Motion Part Regularization: Improving Action Recognition via Trajectory Group Selection

Bingbing Ni[1], Pierre Moulin[2], Xiaokang Yang[3], Shuicheng Yan[4]
[1]Advanced Digital Sciences Center, Singapore. [2]University of Illinois at Urbana-Champaign. [3]Shanghai Jiaotong University. [4]National University of Singapore.

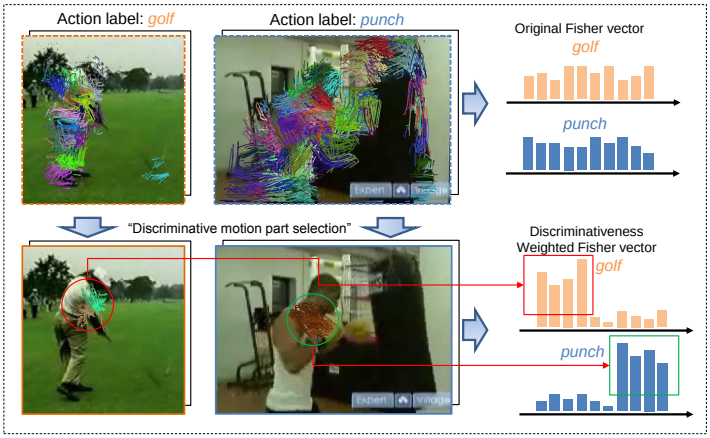

Figure 1: Motivation of our work. Note that our motion part regularization framework generate discriminativeness weighted Fisher vector representation, which is more discriminative than the un-weighted traditional Fisher vector.

Dense local trajectories have been successfully used in action recognition. However, for most actions only a few local motion features (e.g., critical movement of hand, arm, leg etc.) are *responsible* for the action label. Therefore, highlighting the local features which are associated with important motion parts will lead to a more discriminative action representation.

Inspired by recent advances in sentence regularization for text classification, we introduce a *Motion Part Regularization* framework (as shown in Figure 1) to mine for discriminative groups of dense trajectories which form important motion parts. For such a purpose, we first cluster dense trajectories into spatio-temporal groups, which are called *motion parts* in this work. Then, we propose a simple yet effective learning approach which can select discriminative motion parts in a soft manner, *i.e.*, to assign a weight to each motion part to indicate its discriminativeness and use these learned weights for more discriminative action representation by attenuating the of effect of irrelevant motion parts. Our method is inspired by the recent work called *Sentence Regularization* for document classification [1]. In [1], the key observation is that the text words in only a few sentences are relevant to the document label, which is very similar to the action recognition scenario: only a few motion parts that are associated with important moving body parts such as hand, arm, leg, etc., convey high discriminative information. In this sense, each motion part can be regarded as a *sentence* that contains a set of local trajectory features which are indexed into some visual words. For an action video, this results in tens of thousands of motion parts (*sentences*), and the visual words within these sentences are shared, i.e., we have overlapping groups of visual words. To select discriminative local motion part/trajectory group, for each local motion feature in each group we introduce an auxiliary variable, which can be regarded as a local copy of the global weight of the visual word it belongs to. The introduction of these local copies helps to convert the overlapping group lasso feature selection problem (which is not efficient to solve) to a non-overlapping group lasso problem. We thus introduce a simple yet effective alternative optimization scheme to simultaneously optimize the global classifier model weights associated with the visual words and the local copies of these weights. Our learning model is derived as follows.

First, we use the logistic regression function:

$$\mathcal{L}(\mathbf{w},b) = -\sum_{d=1}^{D} \log\Big(1+\exp\big(-y_d(\mathbf{w}^T\mathbf{x}_d+b)\big)\Big). \tag{1}$$

We assume the entries of $\mathbf{w}$ form G groups (*i.e.*, trajectory groups). A key observation is that our trajectory groups are formed *locally* and therefore they are heavily overlapping. In other words, as the visual word vocabulary is globally defined, each visual word may occur in many motion parts (trajectory groups). Mathematically, we use d to index over video samples and p to index over motion parts (trajectory groups) within a video sample. We further denote by P_d the number of motion parts in action video d. Group lasso on $\mathbf{w}$ can be expanded as:

$$\Omega_{gl}(\mathbf{w}) = \sum_{d=1}^{D}\sum_{p=1}^{P_d} \|\mathbf{w}_{d,p}\|_2, \tag{2}$$

where $\mathbf{w}_{d,p}$ corresponds to the sub-vector of $\mathbf{w}$ such that the corresponding features (visual words) are present in motion part p of video d, *i.e.*, different $\mathbf{w}_{d,p}$ vectors may have heavy overlap. The key idea is to introduce a set of auxiliary variables $\mathbf{v}$ to *de-overlap* the groups $\{\mathbf{w}_{d,p}\}$. Each entry of $\mathbf{v}$ defines a weight for each local trajectory feature, thus the length of the vector $\mathbf{v}$ is the total number (denoted by N) of dense trajectories extracted over the entire training video set. In other words, each v_j ($j \in \{1,\cdots,N\}$) can be regarded as a local copy of the associated entry in $\mathbf{w}$ according to the visual word that the j-th (of the entire training trajectory set) trajectory feature is indexed to. $\mathbf{v}$ can be also decomposed into $\{\mathbf{v}_{d,p}\}$. Each $\mathbf{v}_{d,p}$ is associated with the trajectory features in the p-the motion part (trajectory group) of the d-th video, in the similar way that $\mathbf{w}_{d,p}$ is defined. Namely, each $\mathbf{v}_{d,p}$ can also be regarded as a local copy of its corresponding $\mathbf{w}_{d,p}$. The dimensionality of $\mathbf{v}_{d,p}$ will be identical to the size (number of trajectories) of the motion part (d,p), with one dimension per word token. Using the auxiliary variable $\mathbf{v}$, sparse group (motion part) selection could be enforced by the following regularizer:

$$\Omega_{gl}(\mathbf{v}) = \sum_{d=1}^{D}\sum_{p=1}^{P_d} \|\mathbf{v}_{d,p}\|_2, \tag{3}$$

and since each $\mathbf{v}_{d,p}$ is just a local copy of $\mathbf{w}_{d,p}$ and its elements are not shared by other $\mathbf{v}_{d',p'}$ ($d \neq d', p \neq p'$), the original overlapping lasso problem is converted into a non-overlapping one. What remains is to enforce each v_j ($j \in \{1,\cdots,N\}$) to *agree* with its corresponding entry in the global model coefficient vector $\mathbf{w}$. To achieve this, we introduce an assignment matrix $\mathbf{M}$. $\mathbf{M}$ is a $N \times V$ binary matrix, such that $M_{i,j} = 1$ if the local trajectory feature i is indexed to visual word j and 0 otherwise. The integrated learning objective for motion part selection is formulated as:

$$\min_{\mathbf{w},b,\mathbf{v}} \mathcal{L}(\mathbf{w},b) + \lambda_l\|\mathbf{w}\|_1 + \lambda_{gl}\sum_{d=1}^{D}\sum_{p=1}^{P_d}\|\mathbf{v}_{d,p}\|_2 + \beta\|\mathbf{v}-\mathbf{M}\mathbf{w}\|_2^2, \tag{4}$$

where the four terms refer to discriminativeness, sparsity, group sparsity and global-local agreement terms, respectively.

Then, we propose an alternative optimization algorithm to efficiently solve this objective function by introducing a set of auxiliary variables which correspond to the discriminativeness weights of each motion part (trajectory group). These learned motion part weights are further utilized to form a discriminativeness weighted Fisher vector representation for each action sample for final classification. The proposed motion part regularization framework achieves the state-of-the-art performances on several action recognition benchmarks.

[1] Dani Yogatama and Noah A. Smith. Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. In *ICML*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.