Action Recognition with Actons

Jun Zhu [1,2], Baoyuan Wang [3], Xiaokang Yang [1,2], Wenjun Zhang [1,2], Zhuowen Tu [4]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2]Shanghai Key Lab of Digital Media Processing and Transmission, [3]Microsoft Research Asia
[4]Department of Cognitive Science, University of California, San Diego
{zhujun.sjtu,zhuowen.tu}@gmail.com, baoyuanw@microsoft.com, {xkyang,zhangwenjun}@sjtu.edu.cn

Abstract

With the improved accessibility to an exploding amount of video data and growing demands in a wide range of video analysis applications, video-based action recognition/classification becomes an increasingly important task in computer vision. In this paper, we propose a two-layer structure for action recognition to automatically exploit a mid-level "acton" representation. The weakly-supervised actons are learned via a new max-margin multi-channel multiple instance learning framework, which can capture multiple mid-level action concepts simultaneously. The learned actons (with no requirement for detailed manual annotations) observe the properties of being compact, informative, discriminative, and easy to scale. The experimental results demonstrate the effectiveness of applying the learned actons in our two-layer structure, and show the state-of-the-art recognition performance on two challenging action datasets, i.e., Youtube and HMDB51.

1. Introduction

There have been growing demands in robust video analysis systems for action recognition/classification, ranging from content-based video retrieval, sports video analysis, surveillance event detection, to smart human-machine interface and gaming. Continuous efforts are expected in the computer vision field to deliver working systems practical enough to deal with challenging real-world videos. The current state-of-the-art performances [27, 6] have been observed by Bag-of-Features (BoF) model combined with spatial-temporal pyramid (STP) [11, 12]. Developing robust low-level feature descriptors [4, 11, 27] has a critical impact on the success of BoF-STP approaches. Nevertheless, the problem of *video representation* remains to be the center issue in action recognition.

More recently, along the line of research on attributes for 2D static images [25], dedicated efforts have been given to designing/learning mid-level representations [15, 23, 21] for action recognition in dynamic videos, and promising results greatly inspire researchers to explore this direction. From another aspect, hierarchical structures with deep layers [13, 9] have been advocated to a large audience with a big leap on several grand challenge tasks. However, it is a computationally and manually intensive job to learn appropriate models in deep layered structures. Improvement has also been observed in not-so-deep but still hierarchically layered models for object detection [36].

Inspired by these previous works, we propose a new approach to learn mid-level action representation, named "acton" in this paper, based on a weakly-supervised learning strategy. Besides, we develop a two-layer structure for action recognition, with the goal of leveraging the benefits of low-level, mid-level as well as the layered representation for knowledge abstraction. Specifically, the first layer builds a low-level representation using classical BoF-STP model, while the second layer automatically exploits semantically meaningful mid-level representation via a new weakly-supervised learning strategy. Our acton representation in the second layer is built directly on top of the first layer, with the goal of efficient knowledge abstraction and aggregation. More specifically, we use those learned actons as a mid-level dictionary of intermediate concepts to characterize the semantic properties for each *volume of interest* (VOI). In Sec. 4, we present the details of learning effective actons for VOIs, which is the primary focus of this paper. Since the BoF features of the first layer and the acton response features of the second layer both tend to be highly nonlinear, we simply stack the representations of both layers and adopt a linear classifier for video-level action classification.

To achieve this goal, we learn the actons based on a novel *max-margin multi-channel multiple instance learning* (abbreviated by M^{4}IL) method. Multiple linear models are simultaneously learned to discover the actons, which are compact, informative, discriminative, and easy to scale. These actons, each of which corresponds to an underlying cluster/sub-modality of training data, are built on basis of a short sequence of subvolumns (i.e., VOIs) in video clips.

The existing MIL literatures either assume single class [1] (not suitable for building knowledge representation) or lack of explicit competition among different clusters [34, 31]. In practice, our method naturally generalizes multiple instance support vector machine (MI-SVM) [1] and maximum margin multiple instance clustering (M^3IC) [34], and retains complementary advantages of both these two methods. Besides, through making relaxations on the original loss function, we can apply the concave-convex procedure (CCCP) [33] to solve the optimization problem of M^4IL, which can theoretically guarantee the convergence of algorithm. Our experimental results show that adding the second-layer acton representation can provide complementary information w.r.t. the first-layer representation for all the feature descriptor types as well as their different combinations, indicating the generalization ability facilitated by the learning of actons in the second layer. Besides, the proposed two-layer representation can achieve superior classification performance than the state-of-the-art results on two benchmark action datasets (i.e., Youtube and HMDB51).

The contributions of this paper are summarized as follows: (1) We propose a new mid-level acton representation for action recognition, which is automatically exploited through a weakly-supervised learning strategy. (2) By generalizing the single-layer SPM pipeline successfully applied in image classification, we present a two-layer structure on video representation for action recognition task. (3) In the second layer, we propose a novel M^4IL method for learning the actons in a weakly-supervised manner. It can capture multiple mid-level action concepts simultaneously for producing a discriminative and compact representation on action classification.

2. Related Work

As in recent static image classification literature, the BoF-STP approaches with local spatial-temporal features [11, 7, 27] show its significance in action recognition on many challenging datasets [16, 19, 10]. In action recognition literature, typical spatial-temporal feature descriptors include histogram of oriented gradients (HOG) [11], histogram of optical flow (HOF) [11] and motion boundary histogram (MBH) [4, 27] etc., which are computed on the local 3D patches extracted by spatial-temporal interesting point (STIP) detectors [11] or dense sampling strategies [27, 24]. Moreover, recent work [27, 7, 6] demonstrate that leveraging the trajectory information [20, 18] leads to more discriminative feature representation and makes for recognition performance.

Besides those low-level features, mining mid-level feature representation (e.g., discriminative parts/patches [22, 23, 29, 28] or semantical attributes [25, 15, 14]) for image/video recognition has been a recent active area in computer vision. However, most of these work [25, 23] rely on strong supervision in training with detailed manual annotations. It is even more labor intensive to label the videos in action recognition [15, 23]. Our method, instead, only asks for the video-level class information of training video clips without any additional annotations, and thus is much easier to scale to deal with a large amount of training data.

There are some existing work [21, 14, 29] using weakly-supervised learning for visual recognition. The most relevant "MIL-BoF" work [21] directly adopts the mi-SVM algorithm [1] on top of the Bag-of-Features subvolume representation for video action classification. In contrast, we focus on learning the actons that serve as a dictionary of intermediate action concepts to describe each VOI. With a spatial-temporal pooling step on the resultant acton responses of VOIs, we obtain a mid-level video representation and feed it into the final classifier. Our experimental results demonstrate that with multiple actons, this mid-level representation tends to be more discriminative and diverse for action recognition. The very recent work of [14, 29] exploit the MIL methods to learn multiple mid-level visual concepts for each class in a weakly-supervised manner. In [14], the visual concepts are discovered via a successive two-stage method of using the mi-SVM algorithm followed by K-means clustering. In [29], Wang *et al*. present an iterative EM algorithm for visual dictionary learning, which alternates between two steps of sampling positive instances and training off-the-shell multi-class SVM classifiers. By contrast, our M^4IL method can simultaneously explore multiple mid-level action concepts in a unified learning formulation, which could be readily solved by the CCCP algorithm. Besides, we present a two-layer structure on video representation for action recognition, and show its superiority w.r.t. either of the single-layer ones.

3. A two-layer representation of videos for action recognition

In this section, we elaborate the proposed two-layer framework of video representation for action recognition. The first-layer representation is built by the codes of local spatial-temporal feature points, and the second-layer representation is constructed based on the acton responses of VOIs.

3.1. The first-layer representation

As illustrated in Fig. 1, we adopt the linear SPM pipeline [32, 17, 35] to build the first-layer representation for a video clip. It generally includes three common steps: *local feature extraction*, *feature coding* and *spatial-temporal pooling*.

Local feature extraction: For a video clip $\mathbf{V}$, we assume there are P local STIPs detected. Then each STIP is represented by a feature descriptor $\mathbf{a} \in \mathbb{R}^D$ to capture its local visual cues (e.g., appearance, motion) [11, 27]. Thus,

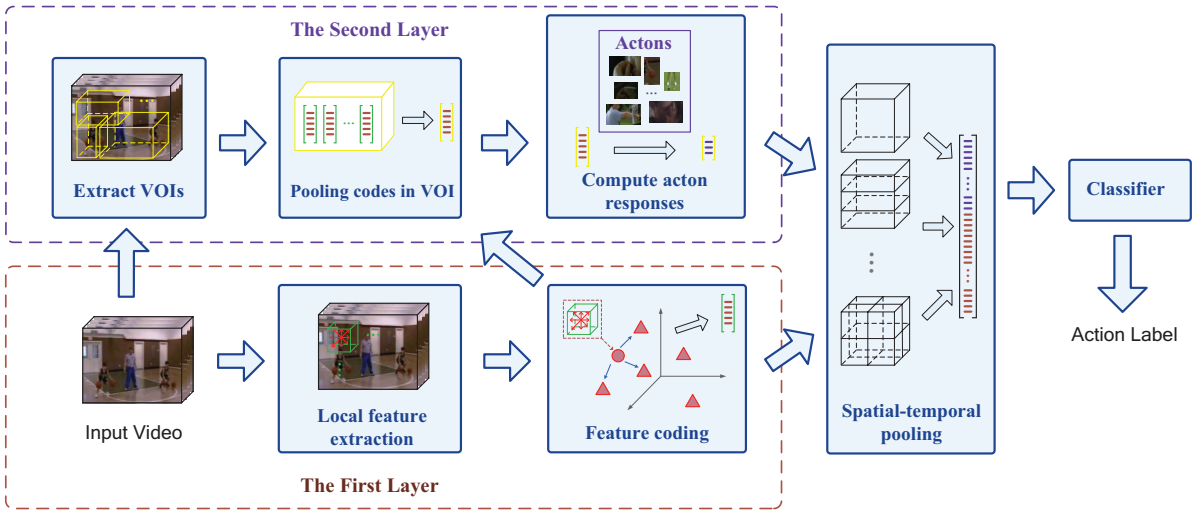

Figure 1. Illustration of the proposed two-layer structure on action classification. (Best viewed in color)

a set of local spatial-temporal features $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_P]$ can be extracted from $\mathbf{V}$.

Feature coding: Given a codebook (denoted by $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_M] \in \mathbb{R}^{D\times M}$) with M codewords, each STIP feature $\mathbf{a}_p$ is encoded into a code vector $\mathbf{c}_p = [c_{p,1}, c_{p,2}, \cdots, c_{p,M}]^{\mathrm{T}}$, where $c_{p,u}\,(u \in \{1, 2, \cdots, M\})$ is its code of the u-th codeword. Rather than using the vector quantization in standard BoF models [11, 21, 7, 27], we adopt the localized soft-assignment quantization (LSAQ) method [17] for feature coding. It computes a L1-normalized code vector based on the L2-distance between $\mathbf{a}_p$ and its n-nearest-neighbour codewords.

Spatial-temporal pooling: For capturing informative statistics and achieving invariance properties (e.g., transformation invariance) in spatial-temporal domain, a maximum pooling step is exploited to form the video-level representation of $\mathbf{V}$, via statistical summarization over the codes $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_P]$ in L different subvolumes. In this paper, we use a volumetric spatial-temporal pyramid as in [27], which includes six different spatial-temporal grids leading to a total of $L = 24$ subvolumes. Let Λ_l denote spatial-temporal domain of the l^{th} subvolume used for pooling. We define an element-wise maximization operation OP_{max} for mapping the codes located in Λ_l into a M-dimensional vector $\gamma_l^{(1)} = [\gamma_{l,1}^{(1)}, \gamma_{l,2}^{(1)}, \cdots, \gamma_{l,M}^{(1)}]^{\mathrm{T}} = OP_{max}(\mathbf{C}; \Lambda_l)$. For any visual word u, the pooled signature is obtained by $\gamma_{l,u}^{(1)} = \max_{p\in\Omega(\Lambda_l)} c_{p,u}$, where $\Omega(\Lambda_l)$ refers to the set of STIPs located in Λ_l. Thus, the first-layer representation of $\mathbf{V}$ is denoted by a $(M \times L)$-dimensional vector $\mathbf{\Gamma}^{(1)} = [\gamma_1^{(1)}; \gamma_2^{(1)}; \cdots; \gamma_L^{(1)}]$.

3.2. The second-layer representation

Besides the first-layer representation based on local STIPs, we construct a higher-level one by the acton responses of VOIs for action recognition in videos. In this layer, a video clip is decomposed into a set of VOIs, which potentially correspond to action parts or relevant objects. Generally, the VOIs can be extracted by saliency detector or densely sampling from a regular grid in spatial-temporal domain.

Feature description for VOIs: Assuming there are J VOIs extracted from $\mathbf{V}$, we represent it by a bag of VOI features $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_J]$. Let $\Lambda_j^{'}$ denote the 3D spatial-temporal bounding box of the j^{th} VOI, then its feature descriptor $\mathbf{x}_j$ can be computed by pooling the codes of STIPs located in $\Lambda_j^{'}$ (i.e., $\mathbf{x}_j = OP_{max}(\mathbf{C}; \Lambda_j^{'})$). Besides, the VOI descriptor $\mathbf{x}_j$ is further normalized by its L2-norm.

A mid-level representation based on acton responses: Given K actons, we construct the second-layer representation of $\mathbf{V}$ via their responses on extracted VOIs. Let $\mathbf{w}_k$ denote model parameter of the k^{th} acton. For VOI j, we obtain a K-dimensional response vector $\mathbf{r}_j = [r_{j,1}, r_{j,2}, \cdots, r_{j,K}]^{\mathrm{T}}$, each component of which is computed by $r_{j,k} = \mathcal{S}(\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_j), \forall\, k = 1, 2, \cdots, K$. $\mathcal{S}(r) = \frac{1}{1+\exp(-\rho r)}$ is the sigmoid function to produce a probabilistic form on the resultant response value, where ρ is a saturation parameter for controlling its sharpness. Let $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_J]$ denote the set of responses on all VOIs in $\mathbf{V}$. Similar as the first-layer representation, we further pooling $\mathbf{R}$ for each subvolume of spatial-temporal pyramid (i.e., $\gamma_l^{(2)} = OP_{max}(\mathbf{R}; \Lambda_l), \forall\, l = 1, 2, \cdots, L$), and obtain a $(K \times L)$-dimensional vector $\mathbf{\Gamma}^{(2)} = [\gamma_1^{(2)}; \gamma_2^{(2)}; \cdots; \gamma_L^{(2)}]$ as the second-layer representation of $\mathbf{V}$.

Given the vectors $\mathbf{\Gamma}^{(1)}$ and $\mathbf{\Gamma}^{(2)}$, we first perform L2-normalization on them respectively. Then, we concatenate them into a single vector $\mathbf{\Gamma} = [\mathbf{\Gamma}^{(1)}; \mathbf{\Gamma}^{(2)}]$, which is the proposed two-layer representation of $\mathbf{V}$. Finally, we can apply

any off-the-shell classifier on $\mathbf{\Gamma}$ for action recognition.

4. Maximum margin multi-channel multiple instance learning (M^4IL) for weakly-supervised actons

In this paper, the mid-level intermediate concepts (i.e., actons) for the second-layer representation are not predefined and learned in a data-driven manner. Although the acton labels of VOIs are unknown in training, the class labels of whole video clip are available and can provide informative cues for weakly-supervised learning of actons on VOIs. This inherently coincides the assumption of multiple instance learning [1, 34]. In this section, we formally introduce the proposed M^4IL algorithm for learning actons.

4.1. Learning formulation

Suppose we are given a set of training video clips (bags) with binary class labels $\{(\mathbf{X}_i, y_i)\}_{i=1}^N$ ($y_i \in \{-1, +1\}$), and each bag is composed of several instances (VOIs): $\mathbf{X}_i = \{\mathbf{x}_{i,j} | j \in \mathcal{B}_i\}$ ($\mathcal{B}_i = \{1, 2, \cdots, J_i\}$), where J_i refers to the total number of instances for $\mathbf{X}_i$. Furthermore, we assume there are K underlying channels (modalities), each of which corresponds to a candidate acton model $\mathbf{w}_k$ ($k \in C, C = \{1, 2, \cdots, K\}$), for explaining the instances in positive bags. Based on the maximum margin principle [1, 34], we present a method based on M^4IL formulation for weakly-supervised actons:

$$\min_{\mathbf{w}_k, \xi_i \geqslant 0, \zeta_i \geqslant 0} \frac{1}{2} \sum_{k=1}^{K} ||\mathbf{w}_k||^2 + C_1 \sum_{i=1}^{N} \xi_i + C_2 \sum_{y_i=+1} \zeta_i \qquad (1)$$

$$\text{s.t.} \quad \forall\, i = 1, 2, \cdots, N : y_i \max_{j \in \mathcal{B}_i} \max_{k \in C} \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_{i,j} \geqslant 1 - \xi_i,$$

$$\forall\, y_i = +1 : \max_{j \in \mathcal{B}_i^+} [\mathbf{w}_{\hat{k}}^{\mathrm{T}} \mathbf{x}_{i,j} - \frac{1}{K-1} \sum_{k \neq \hat{k}} \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_{i,j}] \geqslant 1 - \zeta_i,$$

$$\forall\, k_1, k_2 \in C : | \sum_{y_i=+1} \frac{1}{J_i} \sum_{j=1}^{J_i} (\mathbf{w}_{k_1}^{\mathrm{T}} \mathbf{x}_{i,j} - \mathbf{w}_{k_2}^{\mathrm{T}} \mathbf{x}_{i,j})| \leqslant \eta,$$

where $\hat{k} = \arg\max_{k \in C} \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_{i,j}$ and $\mathcal{B}_i^+ = \{j \,|\, \exists\, k \in C : \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_{i,j} > 0\}$. The user-defined parameters C_1 and C_2 trade off the terms of hinge loss and model regularization.

In Equ. (1), the first set of constraints corresponds to a multiple-instance-based margin (i.e., $y_i \max_{j \in \mathcal{B}_i} \max_{k \in C} \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_{i,j}$) for each bag $\mathbf{X}_i$: There should be at least one instance explained well by someone of the K actons for a positive bag, and a negative bag is preferable to have none of instance belonging to any sub-modality (i.e., a channel) related to positive class. Actually, it generalizes MI-SVM [1] with the multiple-channel assumption of VOIs in a video. Similar to M^3IC [34], the second set of constraints is induced to impose diversity among the sub-modalities each other. The difference between Equ. (1) and [34] for this constraint lies in that we only consider the instances regarded as positive ones by the actons to contribute the margin between different sub-modalities. As in [30, 34], the third set of constraints enforces the class balance to avoid a trivial solution that all of the instances are assigned to one channel.

In the following discussion, we rewrite Equ. (1) to an equivalent formulation. For each instance $\mathbf{x}_{i,j}$, we induce a latent variable $z_{i,j} \in C$ to indicate which sub-modality can explain it best, then a concatenated $M \times K$-dimensional vector can be defined by $\psi(\mathbf{x}_{i,j}, z_{i,j}) = [(0, 0, \cdots,), \cdots, \mathbf{x}_{i,j}^{\mathrm{T}}, \cdots, (0, 0, \cdots,)]^{\mathrm{T}}$, where the $z_{i,j}^{\text{th}}$ stack is equal to $\mathbf{x}_{i,j}$ and the other entries are zero. Let $\mathbf{W} = [\mathbf{w}_1; \mathbf{w}_2; \cdots; \mathbf{w}_K]$ assemble the model parameters of K actons, then the problem of Equ. (1) can be rewritten as below:

$$\min_{\mathbf{W}, \xi_i \geqslant 0, \zeta_i \geqslant 0} \frac{1}{2} ||\mathbf{W}||^2 + C_1 \sum_{i=1}^{N} \xi_i + C_2 \sum_{y_i=+1} \zeta_i \qquad (2)$$

$$\text{s.t.} \ \forall\, i = 1, 2, \cdots, N : y_i \max_{(j, z_{i,j}) \in \mathcal{B}_i \times C} \mathbf{W}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, z_{i,j}) \geqslant 1 - \xi_i,$$

$$\forall\, y_i = +1 :$$

$$\frac{K}{K-1} \max_{j \in \mathcal{B}_i^+} [\max_{z_{i,j} \in C} \mathbf{W}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, z_{i,j}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{W}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, k)] \geqslant 1 - \zeta_i,$$

$$\forall\, k_1, k_2 \in C : | \sum_{y_i=+1} \frac{1}{J_i} \sum_{j=1}^{J_i} \mathbf{W}^{\mathrm{T}} [\psi(\mathbf{x}_{i,j}, k_1) - \psi(\mathbf{x}_{i,j}, k_2)]| \leqslant \eta,$$

4.2. Solving by CCCP

From Equ. (2), We observe that the first and second sets of soft-margin constraints for positive bags are not convex. However, both of them have a form of the difference between two convex functions, and thus the problem of Equ. (2) can be solved by the CCCP [33, 34].

Starting from an initial model $\mathbf{W}_0$, the CCCP iteratively constructs a series of relaxed convex optimization problems for Equ. (2), each of which defines an upper-bound loss function to approximate original one at the solution obtained in the last iteration. Besides, it guarantees that a local optima can be always found when achieving convergence [33]. In the t^{th} iteration of CCCP, we replace the left terms of constraints on positive bags by the first-order Taylor expansions at $\mathbf{W}_{t-1}$. For the left term of the first constraint on a positive bag $\mathbf{X}_i$, let $f(\mathbf{W}, \mathbf{X}_i) = \max_{(j, z_{i,j}) \in \mathcal{B}_i \times C} \mathbf{W}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, z_{i,j})$, and its subgradient at $\mathbf{W}_{t-1}$ can be computed by

$$\nabla f(\mathbf{W}_{t-1}, \mathbf{X}_i) = \frac{\partial f(\mathbf{W}, \mathbf{X}_i)}{\partial \mathbf{W}} |_{\mathbf{W}=\mathbf{W}_{t-1}} = \psi(\mathbf{x}_{i,\hat{j}}, \hat{z}_{i,\hat{j}}), \qquad (3)$$

where $(\hat{j}, \hat{z}_{i,\hat{j}}) = \arg \max_{(j, z_{i,j}) \in \mathcal{B}_i \times C} \mathbf{W}_{t-1}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, z_{i,j})$. Likewise, for the left term of the second constraint on $\mathbf{X}_i$, let $g(\mathbf{W}, \mathbf{X}_i) = \frac{K}{K-1} \max_{j \in \mathcal{B}_i^+} h(\mathbf{W}, \mathbf{x}_{i,j})$ and $h(\mathbf{W}, \mathbf{x}_{i,j}) = \max_{z_{i,j} \in C} \mathbf{W}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, z_{i,j}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{W}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, k)$. Then we can compute the subgradient

Algorithm 1: The CCCP Algorithm for M^{4}IL

Input: a set of bags $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N\}$ with binary labels $\{y_1, y_2, \cdots, y_N\}$, ε, $\mathcal{N}_{max}$

Output: the solution of model parameters $\mathbf{W}^*$

Initialize $\mathbf{W} \leftarrow \mathbf{W}_0$
for $t = 1$ *to* $\mathcal{N}_{max}$ **do**
  foreach *positive bag* $\mathbf{X}_i$ **do**
    Compute the subgradients $\nabla f(\mathbf{W}_{t-1}, \mathbf{X}_i)$ and $\nabla g(\mathbf{W}_{t-1}, \mathbf{X}_i)$ by Equ. (3) and (4), respectively;
  end
  Get a new model $\mathbf{W}_t$ by solving Equ. (5);
  if $||\mathbf{W}_t - \mathbf{W}_{t-1}|| < \varepsilon$ *or* $t == \mathcal{N}_{max}$ **then**
    The algorithm stops, and return $\mathbf{W}^* \leftarrow \mathbf{W}_t$.
  end
end

as follow [34]:

$$\begin{aligned}\nabla g(\mathbf{W}_{t-1}, \mathbf{X}_i) &= \frac{\partial g(\mathbf{W}, \mathbf{X}_i)}{\partial h(\mathbf{W}, \mathbf{x}_{i,j})} \times \frac{\partial h(\mathbf{W}, \mathbf{x}_{i,j})}{\partial \mathbf{W}}|_{\mathbf{W}=\mathbf{W}_{t-1}} \\ &= \frac{K}{K-1}[\psi(\mathbf{x}_{i,\tilde{j}}, \tilde{z}_{i,\tilde{j}}) - \frac{1}{K}\sum_{k=1}^{K}\psi(\mathbf{x}_{i,\tilde{j}}, k)],\end{aligned} \tag{4}$$

where $\tilde{j} = \arg\max_{j \in \mathcal{B}_i^+} h(\mathbf{W}_{t-1}, \mathbf{x}_{i,j})$ and $\tilde{z}_{i,\tilde{j}} = \arg\max_{k \in C} \mathbf{W}_{t-1}^{\mathrm{T}} \psi(\mathbf{x}_{i,\tilde{j}}, k)$. By replacing $f(\mathbf{W}, \mathbf{X}_i)$ and $g(\mathbf{W}, \mathbf{X}_i)$ in Equ. (2) with Equ. (3) and (4) respectively, we obtain the relaxed problem for the t^{th} iteration:

$$\begin{aligned}&\min_{\mathbf{W}, \xi_i \geqslant 0, \zeta_i \geqslant 0} \tfrac{1}{2}||\mathbf{W}||^2 + C_1 \sum_{i=1}^{N} \xi_i + C_2 \sum_{y_i=+1} \zeta_i \\ &\text{s.t. } \forall\, i \text{ and } y_i = -1 : -\max_{(j, z_{i,j}) \in \mathcal{B}_i \times C} \mathbf{W}^{\mathrm{T}} \psi(\mathbf{x}_{i,j}, z_{i,j}) \geqslant 1 - \xi_i, \\ &\forall\, i \text{ and } y_i = +1 : \mathbf{W}^{\mathrm{T}} \nabla f(\mathbf{W}_{t-1}, \mathbf{X}_i) \geqslant 1 - \xi_i, \\ &\forall\, i \text{ and } y_i = +1 : \mathbf{W}^{\mathrm{T}} \nabla g(\mathbf{W}_{t-1}, \mathbf{X}_i) \geqslant 1 - \zeta_i, \\ &\forall\, k_1, k_2 \in C : |\sum_{y_i=+1} \frac{1}{J_i} \sum_{j=1}^{J_i} \mathbf{W}^{\mathrm{T}}[\psi(\mathbf{x}_{i,j}, k_1) - \psi(\mathbf{x}_{i,j}, k_2)]| \leqslant \eta,\end{aligned} \tag{5}$$

Thus, Equ. (5) is a convex quadratic programming problem with potentially large number of constraints[1], which can be efficiently solved by the cutting-plane algorithm [8, 34]. We summarize our algorithm of M^{4}IL in Alg. 1.

5. Experiments

In this section, we evaluate performance of the proposed method for action classification on three action datasets, and compare it with previous methods in literature. Besides, we give detailed analysis and discussion on the effect of acton representation and M^{4}IL learning method.

[1]The first constraint for each negative bag $\mathbf{X}_i$ is equivalent to a set of $J_i \times K$ linear inequality constraints in practice. i.e, $-\mathbf{W}^{\mathrm{T}}\psi(\mathbf{x}_{i,j}, z_{i,j}) \geqslant 1 - \xi_i$, $\forall\, j = 1, 2, \cdots, J_i$, $z_{i,j} = 1, 2, \cdots, K$.

5.1. The datasets

In this paper, our method is evaluated on three standard video datasets in action recognition literation: Youtube [16], Hollywood2 [19] and HMDB51 [10]. They provide diverse benchmarks on realistic scenarios with challenging conditions (e.g., object appearance and pose, camera motion, background clutter, illumination and viewpoint variation). We summarize the experimental settings of these datasets as follows:

- **Youtube**: This dataset includes 1,168 video clips collected from Internet, which are divided into 11 action categories. Following [16], we use a 25-fold leave-one-out (LOO) cross-validation for training and testing. The performance is measured by the average of per-class classification accuracy as in [16].

- **Hollywood2**: This dataset is collected from 69 realistic movies with significant background clutter. It contains 12 action classes and 1707 videos, which are divided into a training set and a testing set for evaluation. According to [19], we compute the average precision (AP) for each action class, and the performance is measured by the mean AP (mAP) over all classes.

- **HMDB51**: This large-scale action dataset is collected from a variety of realistic video sources such as movies and Internet videos. It includes 6,766 video clips and 51 action categories. We use the same three training/testing splits released by [10] for fair comparison. Following [10], the performance is evaluated by average accuracy over these three round splits.[2]

5.2. Experimental setup

Throughout all of the experiments, we extract local features by computing the dense trajectories [27] and adopt the same four types of feature descriptors used in [27][3]: trajectory shape descriptor (TrajShape) [27], MBH [4], HOG and HOF [11, 27]. In the step of feature coding, we randomly select 100,000 local spatial-temporal features and build a visual dictionary via standard K-means algorithm. Following common settings in literature [21, 27], the number of visual words is set by $M = 4000$ for each feature type. The parameter setting of LSAQ coding is the same as [17] (i.e., $\beta = 10$ and $n = 5$).

For the second-layer representation, the VOIs are densely sampled from a regular grid like [21], with a spacing of 50% overlapping in each direction of spatial-temporal domain. There are 8 different configurations of cuboid VOIs used in this paper, where the size in x/y direction is 80

[2]Note that the dataset includes both the original videos and corresponding stabilized ones. We report experimental results on the original version.

[3]We implement dense trajectory features by using the code of [27] from http://lear.inrialpes.fr/people/wang/dense_trajectories.

Youtube		Hollywood2		HMDB51	
Ikizler-Cinbis and Sclaroff [5]	75.21	MIL-BoF *et al.* [21]	48.73	MIL-BoF [21]	31.53
Le *et al.* [13]	75.8	Ullah *et al.* [26]	53.2	TrajMF [7]	40.7
Bhattacharya *et al.* [2]	76.5	Le *et al.* [13]	53.3	Motionlet [28]	42.1
Human Postures [3]	77.8	TrajMF [7]	59.5	Shi *et al.* [24]	47.6
MIL-BoF [21]	80.39	Dense Traj. [27]	59.9	Dense Traj. [27]	48.3
Dense Traj. [27]	85.4	Jain *et al.* [6]	**62.5**	Jain *et al.* [6]	52.1
Our method	**89.4**	Our method	61.4	Our method	**54.0**

Table 1. Performance evaluation on the proposed two-layer representation and comparison to the state-of-the-art methods in action classification literature. We report per-class classification accuracy (%) for HMDB51 and Youtube, mAP (%) for Hollywood2.

or 120 pixels and the size in temporal domain is 40 or 60 frames. We learn the actons in "one-vs-rest" manner. It means that the training video clips of each class are used as positive bags in turn, while the ones of other classes serve as negative bags. After that, we collect all the actons learned from various classes for computing the responses. For performance evaluation in Sec. 5.3, we train $K = 3$ actons for each class. Following [32, 17], we use the linear SVM classifier in experiments, and the "one-vs-rest" criterion is applied to multi-class classification for Youtube and HMDB51 datasets.

5.3. Performance evaluation

In this experiment, we evaluate the classification performance of our two-layer video representation for different individual feature types as well as the combined features, and compare the results with previous methods in action classification literature. Similar as [27], our results are obtained by combining all these four descriptor types. From table 1, we can see that the proposed two-layer representation achieves superior performance than previous action classification methods on Youtube and HMDB51 datasets.

On Youtube dataset, our method achieves the classification accuracy of 89.4%, which is superior than the state-of-the-art result (85.4% [27]) by 4%. On Hollywood2, we report the mAP of 61.4%. It is slightly lower than the state-of-the-art result (62.5 [6]) by 1.1%, which is obtained in [6] by using improved trajectory features and more advanced coding method. On the larger HMDB51 dataset, the proposed method obtains the accuracy of 54.0% and outperforms the state-of-the-art result (52.1% [6]) by 1.9%. Particularly, noting that we use exactly the same dense trajectory features and STP as in [27], which use common BoF approach with a non-linear RBF-χ^2 kernel SVM classifier instead, the comparison to [27] clearly indicates the superiority of our method w.r.t. BoF-STP representation. Besides, compared with the MIL-BoF method [21] which also employs the VOIs and weakly-supervised learning for video representation, our method obtains superior result with a large margin for each of the three datasets.

	Only 1st	Only 2nd	1st + 2nd
Accuracy (%)	52.1	50.7	54.0
Number of Dim.	384,000	14,688	398,688

Table 2. Analysis on the classification performance and the number of dimensions for the proposed two-layer representation. From the second to the fourth columns: using the first layer only, using the second layer only, combination of the two layers.

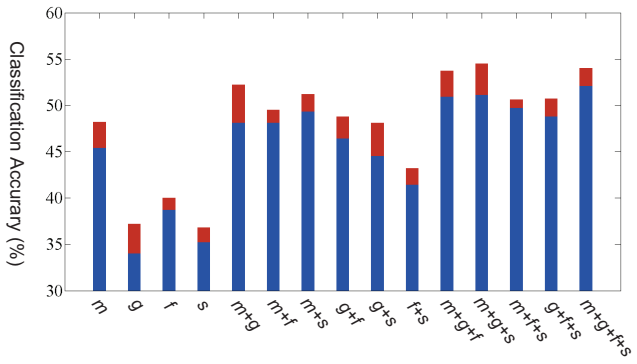

Figure 2. Illustration on the effect of our acton representation. Each bar corresponds to one of the four feature descriptors or possible feature combinations. The blue part represents the accuracy obtained by using only the first-layer representation, and stacked red part indicates the performance gain of adding the second-layer representation. For notation clarity, we use the following characters for abbreviating the name of descriptor types: "m" - MBH, "g" - HOG, "f" - HOF, "s" - TrajShape. For instance, "m+g+f" corresponds to the combination of MBH, HOG and HOF. (Best viewed in color)

5.4. Analysis and discussion

In this subsection, we provide comprehensive analysis on the significance of the second-layer (acton) representation for action recognition, and discuss the proposed M^{4}IL method of learning actons.

Analysis on the acton representation: Based on the learned actons, we produce a compact and discriminative feature representation for the second layer, which can provide complementary information w.r.t. the first-layer repre-

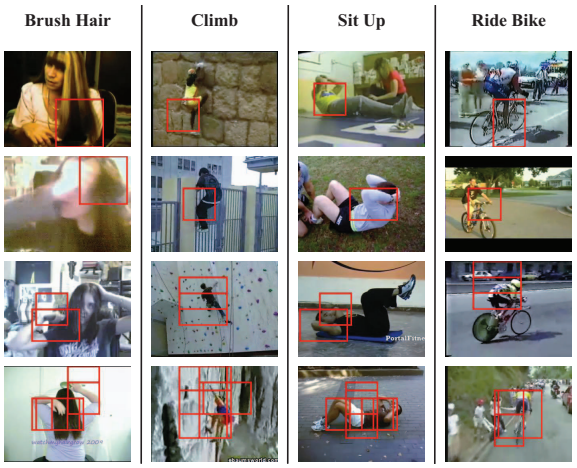

Figure 3. Visualization on examples of positive instances found by our method. The red bounding-boxes illustrate positive instances/VOIs (i.e., with the response values larger than 0.5) according to the actons learned by our M^{4}IL method. We can see there are meaningful VOIs found for corresponding action classes, which tend to be diagnostic action parts or relevant objects (e.g., the waving hand in "Brush Hair", the bicycle wheel in "Ride Bike"). Besides, there are some examples of capturing different intermediate concepts (i.e., the actons) simultaneously, which coincides with the multi-channel property of M^{4}IL. The results are obtained on testing videos from HMDB51 dataset. (Best viewed in color)

sentation and contribute to performance improvement. Table 2 compares the classification accuracy and resultant dimensions for the representations of individual layers as well as the combined representation of the two layers on HMDB51 dataset.

When using the first layer only, we obtain the accuracy of 52.1% and outperforms the result (48.3%) of [27] by a margin of 3.8%. Considering that we adopt identical local features as in [27], it accounts for the superiority of using LSAQ coding and max pooling relative to classical BoF model in [27]. If cooperated with the second-layer representation, the classification accuracy of our method can further boost to 54.0%. Moreover, Fig. 2 shows that adding the second-layer representation can boost performance for every individual feature descriptor as well as all possible combinations of them (The average performance gain over all the 15 kinds of feature type configurations is $2.3 \pm 0.9\%$). This demonstrates that the second-layer representation can provide complementary information w.r.t. the first-layer one to benefit action recognition.

Although the second-layer representation has merely a fraction of dimensions relative to that of the first layer (as shown in table 2), we can see that the performance gap between them is less than 1.5% in practice. This validates the compactness and effectiveness of our acton representation

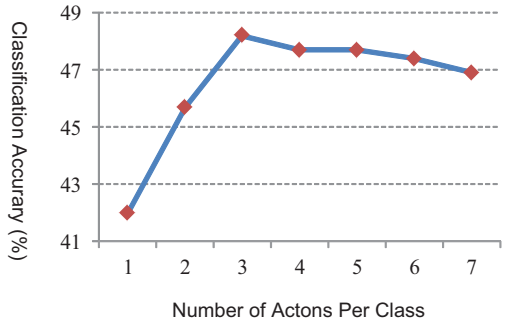

Figure 4. The classification performance w.r.t. different number of actons (i.e., parameter K) learned in the second layer. The results are obtained with the MBH feature on HMDB51 dataset.

learned from data. Fig. 3 visualizes some examples of representative positive instances (VOIs) for different classes. It shows that the learned actons in the second layer can capture diagnostic action parts or relevant objects and thus produce semantically richer representation than the first-layer one.

Discussion on the learning method of M^{4}IL: For the learning of actons, we compare the proposed M^{4}IL with other classical MIL methods such as MI-SVM [1] and M^{3}IC [34]. From table 3, we can see that our method consistently outperforms the competing MIL methods [1, 34]. On one hand, compared to MI-SVM, our method can learn multiple sub-modalities for positive class simultaneously, by discovering more and diverse intermediate concepts for facilitating action recognition. On the other hand, the M^{4}IL induces negative bag examples to utilize the video-level class annotation information from training data, and thus produces a more discriminative acton representation than M^{3}IC.

Besides, we investigate the classification performance w.r.t. different number of actons per class (i.e., K) on HMDB51 dataset. As shown in Fig. 4, we observe that the performance boosts along with increasing number of actons[4] used in our method, and achieves the optimum value at $K = 3$. Continued increase of the actons will not improve accuracy further and even deteriorate performance. This implies that the M^{4}IL can learn actons to produce a rather compact representation for the second layer, and adding excessive actons may lead to some trivial intermediate concepts and does not make for discrimination.

M^{4}IL	MI-SVM [1]	M^{3}IC [34]
48.2	42.0	45.4

Table 3. Performance comparison on other MIL methods for learning actons in the second layer. The results are obtained with the MBH feature descriptor on HMDB51 dataset.

[4]For the case of $K = 1$, our M^{4}IL is actually equivalent to the MI-SVM. Hence we use the result of MI-SVM as the accuracy of $K = 1$ in Fig. 4.

6. Conclusion

In this paper, we present a two-layer representation for action recognition in videos. For the second layer, we propose a M^{4}IL method for learning the actons. It discovers multiple mid-level concepts simultaneously to produce a discriminative and compact representation for action classification. The experimental results demonstrate the effectiveness of our two-layer representation and show the state-of-the-art performance on Youtube and HMDB51.

Acknowledgement: This work was supported by Microsoft Research Asia, NSF IIS-1216528 (IIS-1360566), NSF CAREER award IIS-0844566 (IIS-1360568). Part of this work was done while the first author was an intern at Microsoft Research Asia. It was also in part supported by NSFC (61025005, 61071155), STCSM (12DZ2272600) and 973 Program (2010CB731401).

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2003. 2, 4, 7

[2] S. Bhattacharya, R. Sukthankar, R. Jin, and M. Shah. A probabilistic representation for efficient large scale visual recognition tasks. In *CVPR*, 2011. 6

[3] W. Brendel and S. Todorovic. Activities as time series of human postures. In *ECCV*. 2010. 6

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*. 2006. 1, 2, 5

[5] Ikizler-Cinbis and Sclarof. Object, scene and actions: combining multiple features for human action recognition. In *ECCV*, 2010. 6

[6] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013. 1, 2, 6

[7] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *ECCV*, 2012. 2, 3, 6

[8] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural svms. *Machine Learning*, 2009. 5

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 2, 5

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2, 3, 5

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1

[13] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011. 1, 6

[14] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013. 2

[15] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 1, 2

[16] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *ICCV*, 2009. 2, 5

[17] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*, 2011. 2, 3, 5, 6

[18] X. Liu, L. Lin, S.-C. Zhu, and H. Jin. Trajectory parsing by cluster sampling in spatio-temporal graph, 2009. 2

[19] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2, 5

[20] R. Messing, C. Pal, and H. A. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009. 2

[21] F. C. Michael Sapienza and P. H. Torr. Learning discriminative space-time actions from weakly labelled videos. In *BMVC*, 2012. 1, 2, 3, 5, 6

[22] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012. 2

[23] S. Sadanand and J. Corso. Action bank: a high-level representation of activity in video. In *CVPR*, 2012. 1, 2

[24] F. Shi, E. Petriu, and R. Laganiere. Sampling strategies for real-time action recognition. In *CVPR*, 2013. 2, 6

[25] Y. Su, M. Allan, and F. Jurie. Improving object classification using semantic attributes. In *BMVC*, 2010. 1, 2

[26] M. M. Ullah, S. N. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, 2010. 6

[27] H. Wang, A. Klser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 2013. 1, 2, 3, 5, 6, 7

[28] L. Wang, Y. Qiao, and X. Tang. Motionlets: mid-level 3D parts for human motion recognition. In *CVPR*, 2013. 2, 6

[29] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu. Max-margin multiple instance dictionary learning. In *ICML*, 2013. 2

[30] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *NIPS*, 2005. 4

[31] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, and Z. Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *CVPR*, 2012. 2

[32] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009. 2, 6

[33] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 2003. 2, 4

[34] D. Zhang, F. Wang, L. Si, and T. Li. Maximum margin multiple instance clustering with applications to image and text clustering. *IEEE Transactions on Neural Networks*, 2011. 2, 4, 5, 7

[35] J. Zhu, W. Zou, X. Yang, R. Zhang, Q. Zhou, and W. Zhang. Image classification by hierarchical spatial pooling with partial least squares analysis. In *BMVC*, 2012. 2

[36] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010. 1