Recently, bag of visual words (BoVW) representation, in which the image patches or video cuboids are quantized into visual-words (VWs) based on their appearance similarity, has been widely and successfully explored. The advantages of this model are that no explicit detection of object parts and their tracking are required, and it is efficient for matching. But, the performance of BoVW is sensitive to the vocabulary size. This limitation is partially due to the fact that the VWs are not semantically meaningful. To overcome this shortcoming, in this thesis we present a principled approach to learn a semantic vocabulary (SV) from a large amount of VWs. In this context, the thesis makes two major contributions.

First, we have developed an algorithm to discover a compact yet discriminative SV. This SV is obtained by grouping the VWs based on their distribution in videos (images) into VW clusters (VWCs). The mutual information (MI) between the VWCs and the videos depicts the discriminative power of the SV, while the MI between VWs and VWCs measures the compactness of SV. We apply the information bottleneck (IB) algorithm to find the optimal number of VWCs by finding a good tradeoff between discriminative power and compactness. Our approach achieved average accuracy of 94.2% (cs.ucf.edu/~liujg/mmi_action.html) on the KTH dataset. However, this approach performs one-side clustering, because only VWs are clustered regardless of which video they appear in. In order to leverage the co-occurrence of VWs and images, we have developed the co-clustering algorithm to simultaneously group the VWs and images. We tested our approach on the publicly available fifteen scene dataset and have obtained about a 4% increase in the average accuracy compared to the one side clustering approaches (cs.ucf.edu/~liujg/co_clustering.html).

Second, instead of grouping the VWs, we first embed them into a low-dimensional semantic space by manifold learning, and then perform the clustering. We then apply Diffusion Maps (DM) to capture the local geometric structure of the VWs space. The DM can preserve the explicitly defined diffusion distance, which reflects the semantic similarity between any two VWs. Furthermore, the DM provides multi-scale analysis capability by adjusting the time steps. The experiments on the KTH dataset show that DM can perform much better (about 3% ~ 6% improvement in average accuracy) than other manifold learning approaches and IB method (cs.ucf.edu/~liujg/dm.html). Previous methods use only single types of features. In order to combine multiple heterogeneous features for visual recognition, we further propose the Fielder Embedding to capture the complicated semantic relationships between all entities (i.e., videos, images, heterogeneous features). The discovered relationships are then employed to further increase the recognition rate. We tested our approach on the Weizmann dataset, and achieved about 17% ~ 21% improvements in the average accuracy (cs.ucf.edu/~liujg/action_multiple_features.html).
The public is welcome to attend.