# Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition

Liangjie Cao

Aug. 5, 2018

## Abstract

*Motion representation plays a vital role in human action recognition in videos. In this paper, the authors introduce a novel compact motion representation for video action recognition, named Optical Flow guided Feature (OFF), which enables the network to distill temporal information through a fast and robust approach. The OFF is derived from the definition of optical flow and is orthogonal to the optical flow. The derivation also provides theoretical support for using the difference between two frames. By directly calculating pixel-wise spatio-temporal gradients of the deep feature maps, the OFF could be embedded in any existing CNN based video action recognition framework with only a slight additional cost. It enables the CNN to extract spatiotemporal information, especially the temporal information between frames simultaneously. This simple but powerful idea is validated by experimental results. The network with OFF fed only by RGB inputs achieves a competitive accuracy of 93.3% on UCF-101, which is comparable with the result obtained by two streams (RGB and optical flow), but is 15 times faster in speed. Experimental results also show that OFF is complementary to other motion modalities such as optical flow. When the proposed method is plugged into the state-of-the-art video action recognition framework, it has 96.0% and 74.2% accuracy on UCF-101 and HMDB-51 respectively. The code for this project is available at:* `https://github.com/kevin-ssy/Optical-Flow-Guided-Feature`.

## 1. Introduction

Video action recognition has received longstanding attentions in the community of computer vision for decades. It aims at automatically recognizing human action from video sequences. Since CNNs have achieved great successes in image classification and other related tasks [2, 3, 4, 5], lots of CNN based methods have been proposed by considering video action recognition as a classification task. Compared to the image classification methods, temporal information is the key ingredient of video action recognition.

In this paper, they define a new feature representation from the orthogonal space of optical flow on the feature level [1]. Such definition brings the guidance from optical flow here to the representation, therefore, we name it as the Optical Flow guided Feature (OFF). The feature consists of spatial gradients of feature maps in horizontal and vertical directions, and temporal gradients obtained from the difference between feature maps from different frames. Since all the operations in OFF are differentiable, the whole process is end-to-end trainable when OFF is plugged into one CNN architecture. Actually the OFF unit only consists of pixel-wise operators on CNN features. These operators are fast to apply, and enable the network with RGB input to capture spatial and temporal information simultaneously.

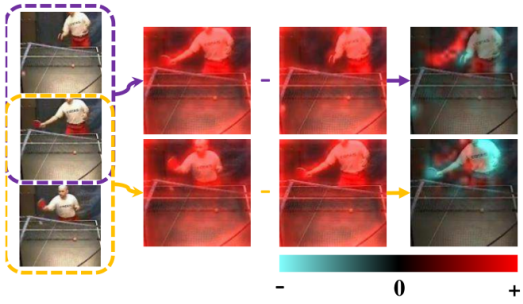


Figure 1. EThe Optical Flow guided Feature

One vital component in OFF is the difference between features from different images/segments. As shown in Fig 1, the difference between the features from two images provides representative motion information that can be conveniently employed by CNNs. The negative values in the difference image depict the locations where the body parts/objects disappear, while the positive values represent where they emerge. This pattern of disappearing at one location and emerging at another location can be easily treated as a specific motion pattern and captured by later CNN layers. The temporal difference could be further combined

with the spatial gradients such that the constituted OFF is guided by the optical flow on feature level according to their derivation in later section. Moreover, calculation of the motion dynamics at the feature level is faster and also more robust because 1) it enables the spatial and temporal networks with the capability of weight sharing and 2) deeply learned features convey more semantic and discriminative representations with reliable elimination of local and background noises in the raw frames.

## 2. Optical Flow Guided Feature

Their proposed OFF is inspired by the famous brightness constant constraint defined by traditional optical flow [1]. It is formulated as follows:

$$I(x, y, t) = I(x, + \triangle x, y + \triangle y, t + \triangle t) \tag{1}$$

where $I(x, y, t)$ denotes the pixel at the location $(x, y)$ of a frame at time t. For frames t and $(t + \triangle t)$, $\triangle x$ and $\triangle y$ are the spatial pixel displacement in $x$ and $y$ axes respectively. It assumes that for any point that moves from $(x, y)$ at frame t to $(x + \triangle x, y + \triangle y)$ at frame $t + \triangle t$, its brightness keeps unchanged over time. When we apply this constraint at the feature level, they have

$$f(I; W) = (x, y, t) = f(I; W)(x + \triangle x, y + \triangle y, t + \triangle t) \tag{2}$$

Previous works have shown that the temporal difference between frames is useful in video related tasks, however, there is no theoretical evidence to help explain why this simple idea works that well. Here, we can find its correlation to spatial features and optical flow.

## References


[1] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 1981. 1, 2

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1

[4] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, et al. Crafting gbd-net for object detection. *TPAMI*, 2017. 1

[5] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 1