論文の要旨

題目　Spatio Temporal Features and its Possible Extensions for Action Recognition: from Handcrafted to Deep Learning
（動作認識のための時空間特徴とその拡張--職人技からディープラーニングへ--）

氏名　Novanto Yudistira

Action Recognition contains information over space and time because action possibly occurs in arbitrary positions, various scales and temporal dynamics which led to the need of robust yet low computational cost features. The progress of action recognition or video classification as broader topic has largely progressed given abundance of common datasets. However, there are still rooms to improve recent features in which still questionable ranging from handcrafted to learned features such as spatio temporal auto correlation, multi layered wavelet packet, motion superpixel localization, and mixture expert via deep Convolutional Neural Network (CNN). Results show that it is either improving state of the art or computationally efficient compared to the existing features.

In Chapter 2, We propose method for fast action recognition and comparable performance using local autocorrelation of optical flows over time. To capture action movement, dense optical flows is generated along sequence of video. Optical flows sometimes yield noise of motions that distract object of interest from another object motions and background. We suppress this by using edge based optical flow. The HOF vector is extracted from each window resolution and correlate its consecutive flow fields within cycle using local autocorrelation over time. It will gather richer information from movement while also gaining discriminative features than standard histogram methods. Comparison shows that the comparable performance is achieved over state of the arts.

In Chapter 3, Action recognition with dynamic actor and scene has been a tremendous research topic. Recently, spatio temporal features such as optical flows have been utilized to define motion representation over sequence of time. However, to increase accuracy, deep decomposition is necessary either to enrich information under location or time varying actions due to spatio temporal dynamics. To this end, we propose algorithm consists of vectors obtained by applying multi- resolution analysis of motion using Haar Wavelet Packet (HWP) over time. Its computation efficiency and robustness have led HWP to gain popularity in texture analysis but their applicability in motion analysis is yet to be explored. To extract representation, a sequence of bin of Histogram of Flow (HOF) is treated as signal channel. Deep decomposition is then applied by utilizing Wavelet Packet decomposition called Packet Flow to many levels. It allows us to represent action's motions with various speeds and ranges which focuses not only on HOF within one frame or one cuboid but also on the temporal sequence. HWP, however, has translation covariant property that is not efficient in performance because actions occur in arbitrary time and sampling's location is various. To gain translation invariant capability, we pool each respective coefficient of decomposition for each level. It is found that with proper packet selection, it gives comparable results on the KTH action and Hollywood dataset with train-test division without localization. Even if spatiotemporal cuboid sampling is not densely sampled like

of baseline method, we achieve lower complexity and comparable performance on camera motion burdened dataset like UCF Sports that usually motion features such as HOF do not perform well.

In Chapter 4, Superpixels are a representation of still images as pixel grids because of their more meaningful information compared with atomic pixels. However, their usefulness for video classification has been given little attention. In this paper, rather than using spatial RGB values as low-level features, we use optical flows mapped into hue-saturation-value (HSV) space to capture rich motion features over time. We introduce motion superpixels, which are superpixels generated from flow fields. After mapping flow fields into HSV space, independent superpixels are formed by iteration of seeded regions. Every grid of a motion superpixel is tracked over time using nearest neighbors in the histogram of flow (HOF) for consecutive flow fields. To define the temporal representation, the evolution of three features within the superpixel region, namely the HOF, the center of superpixel mass, and the neighborhood correlation, are used as descriptors. The bag of features algorithm is used to quantify final features, and generalized histogram-kernel support vector machines are used as learning algorithms. We evaluate the proposed superpixel tracking on first-person videos and action sports videos.

In Chapter 5, Activity recognition requires visual and temporal cues making it challenging to integrate these important clues. The usual schemes of integration are averaging and fixing the weights of both features for all samples. However, how much weight it needs for each sample and modality, is still an open question. A mixture of experts via gating Convolutional Neural Network (CNN) is one of the promising architectures to adaptively weigh every sample within a dataset. In this paper, rather than just averaging or fixed weights, we investigate how natural associative cortex like network integrates expert's networks to form of gating CNN scheme. Starting from Red Green Blue (RGB) and optical flows, we show that with proper treatment, gating CNN scheme actually works and sheds a light on information integration in future for activity recognition.