Making a Case for Learning Motion Representations with Phase

S. L. Pintea
Computer Vision Lab
Delft University of Technology
S.L.Pintea@tudelft.nl

J. C. van Gemert
Computer Vision Lab
Delft University of Technology
J.C.vanGemert@tudelft.nl

Abstract

This work advocates Eulerian motion representation learning over the current standard Lagrangian optical flow model. Eulerian motion is well captured by using phase, as obtained by decomposing the image through a complex-steerable pyramid. We discuss the gain of Eulerian motion in a set of practical use cases: (i) action recognition, (ii) motion prediction in static images, (iii) motion transfer in static images and, (iv) motion transfer in video. For each task we motivate the phase-based direction and provide a possible approach.

1 Introduction

We propose an Eulerian approach towards motion representation learning. The main difference between Lagrangian and Eulerian motion is that Lagrangian motion (optical flow) focuses on individual points and analyzes their change in location over time. Therefore, Lagrangian motion performs tracking of points over time and for this it requires a unique matching method between point or patches. On the other hand, Eulerian motion considers a set of locations in the image and analyzes the changes at these locations over time. Thus, Eulerian motion does not estimate where a given point moves to, instead, it measures flux properties. Figure 1 depicts this difference between Eulerian and Lagrangian motion. As a specific instance of the Eulerian model, we consider phase-based motion. The phase variations over time of the coefficients of the complex-steerable pyramid are indicatives of motion [9] and form the basis for learning motion representations.

The gain of an Eulerian motion approach is that it avoids the need for hand-crafted optical flow constructions. Phase is an innate property of the image, it does not need to be estimated from explicit patch correspondences. We propose a general-purpose phase-based motion description learning setup that can be used in any task relying on motion. Here we explore four use cases: (i) action recognition, (ii) motion prediction in static images, (iii) motion transfer in static images and, (iv) motion transfer in video. Note that phase-based motion representations are readily applicably to other motion-related task as well, including: human gait analysis, object tracking, action localization, etc.

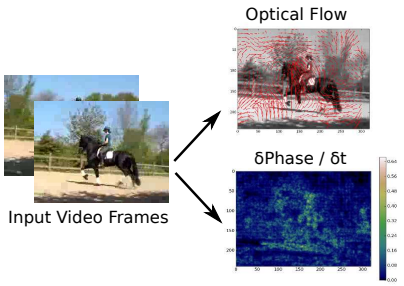


Figure 1: While Lagrangian motion (optical flow) estimates the changes in position over time, it can miss correspondences or find mistaken correspondences. However, in the Eulerian approach (phase variations over time) the number of motion measurements stays constant between frames, as for each input image we analyze the phase variations over time at each image location over multiple orientations and scales.

2 Related Work

2.1 Eulerian Motion

Eulerian motion modeling has shown remarkable results for motion magnification [31] where a phase-based approach significantly improves the quality [28] and broadens its application [1, 16]. A phase-based video interpolation is proposed in [18] and a phase-based optical flow estimation is proposed in [13]. Inspired by the these work, we advocate the use of the Eulerian model as exemplified by phase for learning motion representations.

2.2 Action Recognition

Optical flow-based motion features have been extensively employed for action recognition in works such as [14, 26, 19, 30]. These works, use hand crafted features extracted from the optical flow. Instead, we propose to input phase-based motion measurements to a CNN to reap the benefits of deep feature representation learning methods.

A natural extension of going beyond a single frame in a deep net is by using $3D$ space-time convolutions [15, 25]. $3D$ convolutions learn appearance and motion jointly. While elegant, it makes it difficult to add the wealth of information that is available for appearance-only datasets through pre-training. In our method, we keep the benefit of pre-training by separating the appearance and the phase-based motion streams.

Using pre-trained networks is possible in the two-stream network approaches proposed in [5, 7, 23]. This combines a multi-frame optical flow network stream with an appearance stream and obtains competitive results in practice. The appearance stream can employ a pretrained network. Similarly, we also consider the combination of appearance and motion in a two-stream fashion, but with innate phase information rather than using a hand-crafted optical flow.

The temporal frame ordering is exploited in [8], where the parameters of a ranking machine are used for video description. While in [6, 17, 24] recurrent neural networks are proposed for improving action recognition. In this paper we also model the temporal aspect, although we add the benefit of a two-stream approach by separating appearance and phase variation over time.

2.3 Motion Prediction

In [20], optical flow motion is learned from videos and predicted in static images in a structured regression formulation. In [29] the authors propose predicting optical flow in a CNN from input static images. Where these works predict optical flow, we propose to predict the motion through phase changes, which does not depend on pixel tracking.

Predicting the future RGB frame from the current RGB frame is proposed in [27] in the context of action prediction. Similar to this work, we also start from an input appearance and obtain an output appearance image, however in our case the learning part learns the mapping from input phase information to future phase.

2.4 Motion Transfer

Animating a static image by transferring the motion from an input video is related to the notion of artistic style transfer [11, 12, 21]. The style transfer aims at changing an input image or video such that the artistic style matches the one of a provided target image. Here, instead, we consider the motion transfer — given an input image, transfer the phase-based motion from the video to the image.

Additionally, we also consider video-to-video transfer where the style of performing a certain action is transferred from a target video to the input video. In [2] the authors allow the users to change the video by adding plausible object manipulations in the video. Similar to this work, we also want to change the video motion after the recording is done, by adjusting the style of the action being performed.

3 Learning Motion with Phase

The local phase and amplitude of an image are measured by complex oriented filters of the form: $G^\theta_\sigma + iH^\theta_\sigma$, where θ is the filter orientation and σ the filter scale [10],

$$(G^\theta_\sigma + iH^\theta_\sigma) \otimes I(x,y) = A^\theta_\sigma(x,y)e^{i\phi^\theta_\sigma(x,y)}, \qquad (1)$$

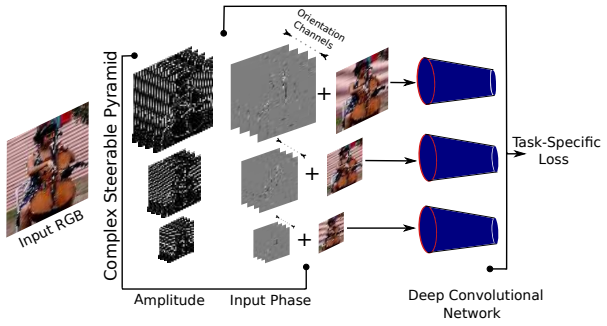


Figure 2: Phase-based representation learning: from an input RGB image we extract phase information over multiple orientations and scales by employing complex steerable filters. For each scale, additional to the RGB input, we add the orientated phases as input to a network stream that optimizes a task-specific loss.

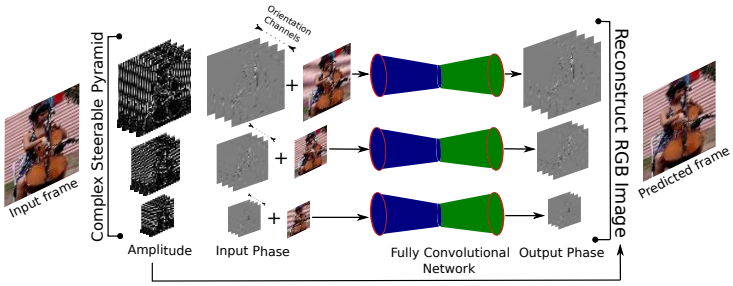
where $\phi^\theta_\sigma(x,y)$ is the local phase at scale σ and orientation θ, and $A^\theta_\sigma(x,y,t_0)$ the amplitude, $I(x,y)$ is the image brightness/input channel, and $\otimes$ the convolution operator, and x, y are image coordinates. The filters have multiple scales and orientations, forming a complex steerable pyramid [22] which captures various levels of image resolution.

There is a direct relation between motion and the change measured in phase over time. The Fourier shift theorem makes the connection between the variation in phase of the subbands over time and the global image motion. Rather than estimating global motion, using a steerable pyramid we can decompose the image into localized subbands and thus, recover the local motion in the phase variations over time. From the above decomposition only the phase, not the amplitude, corresponds to motion. In [9] the authors show that the temporal gradient of phase computed from a spatially bandpassed video over time, directly relates to the motion field. Therefore, here, we focus on local phase at multiple scales and orientations to represent motion.

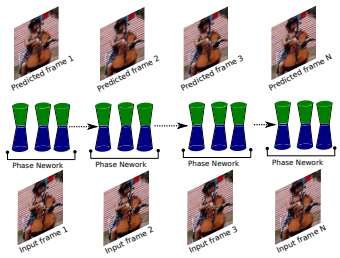
We propose using phase to learn motion representations for solving general motion-related tasks in a deep net. We add phase as an additional motion input channel to a standard appearance (RGB) convolutional deep neural network. Figure 2 shows our proposed general-purpose phase-based pipeline. The input video frame is decomposed using the complex steerable pyramid into amplitude and phase. Both phase and amplitude have multiple corresponding orientations and scales. Since the phase is an indicative of motion, we ignore the amplitude and we use the input phase for the motion representation learning. We treat the orientations as input channels while the scales represent different streams of the network, similar to [4] who use this setup for a different image pyramid.

4 Four Use Cases in Motion Learning

We explore phase-based motion representation learning in four practical use cases. While a thorough in-depth experi-



(a) Phase-motion prediction.



(b) Long term phase-based motion prediction.

Figure 3: (a) Phase prediction in a Phase Network: from an input RGB image, we estimate the phase along multiple scales and orientations. For each scale we train a Fully Convolutional Network that predicts oriented phase at a future time-step. From this we recover the predicted future RGB image. (b) Long-term motion prediction in static images: given the one step convolutional mapping from the input RGB image to the future RGB image, defined in the 'Phase Network', combine multiple of these networks in an Recurrent Neural Network to obtain plausible long-term phase predictions.

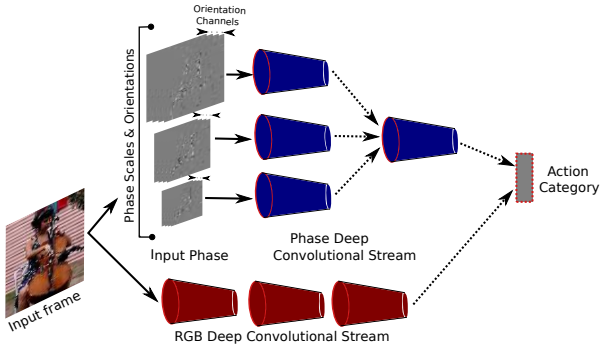


Figure 4: Action recognition approach: two-stream CNN where the first stream receives input RGB frames, while the second stream receives input oriented phases of the video frame over multiple scales that are subsequently combined.

mental investigation is out of scope, we detail the setup of motion representation learning for each use case.

4.1 Phase-based Action Recognition

Separating appearance and motion in two-streams is effective for action recognition [23]. For the appearance stream we follow [23] and use the input RGB frame, which offers the advantage of pre-training features on static images. However, where [23] uses hand-crafted optical flow features, we propose to use Eulerian motion for the second stream with oriented phase over multiple scales, as depicted in figure 4.

For evaluating action recognition, a comparison of our two-stream phase-based motion with the two-stream optical-flow approach of [23] on the two datasets used in their paper — HMDB51 and UCF101 is needed. We expect benefits from a phase-based motion representation because it does not depend on a specific hand-crafted optical flow implementation and does not rely on pixel tracking.

4.2 Phase-based Motion Prediction in Static

The benefit of Eulerian motion for motion prediction is that the prediction locations are fixed over time. This contrasts sharply with Lagrangian motion, as pixels tracked by optical flow may be lost as they move in or out of the frame, or move to the same spatial location. Such lost pixels make it hard to recover long-term relations beyond just the next frame. The fixed prediction locations of a Eulerian motion representation do not suffer from this and offer long-term relation predictions of several frames.

We propose to learn from a given input RGB the output future RGB, by recovering from the RGB the phase scales and orientations, then predicting the multi-scale future phase-orientations and transforming them back into future RGB frames as in figure 3.(a). For long-term motion prediction we propose an RNN (Recurrent Neural Network) version of this phase-based frame prediction, as depicted in figure 3.(b). Thus, predicting motion N timesteps away from the input.

For evaluating motion prediction we use the same datasets as in [29] — HMDB51 and UCF101, where the authors aim at predicting optical-flow based motion in single images. To evaluate the difference between the predicted motion and the actual video motion, we use pixel accuracy, as in our method we recover the appearance of the future frame. For comparison with [29] which reports EPE (End Point Errors), we use their chosen optical flow estimation algorithm to recover optical flow from our predicted RGB.

4.3 Phase-based Motion Transfer in Images

Similar to [12, 21], where the style of a given target painting is transferred to another image, we propose to transfer the short motion of a given video sequence to an input static image. In [12] a combination of two losses is optimized: content loss which ensures that the objects present in the newly generated image remain recognizable and correspond to the ones in the input image, and a style loss which imposes that the artistic style of the new image is similar to the one of the provided target painting. For motion transfer we have an additional requirement, namely that parts of the image that are similar — e.g. horses, people, should move sim-

ilar. For this we use two pretrained network streams, an RGB stream and a phase stream and consider certain convolutional layers along these streams for estimation RGB/ phase responses. Therefore, we first estimate an element-wise correlation between the responses at a given convolutional network layer of the input RGB values of the static image and the target video frame:

$$\mathcal{K}_j^l = \frac{\sum_i^{N_l} C_{ij}^l D_{ij}^l}{\sqrt{\sum_i^{N_l} {C_{ij}^l}^2} \sqrt{\sum_i^{N_l} {D_{ij}^l}^2}}, \quad (2)$$

where N_l is the number of channels in the layer l, C^l and D^l the responses at layer l for the input image and video frame, respectively. Following [12], we subsequently define our motion-style loss by weighting the feature maps in the Gram matrix computation by the appearance correlation. The motion transfer is obtained by enforcing that the phase of objects over time in the input image, should be similar to the phase over time of the same objects present in the target video. The motion-style loss optimization is performed per phase-scale.

$$G_\sigma^l[ij] = \sum_k^{M_l} \mathcal{K}_k^l F_\sigma^l[ik] F_\sigma^l[kj],\ i,j \in \{1,..N_l\}, \quad (3)$$

$$A_\sigma^l[ij] = \sum_k^{M_l} \mathcal{K}_k^l P_\sigma^l[ik] P_\sigma^l[kj],\ i,j \in \{1,..N_l\}, \quad (4)$$

$$\mathcal{L}_l = \sum_\sigma \frac{1}{N_l^2 M_l^2} \sum_{i,j}^{N_l} \mathcal{K}_j^l (G_\sigma^l[ij] - A_\sigma^l[ij])^2, \quad (5)$$

where M_l is the number of elements in one channel of layer l, σ indicates the phase-scale, and G^l is the weighted Gram matrix of the phase-image to be generated, while A^l is the weighted Gram matrix of the current video frame and, F^l and P^l are the responses of the phase-image to be generated and the phase-image of the input video frame, respectively.

Because we want the find similar looking objects by using the element-wise correlations, we expect that the higher convolutional levels of the network will perform better. We additionally also add the content loss term of [12] to avoid large distortions of the image appearance. Due to the input being a static image, only short video motions can be transferred in this case.

For evaluating motion transfer, we perform a two-step evaluation. In the first step, we select an existing video frame and transfer the video motion to the selected frame and compare the transferred motion with the actual video motion. For this we use videos from HMDB51 and UCF101. The second evaluation is transferring the motion to actual static images. For this we select images from the static Willow dataset [3] and transfer the motion of corresponding videos from the HMDB51 and UCF101 datasets containing the same objects. For this we provide the static images animated with the transferred video motion.

4.4 Phase-based Motion Transfer in Videos

We use as a starting point the work of [21], where artistic style is transferred to video. However in our case, the motion of one given video is transferred to another input video. The gain in so doing, is that we can transfer the style of performing a certain action. For example an amateur performing the moonwalk can be lifted to the expert level by transferring the motion of Michael Jackson himself.

The idea of transferring motion in videos is similar to the idea of transferring motion in static images, with the additional constraint that the motion must be temporally coherent. For this, similar to [21], we add a temporal loss term to the motion transfer loss discussed in section 4.3.

For performing motion transfer between videos, we use a set of target videos: the walk of Charlie Chaplin, the moonwalk of Michael Jackson, and the walk of a runway model. We transfer these walking styles to a set of input videos of people walking, and provide the results as a qualitative form of evaluation.

4.5 Preliminary Proof of Concept

Here[1], we show a very simple proof of concept for phase-based motion transfer. We animate a static image by transfering the motion of another semantic related video. Correctly aligning the moving entities between the video frames and the static image is essential for this task. For this proof of concept the alignment was not very good and no learning was used whatsoever. Misalignment errors show up as artifacts in the results and we expect that adding (deep) learning will improve results.

5 Conclusions

We propose an Eulerian –phase-based– approach to motion representation learning. We argue for the intrinsic stability offered by the phase-based motion description. A phase-based approach does not require pixel tracking and directly encodes flux. Phase is an innate property of an image and does not rely on hand-crafted optical-flow algorithms. We explore a set of motion learning tasks in an Eulerian setting: (a) action recognition, (b) motion prediction in static images, (c) motion transfer from a video to a static image and (d) motion transfer in videos. For each one of these tasks we propose a phase-based approach and provide a small proof of concept. We do not offer in-depth experimental results but instead make a case for a brave new motion representation with phase.

[1] Demo: http://silvialaurapintea.github.io/motion_transfer/index.html .

Acknowledgments. This work is part of the research programme Technology in Motion (TIM [628.004.001]), financed by the Netherlands Organisation for Scientific Research (NWO).

References

[1] J G Chen, N Wadhwa, Y J Cha, F Durand, W T Freeman, and O Buyukozturk. Modal identification of simple structures with high-speed video using motion magnification. *Journal of Sound and Vibration*, 345:58–71, 2015.

[2] A Davis, J G Chen, and F Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM-TOG*, 34(6):239:1–239:7, 2015.

[3] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[4] E L Denton, S Chintala, R Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, pages 1486–1494, 2015.

[5] A. Diba, A. Mohammad Pazandeh, and L. Van Gool. Efficient two-stream motion and appearance 3d cnns for video classification. *arXiv preprint arXiv:1608.08851*, 2016.

[6] J Donahue, L A Hendricks, S Guadarrama, M Rohrbach, S Venugopalan, K Saenko, and T Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. *arXiv preprint arXiv:1604.06573*, 2016.

[8] B Fernando, E Gavves, J M Oramas, A Ghodrati, and T Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, pages 5378–5387, 2015.

[9] D J Fleet and A D Jepson. Computation of component image velocity from local phase information. *IJCV*, 5(1):77–104, 1990.

[10] W T Freeman and E H Adelson. The design and use of steerable filters. *PAMI*, 13(9):891–906, 1991.

[11] L A Gatys, M Bethge, A Hertzmann, and E Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.

[12] L A Gatys, A S Ecker, and M Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[13] T Gautama and M Van Hulle. A phase-based approach to the estimation of the optical flow field using spatial filtering. *TNN*, 13(5):1127–1136, 2002.

[14] M Jain, H Jegou, and P Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, pages 2555–2562, 2013.

[15] S Ji, W Xu, M Yang, and K Yu. 3d convolutional neural networks for human action recognition. *PAMI*, 35(1):221–231, 2013.

[16] J FP Kooij and J C van Gemert. Depth-aware motion magnification. In *ECCV*, 2016.

[17] Z Li, E Gavves, M Jain, and C G M Snoek. Videolstm convolves, attends and flows for action recognition. *arXiv preprint arXiv:1607.01794*, 2016.

[18] S Meyer, O Wang, H Zimmer, M Grosse, and A Sorkine-Hornung. Phase-based frame interpolation for video. In *CVPR*, pages 1410–1418, 2015.

[19] D Oneata, J Verbeek, and C Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, pages 1817–1824, 2013.

[20] S L Pintea, J C van Gemert, and A W M Smeulders. Déja vu: Motion prediction in static images. In *ECCV*, pages 172–187, 2014.

[21] M Ruder, A Dosovitskiy, and T Brox. Artistic style transfer for videos. *arXiv preprint arXiv:1604.08610*, 2016.

[22] E P Simoncelli, William T Freeman, E H Adelson, and D J Heeger. Shiftable multiscale transforms. *Transactions on Information Theory*, 38(2):587–607, 1992.

[23] K Simonyan and A Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576. 2014.

[24] N Srivastava, E Mansimov, and R Salakhutdinov. Unsupervised learning of video representations using lstms. *CoRR, abs/1502.04681*, 2, 2015.

[25] D Tran, L Bourdev, R Fergus, L Torresani, and M Paluri. C3d: generic features for video analysis. *CoRR, abs/1412.0767*, 2:7, 2014.

[26] J van Gemert, M Jain, E Gati, and C Snoek. Apt: Action localization proposals from dense trajectories. In *BMVC*, volume 2, page 4, 2015.

[27] C Vondrick, H Pirsiavash, and A Torralba. Anticipating the future by watching unlabeled video. *arXiv preprint arXiv:1504.08023*, 2015.

[28] N Wadhwa, M Rubinstein, F Durand, and W T Freeman. Phase-based video motion processing. *ACM-TOG*, 32(4):80, 2013.

[29] J Walker, A Gupta, and M Hebert. Dense optical flow prediction from a static image. In *ICCV*, 2015.

[30] L Wang, Y Qiao, and X Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015.

[31] H Y Wu, M Rubinstein, E Shih, J Guttag, F Durand, and W T Freeman. Eulerian video magnification for revealing subtle changes in the world. *SIGGRAPH*, 31(4), 2012.