arXiv preprint manuscript No.
(will be inserted by the editor)

Repetition Estimation

Tom F.H. Runia · Cees G.M. Snoek · Arnold W.M. Smeulders

Received: date / Accepted: date

Abstract Visual repetition is ubiquitous in our world. It appears in human activity (sports, cooking), animal behavior (a bee's waggle dance), natural phenomena (leaves in the wind) and in urban environments (flashing lights). Estimating visual repetition from realistic video is challenging as periodic motion is rarely perfectly *static* and *stationary*. To better deal with realistic video, we elevate the static and stationary assumptions often made by existing work. Our spatiotemporal filtering approach, established on the theory of periodic motion, effectively handles a wide variety of appearances and requires no learning. Starting from motion in 3D we derive three periodic motion types by decomposition of the motion field into its fundamental components. In addition, three temporal motion continuities emerge from the field's temporal dynamics. For the 2D perception of 3D motion we consider the viewpoint relative to the motion; what follows are 18 cases of recurrent motion perception. To estimate repetition under all circumstances, our theory implies constructing a mixture of differential motion maps: $\mathbf{F}$, $\nabla\mathbf{F}$, $\nabla\cdot\mathbf{F}$ and $\nabla\times\mathbf{F}$. We temporally convolve the motion maps with wavelet filters to estimate repetitive dynamics. Our method is able to spatially segment repetitive motion directly from the temporal filter responses densely computed over the motion maps. For experimental verification of our claims, we use our novel dataset for repetition estimation, better-reflecting reality with non-static and non-stationary repetitive motion. On the task of repetition counting, we obtain favorable results compared to a deep learning alternative.

Keywords Video Analysis, Motion, Periodicity, Repetition Counting, Wavelets Transform, Motion Segmentation

QUVA Deep Vision Lab, University of Amsterdam
Science Park 904, 1098XH Amsterdam, The Netherlands
Corresponding Author: Tom F.H. Runia, runia@uva.nl

1 Introduction

Visual repetitive motion is common in our everyday experience as it appears in sports, music-making, cooking and other daily activities. In natural scenes, it appears as leaves in the wind, waves in the sea or the drumming of a woodpecker, whereas our encounters of visual repetition in urban environments include blinking lights, the spinning of wind turbines or a waving pedestrian. In this work we reconsider the theory of periodic motion and propose a method for estimating repetition in real-world video.

Improving our ability to estimate repetition in realistic video is important in numerous aspects. In computer vision, periodic motion has proven to be useful for action classification (Goldenberg et al, 2005; Lu and Ferrier, 2004), action localization (Laptev et al, 2005; Sarel and Irani, 2005), human motion analysis (Albu et al, 2008; Ran et al, 2007), 3D reconstruction (Belongie and Wills, 2006), animal behavior study (Davis et al, 2000) and camera calibration (Huang et al, 2016). From a biological perspective, repetition is fascinating as the human visual system relies on rhythm and periodicity to approximate velocity, estimate progress and to trigger attention (Johansson, 1973).

To understand the origin and appearance of visual repetition we rethink the theory of periodic motion inspired by existing work (Pogalin et al, 2008; Davis et al, 2000). We follow a differential geometric approach, starting from the divergence, gradient and curl components of the 3D flow field. From the decomposition of the motion field and its temporal dynamics, we derive three *motion types* and three *motion continuities* to arrive at 3×3 fundamental cases of intrinsic periodicity in 3D. For the 2D perception of 3D intrinsic periodicity, the observer's viewpoint can be somewhere in the continuous range between two viewpoint extremes. Finally, we arrive at 18 fundamental cases for the 2D perception of 3D intrinsic periodic motion.

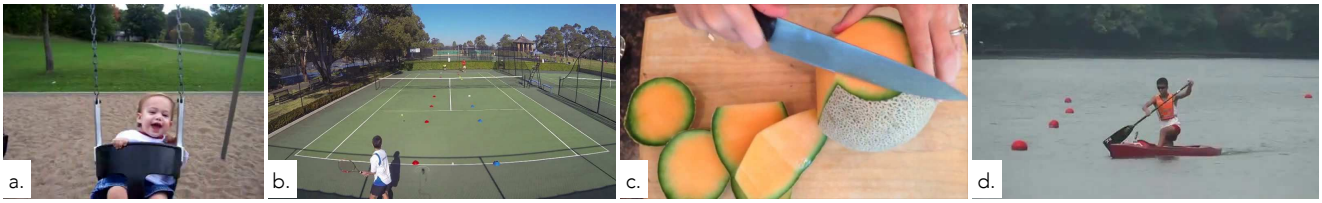


Fig. 1. Examples from our *QUVA Repetition* dataset, containing videos with repetitive motion such as sports, cooking, music making and other daily activities. The videos are challenging in their variety of appearance, non-stationary motion (*e.g.* accelerations or transient phenomena) and non-static appearance induced by camera motion or a changing motion appearance throughout the video. In this paper we focus on dealing with such challenges as they often appear in the real-world.

Estimating repetition in practice remains challenging. First and foremost, repetition appears in many forms due to its diversity motion types and motion continuity (Figure 1). Sources of variation in motion appearance include the action class, origin of motion and the observer's viewpoint. Moreover, the motion appearance is often *non-static* due to a moving camera or as the observed phenomena develops over time. In practice, repetitions are rarely perfectly periodic but rather are *non-stationarity*. Existing literature (Levy and Wolf, 2015; Pogalin et al, 2008) generally assumes static and stationary repetitive motion. As reality is more complex, we here address the challenges involved with non-static and non-stationary by proposing a novel method for estimating repetition in real-world video.

To deal with the diverse and possibly non-static motion appearance in realistic video, our theory implies representing the video with a mixture of first-order differential motion maps. For non-stationary temporal dynamics the fixed-period Fourier transform (Cutler and Davis, 2000; Pogalin et al, 2008) is not suitable. Instead, we handle complex temporal dynamics by decomposing the motion into a time-frequency distribution using the continuous wavelet transform. To increase robustness and to be able to handle camera motion, we combine the wavelet power of all motion representations. Finally, we alleviate the need for explicit tracking (Pogalin et al, 2008) or motion segmentation (Runia et al, 2018) by segmenting repetitive motion directly from the wavelet power. On the task of repetition counting, our method performs well on an existing video dataset and our novel *QUVA Repetition* dataset which emphasizes on more realistic video.

A preliminary version of this work appeared as (Runia et al, 2018). The current manuscript largely maintains the original theory while making significant improvements to the method for repetition estimation. Specifically, we simplify our approach by removing the need for explicit motion segmentation prior to repetition estimation. Instead, we obtain a foreground motion segmentation directly from the wavelet filter responses densely computed over the motion maps. As the most discriminative motion representation is not known *a priori*, our previous work employed a self-quality assessment to select the representation best measurable. However, selecting a single most discriminative representation is inherently unsuitable for handling significant variations due to camera motion or motion evolution over the course of the video. We improve this by combining the wavelet power of all representations for robustness and viewpoint invariance. Together the two improvements simplify our method while improving or giving comparable results on the task of repetition counting. More precisely, the contributions of our work are as follows:

- We rethink the theory of periodic motion to arrive at a classification of periodic motion. Starting from the 3D motion field induced by an object periodically moving through space, we decompose the motion into three elementary components: divergence, curl and shear. From the motion field decomposition and the field's temporal dynamics, we identify 9 fundamental cases of periodic motion in 3D. For the 2D perception of 3D periodic motion we consider the observer's viewpoint relative to the motion. Two viewpoint extremes are identified, from which 18 cases of 2D repetitive appearance emerge.
- Our spatiotemporal filtering method addresses the wide variety of repetitive appearances and effectively handles non-stationary motion. Specifically, diversity in motion appearance handled by representing video as six differential motion maps that emerge from the theory. To identify the repetitive dynamics in the possibly non-stationary video, we use the continuous wavelet transform to produce a time-frequency distribution densely over the video. Directly from the wavelet responses we localize the repetitive motion and determine the repetitive contents.
- Extending beyond the video dataset of Levy and Wolf (2015), we propose a new dataset for repetition estimation, that is more realistic and challenging in terms of non-static and non-stationary videos. To encourage further research on video repetition, we will make the dataset and source code available as download.

The paper proceeds as follows: in Section 2, we provide an overview of related work. Section 3 introduces new theory on periodic motion to arrive at a classification of fundamental motion types and their appearance in video. Our theoretical

insights are at the core of our method for repetition estimation, which is presented in Section 4. The experiments in Section 5 evaluate our method on two challenging video datasets. Section 6 summarizes and concludes the manuscript.

2 Related Work

2.1 Repetition Estimation

Existing approaches for repetition estimation in video typically represent video as one-dimensional signals that preserve the repetitive structure of the motion. Subsequently, frequency information is often extracted by Fourier analysis (Azy and Ahuja, 2008; Cutler and Davis, 2000; Pogalin et al, 2008; Tsai et al, 1994), peak detection (Thangali and Sclaroff, 2005), singular value decomposition (Chetverikov and Fazekas, 2006) or computational topology (Tralie and Perea, 2018). In general, the existing methods perform well under the assumptions of static and stationary videos.

The seminal work of Cutler and Davis (2000) uses normalized autocorrelation to obtain similarity matrices and proceeds by repetition estimation using Fourier analysis. Pogalin et al (2008) estimate the frequency of motion in video by tracking an object, performing principal component analysis over the tracked regions and also employing the Fourier-based periodogram. From the spectral decomposition, the dominant frequencies can be identified by peak detection and non-trivial separation of fundamental and harmonic frequencies. While Fourier-based methods provide a good estimate of strongly periodic motion, they are unsuitable nor intended to deal with more realistic non-stationary repetition, see the accelerating rower in Figure 2.

As strongly periodic motion has received serious attention, less effort has been devoted to non-stationary repetition in video. Briassouli and Ahuja (2007) use the Short-Time Fourier Transform for estimating the time-varying spectral components in video to distinguish multiple periodically moving objects. The filtering-based approach of Burghouts and Geusebroek (2006) uses a time-causal filter bank from Koenderink (1988) to detect quasi-periodic motion in video. Their method works online and shows good results when filter response frequencies are tuned correctly. In this work, we employ the continuous wavelet transform over multiple temporal scales to estimate repetition in complex video.

The deep learning method of Levy and Wolf (2015) is different from all other work but resembles our work in counting-based evaluation over a large video dataset. The general idea is to train a convolutional neural network for predicting the motion period in short video clips. As training data is not available, the network is optimized on synthetic video sequences in which moving squares exhibit periodic motion of four motion types from Pogalin et al (2008). At test time, the method takes a stack of video frames, performs explicit motion localization to obtain a region of interest and then classifies the motion period by forwarding the frame crops through the network. The system is evaluated on the task of repetition counting and shows near-perfect performance on their *YTSegments* dataset. The 100 videos are a good initial benchmark but as the majority of videos have a static viewpoint and exhibit stationary periodic motion, we propose a new dataset. Our dataset better reflects reality by including more non-static and non-stationary examples.

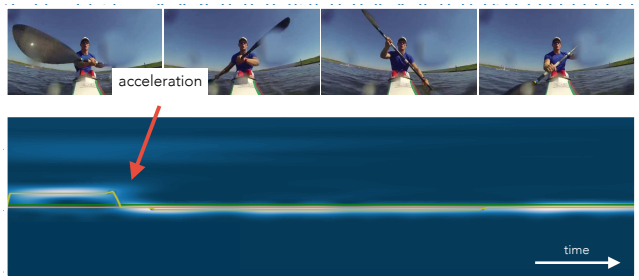


Fig. 2. Non-stationary motion often appears in real-world video. This example shows a rower accelerating as plotted in the time-frequency space. The vertical axis of the spectrum denotes the wavelet scale, inversely proportional to the frequency. The sudden acceleration appears as shift of the maximum power in time-frequency space. The Fourier transform is unable to handle such non-stationary video.

Increased video complexity in terms of motion appearance, scene complexity and camera motion demands intricate spatiotemporal localization of salient motion. While many methods for periodic motion analysis incorporate some form of tracking or motion segmentation (Polana and Nelson, 1997; Pogalin et al, 2008; Levy and Wolf, 2015), few approaches specifically address the challenge of repetitive motion segmentation. Goldenberg et al (2005) estimate the repetitive foreground motion to leverages its center-of-mass trajectory for classifying human behavior. More closely related is the work of Lindeberg (2017) in which scale selection over space and time leads to an effective temporal scale map. Inspired by this, we perform spatial segmentation of repetitive motion directly from the spectral power maps obtained through the continuous wavelet transform. This is appealing, as it connects localization to the temporal dynamics rather than relying on decoupled localization by state-of-the-art motion segmentation, *e.g.* Tokmakov et al (2017).

Instead of considering repetition as their primary goal, various works leverage the presence of periodic motion for auxiliary tasks. Belongie and Wills (2006) exploit periodic human motion for 3D reconstruction of a scene, whereas Laptev et al (2005) uses sequence alignment of periodic motion for stereo-camera correspondence. From a practical point of view, the presence of periodic motion also serves as cue for action classification (Lu and Ferrier, 2004; Goldenberg et al, 2005) and supports camera calibration (Huang et al, 2016).

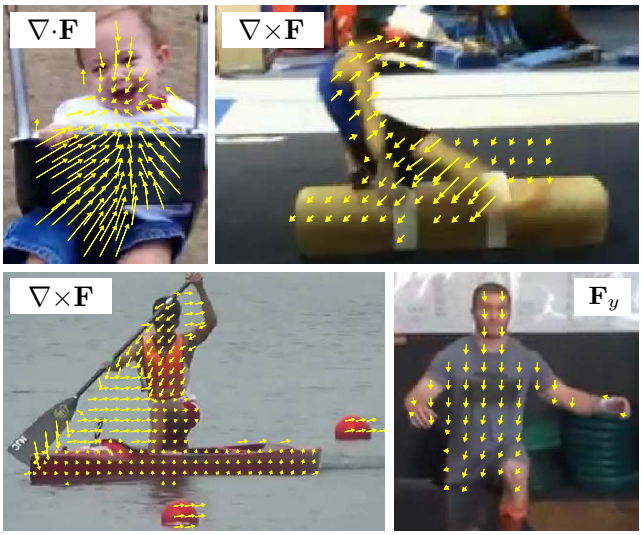


Fig. 3. There is great diversity in appearance of repetitive motion. We decompose the motion field into its fundamental components. Here we visualize the motion fields as optical flow arrows over the foreground motion with the visually dominant motion field component indicated in the white box.

2.2 Categorization of Motion Types

In real-world video, periodic motion emerges in a wide variety of appearances (see Figure 3). We reconsider the theory of periodic motion by proposing a classification of fundamental periodic motion types starting from the 3D motion field tied to a moving object. Using first-order differential analysis, we decompose the motion field into its primitive components. The work of Koenderink and van Doorn (1975) delivered inspiration for our theoretical derivation of repetitive motion types from the flow field. Similar to the Helmholtz-Hodge decomposition (Abraham et al, 1988) into the eigenvalues of the flow field's Jacobian matrix, it finds use in flow field topology for fluid dynamics and electrodynamics. Although our work is similar in differential decomposition of the motion field, we use it to reach a novel classification of periodic motion patterns. We use the insights for establishing our repetition estimation method.

In the context of periodic motion, Davis et al (2000) and Pogalin et al (2008) both propose a categorization of motion patterns. Davis et al (2000) consider a simple sinusoidal model to characterize periodic motion and link each type to animal behavior. In terms of periodic motion categorization, our work bears resemblance to Pogalin et al (2008). The authors identify four visually periodic motion types (translation, rotation, deformation and intensity variation) complemented with three cases of motion continuity (oscillating, constant and intermittent) in the field of view. We take a more principled approach starting from the 3D motion field. Specifically, we show that fundamental periodic motion types emerge from the decomposition of the flow field and the motion direction over time. Moreover, the projection of 3D periodicity on on a 2D image has to take into account the continuous nature of the viewpoint which we address explicitly in theory and experiments.

Although not directly related to our work, first-order differential geometric motion representations have been used extensively as spatiotemporal video descriptors. Klaser et al (2008) proposes a spatial multi-scale motion descriptor based on first-order differential motion and uses integral videos for efficient computation. Along similar lines, MoSIFT (Chen and Hauptmann, 2009) uses spatial interest points and enforces sufficient temporal dynamics to eliminate candidate points. In terms of motion descriptors, our work bears resemblance to the Divergence-Curl-Shear descriptor proposed by Jain et al (2013). Their favorable action classification results associated with the differential-based descriptor support our findings for periodic motion estimation.

3 Repetitive Motion

Visual repetition is defined as a reoccurring pattern over space or time in the 3D world. In this work, we focus on temporally repetitive motion rather than spatially repetitive patterns such as a texture. Consequently, the 3D motion field induced by a moving object is the right starting point for our theoretical analysis.

Let a moving object and observer be positioned in a 3D world specified by the Cartesian coordinates $\mathbf{x} = (x_1, x_2, x_3)$ at time t. Formally, intrinsic periodic motion is defined as the reappearance of the same 3D-flow $\boldsymbol{\mathcal{F}}(\mathbf{x}, t)$ induced by the motion of an object over time.

$$\boldsymbol{\mathcal{F}}(\mathbf{x}, t) = \boldsymbol{\mathcal{F}}(\mathbf{x} + \mathbf{S}, t + T). \tag{1}$$

The parameter T denotes the period over time and $\mathbf{S}$ corresponds to a period over space. We initially exclude the trivial case of a constant flow field inducing periodic appearance due to a reappearing texture on the object's surface. Starting from the motion field, we follow a differential approach to decompose the field into its elementary components. In the end we arrive at nine fundamental cases of intrinsic periodic motion in 3D.

3.1 Motion Field Decomposition

In 3D Cartesian space, the gradient of the flow $\boldsymbol{\nabla}\boldsymbol{\mathcal{F}}(\mathbf{x}, t)$ is described by the Jacobian matrix $\mathbf{J} \in \mathbb{R}^{3\times3}$ containing all first-order partial derivatives of the vector field:

$$(\boldsymbol{\nabla}\boldsymbol{\mathcal{F}})_{ij} = \frac{\partial \mathcal{F}_i}{\partial x_j}, \tag{2}$$

where i and j are dimension indices and we omit the position $\mathbf{x}$ and time t for brevity. From the first-order partial derivatives contained in the Jacobian, three fundamental components of the motion field can be recognized (Abraham et al,

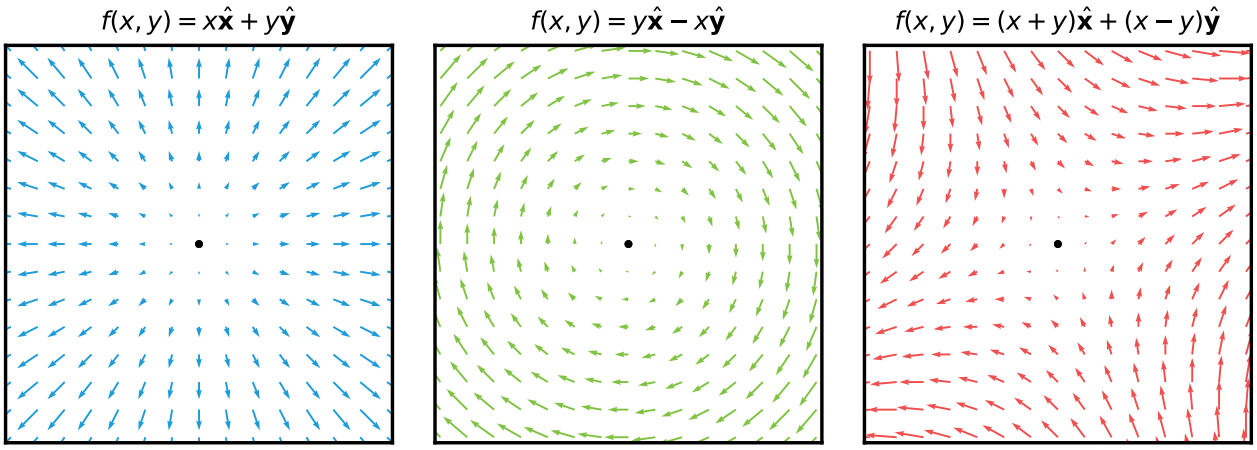


Fig. 4. Three 2D flow fields with fundamentally different characteristics that emerge from the decomposition of the motion Jacobian **J**. *Left*: Pure divergent flow field with outward flux often associated with expansion or depth perception. *Center*: Pure rotational flow field also referred to as vorticity or curl. *Right*: Flow field with a pure shear component related to the deformation tensor. The shear component is divergence- and curl-free as the opposing terms cancel out. In real-world video, shear is generally negligible compared to divergence and curl components.

1988). Specifically, the Jacobian **J** can be decomposed into a sum of a diagonal part **D**, a symmetric part **E** and an anti-symmetric part **R** such that:

$$\nabla\mathcal{F} = \mathbf{D} + \mathbf{R} + \mathbf{E}. \tag{3}$$

This is similar to the Helmholtz-Hodge vector field decomposition well-known from fluid dynamics, which distinguishes divergence-free and curl-free components of a motion field. The diagonal part of the Jacobian **J** is given by:

$$\mathbf{D} = \text{diag}\left(\frac{\partial\mathcal{F}_1}{\partial x_1}, \frac{\partial\mathcal{F}_2}{\partial x_2}, \frac{\partial\mathcal{F}_3}{\partial x_3}\right). \tag{4}$$

The trace of this matrix defines the *divergence* of the field:

$$\nabla \cdot \mathcal{F} = \text{trace}\,(\mathbf{D}). \tag{5}$$

The divergence is a scalar field representing the amount of outward flux from an infinitesimal volume around a given point. Next, the anti-symmetric part **R**, referred to as the spin- or rotation matrix, is given by $\mathbf{R} = \frac{1}{2}(\mathbf{J} - \mathbf{J}^T)$ with elements:

$$\mathbf{R}_{ij} = \frac{1}{2}\left(\frac{\partial\mathcal{F}_i}{\partial x_j} - \frac{\partial\mathcal{F}_j}{\partial x_i}\right). \tag{6}$$

From the elements of the spin matrix we can recognize the *curl* of the flow field. More specifically, the curl of the 3D flow field is defined as:

$$\nabla \times \mathcal{F} = \left[\frac{\partial\mathcal{F}_3}{\partial x_2} - \frac{\partial\mathcal{F}_2}{\partial x_3}, \frac{\partial\mathcal{F}_1}{\partial x_3} - \frac{\partial\mathcal{F}_3}{\partial x_1}, \frac{\partial\mathcal{F}_2}{\partial x_1} - \frac{\partial\mathcal{F}_1}{\partial x_2}\right]^T. \tag{7}$$

This vector field describes the infinitesimal rotation around a given point. Finally, the last fundamental component is given by the symmetric part $\mathbf{E} = \frac{1}{2}(\mathbf{J} + \mathbf{J}^T)$ with elements:

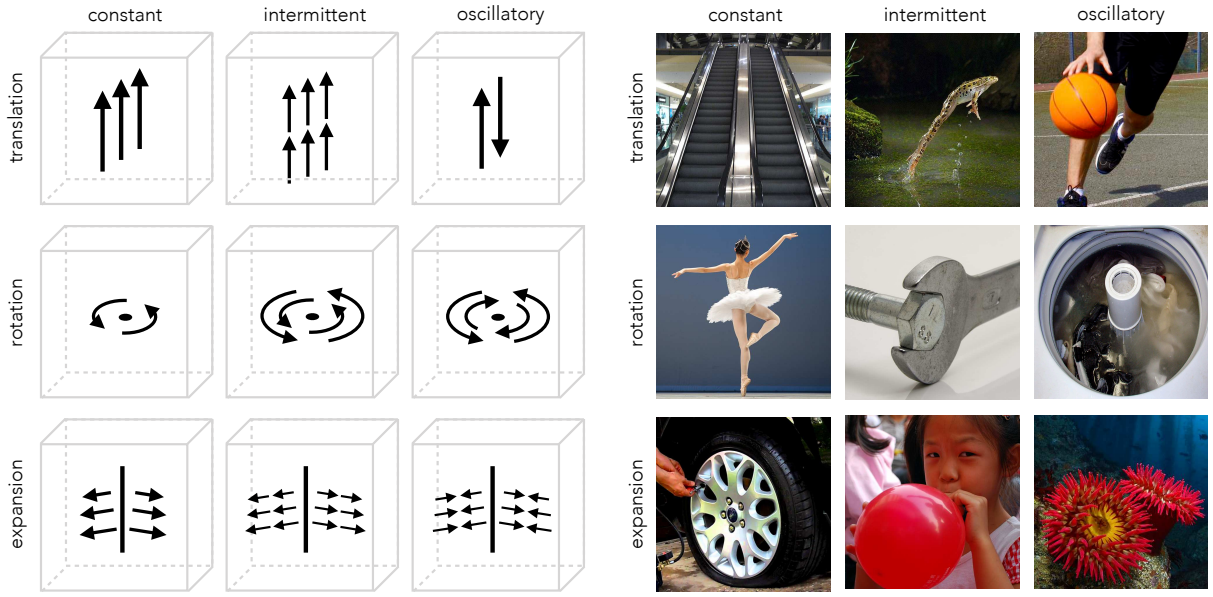
$$\mathbf{E}_{ij} = \frac{1}{2}\left(\frac{\partial\mathcal{F}_i}{\partial x_j} + \frac{\partial\mathcal{F}_j}{\partial x_i}\right). \tag{8}$$

This trace-free matrix is known as the deformation tensor and associated with the *shear* of the flow field. In Figure 4 we illustrate three motion fields with either pure divergent, rotational or shear flow.

3.2 Intrinsic Periodic Motion in 3D

3.2.1 Motion Types

For an object moving periodically through the 3D space, the decomposition of the flow field tied to the object is used to characterize the type of motion. A non-rigid object that is expanding or contracting along one or more axes will produce a purely divergent flow field $\nabla \cdot \mathcal{F}$. Examples include: inflating a balloon or a pulsing anemone. Moreover, a flow field exclusively containing curl $\nabla \times \mathbf{F}$ emerges with rotational motion such as a spinning wheel or tightening a bolt. Finally, shear is associated with deformation or stress on a surface caused by opposing forces parallel to the cross-section of a body. Shear predominantly plays a role for materials with high-elasticity (*e.g.* fluids) or in the presence of large forces (*e.g.* solid mechanics). Generally, the 3D motion field's shear component is negligible as excessive forces are required to deform the material. For softer materials such as foam, paper or plastics, the shear components can be measurable but this is rare in practice. Based on its rare appearance, we therefore leave the shear for what it is.



(a) Flow Abstractions in 3D **(b)** Examples in Real Life

Fig. 5. 3×3 Cartesian table of the *motion type* times the *motion continuity*. These are the basic cases of periodicity in 3D emerging from the motion field decomposition and the temporal dynamics. The examples are: escalator, leaping frog, bouncing ball, pirouette, tightening a bolt, laundry machine, inflating a tire with repetitive texture, inflating a balloon and a breathing anemone.

In particular, three basic 3D motion types emerge depending on the values of divergence and curl as follows:

$$\begin{aligned} \textit{translation}: &\quad \nabla \times \mathcal{F}(\mathbf{x}, t) = \mathbf{0}, \quad \nabla \cdot \mathcal{F}(\mathbf{x}, t) = 0 \\ \textit{rotation}: &\quad \nabla \times \mathcal{F}(\mathbf{x}, t) \neq \mathbf{0}, \quad \nabla \cdot \mathcal{F}(\mathbf{x}, t) = 0 \\ \textit{expansion}: &\quad \nabla \times \mathcal{F}(\mathbf{x}, t) = \mathbf{0}, \quad \nabla \cdot \mathcal{F}(\mathbf{x}, t) \neq 0. \end{aligned}$$

These motion types are tied to a particular 3D motion field of pure form. In practice there may be a mixture types. As we are aiming to handle realistic video, our method employs a combination of first-order differential motion maps from which the dominant 3D periodicity in the object's motion is determined.

3.2.2 Motion Continuities

As periodic motion by its nature contains a temporal component, we here transition to the temporal dynamics of the time-varying motion field. Consecutive measurements of the flow field $\mathcal{F}(\mathbf{x}, t)$ produce a time-varying motion field with particular temporal dynamics. Depending on the type of motion, the motion field needs to satisfy one of the following necessary periodic conditions:

$$\begin{aligned} \nabla \mathcal{F}(\mathbf{x}, t) &= \nabla \mathcal{F}(\mathbf{x} + \epsilon, t + T) \\ \nabla \times \mathcal{F}(\mathbf{x}, t) &= \nabla \times \mathcal{F}(\mathbf{x} + \epsilon, t + T) \\ \nabla \cdot \mathcal{F}(\mathbf{x}, t) &= \nabla \cdot \mathcal{F}(\mathbf{x} + \epsilon, t + T), \end{aligned} \tag{9}$$

where ϵ denotes a translation as the object's periodicity may be superposed on translation. For robustness, our method measures both $\mathcal{F}(\mathbf{x}, t)$ and $\nabla \mathcal{F}(\mathbf{x}, t)$. From the direction and temporal dynamics of motion, three distinct periodic motion continuities can be distinguished: *constant*, *intermittent* and *oscillating* periodicity. In practice the motion continuity may be a mixture between types. For intermittent and oscillating motion repetitive nature is intrinsically in the temporal dynamics whereas for constant motion to appear repetitively, there will be special conditions on the object's texture or albedo.

3.2.3 Categorization of Periodic Motion

The intrinsic periodicity in 3D does not cover all perceived recurrence in an image sequence. For the trivial cases of constant translation and constant expansion in 3D, the perceived recurrence will appear when a repetitive chain of objects (conveyor) or a repetitive appearance (texture on a car tire) on the object is aligned with the motion. In such cases, the recurrence will also be observed in the field of view. For constant rotation, the restriction is that the appearance cannot be constant over the surface, as no motion, let alone recurrent motion would be observed. In the rotational case, any rotational symmetry in appearance will induce a higher order recurrence as a multiplication of the symmetry and the rotational speed.

For the purpose of periodic motion, nine cases organize in a 3×3 Cartesian table of basic *motion type* times *motion continuity*, see Figure 5a. The corresponding examples of these nine cases are given in Figure 5b. This is the list of fundamental cases, where a mixture of types is permitted.

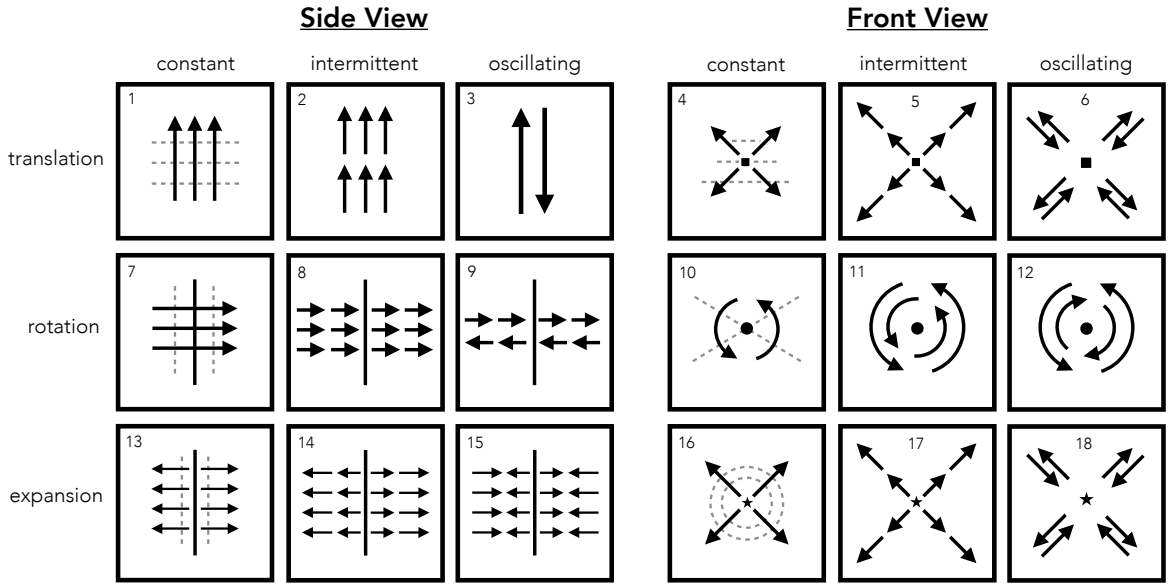


Fig. 6. Observed flow: the 18 fundamental cases for 2D perception of 3D recurrence. The perception follows from the motion pattern (3×), motion continuity (3×) and the viewpoint on the continuous interval between the two extremes: side and front view. ↑ denotes flow direction, ■ denotes a vanishing point, • denotes a rotation point, ★ denotes expansion point. Dashed grey lines for constant motion indicate the need for texture to perceive recurrence. Pairs 4-16, 5-17 and 6-18 appear similar at first sight but vary in their signal profile.

In practice, some cases are ubiquitous, while for others it is hard to find examples at all.

3.3 Visual Recurrence in 2D

So far we have considered the intrinsic periodicity in 3D. We reserve the term *recurrent* for the 2D observation of the 3D periodicity. Recurrence in the field of view is defined by:

$$\mathbf{F}(\mathbf{x}', t) = \mathbf{F}(\mathbf{x}' + \epsilon', t + T) \tag{10}$$

where $\mathbf{F}(\mathbf{x}', t)$ is the perceived flow in 2D image coordinates $\mathbf{x}'$. The observed displacement is denoted by ϵ' and T is the temporal period. *The underlying principle is that the same period length T will be observed in both 3D and 2D for all cases of periodicity.* This permits us to measure 3D motion periodicity T from the 2D flow field. Only in some rare cases, the period of motion may change due to a partial or complete occlusion; or the periodic motion disappears entirely due to lack of texture or albedo from a given viewpoint (*e.g.* a constantly rotating textureless disk). These are exceptional cases as the general principle applies that the temporal period is viewpoint invariant.

The camera position relative to the object's motion has a large influence on the perception of the flow field. There are two fundamentally different viewpoints: the *frontal* view and the *side* view:

frontal view: on the main axis of motion
side view: perpendicular to the main axis of motion.

For translation, there is one main axis and two perpendicular axes, which are both identical for our purpose. There is no distinction between the two perpendicular views as their perception is equivalent. Similarly, for rotation, the two perpendicular cases are also indistinguishable. For expansion there are one, two or three axes of expansion, again leaving us with the frontal case and the perpendicular case as the two fundamental cases. Consequently, for all cases considered, a distinction between frontal view and side view is sufficient. As a result, the perceived recurrence is defined on the continuous range between the two extreme viewpoints. Combining the two viewpoint extremes with the nine cases of periodic motion we arrive at the classification of 18 basic cases as illustrated in Figure 6. The two views are the end of a continuous range of viewpoints. Most of the time an actual viewpoint will be somewhere in between the frontal view and the side view. This leaves the flow field asymmetrical or skewed, either in gradient, curl or divergence. As long as T can be measured from the observed signal, the skewed or asymmetric observation will not affect the recurrent nature nor the period of the 3D motion field.

3.4 Non-Static Repetition

Relative motion between the moving object and the observer adds another dimension of complexity. In particular with recurrent motion (1) the camera may move because the camera is mounted on the moving object itself, or (2) the camera is following the target of interest, or (3) the camera is in motion independent of the motion of the object. For the first two cases, the camera motion reflects the periodic dy-

(a) Frontal view **(b)** In-between view **(c)** Side view

Fig. 7. Example video displaying *girl on a swing* captured from three distinct viewpoints. Moving from one end of the continuous viewpoint spectrum (frontal) to the other (side) results in a dramatic change of motion appearance. The in-between viewpoint leaves the motion measurements either skewed or asymmetrical. In practice, we combine the motion representations to emphasize the one best measurable.

namics of the object's motion. The flow field may be outside the object, but otherwise it displays a complementary pattern in the flow field.

In the first case, the periodically moving camera will produce a global repetitive flow field as opposed to local repetitive flow when the object itself is moving. The third case particularly demands the removal of the camera motion prior to the repetitive motion analysis. In practice, this situation occurs frequently. Therefore, particular attention needs to be paid to camera motion independent of the target's motion. When the viewpoint changes from frontal to side view due to camera motion, the analysis will be inevitably hard. Figure 6 illustrates the dramatic changes in the flow field when the camera changes from one extreme viewpoint (side) to the other (frontal), or vice versa. Our method handles such appearance changes by simultaneously using multiple motion representations and summing temporal filter responses.

In addition, even when object motion and camera are both static, for none of the intrinsic motion types (translation, rotation, expansion), a point on the object will be at the same position in the camera field all the time. Under the double static condition, a point will just return to the same point on the camera field. As the intermediate points on the object or background have an arbitrary albedo and radiate an arbitrary luminance, it will not produce a sinusoidal signal in general. This is noteworthy as previous work (Cutler and Davis, 2000; Liu and Picard, 1998; Pogalin et al, 2008) implicitly assume such a signal by considering the Fourier transform or variants.

3.5 Non-Stationary Repetition

A recurrent signal is said to be stationary when the period length is constant over time. In the initial steps of periodicity analysis, it was assumed the periodic signal was near-stationary. However, decay in frequency or acceleration are common in realistic video. In practice, we have observed that non-stationary is often present, to which we return later with the discussion of our dataset. Therefore, in contrast to Pogalin et al (2008) and Levy and Wolf (2015) we loosen the stationarity assumption, leaving the option of acceleration open. More precisely our method employs the continuous wavelet transform for spectral decomposition of the video.

4 Method

In this section we present our method for estimating repetition in video. The method takes as input a sequence of RGB frames and outputs a frequency distribution densely computed over space and time. Subsequently, the spectral power distribution, which we obtain from the continuous wavelet transform, is used for repetition counting, motion segmentation or other frequency-based measurements. We target the general case in which moving objects may exhibit non-stationary periodicity or have a non-static appearance due to camera motion or repetition superposed on translation. Our method, summarized in Figure 8, comprises motion estimation and two consecutive filtering steps: first we spatially filter the motion fields to arrive at first-order differential geometric motion maps, and then we determine the video's repetitive contents by applying the continuous wavelet transform densely over the motion maps. Task-dependent post-processing steps may give the desired output; here we focus on repetition counting as it enables straightforward evaluation of our method in the presence of non-stationary repetitions.

4.1 Differential Geometric Motion Maps

Given a sequence of video frames, we first estimate the motion between pairs of consecutive frames to obtain the motion field $\mathbf{F}(\mathbf{x}', t) = (F_x, F_y)$ for all timesteps. Next, the theory implies decomposition of the motion field into the primitive first-order differentials. For a moment in time t, we compute the differential motion maps by spatially convolving the flow field with first-order Gaussian derivative filters:

$$G^x(\mathbf{x}'; \sigma) = -\frac{x}{2\pi\sigma^4} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{11}$$

$$G^y(\mathbf{x}'; \sigma) = -\frac{y}{2\pi\sigma^4} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), \tag{12}$$

where σ denotes the spatial scale parameter and image coordinates are given by $\mathbf{x}' = (x, y)$. Through convolution with Gaussian kernels we obtain the first-order spatial derivatives $\nabla_x F_x, \nabla_y F_x, \nabla_x F_y$ and $\nabla_y F_y$ for a moment in time. Given the spatial partial derivatives of the motion, we compute $\boldsymbol{\nabla} \cdot \mathbf{F}$ and $\boldsymbol{\nabla} \times \mathbf{F}$ using the 2D equivalents of Eq. (5) and Eq. (7). For the 2D case, curl is a single-component vector field perpendicular to the image plane whereas the divergence is a scalar field. To effectively handle all cases of repetitive motion (Figure 6), we compute six motion maps for each frame:

$$\left\{\boldsymbol{\nabla} \cdot \mathbf{F}, \boldsymbol{\nabla} \times \mathbf{F}, \quad \nabla_x F_x, \nabla_y F_y, \quad F_x, F_y\right\} \tag{13}$$

Periodicity in $\boldsymbol{\nabla} \cdot \mathbf{F}$ or $\boldsymbol{\nabla} \times \mathbf{F}$ will only occur for the frontal view. For oscillatory or intermittent motion from the side

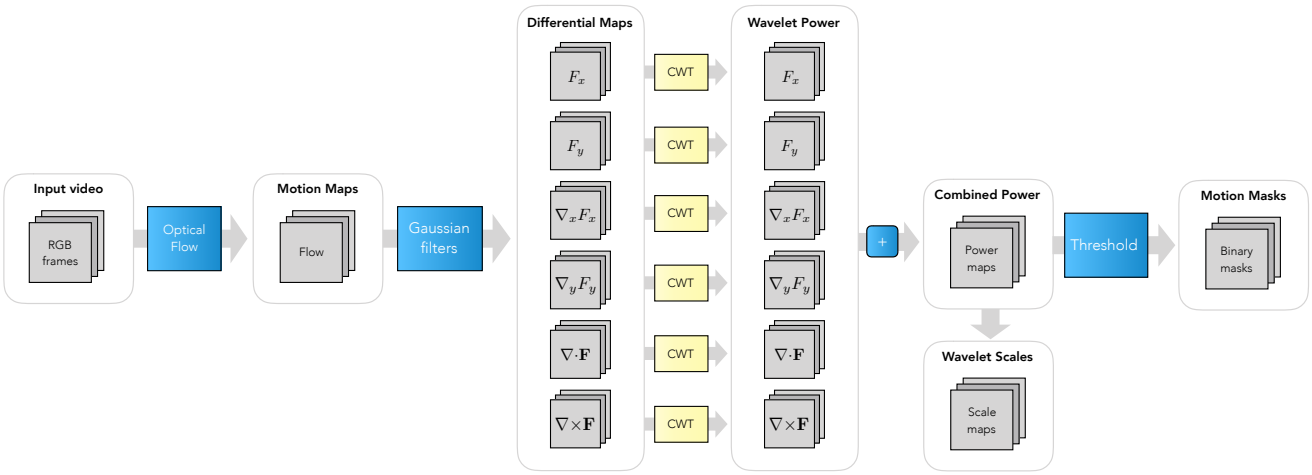


Fig. 8. Overview of our method for repetition estimation in video. Given an input video as RGB frames we first estimate the motion between consecutive frames using optical flow. We perform spatial Gaussian filtering to obtain six (differential) motion representations. Next, we apply the continuous wavelet transform (CWT) through temporal convolution over all six representations individually. We combine all power maps by summation to arrive at a single power map for a moment in time. Finally, we spatially segment repetitive motion by mean-thresholding of the power maps. To estimate repetition, we median-pool the wavelet scales over the motion segmentation producing an instantaneous frequency measurement.

view, $\nabla_x F_x$ and $\nabla_y F_y$ will produce the strongest periodicity while the zeroth-order flow field F_x and F_y will deliver a stronger response for the cases of repetitive periodic appearances at constant motion.

Figure 9 displays an example frame with four of six motion maps (the two are omitted here). The six motion maps represent the video for each moment in time and address the diversity in repetitive motion. In our experiments, we will evaluate the individual and joint representative power associated with the motion maps. *A priori* it is unknown which motion we are dealing with, to which we return later by combining the temporal responses of all motion maps.

4.2 Dense Temporal Filtering

So far we have only considered spatial filtering to obtain the motion maps for a moment in time. Here we include time and proceed by temporal filtering of the motion maps to estimate the video's repetitive motion. This is where the current method diverges from our previous work. In (Runia et al, 2018), we relied on the same motion maps but performed max-pooling over the foreground motion segmentation obtained separately from Papazoglou and Ferrari (2013). The max-pooled values over time construct a one-dimensional signal acting as a surrogate for the dynamics in a particular motion map. Spectral decomposition for each of the signals led to six (possibly contrasting) time-frequency estimates. To select the most discriminative representation, we employed a self-quality assessment based on the spectral power in the signals.

We found two problems with this approach: (1) the decoupled motion segmentation may not be optimal for estimating repetitive motion dynamics, and (2) max-pooling over the foreground motion mask discards most information and is unable to deal with multiple moving parts. We here address these problems by dense temporal filtering over all locations in the motion map instead of operating on the max-pooled signals. Spatially dense estimation of the local spectral power enables us to localize regions likely containing repetitive motion. The temporal filtering can be implemented in several ways, for example, as Fourier transform through temporal convolution. To handle non-stationary video dynamics, we perform the continuous wavelet transform by convolution to obtain a time-varying spectral decomposition.

4.3 Continuous Wavelet Transform

Given a discrete signal h_n for timesteps $n = 1, \ldots, N-1$ sampled at equally spaced intervals δt. Let $\psi_0(\eta)$ be some admissible wavelet function, depending on the non-dimensional time parameter η. The continuous wavelet transform (Grossmann and Morlet, 1984) is defined as the convolution of h_n with a "daughter" wavelet generated by scaling and translating the wavelet function $\psi_0(\eta)$:

$$W_n(s) = \sum_{n'=0}^{N-1} h_{n'} \psi^* \left[\frac{(n'-n)\delta t}{s} \right], \tag{14}$$

where the asterisk represents the complex conjugate. By varying time parameter n and the scale parameter s, the wavelet transform generates a time-scale representation describing how the amplitude of the signal changes with time and scale. We use the Morlet wavelet, a complex exponential carrier modulated by a Gaussian envelope:

$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{\eta^2/2}. \tag{15}$$

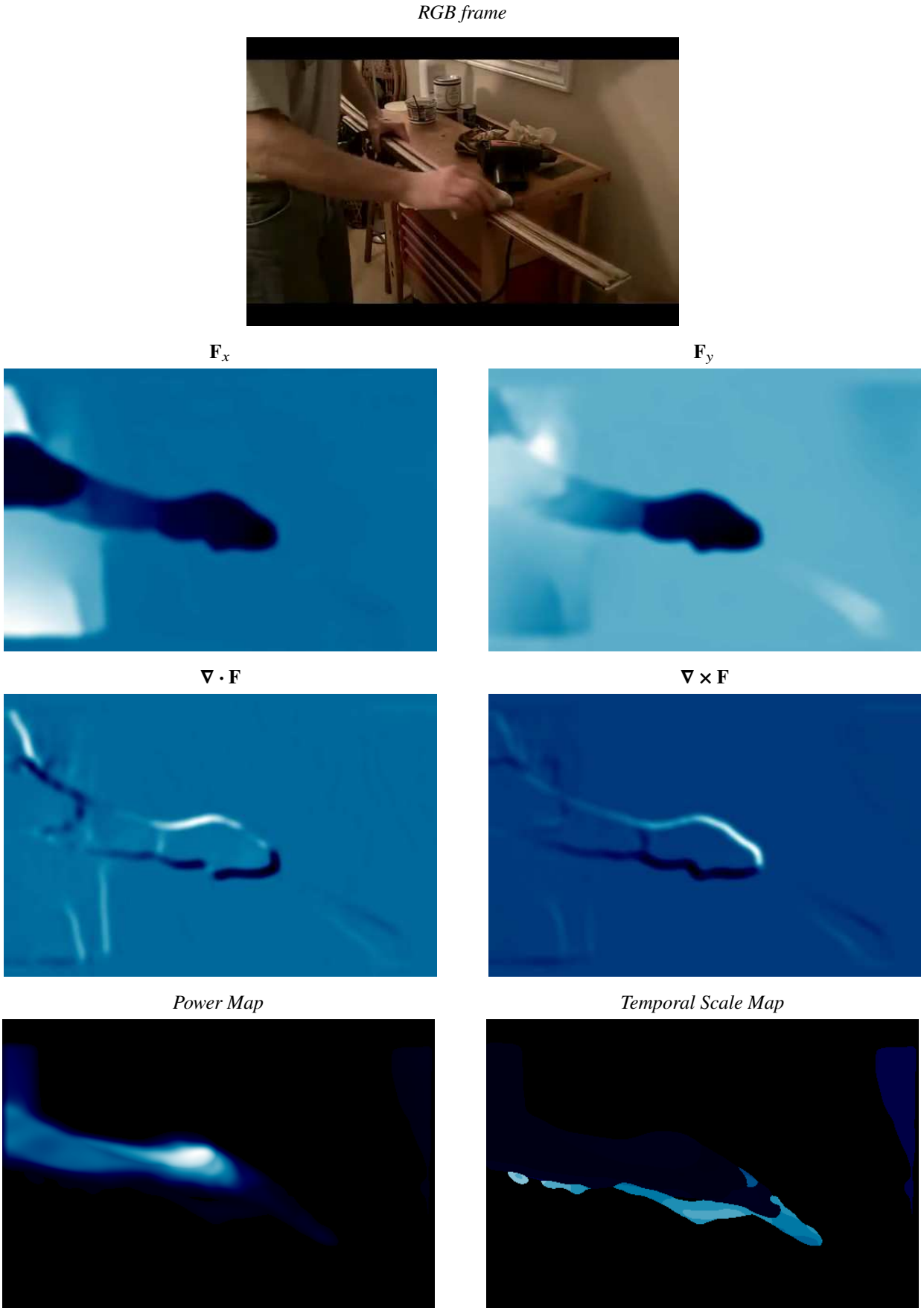


Fig. 9. Intermediate motion maps for a video displaying a *man brushing wood* from the *QUVA Repetition* dataset displaying a brushing motion. We perform wavelet filtering over six motion maps, due to space constraints only four are shown while $\nabla_x \mathbf{F}_x$ and $\nabla_y \mathbf{F}_y$ are omitted. Notice how the regions with repetitive motion appear in the wavelet power maps. By thresholding the wavelet power map with the mean power we obtain a repetitive motion map. The temporal scale maps indicate spatial regions with motion of low- and high-frequency.

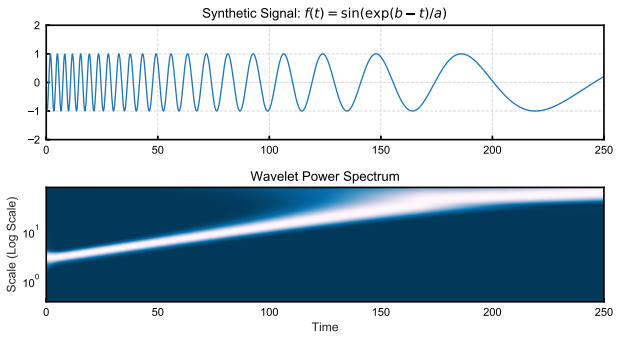


Fig. 10. Exponential chirp signal and the corresponding scalogram obtained from the continuous wavelet transform. Note increasing scale (period) in the scalogram as the signal's frequency decreases.

In all our experiments we set $\omega_0 = 6$ as it provides a good balance between time and frequency localization. Since the Morlet wavelet is complex, the wavelet transform $W_n(s)$ is also complex. Therefore, it is useful to define the wavelet power spectrum or *scalogram* as $|W_n(s)|^2$ representing the time-frequency localized energy. Figure 10 gives a non-stationary signal example and plots its wavelet power. It is clear that the scalogram is effective in revealing the signal's non-stationary repetitive dynamics.

The resolution of the scalogram $|W_n(s)|^2$ is defined by the distribution of scale parameter s. In practice, we use a discrete scale set that is logarithmically distributed:

$$s_j = s_0 2^{j\delta j}, \quad j = 0, 1, \ldots, J \tag{16}$$

$$J = \delta j^{-1} \log_2 (N\delta t / s_0) . \tag{17}$$

The smallest measurable scale s_0 and the number of scales J determines the range of the detectable frequencies. The smallest scale should be chosen such that the Fourier period of the wavelet is approximately $2\delta t$.

For a moment in time, the scalogram's maximum power will give the wavelet scale s producing the strongest filter response. Often the temporal frequency associated with the scale s will be a more convenient measurement. Therefore, the wavelet scale can be converted to a temporal frequency. For a Morlet wavelet, the relationship between scale and wavelength is given by (Torrence and Compo, 1998):

$$\lambda = \frac{4\pi}{\omega_0 + \sqrt{2 + \omega^2}}, \tag{18}$$

where ω_0 corresponds to the non-dimensional frequency. For $\omega_0 = 6$ corresponds to $\lambda = 1.03s$ for the Morlet wavelet, thus having the attractive property of wavelet scale being almost identical to the wavelength. We use (18) to obtain the frequency estimate for each time t and location $\mathbf{x}'$.

4.4 Combining Spectral Power Maps

We compute the time-localized frequency estimates by temporal convolution densely over the six individual motion representations. For each representation this produces a time-varying maximum *power map* and *scale map*. The power map contains the spatial distribution of maximum wavelet power over all temporal scales; the scale map holds the temporal scales corresponding to the wavelets with maximum power. What remains is combining the wavelet responses from all motion representations.

Rather than selecting the single most discriminative representation (Runia et al, 2018), we combine the spectral power maps by summation on a per-frame basis. To illustrate this, we visualize four (out of six) individual power maps and their combined response in Figure 11. Summation of the spectral power maps has a number of attractive properties. Most importantly, the motion maps with the strongest repetitive appearance will contribute most to the final power map whereas weakly-periodic motion maps will have a negligible contribution. This effectively serves as a dynamic selection of the most discriminative motion representation. Moreover, as the spectral power is time-localized, the relative contribution per motion representation will be evolving over time. This is appealing because motion appearance can be non-static in realistic video due to camera motion or gradual change in motion type.

4.5 Spatial Segmentation

The combined wavelet power map gives a time-varying spatial distribution of spectral power over all motion representations, whereas the corresponding effective scale map relates to the temporal scale with maximum spectral power. We propose to use the spatial distribution of spectral power for segmentation of the regions with strongest repetitive appearance. Subsequently, we use the scale map to infer the dominant temporal scale (related to the motion frequency) over the localized region.

The spatial segmentation of repetitive motion is performed in a straightforward manner. For a moment in time, we simply mean-threshold the combined wavelet power map to obtain a binary segmentation mask associated with regions containing significant spectral power. More precisely, the wavelet-based motion segmentation will attend to regions in which the maximum spectral power over all temporal scales is significant. Figure 9 (bottom row) illustrates this by displaying the combined power map and corresponding scale map. In general, performing motion segmentation directly from the spatial distribution of spectral power is appealing as it couples the localization and subsequent frequency measurements. Our experiments will verify this claim and compare them with specialized motion segmentation methods. We would like to mention that our segmentation method leaves the door open for multiple repetitively moving objects whereas most state-of-the-art segmentation methods assume a single dominant foreground motion (Tokmakov et al, 2017).

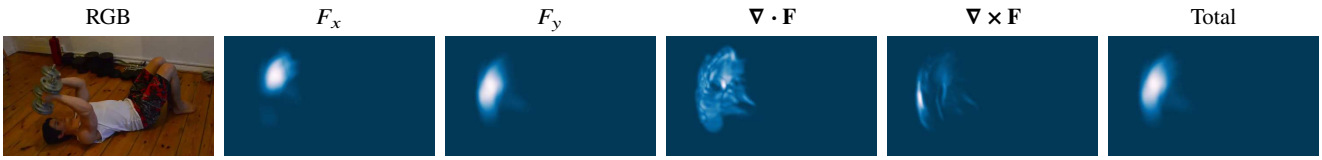


Fig. 11. Video displaying *a man lifting weights* from our video dataset and its corresponding wavelet power maps for individual representations (we omit $\nabla_x F_x$ and $\nabla_y F_y$) . On the right, the total wavelet power obtained through the summation of all six responses. We normalize the power maps for displaying purpose. The vertical flow and curl produce the power maps with the largest norm for this moment in time. Summation of the individual power map combines the responses by emphasizing on the strongest repetitive motion appearance.

4.6 Repetition Counting

To obtain an instantaneous frequency estimate of the salient motion, we median-pool the temporal wavelet scales over the segmentation mask. Median-pooling is preferred over mean-pooling as it relatively robust to outliers and will produce a better estimate of the dominant frequency. The corresponding temporal wavelet scale is then converted to an instantaneous frequency using Eq. 18. For a moment in time, this will deliver a frequency estimate for the salient repetitive motion. Counting the number of repetitions follows temporal integration of the consecutive frequency measurements with the temporal sampling spacing inferred from the video's frame rate.

We emphasize our method's ability to count the number of cycles in non-stationary video. For a stationary periodic signal, the median-pooled temporal scales will be constant over time, while non-stationarity motion produces time-varying frequency estimates. Although the videos considered in our experiments are temporally segmented, the time-localized wavelet responses could also be used for temporal localization of repetitive actions. Moreover, although the current approach performs median-pooling over the motion segmentation mask, the spatial distribution of wavelet power also enables the identification of multiple periodically moving parts.

5 Experiments

We perform experiments to show the effectiveness of our method on the task of counting repetitions in video. Prior to evaluating our full method, we demonstrate the strength of the continuous wavelet transform for estimating repetition in non-stationary signals, show the need for diversified motion maps to deal with the wide variety in motion appearance, and investigate our method's ability to handle dynamic viewpoints. Before discussing the actual experiments, we introduce the video datasets for testing, give implementation details and specify our counting evaluation metrics.

5.1 Datasets and Evaluation

The main experiments consider two video datasets: the existing *YTSegments* and our new *QUVA Repetition* dataset; both collected for the purpose of evaluating repetition estimation in video. Additionally, we perform a controlled experiment on viewpoint estimation with synthetic video that we generated through 3D modeling in Blender.

YTSegments Dataset. For the purpose of evaluating repetition counting in video, Levy and Wolf (2015) introduced a new video benchmark. The 100 videos downloaded from YouTube are purely for evaluation purpose as training the network is performed with synthesized videos. A wide range of actions appears in the videos: several sports, cooking and animal movement. Each video is temporally segmented such that only the repetitive action is covered. The clips are annotated with a total repetition count. While the dataset serves as a good initial benchmark for repetition estimation, it is limited in terms of cycle length variation (non-stationarity), motion appearances and camera motion. As our goal is to evaluate our method on more realistic video, we introduce a new video dataset that is more challenging in terms of non-stationarity, motion appearance, camera motion and background clutter.

QUVA Repetition Dataset. In Runia et al (2018) we introduced a more realistic video benchmark for repetition estimation. The *QUVA Repetition* consists of 100 videos displaying a wide variety of repetitive video dynamics, including various kinds of sport, music-making, cooking, grooming, construction and animal behavior. The videos are collected from YouTube with emphasis on creating a diverse collection of videos suitable for evaluating our method's ability to deal with non-stationary motion, camera motion and significant evolution of motion appearance over the course of a video.

After video collection, we adopt a multi-stage annotation process to obtain the final dataset. First, we asked two human annotators to label the temporal bounds of each interval containing at least four unambiguous repetitions. We found high inter-agreement between the annotators and keep the 100 intervals with the highest overlap to increase clarity. Final video clips are obtained by temporal clipping of the intersection of the two intervals. As a result, some motion cycles may be partial either at the beginning or end of the video. In the last round of annotation, we ask the annotators to mark all individual cycle bounds in the video clips (also producing the final repetition count). We also mark the individual cycle bounds for the videos of the *YTSegments* dataset

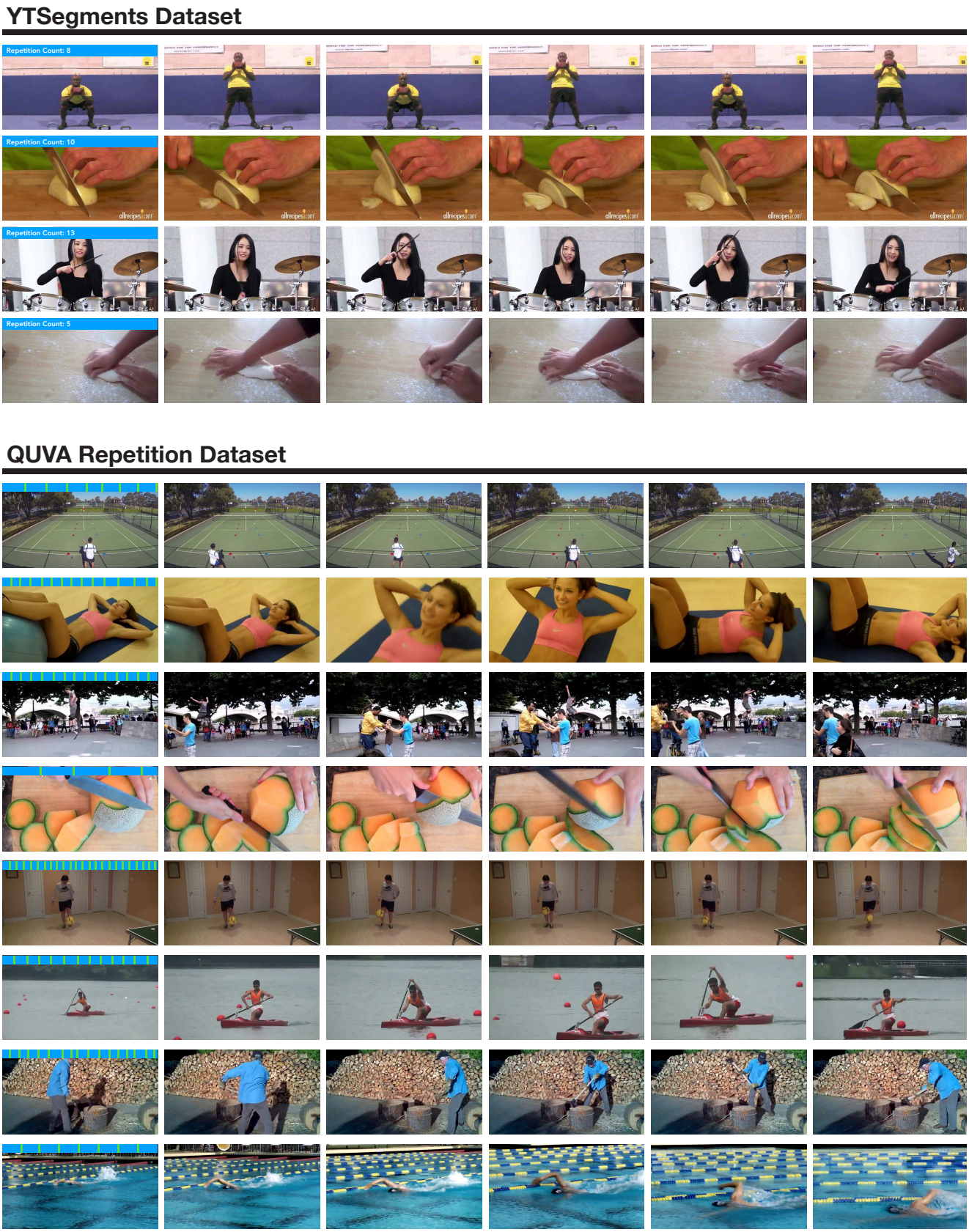


Fig. 12. Four examples from the *YTSegments* dataset (Levy and Wolf, 2015) and eight examples from our *QUVA Repetition* dataset. The *YTSegments* dataset as released by the authors features a final repetition count annotation (indicated); Our dataset is additionally annotated with individual cycle bounds suitable for determining the level of non-stationarity. The blue timeline in the first frame displays the individual cycle annotations for the given video. The final count is determined by summing the number of individual cycles. Note variation in cycle length and the increased difficulty of our dataset due to camera motion, occlusions and background clutter.

Table 1. Dataset statistics of *YTSegments* (Levy and Wolf, 2015) and *QUVA Repetition*. The cycle length variation is defined as the average value of the absolute difference between the minimum and maximum cycle length divided by the average cycle length. To determine this, we annotate all individual cycle bounds for both datasets. The last two rows are also obtained by manual annotation.

	YTSegments	QUVA Repetition
Number of Videos	100	100
Duration Min/Max (s)	2.1/68.9	2.5/64.2
Duration Avg. (s)	14.9 ± 9.8	17.6 ± 13.3
Count Avg.± Std.	10.8 ± 6.5	12.5 ± 10.4
Count Min/Max	4/51	4/63
Cycle Length Variation	0.22	0.36
Camera Motion	21	53
Superposed Translation	7	27

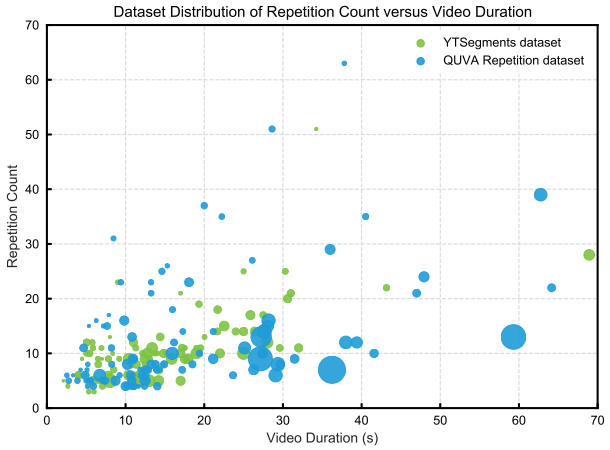


Fig. 13. Distribution of repetition count versus video duration for the *YTSegments* and *QUVA Repetition* dataset. The radius of each datapoint is proportional to the cycle length variation of the video. Note the increased variability in non-stationarity and repetition count of our dataset in comparison to *YTSegments*.

to compare the inter-cycle length variability representing the level of non-stationarity.

The characteristics for both datasets are reported in Table 1. It is apparent that our videos have more variability in cycle length, motion appearance, camera motion and background clutter. The increased difficulty in both appearance and temporal dynamics give a more realistic benchmark for repetition estimation in the wild. Figure 12 displays a number of examples from both datasets. The project page[1] contains the dataset download link and several video previews.

Evaluation Metrics. Given a set of N videos, we evaluate the performance between ground truth count c_i and the count prediction $\widehat{c_i}$ for all videos $i \in \{1, \ldots, N\}$. We report the mean absolute error following prior work (Levy and Wolf, 2015) and also record the off-by-one-accuracy (OBOA) over the entire dataset:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\widehat{c_i} - c_i| \, / c_i \tag{19}$$

$$\text{OBOA} = \frac{1}{N} \sum_{i=1}^{N} \left[|\hat{c}_i - c_i| \leq 1 \right] \tag{20}$$

The mean-absolute error is preferred over the common mean-squared error as it is relative to the true count. To account for rounding errors and possible cycle cut-offs at both ends of the video, the off-by-one-accuracy is more suitable than the traditional accuracy.

5.2 Implementation Details

Optical Flow. Our method takes two consecutive video frames as input and first estimates the motion using optical flow. As the quality of motion estimation may be important, we measure our method's sensitivity to three flow estimation methods. To evaluate a more traditional flow estimation method we choose TV-L[1] (Zach et al, 2007). This variational based method is still competitive with more recent methods. Current state-of-the-art motion estimation methods all use convolutional neural networks for the purpose. We compare the deep learning based methods EpicFlow (Revaud et al, 2015) and FlowNet 2.0 (Ilg et al, 2017). Both deep networks are trained on large (synthetic) video datasets to estimate the motion in complex video. As default we use FlowNet 2.0.

Motion Segmentation. Complex videos with background clutter or camera motion demand segmentation of the foreground motion prior to further analysis. Although our method directly performs localization from the densely computed wavelet power, we also evaluate with state-of-the-art motion segmentation methods. The fast video segmentation method of Papazoglou and Ferrari (2013) is chosen as classical approach and was also used in Runia et al (2018). This approach separates foreground objects from the background in a video by combining motion boundaries followed by segmentation refinement. We also evaluate the more recent deep learning based method of Tokmakov et al (2017). The method trains a two-stream convolutional neural network with a long-short term memory (LSTM) module to capture the evolution over time. The network parameters are optimized using the large FlyingThings 3D dataset (Mayer et al, 2016). To refine the motion masks from the trained networks, a conditional random field is applied for refinement. For both methods we use the official implementations made available by the authors. While both methods generally attain excellent segmentations, we observed that segmentation fails completely for

[1] http://tomrunia.github.io/projects/repetition/

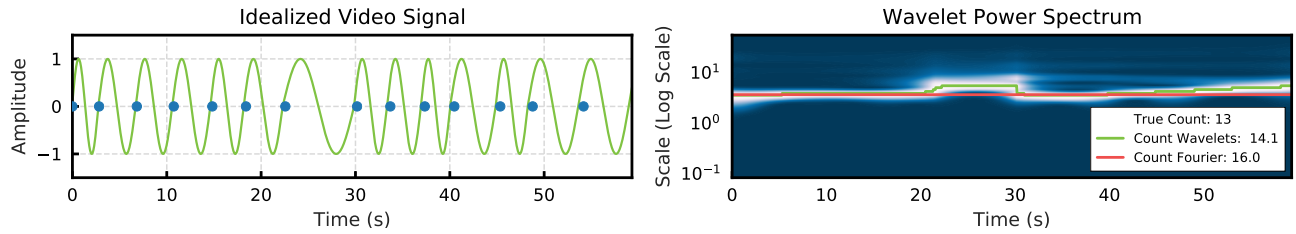


Fig. 14. Idealized signal for a difficult non-stationary video displaying a violin player. The blue markers indicate the cycle bounds, manually annotated for each video in our *QUVA Repetition* dataset. Note how the wavelet scalogram correctly exposes the rhythmic slowdown (around 20 seconds). On the right, the green line corresponds to the local frequency predictions from the scalogram whereas the red (straight) line indicates the stationary Fourier-based frequency measurement. This demonstrates the effectiveness of wavelet analysis for optical non-stationary video signals.

some more difficult frames (either all or none pixels selected as foreground). To remedy incorrect segmentation masks we reuse the segmentation of the previous frame if the fraction of foreground pixels is less than 1% of the entire frame.

Differential Geometric Motion Maps. To compute the motion maps we perform spatial filtering by first-order Gaussian kernels. The filtering is implemented in PyTorch and runs in large batches on the GPU to accelerate computation. Spatial convolution is performed with $\sigma = 4$ for all experiments. We also evaluated $\sigma = \{2, 8, 16\}$ but found only minor variation in performance. In practice, a combination of multiple spatial scales may produce best results. Once the spatial first-order derivatives $\nabla_x F_x$, $\nabla_y F_x$, $\nabla_x F_y$ and $\nabla_y F_y$ have been obtained through convolution, the differential motion maps are computed as specified in Section 4.1.

Continuous Wavelet Transform. We use the continuous wavelet filtering implementation as outlined in Torrence and Compo (1998). In comparison to the previous version of our work, we now also perform temporal filtering on the GPU[2] resulting in a considerable speed-up. This enables us to apply the wavelet transform in large batches over all spatial locations in the video. As previously mentioned, we use a Morlet wavelet ($\omega_0 = 6$) with logarithmic scales ($\delta j = 0.125$, $s_0 = 2\delta t$). We limit the range of J corresponding to a minimum of four repetitions by setting s_{min} and s_{max} accordingly in (16) and (17). Depending on the video length, there are typically between 50 and 60 temporal scales levels. When compute budget is tight, computational efficiency can be improved by pruning the filter bank with scale selection, for example using the maximum response of a Laplacian filter (Lindeberg, 2017).

Repetition Counting. The instantaneous frequency estimates are obtained from the dense wavelet power by pooling over the motion foreground mask. As detailed in Section 4.6, the frequencies are integrated over time to arrive at a final repetition count. To remove frequency estimate outliers inconsistent with adjacent frames, we apply a median filter of 9 timesteps (frames) to enforce local smoothness. This gives a slight improvement on both video datasets. The final Count predictions are not rounded, hence evaluation metrics may be slightly off due to incomplete cycles.

[2] https://github.com/tomrunia/PyTorchWavelets

Reimplementation of Baselines. We compare our method against two existing works for repetition estimation. The method of Pogalin et al (2008) is chosen to represent the class of Fourier-based methods. Our reimplementation uses a more recent object tracker (Henriques et al, 2012) but is identical otherwise. The tracker is initialized by manually drawing a box on the first frame. Converting the frequency to a count is trivial using the video length and frame rate. Additionally, we compare with the deep learning method of Levy and Wolf (2015) using their publicly available code and pretrained model without any modifications.

5.3 Temporal Filtering: Fourier versus Wavelets

Setup. The goal of our first experiment is to demonstrate the effectiveness of the continuous wavelet transform for counting repetitions in non-stationary signals. We compare the stationary Fourier-based periodogram with the time-scale representation given by the wavelet scalogram. To isolate the effect of frequency measurements, we generate idealized signals of the videos in our *QUVA Repetition* dataset. Specifically, we fit sinusoidal signals through the individual cycle bounds for each video to obtain simple 1D waveforms representing the video. Figure 14 shows an idealized signal example and the corresponding wavelet spectrum with count predictions. To compare with the Fourier-based measurement, we compute the periodogram, detect the maximum frequency peak and convert the corresponding frequency to a count using the video's duration. This yields a repetition count prediction for both the stationary and non-stationary measurements that we evaluate over the entire dataset.

Results. From the results in Figure 15 it is clear that wavelet-based counting outperforms the periodogram on idealized signals. As expected, we observe that the Fourier-based mea-

surements generally fail on videos with significant cycle length variation as they give a global frequency prediction. Wavelets naturally handle non-stationary repetition and are less sensitive to cycle length variability. We also tried adding a substantial amount of Gaussian noise ($\sigma = 0.5$) to the signals; this resulted in a minor negative effect on both methods (data not shown). This controlled experiment shows the effectiveness of wavelets for repetition estimation assuming a clear signal can be distilled from the videos.

5.4 Viewpoint Invariance

Setup. The theory of repetition considers two viewpoint extremes (Figure 6). In this experiment we evaluate our method's ability to handle a continuous transition from one viewpoint extreme to the other. The designated mechanism for this is the use of multiple motion representations and the summation of their spectral power obtained from the continuous wavelet transform. To test this, we set-up a controlled experiment in which we synthesize a video clip from 3D modeled data in Blender. This enables full control over the object's motion and the viewpoint. Specifically, we choose to build a simple 3D scene containing a ball periodically bouncing on the floor as displayed in the top row of Figure 16. Initially, the camera captures the bouncing ball from the side view but after a number of full motion cycles, the camera smoothly transitions to frontal view (case 3 to case 6 in Figure 6). We record the median-pooled vertical flow and divergence over the foreground region to obtain two time-varying signals. The spectral power for both signals is individually estimated using the continuous wavelet transform, after which we combine the power by summation.

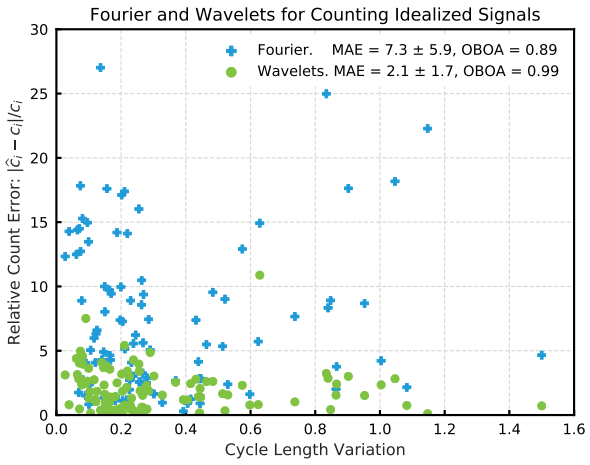


Fig. 15. Fourier- versus wavelet-based repetition counting on idealized signals videos from the *QUVA Repetition* dataset. Our wavelet-based method outperforms a Fourier-based baseline for 83 out of 100 videos. High cycle length variation results in notable error for Fourier measurements, whereas the time-localized wavelets are less sensitive to non-stationary repetition.

Results. Figure 16 plots the two median-pooled flow signals and their joint wavelet power obtained by summation. Initially, as the moving object is captured from the side view, vertical flow is best measurable. Upon the viewpoint transition, vertical flow vanishes while the divergent flow becomes dominant. As a result of the camera motion, the measurement of the spectral power for both individual signals will only give a strong response for either the first or second half of the video. However, the summation of the spectra gives a clear measurement over the complete video as is apparent from the combined wavelet power spectrum. This illustrates our method's ability to handle viewpoint changes by the combination of the wavelet power contained in multiple motion representations. By summation of the spectra, the best measurable motion representation will naturally give the largest contribution to the combined power. Therefore, this mechanism acts as a replacement of the global representation selection used in (Runia et al, 2018) by dynamically leveraging information in all representations.

5.5 Diversity in Motion Maps

Setup. As wavelets prove to be effective for repetition estimation and multiple representations show value on a synthetic video, we now assess the value of a diversified video representation on real videos of our *QUVA Repetition* dataset. We hypothesize that, due to the high variability in motion pattern and viewpoint, no single representation is powerful but their joint diversity is effective. To test this, we perform repetition counting over all individual motion maps listed in Eq. (13). Instead of summing the wavelet power for all representations, we test the performance of the six motion representations individually. For each representation we densely compute te wavelet power and count the number of repetitions as outlined in the method's section. For a fair comparison, we exclude our motion segmentation mechanism based on wavelet power and instead use the motion segmentation proposed by Papazoglou and Ferrari (2013). Again, we evaluate repetition counting on our *QUVA Repetition* dataset. To obtain a lower-bound on the error, we also select the best representation per video in an oracle fashion.

Results. The results in Table 2 reveal that for the wide variability of repetitive appearance there is no one size fits all solution. The individual motion maps are unable to handle the variety of repetitive motion appearances by themselves, resulting in poor count performance over the dataset. However, their joint diversity produces a good lower-bound by oracle selection of the most discriminative motion map. We notice the superiority of vertical flow F_y as it performs best and is selected most often by the oracle. We explain this bias towards vertical flow by the observation that our dataset con-

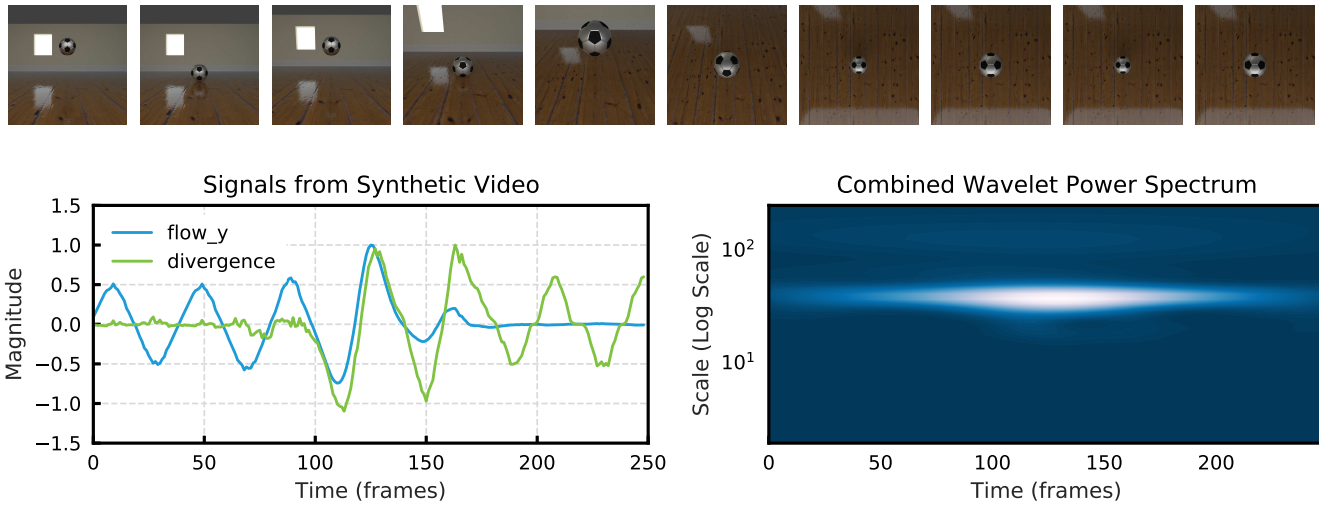


Fig. 16. *Top*: synthetized video sequence for a controlled experiment on the influence of viewpoint relative to the motion. This video clip shows a 3D modeled scene containing a bouncing ball. At the midpoint of the animation, the camera smoothly transitions from side view to frontal view. *Bottom Left*: the time-varying magnitude of vertical flow and divergence measured over the foreground segmentation. Initially, the vertical flow is dominant and divergence is negligible. This reverses with the viewpoint transition. *Bottom Right*: the combined wavelet spectrum of both signals. Notice the spectrum's invariance to viewpoint change as a result of wavelet power summation.

Table 2. Value of diversity in six motion maps for videos from *QUVA Repetition*. The last column denotes how often each signal is selected by the oracle. While the individual signals struggle to obtain good performance by themselves, exploiting their joint diversity is beneficial.

	MAE	OBOA	# Selected
$\nabla \cdot \mathbf{F}$	77.8 ± 90.8	0.21	10
$\nabla \times \mathbf{F}$	53.0 ± 65.5	0.32	11
$\nabla_x F_x$	58.1 ± 63.5	0.29	15
$\nabla_y F_y$	59.5 ± 68.4	0.31	9
F_x	49.6 ± 48.0	0.35	25
F_y	42.0 ± 45.3	0.43	30
Oracle Best	24.1 ± 33.5	0.63	100

tains several sports videos in which the gravity is often used as opposing force.

5.6 Video Acceleration Sensitivity

Setup. In this experiment, we examine our method's sensitivity to acceleration by artificially speeding-up videos. Starting from the *YTSegments* dataset, in which most videos exhibit strong periodic motion, we induce significant non-stationarity by artificially accelerating the videos halfway. More precisely, we modify the videos such that after the midpoint frame, the speed is increased by dropping every second frame. What follows are 100 videos with a 2× acceleration starting halfway. We compare against the deep learning method of Levy and Wolf (2015) which handles non-stationarity by running the period-predicting convolutional neural network in sliding-window fashion over the video. Fourier-based analysis was left out as it will inevitably fail on this task.

Results. The bar chart of Figure 17 presents the mean absolute error in both original and accelerated setting. On their own dataset, the system of Levy and Wolf (2015) slightly outperforms our method. Acceleration reverses the results as our method suffers less and obtains a lower error on the accelerated videos. It reveals their sensitivity to acceleration, whereas our method deteriorates less. This shows the effectiveness of wavelets for dealing with non-stationarity in realistic videos. To illustrate how our method deals with midpoint acceleration, we also plot the count increments and cumulative counts throughout the video; see Figure 18. As is evident from the plot, there is a distinct increase in count increments per timestep when upon enabling acceleration. This is observed for most videos in the dataset. This could be beneficial for detecting acceleration or temporal localization of transient phenomena in video.

5.7 Motion Segmentation

Setup. In this experiment investigate the effectiveness of the motion segmentations obtained directly from the wavelet power for repetition estimation. We visually compare the motion segmentations and test whether replacing our localization mechanism with a state-of-the-art motion segmentation method improves repetition estimation performance. We keep the method identical except for the segmentation method to obtain a motion mask. In addition to our wavelet-based motion segmentation to obtain the discriminative motion mask we compare our method's performance without any localization (full-frame), the video segmentation method

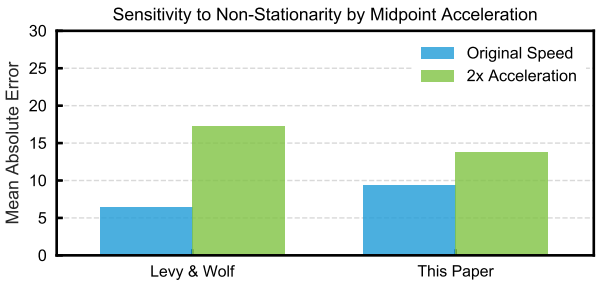


Fig. 17. The effect of midpoint acceleration on the *YTSegments* dataset. Our method increases 4.4 in mean absolute error whereas the method of Levy and Wolf (2015) rises with 10.8 points. The deep learning method has difficulty dealing with non-stationary acceleration, whereas our method is more robust due to the wavelet transform.

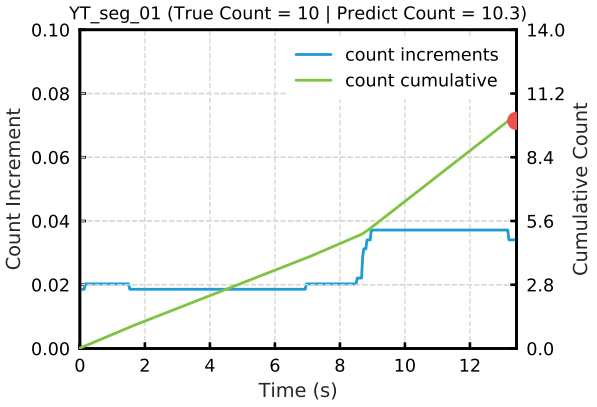


Fig. 18. Count increments and cumulative count over time for the first video of *YTSegments* with midpoint acceleration. The red marker on the right corresponds to the ground truth count. Note how the increase in speed around 9 seconds is clearly reflected in the count increments.

of Papazoglou and Ferrari (2013) and the deep learning approach of Tokmakov et al (2017).

Results. We visually compare the three different motion segmentation methods in Figure 19. For most videos, our method is able to localize the repetitive motion. By all means, the state-of-the-art methods specifically devoted to foreground motion segmentation produce the visually best results. However, our intention is to obtain a motion mask best suitable for repetition estimation which not necessarily overlaps with the foreground motion. By thresholding the wavelet power maps, our method seems to emphasize on regions with most discriminative repetitive motion. This is best recognizable from the bottom two rows where the motion segmentation includes background regions that periodically change due to the motion. In Table 3 we report quantitative results of our method with different motion segmentation methods. Our localization mechanism produces significantly better results than the existing motion segmentation methods. Partially, this might be explained by the temporal delay of the wavelet responses in comparison to the motion segmentation masks. For our method, this convincingly demonstrates that the segmentation directly obtained from the wavelet spectrum are more suitable than decoupled motion segmentation approaches.

5.8 Comparison to the State-of-the-Art

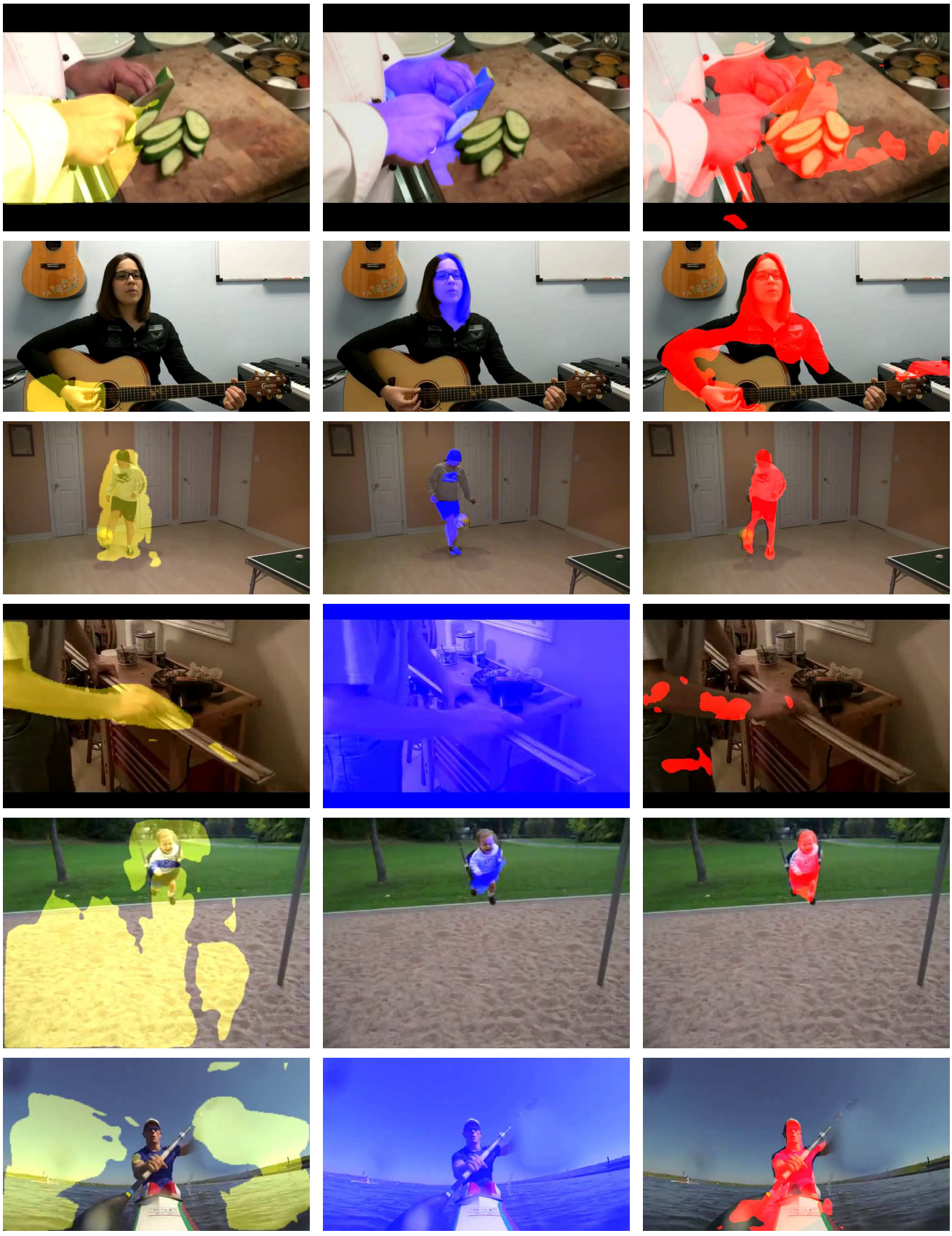
Setup. In this experiment, we perform a full comparison on the task of repetition counting for both video datasets. We compare against the Fourier-based method of Pogalin et al (2008) and the deep learning approach of Levy and Wolf (2015).

Results. The full count evaluation is presented in Table 4. On their own *YTSegments* dataset, the method of Levy and Wolf (2015) performs best with an MAE of 6.5, where our method achieves a comparable error of 9.4 and near-identical off-by-one accuracy. Despite the stationary nature of most videos in this dataset, the Fourier-based approach of Pogalin et al (2008) performs unfavorably compared to all other methods.

The results change dramatically when considering our challenging *QUVA Repetition* dataset; notably the deep learning approach of Levy and Wolf (2015) now performs the worst, with an MAE of 48.2. This could possibly be explained by the fact that their network only considers four motion types during training or the convolutional network's fixed temporal input dimension posing a constraint on the effective motion periods (ranging from 0.2 to 2.33 seconds). Dealing with motion periods outside of this range most likely requires retraining the network. The Fourier-based method of Pogalin et al (2008) scores an MAE of 38.5, whereas we obtain an average error of 26.1. On the *YTSegments* dataset our simplified method slightly improves over the MAE of 10.3 ± 19.8 reported in (Runia et al, 2018), while giving comparable results to previously reported MAE of 23.2 ± 34.4 on the *QUVA Repetition* dataset. The Fourier-based and deep learning-based approaches are unable to effectively handle the increased non-stationarity and motion complexity found in our challenging video dataset. The method proposed here improves the ability to handle such difficult videos without relying on explicit motion segmentation methods.

We also report the repetition count results using TV-L[1] (Zach et al, 2007) and EpicFlow (Revaud et al, 2015) to investigate our method's sensitivity to optical flow quality. The results in Table 5 show the robustness to different flow methods as the algorithm of choice has limited effect on the count performance for both datasets.

To gain a better understanding of our method's characteristics we study success and failure cases. We observe that our wavelet-based motion segmentation struggles with scenes containing water (*e.g.* Figure 12, bottom row) or other dynamic texture. The rippling water produces visual repeti-



(a) This paper **(b)** Papazoglou and Ferrari (2013) **(c)** Tokmakov et al (2017)

Fig. 19. Comparison of different motion segmentation masks. In most cases, our method succeeds to spatially segment the repetitive motion. In comparison to methods specifically devoted to the task of motion segmentation, our masks are less precise. However, as our numerical evaluation shows, our segmentation masks are more suitable for the task of repetition estimation. The most informative repetitive cues do not necessarily overlap with the foreground motion. In the last example, the regions through which the paddles moves produce the strongest repetitive response.

Table 3. Repetition counting results of our method with different motion segmentation mechanism. While the state-of-the-art motion segmentation methods produce visually excellent results, their segmentations are suboptimal for the task of repetition estimation. This is expected as the most discriminative repetitive cues are not always contained in the foreground motion. See Figure 19 for a visual comparison of segmentation masks.

	YTSegments		QUVA Repetition	
Motion segmentation method	MAE ↓	OBOA ↑	MAE ↓	OBOA ↑
Full-frame	46.0 ± 67.2	0.28	60.8 ± 49.4	0.22
Papazoglou and Ferrari (2013)	13.1 ± 20.3	0.78	42.6 ± 49.2	0.44
Tokmakov et al (2017)	21.6 ± 57.2	0.76	38.9 ± 39.2	0.42
Differential geometry (this paper)	**9.4 ± 17.4**	**0.89**	**26.1 ± 39.6**	**0.62**

Table 4. Comparison with the state-of-the-art on repetition counting for the *YTSegments* and our *QUVA Repetition* dataset. The deep learning-based method of Levy and Wolf (2015) achieves good results on their own dataset of relatively clean videos. On our more realistic and challenging dataset, the current method improves considerably over the existing approaches. In comparison to our previous work, our method segments the repetitive motion directly rather than relying on decoupled motion segmentation.

	YTSegments		QUVA Repetition	
	MAE ↓	OBOA ↑	MAE ↓	OBOA ↑
Pogalin et al (2008)	21.9 ± 30.1	0.68	38.5 ± 37.6	0.49
Levy and Wolf (2015)	**6.5 ± 9.2**	**0.90**	48.2 ± 61.5	0.45
This paper	9.4 ± 17.4	0.89	**26.1 ± 39.6**	**0.62**

tive dynamics resulting in a strong wavelet response over its entire surface. Consequently, motion segmentation by mean-thresholding of the spectral power will fail inevitably; and subsequent measurements over the foreground motion mask will be incorrect as well. For such videos, we observe an enormous over-count as the frequency estimates correspond to the high-frequent rippling water. The error associated with these videos explains the limited improvement over our previous method (Runia et al, 2018) which relied on Papazoglou and Ferrari (2013) for motion segmentation, being less prone to such segmentation failures.

We also observe that all methods make a common mistake: over-counting videos with a factor of two. The similarity in these videos is that one full cycle contains the exact same motion first with one arm (or leg) followed by the other (*e.g.* walking lunges or swimming front-crawl). As the perceived motion is almost identical for both limbs, the estimated temporal dynamics are twice as fast. Again, the significant overestimate of the motion frequency produces a large count error for all methods. Solving this problem is not easy, as current repetition estimates in those cases are essentially also a correct prediction; however, the human annotators define salient motion as a full cycle with both limbs.

6 Conclusion

We have categorized 3D intrinsic periodic motion as translation, rotation or expansion depending on the first-order differential decomposition of the motion field. Additionally, we distinguish three periodic motion continuities: constant, intermittent and oscillatory motion. For the 2D perception of 3D periodicity, the camera will be somewhere in the continuous range between two viewpoint extremes. What follows are 18 fundamentally different cases of repetitive motion appearance in 2D. The practical challenges associated with repetition estimation are the wide variety in motion appearance, non-stationary temporal dynamics and camera motion. Our method addresses all these challenges by computing a diversified motion representation, employing the continuous wavelet transform and combining the power spectra of all representations to support viewpoint invariance. Whereas related work explicitly localizes the foreground motion, our method performs repetitive motion segmentation directly from the wavelet power maps resulting in a simplified approach. We verify our claims by improving the state-of-the-art on the task of repetition counting on our challenging new video dataset. The method requires no training and requires only a minimum number of hyper-parameters which are fixed throughout the paper. We envision applications beyond repetition estimation as the wavelet power and scale maps can support localization of low- and high-frequency regions suitable for region pruning or action classification.

Table 5. Sensitivity of our method with respect to different optical flow methods. We report repetition counting results over both datasets. Only slight variation in the performance is observed, demonstrating our method's robustness to optical flow quality.

	YTSegments		QUVA Repetition	
	MAE ↓	OBOA ↑	MAE ↓	OBOA ↑
TV-L[1]	9.8 ± 17.9	0.89	26.5 ± 67.5	**0.67**
EpicFlow	9.7 ± 17.9	0.88	30.8 ± 38.2	0.55
FlowNet 2.0	**9.4 ± 17.4**	**0.89**	**26.1 ± 39.6**	0.62

References

Abraham R, Marsden JE, Ratiu T (1988) Manifolds, Tensor Analysis, and Applications, vol 75. Springer Berlin Heidelberg

Albu AB, Bergevin R, Quirion S (2008) Generic temporal segmentation of cyclic human motion. Pattern Recognition 41(1)

Azy O, Ahuja N (2008) Segmentation of periodically moving objects. In: Proceedings of the IEEE International Conference on Pattern Recognition, pp 1–4

Belongie S, Wills J (2006) Structure from periodic motion. In: Spatial Coherence for Visual Motion Analysis, Springer Berlin Heidelberg

Briassouli A, Ahuja N (2007) Extraction and analysis of multiple periodic motions in video sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(7)

Burghouts GJ, Geusebroek JM (2006) Quasi-periodic spatiotemporal filtering. IEEE Transactions on Image Processing 15(6):1572–1582

Chen My, Hauptmann A (2009) MoSIFT: Recognizing human actions in surveillance videos. Tech. Rep. CMU-CS-09-161, Carnegie Mellon University

Chetverikov D, Fazekas S (2006) On motion periodicity of dynamic textures. In: Proceedings of the British Machine Vision Conference, pp 167–176

Cutler R, Davis LS (2000) Robust real-time periodic motion detection, analysis, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8)

Davis J, Bobick A, Richards W (2000) Categorical representation and recognition of oscillatory motion patterns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol 1, pp 628–635

Goldenberg R, Kimmel R, Rivlin E, Rudzsky M (2005) Behavior classification by eigendecomposition of periodic motions. Pattern Recognition 38(7)

Grossmann A, Morlet J (1984) Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM Journal on Mathematical Analysis 15(4)

Henriques JF, Caseiro R, Martins P, Batista J (2012) Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of the European Conference on Computer Vision

Huang S, Ying X, Rong J, Shang Z, Zha H (2016) Camera calibration from periodic motion of a pedestrian. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2017) FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Jain M, Jegou H, Bouthemy P (2013) Better exploiting motion for better action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2555–2562

Johansson G (1973) Visual perception of biological motion and a model for its analysis. Perception & Psychophysics

Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: Proceedings of the British Machine Vision Conference, pp 275–1

Koenderink J, van Doorn A (1975) Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. Optica Acta: International Journal of Optics 9:773–791

Koenderink JJ (1988) Scale-time. Biological Cybernetics 58(3):159–162

Laptev I, Belongie SJ, Perez P, Wills J (2005) Periodic motion detection and segmentation via approximate sequence alignment. In: Proceedings of the IEEE International Conference on Computer Vision, vol 1, pp 816–823

Levy O, Wolf L (2015) Live Repetition Counting. In: Proceedings of the IEEE International Conference on Computer Vision

Lindeberg T (2017) Dense scale selection over space, time and space-time. Journal on Imaging Sciences 11(1):438–451

Liu F, Picard RW (1998) Finding periodicity in space and time. In: Proceedings of the IEEE International Conference on Computer Vision, pp 376–383

Lu C, Ferrier NJ (2004) Repetitive motion analysis: Segmentation and event classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(2)

Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T (2016) A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4040–4048

Papazoglou A, Ferrari V (2013) Fast object segmentation in unconstrained video. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1777–1784

Pogalin E, Smeulders AWM, Thean AHC (2008) Visual quasi-periodicity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Polana R, Nelson RC (1997) Detection and recognition of periodic, nonrigid motion. International Journal of Computer Vision 23(3)

Ran Y, Weiss I, Zheng Q, Davis LS (2007) Pedestrian detection via periodic motion analysis. International Journal of Computer Vision 71(2)

Revaud J, Weinzaepfel P, Harchaoui Z, Schmid C (2015) EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Runia TFH, Snoek CGM, Smeulders AWM (2018) Real-world repetition estimation by Div, Grad and Curl. In:

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Sarel B, Irani M (2005) Separating transparent layers of repetitive dynamic behaviors. In: Proceedings of the IEEE International Conference on Computer Vision

Thangali A, Sclaroff S (2005) Periodic motion detection and estimation via space-time sampling. In: Proceedings of the IEEE Workshops on Application of Computer Vision

Tokmakov P, Alahari K, Schmid C (2017) Learning motion patterns in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Torrence C, Compo GP (1998) A practical guide to wavelet analysis. Bulletin of the American Meteorological society 79(1)

Tralie CJ, Perea JA (2018) (quasi) periodicity quantification in video data, using topology. SIAM Journal on Imaging Sciences 11(2):1049–1077

Tsai PS, Shah M, Keiter K, Kasparis T (1994) Cyclic motion detection for motion based recognition. Pattern Recognition 27(12)

Zach C, Pock T, Bischof H (2007) A duality based approach for realtime TV-L 1 optical flow. In: Pattern Recognition, LNCS, vol 4713, Springer Berlin Heidelberg, pp 214–223