Feature Extraction and Human-Human Interaction Recognition for Video Surveillance

Nguyen Thuy Ngoc
School of Information Science,
Japan Advanced Institute of Science and Technology

September, 2016

Abstract

Human interaction recognition has been widely studied because it has great scientific importance and many potential practical applications. However, this problem is very challenging especially in realistic environments where background is dynamic and has varying lighting conditions. This dissertation addresses human activity recognition, especially human-human interactions in realistic video material, such as movies, surveillance videos. For classification problem, most existing methods rely on either spatio-temporal local features (i.e. SIFT) or human poses, or human joints to model human interactions. As a result, they are not fully unsupervised processes because they require either hand-designed features or human detection results.

Motivated by the recent success of deep learning networks, we investigate a three-layer convolutional network which uses the Independent Subspace Analysis (ISA) algorithm to learn hierarchical invariant features. The ISA algorithm is a generalization of the Independent Component Analysis (ICA), which is very well-known in natural image statistics. Compared to the ICA algorithm, the most notable advantage of the ISA is that it can learn features which are invariant to phase while being selective to orientation and frequency. However, the ISA algorithm becomes slow when applying it on video data. In order to solve this computational problem, we combine the idea of convolutional neural network with the ISA algorithm. Specifically, instead of training the ISA algorithm directly on raw video data, we first train it on small video blocks extracted by our procedure. The obtained features are then convolved with larger video blocks. The outputs of this convolution step are fed into the next layer, which is implemented by another ISA algorithm. This organization enables the three-layer convolutional ISA network to learn hierarchical invariant features. Furthermore, we introduce a pooling layer to reduce the contributions of features learned in lower layers while still achieving translation invariant. Using the invariant features learned by the three-layer convolutional ISA network, we build a bag-of-features representation for videos. Finally, we apply Support Vector Machine (SVM) to classify human interactions. For temporal localization, we slide temporal detection windows with different durations over a continuous video sequence with a stride of 10 frames. For each temporal window, our convolutional ISA network extracts hierarchical invariant features on a dense grid. After scoring the temporal detection windows, a non-maximum suppression is applied to enforce that non of the retained windows are overlapping.

In all two cases, we conduced thorough experiments on realistic videos from challenging benchmarks used by activity recognition community. We show that our three-layer convolutional ISA network is effective to represent complex activities such as human interactions in realistic environments. Besides, we believe that our temporal localization method is the first work which reports experimental results on the continuous video sequences of human interactions. Although temporal localization results are insufficient for real applications, it is a first step for further research in localization of human interactions.

Key Words: **temporal localization, classification, independent subspace analysis, human-human interactions, convolutional network, pooling**