Experimenting Motion Relativity for Action Recognition with a Large Number of Classes

Feng Wang, Xiaoyan Li, and Wenmin Shu
East China Normal University
500 Dongchuan Rd., Shanghai, China

Abstract

In this paper, we present our approach and experiments for human action recognition on UCF101 dataset. In our previous work [9], we have proposed motion relativity for video event detection. In this work, we mainly experiment the performance of motion relativity feature for recognizing a specific set of events, i.e. human actions.

1. Feature Extraction

We submitted one run to the action recognition task, where three features are used: SIFT, STIP and ERMH (Expanded Relativity Motion Histogram of Bag-of-Visual-Words) proposed in [9]. For SIFT descriptors, Difference of Gaussian (DoG) [1] and Hessian Affine [2] detectors are used to detect local interest points, and 128-dimension SIFT feature [5] is extracted to describe each local image patch. A visual vocabulary of 5000 visual words is generated by clustering the SIFT descriptors with k-means algorithm. Given an image, each detected keypoint is then mapped to the three nearest visual words to form the BoW histogram. For STIP descriptors, we directly make use of the features provided by UCF101 dataset [3], where 4000 words are used to calculate the histogram for each video.

In our previous work [9] [10], ERMH-BoW was used for video event detection and achieved encouraging results. The motivation of ERMH-BoW feature is to employ motion relativity to describe the behaviors and interactions between different objects/scenes. Considering that object segmentation remains difficult in unconstrained videos, we employ visual words to capture what objects/scenes are present in the video (or event). The relative motion between different visual words are then computed to capture the interactions between objects/scenes. In the following, we first briefly recall the algorithm for extracting ERMH-BoW feature and then employ it for human action recognition.

Given two visual words a and b, the relative motion histogram between them is calculated as

$$R_i(a, b) = \sum_{r \in N_a, t \in N_b} D_i(m_r - m_t)$$  \hspace{1cm} (1)

where r and t are keypoints mapped to visual words a and b respectively, $m_r - m_t$ is the relative motion of r with reference to t, and $D_i(.)$, $i = 1, 2, 3, 4$ decomposes the relative motion vector to the four directions (left, right, up and down) to generate a 4-directional histogram between visual words a and b. By Equation (1) the motion information in a video clip is represented as an $S \times S$ matrix $R$, where $S$ is the size of the visual vocabulary, and each element $R(a, b)$ is a relative motion histogram between the two visual words a and b. We call this feature matrix Relative Motion Histogram of BoW (RMH-BoW).

To further address the well-known visual word ambiguity problem with BoW approach [6] [7], we estimate the visual relatedness between each pair of visual words by employing the approach in [7]. RMH-BoW is then expanded by diffusing the relative motion histograms between two visual words to their correlated visual words. The Expanded Relative Motion Histogram of BoW (ERMH-BoW) is calculated as

$$E(a, b) = R(a, b) + \sum_{s_a, s_b} J(s_a, a) \times R(s_a, s_b) \times J(s_b, b)$$  \hspace{1cm} (2)

where $J(\cdot)$ is the relatedness between two words, $\{s_a\}$ and $\{s_b\}$ are the sets of visual words that are correlated to the words a and b respectively. The aim of RMH-BoW expansion is to alleviate the problem of visual word ambiguity. More specifically, the relative motion between two words are diffused by the influence of other words that are related to them. The diffusion inherently results in the expansion of RMH-BoW to facilitate the utilization of word-to-word correlation for video clip comparison. In our experiments, for each visual word, we empirically choose the five most similar words for diffusion in Equation (2). On one hand, this guarantees the efficiency of the RMH-BoW expansion process; on the other hand, diffusing with more visual words...
does not promise better performance. Similar to BoW feature representations, we call each element in ERMH-BoW feature a motion word which encodes the relative motion pattern between two visual words. The resulting ERMH-BoW feature counts the presence of different motion words in the given video clip.

In RMH-BoW, to employ the motion relativity for event description, the relative motion between each pair of visual words is computed. This results in a sparse matrix with high dimensionality. By word expansion, the problem becomes even worse since the sparsity of the feature is reduced. Eventually, it is space-consuming to store the ERMH-BoW features and time-consuming to train and test classifiers due to the curse of dimensionality. In our approach, we then employ Information Gain and another approach in [11] to weight the informativeness of each motion for action recognition. Eventually, the less useful motion words are removed and a cleaner set of features are obtained for classifier training.

For ERMH-BoW feature, a video is segmented into a few 5-second clips and a feature vector is computed over each clip. This results in a sequence of vectors along the timeline. The EMD kernel in [8] is then employed for SVM classification so as to capture the sequence information in the video.

2. Experiments

For experiment setup, we use the three train/test split of the UCF101 dataset [12]. Table 1 shows the recognition accuracies for 5 types of actions during our training stage. The overall accuracy of recognition of all 101 action classes are 50.07%. Among all classes, the best performance is achieved for Sports actions as these videos are the most motion intensive and motion plays the key role in recognizing this kind of actions. For Human-Object Interaction actions, some actions are quite difficult to recognize (such as Apply Eye Makeup, Apply Lipstick and Blow Dry Hair) since only little relative motion between human and objects is observed, which could be easily buried in the large amount of other action-irrelevant motion.

### Table 1. The recognition accuracies of five types of actions.

<table>
<thead>
<tr>
<th>Action Type</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Object Interaction</td>
<td>40.63</td>
</tr>
<tr>
<td>Body Motion Only</td>
<td>46.40</td>
</tr>
<tr>
<td>Human-Human Interaction</td>
<td>50.24</td>
</tr>
<tr>
<td>Playing Musical Instruments</td>
<td>45.74</td>
</tr>
<tr>
<td>Sports</td>
<td>55.87</td>
</tr>
</tbody>
</table>

3. Conclusion

In this work, we have presented and experimented our approach for action recognition on the UCF101 dataset. Although the ERMH-BoW feature was originally proposed for video event detection, our experiments show that it can be successfully applied to the specific set of event, i.e. human actions and achieve acceptable results.

Acknowledgement

The work described in this paper is partially sponsored by the National Natural Science Foundation of China (No. 61103127 and 61375016), Shanghai Pujiang Program (No. 12PJ1402700), SRF for ROCS, SEM, and the Fundamental Research Funds for the Central Universities.

References