial band-pass energy at a suitable frequency. The flippers of a scuba diver normally oscillate at frequencies of between 1 and 2 Hz. Any region of the image that exhibits high energy responses in those frequencies is a potential location of a flipper.

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. 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 2D region of the image extended in time). In principle, one could directly employ color information at this stage as well, but both due to the need to limit computational cost as well as the low mutual information content between color channels (especially underwater), we perform the frequency analysis on luminance signals only. The algorithm is explained in further detail in the following subsections.

A. Fourier Tracking

The core concept of 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 equation relating the spatial and frequency domain is as follows.

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

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

where $x[n]$ is a discrete aperiodic function, and $X(e^{j\omega})$ is periodic with length 2π . Equation 1 is referred to as the *synthesis* equation, and Eq. 2 is the *analysis* equation where $X(e^{j\omega})$ is often called the *spectrum* of $x[n]$. 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 discriminate between specific individuals. This particular work does not differentiate between the individual flippers during tracking. This also speeds up the detection of high energy responses, at the expense of sacrificing relative phase information.

In this paper we will consider only purely temporal gait signatures. That is, we approximate the person's motion as directly away from or towards the camera over the sampling interval used to compute the gait signal. In practice, the diver may often have a small lateral motion as well, but in the viscous underwater medium it appears this can be ignored. We comment further on this later on.

To detect an oscillating object close to the frequency of a scuba diver's flippers, we search in small regions of the image and compute the amplitude spectrum using a temporal window (since the diver may not remain in a region of the image for very long). Spatial sampling is accomplished using a Gaussian windowing function at regular intervals over the image. The Gaussian is appropriate since it is well-known to simultaneously optimize localization in both space and frequency space. It is also a separable filter, making it computationally efficient. Note, as an aside, that some authors have considered tracking using box filter for sampling and these produce undesirable ringing in the frequency domain, which can lead to unstable tracking. Since we need a causal filter in the temporal domain, we employ an exponential weighting kernel. This filter has good frequency domain properties and it can be computed recursively making it exceedingly efficient.

Since we are computing a purely temporal signal for the amplitude computation, we use a Gaussian kernel to compute a weighted-mean intensity value at each sample location as a function of time. We have one such signal going backwards over time in each of these rectangular subwindows. This local signal is windowed with an exponentially decaying window (decaying backwards in time) to produce the windowed signal used for frequency analysis. Each such signal provides an amplitude spectrum that can be matched to a profile of a typical human gait. In principle matching these amplitude spectra to human gaits would be an ideal application for a statistical classifier trained on a large collection of human gait signals. In practice however, these human-associated sig-

nals appear to be easy to identify and an automated classifier is not currently being used. (In addition, the acquisition of sufficient training data is a substantial challenge.)

B. Using color cues

Like any feature-based detector or tracker, the spatio-temporal features used in our filter can sometimes provide false detections, either responding to multiple cues in the environment or simply to the wrong cue. This can be particularly true if the robot is swimming near the bottom of the ocean floor inhabited by coral reefs. Periodic motion of the robot (due to the robot's propulsion system, strong surge or current underwater) can also confuse the tracker by generating high responses in low-frequency components. Likewise when used in terrestrial applications there may be periodic structures in the environment (such as fences) or periodic motion as the robot moves (especially for a walking robot like ours). To address such issues, we combine the output of the Fourier tracker with supplementary tracking system for intensity targets, specifically a blob tracker [6] tuned to detect the color of the divers flippers. The blob tracker uses precomputed color threshold values to segment a portion of the image that falls within these thresholds. Regions that are common to both the color threshold tracker and the Fourier tracker are chosen as the diver's location. Conversely, the Fourier tracker can be used as a weighting function for the blob tracker for flipper tracking, by helping the blob tracker track the blob with the proper frequency characteristics.

IV. EXPERIMENTAL RESULTS

We applied the proposed algorithm on video sequences of divers swimming underwater in both open-water (ocean) and closed-water (pool) environments. Both types of video sequence are challenging due to the unconstrained motion of the target and the diver, and the poor imaging conditions (in the open-water footage). The success rate of the tracker was measured (successful versus failed tracking sequences) both with and without the aid of color cues for tracking flippers. Since the Fourier tracker looks backward in time every N frames to find the new location of the diver, the output of the computed locations are only available every N frames, unlike the color blob tracker which finds the location of color blobs matching the tuned parameters in every frame.

For the open-sea video footage, we have tracked a diver swimming in front of the robot for a duration of approximately 3 minutes, or 180 seconds, at a frame rate of 29 frames per second. The time window for the Fourier tracker for this experiment is 15 frames, corresponding to the 0.5 seconds of footage. Each frame has dimensions 720×480

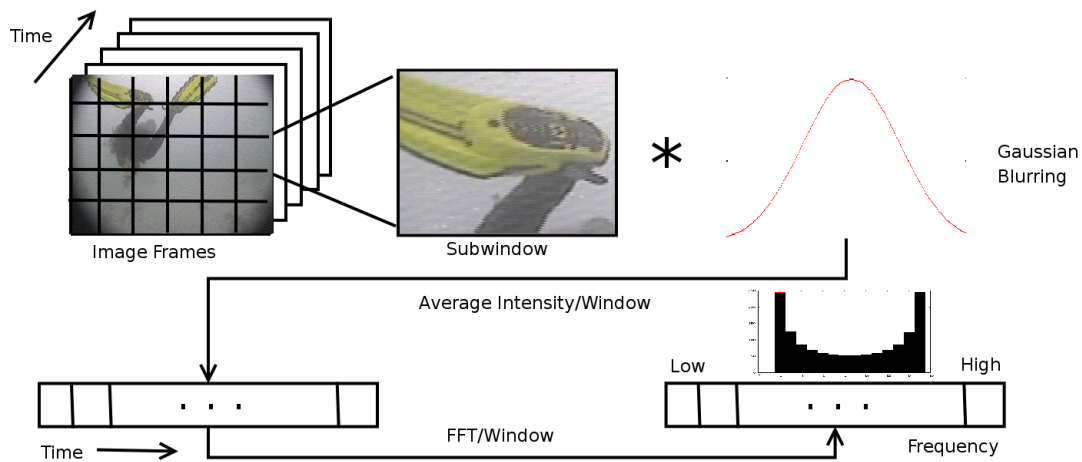


Fig. 2. Fourier tracking process outline.



Fig. 3. Flipper tracker tracking diver's flippers: Sequence 1. The circular mark on the heel of the left flipper is the target location determined by the system.

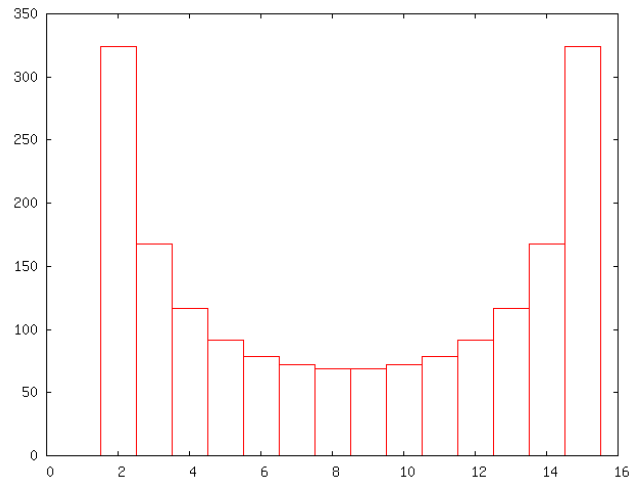


Fig. 4. FFT for the intensities in Fig. ?? Note the large low frequency response.

pixels and each rectangular subwindow is 180×120 pixels in size (one-fourth in each dimension). The subwindows overlap each other by half the width and the height.

Each subwindow is first blurred with a Gaussian having 0 mean and 0.5 variance to remove high frequency noise and artifacts. The average of the intensity values of each subwindow is calculated and saved in a time-indexed vector. At the end of frame 15, Fourier transforms are performed on each these vectors. The resulting FFT output with the maximum lowest frequency energy is chosen as the probable location of the diver.

Figure 3 shows output of the Fourier tracker on one frame of a swimming diver sequence. The output of the tracker

is shown by the red circular blob on the diver's flippers. The absolute values of the amplitude spectrum from this sequence is shown in Fig. 4. Observe that the FFT output is symmetric around the Nyquist frequency, and the DC component of the amplitude spectrum has been eliminated. The low frequency components of the amplitude spectrum exhibit very high energy as expected around the region of the oscillating flippers. Since the video was shot at approximately 30 frames per second, but sub-sampled at 15 frame per second, the lowest frequency components correspond roughly to the frequency of 1Hz. This matches exactly with the requirements for tracking flippers, and the FFT response

TABLE I
RESULTS OF STANDALONE FOURIER TRACKING

Total Frames	Tracker outputs	Successful Tracks	Error Rate
5400	360	288	20%

TABLE II
RESULTS OF FOURIER TRACKING WITH COLOR CUES

Total Frames	Tracker outputs	Successful Tracks	Error Rate
5400	360	352	2.2%

for the shown sequence consolidates that concept.



Fig. 5. Blob tracker output for the sequence in Fig. 3.

The result of using the blob tracker to aid the Fourier tracker narrows down the probable positions of the flippers. The output of the blob tracker (tuned to yellow) can be seen in the binary image of Fig. 5. The white segments of the image are the portions with color signature that falls within the thresholds of the flipper’s color. As can be seen, the blob tracker outputs two different blobs for both flippers, and one of them correspond exactly to the output from the Fourier tracker. The results of the standalone Fourier tracker and combining the Fourier tracker with the color tracker are shown in Tab.I and Tab.II.

V. DISCUSSION AND CONCLUSIONS

In this paper we propose a mechanism for tracking humans in applications where they are being followed by a robot and, as a result, their motion is largely along the viewing direction of the robot. This configuration is especially ill-suited to existing gait tracking mechanisms. In particular, we

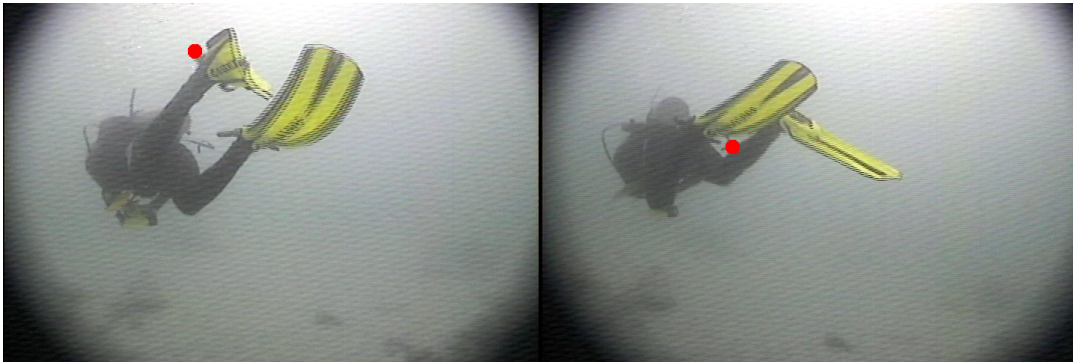
are interested in the specific problem of following a scuba diver with a mobile robot.

Our approach is to exploit the periodic motion of human locomotion, and specifically kicking during swimming, to identify a diver. This signal is extracted using a local space-time filter in the frequency (Fourier) domain and, as a result, we call it a Fourier-tracker. It allows a human diver to be detected and followed and, in practice, it is combined with a color blob tracker. By using a color/appearance signal in combination with the frequency-domain signal, we gain some robustness to moments when the diver may stop kicking or in some way disturb the periodic signal. Likewise, the Fourier-tracker makes the appearance signal much more robust to objects in the environment that might cause errors (which is a particular problem on a coral reef where very diverse color combinations can occur).

The work reported here suggests that this technique works well, although the experiments have been conducted on stored data and have not yet been tested in a full closed-loop tracking situation underwater (due to the serious complications and risks involved in a full experiment). Based on these experimental results the system is being deployed for real use now. For other applications it might be interesting to examine cases where the target has a larger lateral motion. It appears that this could be naturally implemented using a convolution-based filtering mechanism applied directly to the image, as has been used in visual motion computation, but the computational overhead of such a procedure might be greater than what we propose here (and would thus probably be outside the scope of what we can achieve on our test vehicle at present).

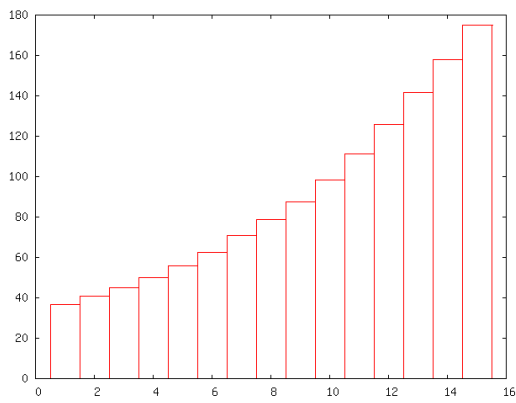
REFERENCES

- [1] G. Dudek, M. Jenkin, C. Prahacs, A. Hogue, J. Sattar, P. Giguère, A. German, H. Liu, S. Saunderson, A. Ripsman, S. Simhon, L. A. Torres-Mendez, E. Milios, P. Zhang, and I. Rekleitis, “A visually guided swimming robot,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, Alberta, Canada, August 2005.
- [2] W. T. Freeman and E. H. Adelson, “The design and use of steerable filters,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, no. 9, pp. 891–906, 1991.
- [3] S. A. Niyogi and E. H. Adelson, “Analyzing and recognizing walking figures in xyt,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994, pp. 469–474.
- [4] M. S. Nixon, T. N. Tan, and R. Chellappa, *Human Identification Based on Gait*, ser. The Kluwer International Series on Biometrics. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2005.
- [5] J. Sattar, P. Giguere, G. Dudek, and C. Prahacs, “A visual servoing system for an aquatic swimming robot,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, Alberta, Canada, August 2005.
- [6] J. Sattar and G. Dudek, “On the performance of color tracking algorithms for underwater robots under varying lighting and visibility,” Orlando, Florida, May 2006.

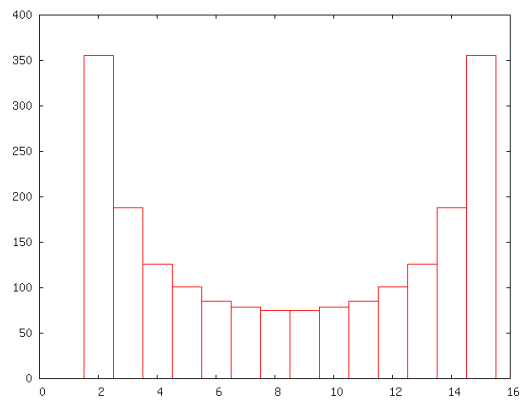


(a) Fourier tracker output, sequence 2

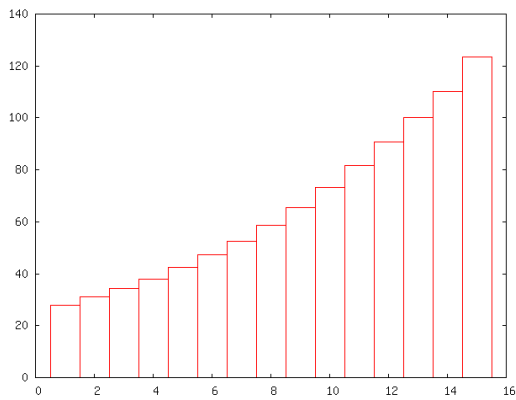
(b) Fourier tracker output, sequence 3



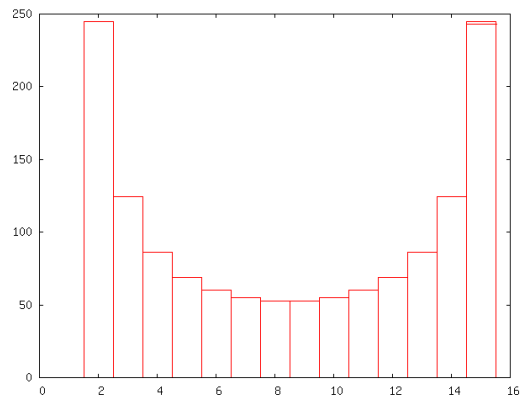
(c) Spatio-temporal intensity variations, sequence 2



(d) Spatio-temporal intensity variations, sequence 3



(e) FFT output, sequence 2



(f) FFT output, sequence 3

Fig. 6. Fourier tracker operations for two consecutive sequences with intensity and corresponding frequency responses.