Learning Filter-Based Motion Features for Dynamic Scene Analysis

Damien Teney dteney@andrew.cmu.edu
Martial Hebert hebert@ri.cmu.edu
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA

Context and motivation The analysis and recognition of videos of dynamic scenes involves the processing of the spatial and temporal parts of the input. The former correspond to the appearance of contents of the scene, and the latter to their *changes* of appearance over time, *e.g.* due to their motion. The typical approach for separating these two parts is to use an optical flow algorithm, which models the temporal part as a map of pixelwise displacements. This representation presents limitations when dealing with complex scenes, since a pixel may not correspond to a single displacement (*e.g.* with transparencies, semi-transparencies, or dynamic occlusions), and the assumption of the conservation of an image quantity (brightness, typically), may not hold when the changes of appearance do not involve simple translations, as in dynamic textures (*e.g.* water, smoke, or foliage). We are thus interested in extracting descriptors of the temporal component of videos of wider applicability than optical flow.

The second motivation for this work comes from the success of convolutional neural networks (CNNs) on image-based tasks, from object recognition to semantic segmentation or geometry prediction. This success is inspiring a series on similar developments on video input, but the deep learning paradigm has not proven as successful in identifying the low-level features of the temporal component of videos. Recognizing that the spatial and temporal parts should benefit from separate processing, the current most successful approach [7] consists in feeding the network, on one side with the raw input pixels, and on the other with precomputed optical flow or dense trajectories. The need for a separate algorithm to extract the dynamic information is less than satisfactory, since the spatial part of the network, in comparison, is fed directly with pixel values. Other works have use a single network for the video volume of pixels, thus processing the spatial and temporal components together and indifferently (*e.g.* [4, 5]), but this approach seemed less successful.

Contribution In the following, we present a shallow convolutional network architecture that maps a volume of pixels (a pair or sequence of stacked frames) to a high-dimensional motion descriptor. This motion descriptor can be projected onto a tradition flow map, but can itself represent a much wider range of phenomena, including multiple, overlapping motions a single location due to transparencies. The network is trained using sequences with ground truth optical flow, and is intended to serve as building block in deeper architectures for the analysis of videos.

Technical approach We perform motion estimation in a convolutional network framework by revisiting the classical approach of Heeger *et al.* [3]. The idea behind filter-based motion estimation is to decompose the input video in the frequency domain. Remember that our objective is to identify motion regardless of texture and appearance, and we thus want to retain the temporal part of such a decomposition, independently of the spatial part. The convolution of a signal with a given kernel corresponds to a multiplication of their respective spectra in the frequency domain. Convolutions with a bank of bandpass filters thus allow one to measure energies in these bands, which are then suitable for frequency analysis. In the case of videos, we use 3-dimensional kernels, applied on volume of pixels of the video. Whereas previous work studied the engineering of appropriate bandpass filters to decompose video signals, our work seeks to *learn* such filters. The motivation behind a learning approach is to allow optimizing the filters to particular domains. Note that, although filter-based motion estimation has not been historically popular compared to the Horn and Schunk approach, a number of recent works have used spatiotemporal filters to describe dynamic scenes and dynamic textures [2, 9]. Note also that filter-based motion estimation has a long history as a computational model of the human perception of motion [8, 6].

A pattern moving in a video with a constant speed and orientation manifests itself as a plane in the frequency domain [1]. The energy of the signal will then entirely lie in this plane, which passes through the origin, and the orientation and tilt of which correspond respectively to the orientation and speed of the image translation. Interestingly, transparent signals in an image, moving with different speeds or directions, will correspond to distinct planes in the frequency domain. One can readily see how frequency analysis is conducive to the identification of complex motions.

The proposed architecture is summarized in Fig. 1. Instead of relying only on training data to learn to enforce desired invariances (to spatial phase, contrast, gradient orientation, etc.), we use a few signal processing principles to build them into the network architecture. First, we ensure invariance to additive brightness changes by convolv-

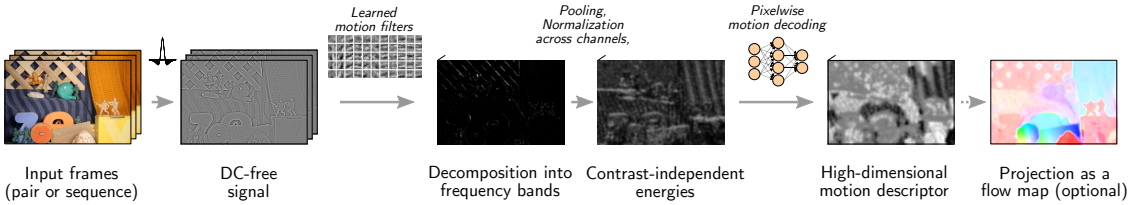

Figure 1. Our method is implemented as a shallow convolutional network. Spatiotemporal filters decompose the signal into energies of different frequency bands, then mapped to a motion descriptor with a pixelwise neural network. A number of operations and architecture choices are critical to handling the desired invariances, ensuring estimation of motion independent of image appearance and texture.

ing each input frame with a fixed center-surround filter. This removes the local DC and low-frequency components, which is desirable since the subsequent, learned, spatiotemporal filters may not have a null DC response. This bank of spatiotemporal filters is the core of the method. They decompose the video signal into frequency bands. In other words, they sparsely sample the spectrum of the Fourier transform of the video. We then enhance invariance to local image structure and contrast by L1-normalizing the filter responses. We emphasize that this simple operation is crucial to obtain reliable motion estimates, and could not easily be learned if not provisioned for in the network architecture. We also pool the measured energies spatially to account for invariance to phase and gradient orientation. Finally, we learn a mapping from these energies to a high dimensional motion descriptor, as pixelwise neural network. It uses a single hidden layer, and is trained together with the motion filters through backpropagation. The output layer comprises M units that form our final motion feature. Each of them outputs a non-negative speed in one orientation in $[0, 2\pi[$. This representation can thus capture multiple motions in different orientations, but can also be trivially projected onto a traditional flow map (during training or to recover unimodal flow maps at test time).

Experiments We trained our network on the Middlebury dataset with known optical flow, using sequences of 5 frames. Most of the learned spatiotemporal filters (Fig. 1) are similar to those used in hardcoded filter-based motion estimation methods (*e.g.* 3D Gabors). We show three of our preliminary experiments in Fig. 2. (1) We can accurately estimate rigid motions (compared as a flow map with the ground truth, on a scene of the Middlebury dataset). (2) We can detect the motion of phenomena that violate the assumptions of optical flow, as shown with a transparent, non-rigid cloud of steam. (3) The proposed high-dimensional descriptor can capture motions that go beyond translations and image displacements, as in the well-known "ocean-fire" sequence. We simply apply K-means (K=2) on the pixelwise feature vector, which reliably segments the two dynamic textures.

Perspectives The proposed convolutional architecture was validated on the estimation of simple, rigid motions, and shows promising potential for the analysis of complex scenes with transparencies and dynamic textures, which violate common assumptions of optical flow. The next step is to use the proposed network as a building block for deeper CNN architectures, addressing higher-level tasks such as activity recognition, scene recognition or semantic segmentation of dynamic textures. This design will allow to fine-tune the motion extraction together with the remaining of the network. It will be worth investigating whether retraining the motion estimation for different end-tasks may lead to different intermediate motion representations.

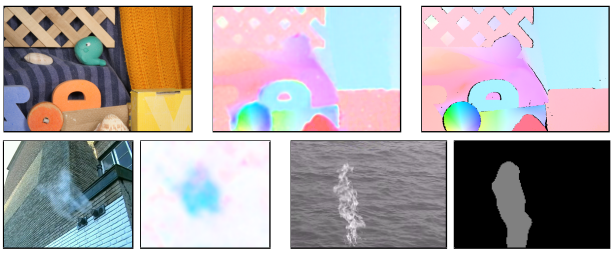

Figure 2. Preliminary experiments show reliable motion estimation in a wide range of scenes, including rigid motions (top), transparent phenomena (steam, bottom-left) and complex dynamic textures (fire/water, bottom-right).

[1] E. H. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A.*, 1985.

[2] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *PAMI*, 2012.

[3] D. J. Heeger. Model for the extraction of image flow. *J. Opt. Soc. Am. A*, 1987.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[5] B. Olshausen. Learning sparse, overcomplete representations of time-varying natural images. In *ICIP*, 2003.

[6] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How mt cells analyze the motion of visual patterns. *Nature Neuroscience*, 9, 2006.

[7] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *NIPS Spotlight*, 2014.

[8] F. Solari, M. Chessa, and P. Medathati, N. Kornprobst. What can we expect from a V1-MT feedforward architecture for optical flow estimation ? *Sig. Proc.: Imag. Comm.*, 2015.

[9] D. Teney and M. Brown. Segmentation of dynamic scenes with distributions of spatiotemporally oriented energies. In *BMVC*, 2014.