Primitive Motion Types for Learning from Instructional Video

Runia, T.F.H.; Snoek, C.G.M.; Smeulders, A.W.M.

Publication date
2018
Document Version
Final published version
Published in
FIVER @ CVPR 2018

Primitive Motion Types *for* Learning from Instructional Video

Tom F. H. Runia Cees G. M. Snoek Arnold W. M. Smeulders
QUVA Deep Vision Lab, University of Amsterdam
{runia,cgmsnoek,a.w.m.smeulders}@uva.nl

1. Introduction

Instructional videos are abundant on the internet and serve as primary source of information for accomplishing non-trivial tasks such as "replacing a phone battery", "rolling sushi" or "grinding calligraphy ink" (Figure 1). As the popularity of instructional video indicates, the visual image in addition to the linguistic narrative is often decisive for successful completion of complex tasks. Understanding the visual stream of instructional video can benefit robotics research (*e.g.* through imitation learning) and decomposing complex tasks into primitive subtasks. In this abstract we focus on this visual stream. Specifically, we propose to model the salient motion in instructional videos by measuring the time-varying first-order differentials of the flow field. This approach is an application of our conference paper [3].

Existing work on learning from instructional videos [1, 2] leverages the supervision of the spoken narrative available through YouTube's automatic speech recognition system. Alayrac *et al.* [1] propose to learn primitive subtasks for instructional video classes through unsupervised learning. Specifically, textual and visual concepts are independently clustered followed by a joint sequence alignment. Huang *et al.* [2] focus on instructional cooking videos. Their unsupervised approach combines linguistic and visual input to temporally link entities (*e.g.* "dressing") to the action that produced it (*e.g.* "mix vinegar"). Those approaches emphasize on the unsupervised alignment of linguistic and visual streams rather than focusing on the visual representation of motion.

Focusing on the visual representation, we note that in comparison to other computer vision tasks (*e.g.* image classification, action recognition and video segmentation), the *context* of the scene is often irrelevant for understanding the instructions. We believe that in order to understand instructional video, successful approaches need to emphasize on fine-grained motion patterns, object interaction and procedural knowledge parsing. Here, we focus on modeling the motion by deriving primitive motion types from the 3D flow field and then measuring the time-varying observed flow in the image plane.

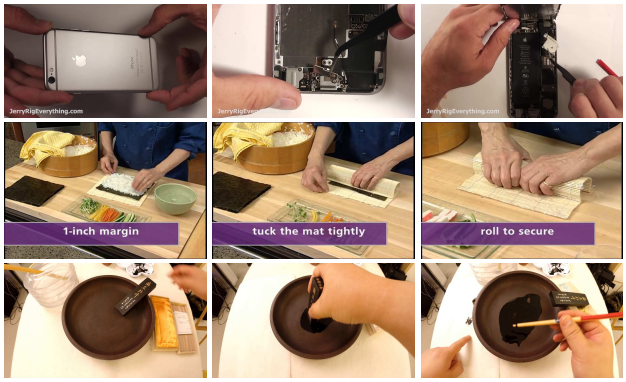


Figure 1. Examples of instructional videos found on YouTube. Top: replacing a phone battery. Center: preparing Maki sushi rolls. Bottom: grinding Chinese calligraphy ink. Each instructional video is a chain of primitive motion types; identifying these can aid the visual understanding of such videos.

2. Proposed Method

The perceived motion in instructional videos exists on the 2D image plane but the correct starting point for modeling the motion is the 3D world. For a moment in time t, we denote the 3D flow field tied to an object by $\boldsymbol{\mathcal{F}}_t(\mathbf{x})$. The flow field can be decomposed into its directional components: $\boldsymbol{\mathcal{F}}_t = (\mathcal{F}_x, \mathcal{F}_y, \mathcal{F}_z)$. From differential geometry, we have the three operators acting on the flow field:

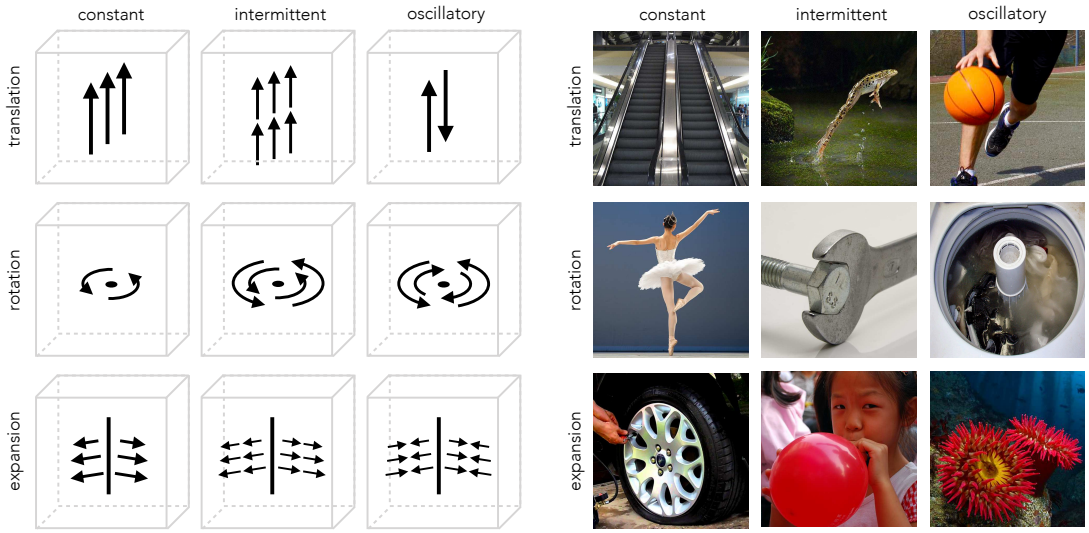
$$\boldsymbol{\nabla}\boldsymbol{\mathcal{F}}_t = \frac{\partial \mathcal{F}_k}{\partial x_j}\widehat{\mathbf{e}}_j \otimes \widehat{\mathbf{e}}_k \tag{1}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{F}}_t = \frac{\partial \mathcal{F}_x}{\partial x} + \frac{\partial \mathcal{F}_y}{\partial y} + \frac{\partial \mathcal{F}_z}{\partial z} \tag{2}$$

$$\boldsymbol{\nabla}\times\boldsymbol{\mathcal{F}}_t = \left(\frac{\partial \mathcal{F}_z}{\partial y} - \frac{\partial \mathcal{F}_y}{\partial z}, \frac{\partial \mathcal{F}_x}{\partial z} - \frac{\partial \mathcal{F}_z}{\partial x}, \frac{\partial \mathcal{F}_y}{\partial x} - \frac{\partial \mathcal{F}_x}{\partial y}\right). \tag{3}$$

The equations define the gradient, divergence and curl of the flow field [4]. Three basic 3D-motion types emerge depending on the values of divergence and curl as follows:

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



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

Figure 2. 3×3 Cartesian table of the *motion type* times the *motion continuity*. Following from the differential operators acting on the flow, these are the basic cases of (periodic) motion in 3D. The examples are: escalator, leaping frog, bouncing ball, pirouette, tightening a bolt, laundry machine, inflating a tire, inflating a balloon and a breathing anemone. Figure originally appeared in [3].

In addition, the temporal gradient of the time-varying 3D flow field results in three motion continuities: constant, oscillatory and intermittent motion. Jointly, the motion types and motion continuities organize in a 3×3 Cartesian table of fundamental motion cases (Figure 2). For 2D video, the 3D flow field is projected on the image plane based on the observer's viewpoint. Considering the two distinct viewpoint extremes (frontal and side view), this gives rise to 18 atomic motion types to be measured with divergence, gradient and curl and operators (see [3] for an illustration of these cases).

Instructional videos typically consist of an ordered chain of primitive subtasks. Each of these primitive subtasks will produce a characteristic flow field to be measured using the first-order differential operators. Specifically, in [3] we propose to segment the foreground motion and obtain a max-pooled representation of the differential measures using Gaussian derivative filers over the foreground mask. In effect, the video is represented by time-varying signals that encode the presence of primitive motion types. The motion types can be related to instructional subtasks and their ordering over time contains valuable information for understanding the instructional video. To distill repeated motion or transient phenomena from the signals one can rely on the continuous wavelet transform, correlation-based methods or recurrent neural networks.

An example, in the context of understanding instructional video, is the consecutive measurements of oscillatory translation ("cutting") and constant rotation ("stirring") that establishes a strong cue for the task "making soup". The successive identification of those primitive actions to be measured by the divergence and gradient pairs underlies our proposed approach to a visual understanding of instructional video. One challenge that remains is learning the ordering of primitive subtasks collectively define the full instruction set. To alleviate this difficulty, [1] assumes that each task has the same order of primitive subtasks. In practice, this assumption does not hold as tasks can be completed in many different ways and steps may be omitted. Learning complex ordering seems feasible given the vast number of videos available with weak supervision in the form of textual narrative obtained through automatic speech recognition.

3. Conclusion

In this brief abstract we have proposed a method for visual identification of primitive subtasks in instructional video. Specifically, we propose to represent motion by measuring the divergence, gradient and curl of the flow field. On top of this, the order in the chain of primitive subtasks can be associated with a complete tasks through temporal modeling.

References

[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, July 2016. 1, 2

[2] D.-A. Huang, J. J. Lim, L. Fei-Fei, and J. Carlos Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *CVPR*, July 2017. 1

[3] T. F. H. Runia, C. G. M. Snoek, and A. W. M. Smeulders. Real-world repetition estimation by div, grad and curl. In *CVPR*, June 2018. 1, 2

[4] H. M. Schey. *Div, grad, curl, and all that: an informal text on vector calculus*. WW Norton, 2005. 1