Our system performs late fusion of several features whose weights have been optimized on UCF50 dataset. The fusion is done over the following features:

1) Our newly developed fast local descriptors for HoG and HoF, both grey-scale and RGB. In RGB-HoG/HoF we compute the dense HoG and HoF descriptors for all color channels and concatenate them. To obtain a single vector per video, we use the Fisher kernel with 256 centers using power normalization and L2 norm.

2) Global HoF features extracted for each video frame. The temporal variation has been modeled by using the Fisher Kernel representation, shown to be beneficial in [1].

Finally, we apply the late fusion over the outputs of these two methods on the UCF101 dataset.

1. Introduction

Nowadays, with the advent of large action recognition datasets, computational efficiency has become as important as the classification performance. In our work we address both the problems of computational efficiency and the performance accuracy. Our work is a late fusion of the outputs of two different methods. Each of them has its own novelty and advantages compared to the literature. The weights for our late fusion were determined on the UCF50 dataset [3].

2. Method explanation

2.1. Bag-of-words

We build our first contribution on top of the Bag-of-Words (BoW) representation using local visual descriptors. Typical Bag-of-words approaches for video action recognition rely on a two-steps paradigm. The first step concerns the generation of feature vectors: features are extracted and assigned according to a pre-defined codebook to form the so called bag-of-words (BoW). The second step takes these bags-of-words as input and learns how to classify the different actions. In the Action recognition area usually Histograms of Oriented Gradient (HoG) and Histograms of Optical Flows (HoF) descriptors are among the common popular descriptors used in this pipeline.

In the first step we extract the 3D volumes of Histograms of Oriented Gradients (HoG) and Histograms of Optical Flow (HoF). In this phase we significantly speed up the descriptor extraction phase taking inspiration from Uijlings et al [5]: we exploit the nature of densely sampled descriptors in order to speed up their computation[1]. In the second step, for the visual word assignment instead of applying Kmeans, the typical method used in the literature, we apply the Fisher Kernel representation [2] as a more accurate alternative. In order to reduce the feature dimension for speeding up the computation, we apply the PCA to our descriptors. Specifically, we use the following descriptors: 3D-(HoG, HoF, RGB-HoG and RGB-HoF), which we combine in a late fusion fashion.

2.2. Modeling Temporal Variation

Global Features in videos are usually used for computational efficiency, where each global feature captures information of a single video frame. Typically, the representations in the state of the art ignore the temporal variation and in our work, we propose a solution to this problem. Based on [1], we meaningfully aggregate the global features taken from single frames of videos over time. We do this by using the Fisher Kernel. This way frame-based features are aggregated while their variation in time is preserved: (1) dissimilar frames will be represented by different mixture components (i.e. clusters), preventing blending of unrelated features while enabling them to co-exist in a single representation. This enables representing videos which consist of dissimilar parts (which may not even have a fixed temporal order). Furthermore, (2) similar frames that fall in the same mixture component will be modeled with respect to the general distribution of that component, capturing subtle variations in time (e.g., the different appearances of a person walking) [1].

Modeling the temporal variation using the Fisher Kernel has been shown to be successful in video classification: improvements were made over a wide range of ap-

1The code for our fast HoG and HoF descriptors will be made publicly available in the nearby future.
<table>
<thead>
<tr>
<th>Method</th>
<th>PCA dimension</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW using Local Features</td>
<td>64</td>
<td>80</td>
</tr>
<tr>
<td>Fisher-Kernel using Global Features</td>
<td>64</td>
<td>256</td>
</tr>
</tbody>
</table>

Table 1. Experimental set-up

applications (genre retrieval, daily activity recognition, action recognition) and over a wide range of features (global HOG, global HOF, human-part based descriptors, audio descriptors) [1, 3]. We use the global HOG and global HOF features and we optimize their weights for the late fusion on the UCF50 dataset [1].

2.3. Result

The details on our descriptors and the number of cluster used for each method are presented in Table 1. The final result of our approach on the UCF101 dataset is 73.02%.

References