Exploiting Motion Information from Unlabeled Videos for Static Image Action Recognition

Yiyi Zhang, Li Niu,* Ziqi Pan, Meichao Luo, Jianfu Zhang, Dawei Cheng, Liqing Zhang*
MoE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{yi95yi, ustcnewly, panziqi_ai, luomcmxg, c.sis, dawei.cheng}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract

Static image action recognition, which aims to recognize action based on a single image, usually relies on expensive human labeling effort such as adequate labeled action images and large-scale labeled image dataset. In contrast, abundant unlabeled videos can be economically obtained. Therefore, several works have explored using unlabeled videos to facilitate image action recognition, which can be categorized into the following two groups: (a) enhance visual representations of action images with a designed proxy task on unlabeled videos, which falls into the scope of self-supervised learning; (b) generate auxiliary representations for action images with the generator learned from unlabeled videos. In this paper, we integrate the above two strategies in a unified framework, which consists of Visual Representation Enhancement (VRE) module and Motion Representation Augmentation (MRA) module. Specifically, the VRE module includes a proxy task which imposes pseudo motion label constraint and temporal coherence constraint on unlabeled videos, while the MRA module could predict the motion information of a static action image by exploiting unlabeled videos. We demonstrate the superiority of our framework based on four benchmark human action datasets with limited labeled data.

1 Introduction

Video action recognition has achieved competitive performance (Tran et al. 2015; Carreira and Zisserman 2017), but analyzing videos is computationally expensive. As an alternative approach, static image action recognition is becoming an active research topic (Sharma, Jurie, and Schmid 2012; Zheng et al. 2012; Mezghiche, Melkemi, and Foufou 2014; Vondrick, Pirsiavash, and Torralba 2016; Zhang et al. 2016; Gao, Xiong, and Grauman 2018).

However, image action recognition usually requires a large number of labeled action images. Although some image action datasets such as Pascal VOC (Everingham et al. 2010) and Stanford40 (Yao et al. 2011) are available, but they can only cover a limited number of action categories. In reality, not all categories have ample training samples, especially for those ”long-tail” catgories (Jayaraman and Grauman 2016; Mobahi, Collobert, and Weston 2009). Moreover, deep action image recognition approaches (Zhang et al. 2016; Gao, Xiong, and Grauman 2018) generally require CNN models pretrained on large-scaled labeled datasets (*e.g.*, ImageNet (Deng et al. 2009)) to obtain advanced visual representations of static images, while the collection of such datasets is very expensive and time-consuming. Instead, unsupervised visual representation learning is gaining momentum and attracting many researchers (Zhang, Isola, and Efros 2017; Doersch, Gupta, and Efros 2015; Noroozi and Favaro 2016; Gidaris, Singh, and Komodakis 2018; Jayaraman and Grauman 2016; Fernando et al. 2017; Misra, Zitnick, and Hebert 2016; Lee et al. 2017).

In this paper, we focus on static image action recognition with minimum supervision, *i.e.*, only a few labeled action images and no pretrained CNN model. Despite the difficulty of this task, we humans are able to recognize the action in static images even after seeing only a few labeled action images, because our visual cortex can hypothesize motion signals from static image (Schmid et al. 2013). Inspired by neuropsychology, several works have attempted to learn motion cues from unlabeled videos to help image action recognition, because a large amount of unlabeled videos are freely available. These works can be divided into two research lines: (a) Enhance visual representations of action images with a proxy task based on unlabeled videos, which can capture supervisory signals in unlabeled videos such as temporal coherence (Mobahi, Collobert, and Weston 2009; Hadsell, Chopra, and LeCun 2006a; Jayaraman and Grauman 2016) and temporal order (Misra, Zitnick, and Hebert 2016; Lee et al. 2017; Srivastava, Mansimov, and Salakhutdinov 2015). This group of approaches are located in the realm of self-supervised learning, in which the main task is image action recognition and different proxy tasks are designed based on unlabeled videos; (b) Generate auxiliary representations for action images with the generator learned from unlabeled videos. For example, future visual representation and optical flow map are generated from action images in (Vondrick, Pirsiavash, and Torralba 2016) and (Gao, Xiong, and Grauman 2018; Walker, Gupta, and Hebert 2015) respectively.

One question is whether the above two research lines can be unified to fully exploit unlabeled videos. In this work,

[*] Corresponding author

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

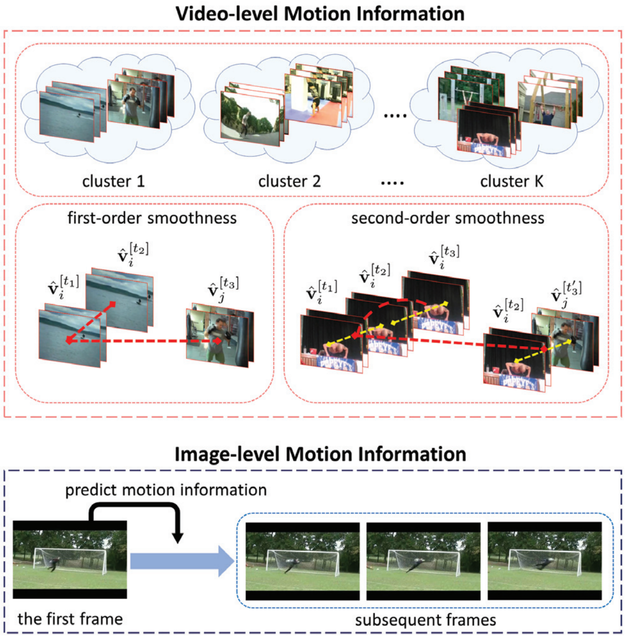
Video-level Motion Information
cluster 1
cluster 2
....
cluster K
first-order smoothness
$\hat{\mathbf{v}}_i^{[t_1]}$
$\hat{\mathbf{v}}_i^{[t_2]}$
$\hat{\mathbf{v}}_j^{[t_3]}$
second-order smoothness
$\hat{\mathbf{v}}_i^{[t_1]}$
$\hat{\mathbf{v}}_i^{[t_2]}$
$\hat{\mathbf{v}}_i^{[t_3]}$
$\hat{\mathbf{v}}_i^{[t_2]}$
$\hat{\mathbf{v}}_j^{[t_3']}$
Image-level Motion Information
predict motion information
the first frame
subsequent frames

Figure 1: Our framework exploits both video-level motion information and image-level motion information from unlabeled videos. On one hand, we exploit the discriminability and temporal coherence (*i.e.*, first-order and second-order smoothness) of video-level motion information (see Section 3.1 for the definition of $\hat{\mathbf{v}}_i^{[t]}$). On the other hand, we exploit the image-level motion information by predicting the motion information of first frames in unlabeled videos.

we explore this question by proposing a framework which can enhance the visual representations of action images and also generate auxiliary motion representations for action images. Our framework is composed of Visual Representation Enhancement (VRE) module and Motion Representation Augmentation (MRA) module: (a) On one hand, our VRE module leverages video-level motion information from unlabeled videos to enhance visual representations of action images. Specifically, we design a proxy task based on the discriminability of video-level motion information, that is, unlabeled video clips could be classified into different clusters based on their pseudo motion labels (see Figure 1). Inspired by (Jayaraman and Grauman 2016), we also utilize the temporal coherence of video-level motion information, *i.e.*, first-order and second-order smoothness between neighboring video clips (see Figure 1), as supervisory signals; (b) On the other hand, our MRA module generates auxiliary motion representations for action images by exploiting the image-level motion information in unlabeled videos. Specifically, through predicting the motion information of the first frames in unlabeled videos (see Figure 1), our MRA module is enabled to predict the motion information of static action images, which is used as auxiliary motion representations. Unlike (Gao, Xiong, and Grauman 2018; Walker, Gupta, and Hebert 2015) that generate low-level optical flow from action images, we generate more robust high-level motion information from action images. Finally, we concatenate the enhanced visual representations and the generated motion representations of action images for the final classification.

Our contributions are threefold: 1) We propose a novel framework with VRE and MRA modules for static image action recognition, which unifies two research lines of exploiting unlabeled videos; 2) Our VRE and MRA module can take full advantage of video-level and image-level motion information from unlabeled videos; 3) Our framework outperforms the state-of-the-art image action recognition methods which exploit unlabeled videos, when minimum supervision is available.

2 Related Work

In the section, we discuss the existing works which use unlabeled videos to help static image action recognition.

2.1 Enhancing Visual Representation via Self-supervised Learning

When using unlabeled videos to help image action recognition, many methods belong to self-supervised learning (Jayaraman and Grauman 2016; Fernando et al. 2017; Misra, Zitnick, and Hebert 2016; Lee et al. 2017), which consists of a main task and a proxy task. Particularly, the main task is image action recognition and the proxy task is exploiting supervisory signals in unlabeled videos. Because the main task and the proxy task share the same visual representations, the supervisory signals of proxy task can help learn better visual representation in the main task.

A diversity of proxy tasks have been designed to leverage unlabeled videos. Several methods (Wiskott and Sejnowski 2002; Mobahi, Collobert, and Weston 2009; Hadsell, Chopra, and LeCun 2006a; Jayaraman and Grauman 2016) investigated first-order or second-order smoothness, which means smooth transition between neighboring frames. Temporal order of video frames has also been studied in prior research (Fernando et al. 2015; Wang and Gupta 2015; Fernando et al. 2017; Misra, Zitnick, and Hebert 2016; Lee et al. 2017). Besides, there are some other types of proxy tasks using unlabeled videos. To name a few, the proxy task in (Purushwalkam and Gupta 2016) is to verify whether an optical flow map corresponds to a pair of starting and ending frames, while the proxy task in (Srivastava, Mansimov, and Salakhutdinov 2015) is using an encoder LSTM and a decoder LSTM to reconstruct video sequences.

Different from the above methods, video motion classification based on pseudo motion labels, i.e. the proxy task in our Visual Representation Enhancement (VRE) module, has not been explored in self-supervised learning previously. Although our proxy task also utilizes smoothness regularization similar to (Jayaraman and Grauman 2016), our smoothness regularizer is imposed on video clips instead of frames to eliminate the disturbance of camera movement. Moreover, our framework can also generate auxiliary representations for action images to further improve the performance, which is out of the scope of self-supervised learning.

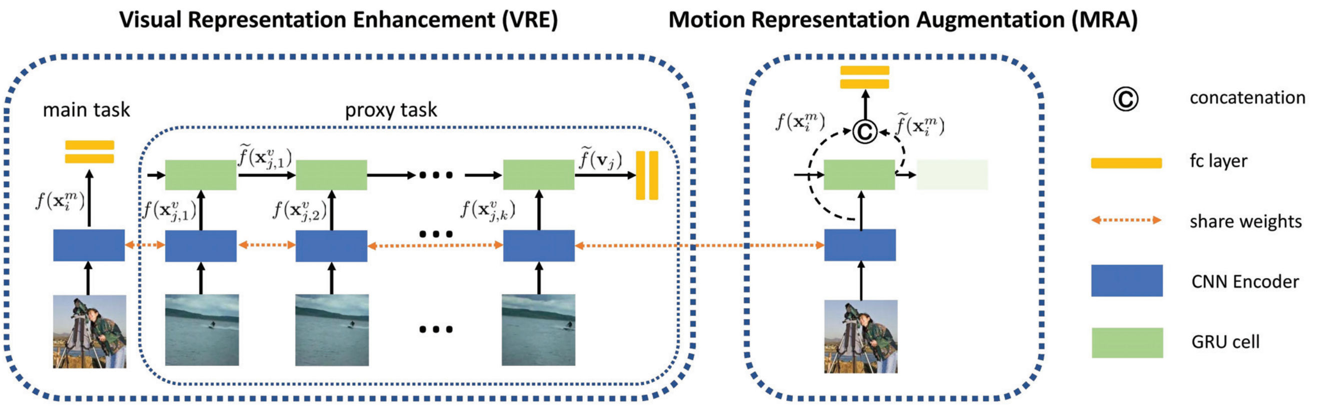

Figure 2: Left: Our Visual Representation Enhancement (VRE) module is composed of a main task and a proxy task. The main task is image action recognition based on visual features $f(\mathbf{x}_i^m)$. The proxy task is video motion classification based on video-level motion features $\widetilde{f}(\mathbf{v}_j)$ with pseudo motion labels. All Encoders share the same model parameters and so do all GRU cells. Right: Our Motion Representation Augmentation (MRA) module generates image-level motion features $\widetilde{f}(\mathbf{x}_i^m)$ for static action images, which are concatenated with visual features $f(\mathbf{x}_i^m)$ for the final classification.

2.2 Generating Auxiliary Information from Action Images

As another branch of using unlabeled videos for image action recognition, some methods aim to learn a generator based on unlabeled videos, and use the learned generator to produce auxiliary information from action images. For example, Carl *etal.* (Vondrick, Pirsiavash, and Torralba 2016) proposed to anticipate future visual representation of a static image. More works tended to predict motion information (*e.g.*, optical flow, trajectory) of a static image (Walker, Gupta, and Hebert 2015; Gao, Xiong, and Grauman 2018; Pintea, van Gemert, and Smeulders 2014).

The closest related work is (Walker, Gupta, and Hebert 2015; Gao, Xiong, and Grauman 2018), in which motion information can be predicted from a static image. Their methods produce low-level optical flow map, while our Motion Representation Augmentation (MRA) module generates high-level motion representation that is more robust and discriminative. Moreover, our framework can also enhance visual representations of action images, which cannot be achieved by (Walker, Gupta, and Hebert 2015; Gao, Xiong, and Grauman 2018).

3 Our Method

For ease of presentation, in the remainder of this paper, we use a letter in boldface to denote a vector (*e.g.*, $\mathbf{a}$) and a plain letter to denote a scalar (*e.g.*, a). Besides, we use $\mathbf{a}^T$ to denote the transpose of vector $\mathbf{a}$. Assume we have a labeled action image dataset $\mathcal{D}^m = \{(\mathbf{x}_i^m, \mathbf{y}_i)|_{i=1}^{N_m}\}$, in which N_m is the number of images and $\mathbf{y}_i$ is the one-hot label vector of $\mathbf{x}_i^m$ with a single entry corresponding to its category label being 1. Besides the labeled image dataset, we also have a set of unlabeled videos. For each unlabeled video, we sample the first k key frames at the interval of 3 frames to form a new video clip, in order to alleviate the influence of camera motion. Thus, we obtain a video clip dataset $\mathcal{D}^v = \{\mathbf{v}_i|_{i=1}^{N_v}\}$ with $\mathbf{v}_i = \{\mathbf{x}_{i,1}^v, \ldots, \mathbf{x}_{i,k}^v\}$. In the training stage, we tend to use both $\mathcal{D}^m$ and $\mathcal{D}^v$ to learn a robust action image classifier. As illustrated in Figure 2, our framework is composed of two modules: Visual Representation Enhancement (VRE) module and Motion Representation Augmentation (MRA) module.

3.1 Visual Representation Enhancement (VRE) Module

In our VRE module, we enhance visual representations of action images by exploiting discriminative and smooth motion information of unlabeled videos. Following the terminology of self-supervised learning, our main task is image action recognition and our proxy task is video motion classification with pseudo motion labels.

Image Action Recognition as Main task: Our main task is static image action recognition and we use CNN model (*e.g.*, AlexNet) with random initialization as our backbone network. By taking AlexNet (Krizhevsky, Sutskever, and Hinton 2017) as an example, we denote AlexNet without the last two fully connected (fc) layers as a visual feature extractor $f(\cdot)$ with model parameters $\boldsymbol{\theta}_n$, which is also referred to as CNN Encoder in the remainder of this paper. Thus, the visual feature of action image $\mathbf{x}_i^m$ can be represented by $f(\mathbf{x}_i^m)$. Besides, we denote the last two fc layers of AlexNet as a visual classifier $p_c(\cdot)$ with model parameters $\boldsymbol{\theta}_c$. Then, the cross-entropy classification loss of main task can be written as

$$L_{main} = -\sum_{i=1}^{N_m} \mathbf{y}_i^T \log p_c(f(\mathbf{x}_i^m)). \qquad (1)$$

Video Motion Classification as Proxy Task: Our proxy task is video motion classification with pseudo motion labels, where the discriminability of video-level motion information of unlabeled videos can be exploited. We cluster unlabeled videos using K-means based on their hand-crafted features ($K = 16$ in our experiments) and treat the cluster label of each video clip as its pseudo motion label. The details of hand-crafted features are left to Supplementary. To perform video motion classification based on the obtained pseudo motion labels, we employ Recurrent Neural Network (RNN) architecture with multiple cells sharing the same model parameters, as shown in Figure 2. Specifically, we adopt Gated Recurrent Unit (GRU) cell (Chung et al. 2015) with model parameters $\boldsymbol{\theta}_g$ in RNN architecture. Given a video clip $\mathbf{v}_j$, the t-th GRU cell takes the output from the $(t-1)$-th GRU cell which is denoted as $\widetilde{f}(\mathbf{x}^v_{j,t-1})$ and the visual feature $f(\mathbf{x}^v_{j,t})$ of the t-th frame as input. Video motion classification is based on the output $\widetilde{f}(\mathbf{x}^v_{j,k})$ from the last GRU cell, which is also denoted as $\widetilde{f}(\mathbf{v}_j)$. $\widetilde{f}(\mathbf{v}_j)$ is supposed to represent the video-level motion feature because of the supervision by pseudo motion labels.

By using $\hat{\mathbf{y}}_i$ to denote the one-hot pseudo motion label vector of $\mathbf{v}_i$, the cross-entropy classification loss can be written as

$$L_{motion} = -\sum_{i=1}^{N_v} \hat{\mathbf{y}}_i^T \log p_m(\widetilde{f}(\mathbf{v}_i)), \tag{2}$$

in which $p_m(\cdot)$ is the motion classifier consisting of two fully connected (fc) layers with model parameters $\boldsymbol{\theta}_m$.

The proxy task shares the same CNN Encoder with the main task, so the supervision information from the proxy task can contribute to better visual representation in the main task.

Video-level Smoothness Regularization: Inspired by (Jayaraman and Grauman 2016), we additionally exploit the temporal coherence of unlabeled videos to further enhance visual representations. Unlike (Jayaraman and Grauman 2016), we impose smoothness constraints on video-level motion features in lieu of image-level visual features, to mitigate the noise induced by camera motion.

First, we segment some shorter video clips based on $\mathcal{D}^v = \{\mathbf{v}_i|_{i=1}^{N_v}\}$. Specifically, we use $\hat{\mathbf{v}}_i^{[t]} = \{\mathbf{x}^v_{i,t}, \cdots, \mathbf{x}^v_{i,t+\Delta t-1}\}$ to denote a shorter video clip with the starting frame t and duration Δt, which is segmented from $\mathbf{v}_i$. We collect the set of temporal neighbor pairs $\mathcal{S}_1 = \{(\hat{\mathbf{v}}_i^{[t_1]}, \hat{\mathbf{v}}_i^{[t_2]}) \;|_{i=1}^{N^v} \; |t_2 - t_1| \leq \lceil \Delta t/2 \rceil\}$, which means that in a temporal neighbor pair, two video clips should be from the same $\mathbf{v}_i$ and the overlap between them should be at least half of their duration Δt. We impose first-order smoothness regularization on $\mathcal{S}_1$, which can be formulated as contrastive loss (Hadsell, Chopra, and LeCun 2006b) as follows,

$$L_{smooth1} = \sum_{(\hat{\mathbf{v}}_i^{[t_1]}, \hat{\mathbf{v}}_i^{[t_2]}) \in \mathcal{S}_1} d\left(\widetilde{f}(\hat{\mathbf{v}}_i^{[t_1]}), \widetilde{f}(\hat{\mathbf{v}}_i^{[t_2]})\right)$$
$$+ \sum_{(\hat{\mathbf{v}}_i^{[t_1]}, \hat{\mathbf{v}}_j^{[t_2]}) \notin \mathcal{S}_1} \max\left(\delta - d\left(\widetilde{f}(\hat{\mathbf{v}}_i^{[t_1]}), \widetilde{f}(\hat{\mathbf{v}}_j^{[t_2]})\right), 0\right),$$

in which $d(\cdot,\cdot)$ measures the Euclidean distance and the margin δ is set to 1 in our experiments. $\widetilde{f}(\mathbf{v})$ is the video-level motion feature of $\mathbf{v}$, which is extracted by the RNN architecture consisting of Δt GRU cells.

In analogy to temporal neighbor pairs, we collect the set of temporal neighbor tuples $\mathcal{S}_2 = \{(\hat{\mathbf{v}}_i^{[t_1]}, \hat{\mathbf{v}}_i^{[t_2]}, \hat{\mathbf{v}}_i^{[t_3]}) \;|_{i=1}^{N^v} \; |t_2 - t_1| \leq \lceil \Delta t/2 \rceil, |t_3 - t_2| \leq \lceil \Delta t/2 \rceil\}$, which means that in a temporal neighbor tuple, three video clips should be from the same $\mathbf{v}_i$, and they should be overlapped by at least half of their duration Δt. Besides, for each tuple in $\mathcal{S}_2$, we replace the third element with a video clip $\hat{\mathbf{v}}_j^{[t_3]}$ from another $\mathbf{v}_j$, leading to another set of tuples $\hat{\mathcal{S}}_2 = \{(\hat{\mathbf{v}}_i^{[t_1]}, \hat{\mathbf{v}}_i^{[t_2]}, \hat{\mathbf{v}}_j^{[t_3]}) \;|_{i,j=1}^{N^v} \; i \neq j, |t_2 - t_1| \leq \lceil \Delta t/2 \rceil\}$.

By denoting $\hat{d}(t_1, t_2) = \widetilde{f}(\hat{\mathbf{v}}_i^{[t_1]}) - \widetilde{f}(\hat{\mathbf{v}}_i^{[t_2]})$, we impose second-order smoothness regularization on $\mathcal{S}_2$ and $\hat{\mathcal{S}}_2$ by using contrastive loss:

$$L_{smooth2} = \sum_{(\hat{\mathbf{v}}_i^{[t_1]}, \hat{\mathbf{v}}_i^{[t_2]}, \hat{\mathbf{v}}_i^{[t_3]}) \in \mathcal{S}_2} d\left(\hat{d}(t_1, t_2), \hat{d}(t_2, t_3)\right)$$
$$+ \sum_{(\hat{\mathbf{v}}_i^{[t_1]}, \hat{\mathbf{v}}_i^{[t_2]}, \hat{\mathbf{v}}_j^{[t_3]}) \in \hat{\mathcal{S}}_2} \max\left(\delta - d\left(\hat{d}(t_1, t_2), \hat{d}(t_2, t_3)\right), 0\right),$$

which has similar explanation to $L_{smooth1}$. After incorporating smoothness regularization, the loss of the proxy task in our VRE module can be written as

$$L_{proxy} = L_{motion} + L_{smooth1} + L_{smooth2}. \tag{3}$$

Then, we can arrive at the total loss of our VRE module

$$L_{VRE} = L_{main} + \lambda L_{proxy}, \tag{4}$$

where λ is a trade-off parameter and set to 0.1 in our experiments.

3.2 Motion Representation Augmentation (MRA) Module

After the RNN architecture is trained in our VRE module, we assume that the output from the last GRU cell represents the video-level motion information of the input video clip due to the supervision of pseudo motion labels. In our MRA module, we only use one GRU cell learned in our VRE module to generate motion representation for a static action image. Note that in the proxy task of our VRE module, the first GRU cell uses a constant vector (*i.e.*, all-zero vector) to supersede the output from previous GRU cell, which means that the first frame in a video does not have prior motion information from previous frames. Actually, a static action image $\mathbf{x}_i^m$ can be treated as the first frame of an imaginary video clip with its prior motion information unknown. Therefore, we feed the same constant all-zero vector and $f(\mathbf{x}_i^m)$ into a GRU cell, as illustrated in Figure 2. We assume that the output $\widetilde{f}(\mathbf{x}_i^m)$ from this GRU cell could represent the high-level motion information of $\mathbf{x}_i^m$, which is visualized and verified in our experiments.

Although CNN Encoder is also supervised by unlabeled videos in the proxy task, it focuses on the main task of image action recognition. So we claim that CNN Encoder produces visual representations which are enhanced by the proxy task. In contrast, GRU cell is only supervised by unlabeled videos

Algorithm 1: Training procedure of our framework.

```
Input: Labeled images $\mathcal{D}^m = \{(\mathbf{x}_i^m, \mathbf{y}_i)|_{i=1}^{N_m}\}$ and
       unlabeled video clips $\mathcal{D}^v = \{\mathbf{v}_i|_{i=1}^{N_v}\}$.
Acquire pseudo motion labels for $\mathcal{D}^v = \{\mathbf{v}_i|_{i=1}^{N_v}\}$.
Initialize $\{\boldsymbol{\theta}_c, \boldsymbol{\theta}_n, \boldsymbol{\theta}_g, \boldsymbol{\theta}_m, \boldsymbol{\theta}_a\}$ randomly.
/* Visual Representation Enhancement
   Module                                  */
Update $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_c\}$ based on $L_{main}$ in (1).
Update $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_c, \boldsymbol{\theta}_g, \boldsymbol{\theta}_m\}$ based on $L_{VRE}$ in (4).
Update $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_c\}$ based on $L_{main}$ in (1).
/* Motion Representation Augmentation
   Module                                  */
Update $\{\boldsymbol{\theta}_g, \boldsymbol{\theta}_m\}$ based on $L_{proxy}$ in (3).
Update $\boldsymbol{\theta}_a$ based on $L_{MRA}$ in (5).
return $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_g, \boldsymbol{\theta}_a\}$
```

and focuses on learning transferable motion representation. So the generated motion representations of action images are complementary with their visual representations to some extent. Therefore, the combination of both is likely to boost the performance in action recognition.

For each action image, we concatenate its visual representation $f(\mathbf{x}_i^m)$ and motion representation $\widetilde{f}(\mathbf{x}_i^m)$ as $[\widetilde{f}(\mathbf{x}_i^m); f(\mathbf{x}_i^m)]$, followed by two fully connected (fc) layers with model parameters $\boldsymbol{\theta}_a$ which play the role as the final classifier $p_a(\cdot)$. The cross-entropy classification loss can be written as

$$L_{MRA} = -\sum_{i=1}^{N_m} \mathbf{y}_i^T \log p_a([\widetilde{f}(\mathbf{x}_i^m); f(\mathbf{x}_i^m)]), \qquad (5)$$

which resembles L_{main} in (1) except that (5) uses $[\widetilde{f}(\mathbf{x}_i^m); f(\mathbf{x}_i^m)]$ instead of $f(\mathbf{x}_i^m)$.

Training Procedure: After randomly initializing all model parameters, our whole training procedure has five steps: (1) Initialize CNN model (*i.e.*, $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_c\}$) based on the main task in our VRE module, in which only labeled action images are used; (2) Train CNN model (*i.e.*, $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_c\}$), GRU cell (*i.e.*, $\boldsymbol{\theta}_g$), and motion classifier (*i.e.*, $\boldsymbol{\theta}_m$) based on the main task and the proxy task in our VRE module, in which both labeled action images and video clips with pseudo motion labels are used; (3) Fine-tune CNN model (*i.e.*, $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_c\}$) based on the main task in our VRE module, which is similar to the first step. Up to now, we hope to obtain satisfactory visual representations for action images; (4) Fine-tune GRU cell (*i.e.*, $\boldsymbol{\theta}_g$) and motion classifier (*i.e.*, $\boldsymbol{\theta}_m$) based on the proxy task in our VRE module, in which only video clips with pseudo motion labels are used. Up to now, we expect to obtain satisfactory motion representations for action images; (5) Train the classifier (*i.e.*, $\boldsymbol{\theta}_a$) based on the concatenated visual and motion representations, in which only labeled action images are used. In the testing stage, given a test image, we pass it through the CNN Encoder and GRU cell to obtain its visual representation and motion representation, which are concatenated for the final prediction.

4 Experiments

We evaluate our framework through extensive experiments on two action image datasets and two unlabeled video datasets.

4.1 Experimental Setup

We use two action image datasets (*i.e.*, PASCAL VOC and Stanford40) and two video datasets (*i.e.*, UCF101 and HMDB), which can form four pairs of video→image datasets corresponding to four experimental settings (*i.e.*, UCF101→VOC, UCF101→Stanford40, HMDB→VOC, and HMDB→Stanford40).

Action image datasets: PASCAL VOC (Everingham et al. 2010) is an action image dataset with 10 categories and Stanford40 (Yao et al. 2011) contains 40 daily human actions. Since this work focuses on minimum supervision, we follow (Jayaraman and Grauman 2016) to use only a few labeled training images per category. On VOC dataset, we use 10 labeled training images per category and 2000 images for testing. On the Stanford40 dataset, we use 30 labeled training images per category and standard 5532 test images specified in (Yao et al. 2011). Besides using minimum supervision which is the focus of this paper, we also investigate using more training samples per category in Section 4.6.

Video datasets: HMDB (Kuehne et al. 2011) contains 6849 short video clips distributed in 51 actions and UCF101 (Soomro, Zamir, and Shah 2012) contains 13320 videos from 101 action categories. Note that we do not use the provided action labels of videos. Following (Jayaraman and Grauman 2016), we randomly sample 1000 video clips from both datasets, *i.e.*, $N_v = 1000$. We extract the first $k = 12$ (*resp.*, $k = 7$) key frames from each video at the interval of 3 frames on the UCF101 (*resp.*, HMDB) dataset considering the difference of video lengths between these two datasets. For video motion classification, we split each k-frame video clip into three $(k-2)$-frame video clips to augment training data. For smoothness regularization, Δt is set as 10 (*resp.*, 5) on the UCF101 (*resp.*, HMDB) dataset.

Implementation details: We use AlexNet as backbone by default. The visual representation from CNN Encoder is 4096-dim and the motion representation from GRU cell is 512-dim. The motion classifier $f_m(\cdot)$ in our VRE module has two fully connected layers with intermediate 512-dim output. The classifier $f_a(\cdot)$ based on concatenated 4608-dim features in MRA module also has two fully connected layers with intermediate 4608-dim output. During training, we use Adam optimizer, in which the learning rate starts with 0.0001 and decays by 0.1 every 1800 iterations.

4.2 Baselines

We compare with two groups of state-of-the-art baselines: 1) The first group of baselines design different proxy tasks based on unlabeled videos to enhance visual representations, which falls into the scope of self-supervised learning. Specifically, we compare with O3N (Fernando et al.

Table 1: Accuracies (%) of different methods in four experimental settings. The best results are denoted in boldface.

Datasets	UCF101→VOC	UCF101→S40	HMDB→VOC	HMDB→S40
UNREG	16.22	15.42	16.22	15.42
O3N (Fernando et al. 2017)	16.28	16.36	16.28	16.79
Lee et al. (2017)	18.00	17.19	16.89	16.52
OPN (Misra, Zitnick, and Hebert 2016)	18.44	16.83	18.39	16.63
Srivastava, Mansimov, and Salakhutdinov (2015)	16.42	15.99	18.39	16.42
Purushwalkam and Gupta (2016)	16.55	16.06	16.55	16.89
SSFA (Jayaraman and Grauman 2016)	15.11	16.27	17.06	17.66
Vondrick, Pirsiavash, and Torralba (2016)	19.28	17.42	17.94	16.87
Gao, Xiong, and Grauman (2s) (2018)	17.94	15.29	16.84	14.67
Gao, Xiong, and Grauman (2018)	18.39	16.87	17.94	15.89
Pintea, van Gemert, and Smeulders (2014)	18.39	15.93	19.17	17.08
Walker, Gupta, and Hebert (2015)	19.11	17.44	17.56	15.69
Ours	**21.06**	**18.76**	**22.94**	**19.11**

Table 2: Accuracies (%) of our special cases in two experimental settings.

Datasets	UCF101→VOC	HMDB→VOC
UNREG	16.22	16.22
UNREG+motion	19.39	19.72
UNREG+smooth1	18.89	18.61
UNREG+smooth2	18.72	18.11
Ours (w/o MRA)	20.06	20.44
only MR	13.72	18.17
Ours	21.06	22.94
SUP-FT	28.72	28.72

2017), OPN (Misra, Zitnick, and Hebert 2016), Misra *etal.* (Lee et al. 2017), SSFA (Jayaraman and Grauman 2016), Srivastava *etal.* (Srivastava, Mansimov, and Salakhutdinov 2015), and Purushwalkam *etal.* (Purushwalkam and Gupta 2016); 2) The second group of baselines generate auxiliary representations for action images, with the generator learned based on unlabeled videos. We compare with Carl *etal.* (Vondrick, Pirsiavash, and Torralba 2016), which can generate future visual representation given a static action image. For each image, we concatenate its current visual representation and future visual representation, followed by two fully connected layers for classification. Moreover, we compare with Jacob *etal.* (Walker, Gupta, and Hebert 2015), Pintea *etal.* (Pintea, van Gemert, and Smeulders 2014), and Gao *etal.* (Gao, Xiong, and Grauman 2018), which can generate optical flow map given a static action image. We extract features from their generated optical flow maps in a similar way to our hand-crafted video features, and then concatenate them with visual representations for the final classification. For Gao *etal.* (Gao, Xiong, and Grauman 2018), we additionally adopt two-stream network (Simonyan and Zisserman 2014a) by combining spatial stream CNN and temporal stream CNN for prediction score fusion, which is referred to as Gao *etal.* (2s) in Table 1.

Additionally, we include the baseline UNREG, which stands for UNREGularized network. Specifically, we train a CNN model from scratch by using labeled training images based on L_{main} in (2), without the help of unlabeled videos.

4.3 Experimental Results

The results of different methods are reported in Table 1, from which we observe that the first group of baselines are generally better than UNREG, which proves that it is helpful to use various proxy tasks based on unlabeled videos to enhance visual representations. The second group of baselines also generally outperform UNREG, which indicates the advantage of future visual representation or auxiliary motion information. Finally, our framework outperforms two groups of baselines and achieves the best results in all four experimental settings, which corroborates the effectiveness of enhancing visual representations by video motion classification and generating auxiliary motion representations.

4.4 Ablation Studies

To validate each component of our framework, we perform extensive ablation studies on our special cases in this section. Firstly, we focus on our VRE module and verify the effectiveness of each loss term (*i.e.*, L_{motion}, $L_{smooth1}$, $L_{smooth2}$). In particular, we perform the first three steps in the training procedure and replace L_{proxy} in the second step with only L_{motion} (*resp.*, $L_{smooth1}$, $L_{smooth2}$), which is referred to as UNREG+motion (*resp.*, UNREG+smooth1, UNREG+smooth2) in Table 2. Similarly, we perform the first three steps in the training procedure and use full L_{proxy} in the second step, which is referred to as Ours (w/o MRA) in Table 2 because this special case does not use the MRA module. Secondly, we focus on the MRA module and test the performance only using auxiliary Motion Representation (MR), which is referred to as "only MR" in Table 2. In this special case, we need to learn a classifier based on auxiliary motion representations.

We also include the baseline SUP-FT, which stands for SUPervised-pretrained and Fine-Tune. Specifically, we finetune the CNN model pretrained on the ImageNet dataset using the labeled action training images based on L_{main} in (2), without the help of unlabeled videos.

By taking UCF101→VOC and HMDB→VOC as examples, the results of all special cases are summarized in Table 2. By comparing UNREG+motion with UNREG, we observe that the proxy task of video motion classification with pseudo motion labels plays a crucial role in our

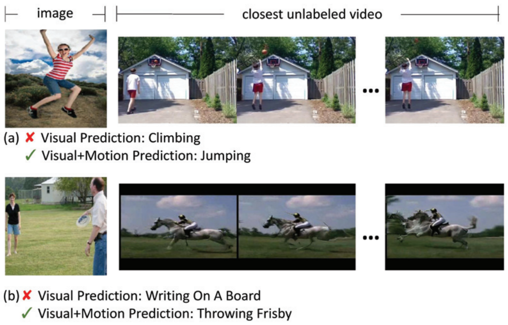
image
closest unlabeled video
...
(a) ✗ Visual Prediction: Climbing
✓ Visual+Motion Prediction: Jumping
...
(b) ✗ Visual Prediction: Writing On A Board
✓ Visual+Motion Prediction: Throwing Frisby

Figure 3: Images are misclassified based on visual representation, but are correctly classified based on both visual representation and motion representation.

VRE module, which achieves 2.86% improvement over UNREG on average. Besides, first-order (*resp.*, second-order) smoothness regularization are also helpful when comparing UNREG+smooth1 (*resp.*, UNREG+smooth2) with UNREG. The performance only using auxiliary motion representations is better than random guess, and even better than UNREG in HMDB→VOC setting, suggesting that the GRU cell could generate meaningful motion representations. It can also been seen that our full-fledged method outperforms Ours (w/o MRA), which demonstrates that it is beneficial to augment visual representations with motion representations.

Based on Table 2, there is a performance gap between our method and SUP-FT. However, SUP-FT is pretrained on ImageNet with millions of labeled images which requires sheer expense of human annotation, while our method only utilizes minimum supervision.

4.5 Qualitative Analyses of Motion Representation

To visualize auxiliary motion representation (MR) of action images, for each action image $\mathbf{x}_i^m$, we show the video clip whose first frame has the closest MR to this action image. Specifically, after comparing $\widetilde{f}(\mathbf{x}_i^m)$ of image $\mathbf{x}_i^m$ with $\widetilde{f}(\mathbf{x}_{j,1}^v)$ of all unlabeled video clips $\mathbf{v}_j$ in $\mathcal{D}^v$, we retrieve the video clip $\mathbf{v}_{j^*}$ with minimum $||\widetilde{f}(\mathbf{x}_i^m) - \widetilde{f}(\mathbf{x}_{j^*,1}^v)||_2$. The retrieved video clips are assumed to visually represent the motion information encoded in MR.

In Figure 3, we observe that auxiliary MR can rectify wrong prediction. For example, the image in (b) is misclassified as “writing on a board” based on visual representation, probably because the frisby visually resembles a white board. However, the motion information in MR, *i.e.*, similar hand movement in “horse riding”, helps classify this image as “throwing frisby” correctly. The experimental results show that MR can capture the motion information of static action images and compensate the drawback of visual representation. More results are provided in Supplementary.

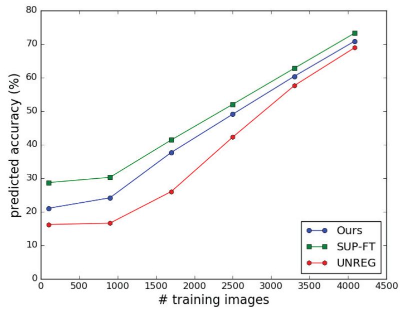

Figure 4: Accuracies (%) of UNREG, SUP-FT, and our framework in UCF101→VOC setting with different numbers of labeled training images.

4.6 Impact of Different Numbers of Labeled Action Images

We have demonstrated that exploiting motion information from unlabeled videos is beneficial when the number of labeled training images is limited. It is interesting to investigate whether exploiting motion information from unlabeled videos is still useful when the number of labeled training images increases. By taking UCF101 → VOC as an example, we vary the number of labeled training images in [100, 900, 1700, 2500, 3300, 4087], in which 4087 is the maximum number of available training images. The performance of UNREG, SUP-FT, and our framework is reported in Figure 4, from which it can be seen that our framework is consistently effective when using different numbers of labeled training images. When the number of labeled training images is very large (*e.g.*, 4087), the performance of UNREG is approaching SUP-FT, which means that labeled action images already have sufficient supervision information. In this case, the performance gain brought by our framework is also becoming marginal.

4.7 Impact of Different Network Structures

To demonstrate the generalizability of our framework, we replace the AlexNet backbone with VGG16 (Simonyan and Zisserman 2014b). Experimental results are left to Supplementary due to space limitation.

5 Conclusion

In this paper, we proposed a novel framework, which unifies two research lines of exploiting unlabeled videos to help static image action recognition: Visual Representation Enhancement (VRE) module and Motion Representation Augmentation (MRA) module. Comprehensive experimental results have revealed the superiority of our framework.

6 Acknowledgement

This work is supported by the National Key R&D Program of China (2018AAA0100704) and is partially sponsored by Shanghai Sailing Program (BI0300271) and Startup Fund for Youngman Research at SJTU (WF220403041).

References

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2015. Gated feedback recurrent neural networks. In *ICML*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*.

Everingham, M.; Gool, L. J. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2010. The pascal visual object classes (VOC) challenge. *IJCV* 88(2):303–338.

Fernando, B.; Gavves, E.; M., J. O.; Ghodrati, A.; and Tuytelaars, T. 2015. Modeling video evolution for action recognition. In *CVPR*.

Fernando, B.; Bilen, H.; Gavves, E.; and Gould, S. 2017. Self-supervised video representation learning with odd-one-out networks. In *CVPR*.

Gao, R.; Xiong, B.; and Grauman, K. 2018. Im2flow: Motion hallucination from static images for action recognition. In *CVPR*.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *ICLR*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006a. Dimensionality reduction by learning an invariant mapping. In *Computer Society Conference on Computer Vision and Pattern Recognition*, 1735–1742.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006b. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

Jayaraman, D., and Grauman, K. 2016. Slow and steady feature analysis: Higher order temporal coherence in video. In *CVPR*, 3852–3861.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60(6):84–90.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T. A.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *ICCV*.

Lee, H.; Huang, J.; Singh, M.; and Yang, M. 2017. Unsupervised representation learning by sorting sequences. In *ICCV*, 667–676.

Mezghiche, K. M.; Melkemi, K. E.; and Foufou, S. 2014. Matching with quantum genetic algorithm and shape contexts. In *AICCSA*.

Misra, I.; Zitnick, C. L.; and Hebert, M. 2016. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*.

Mobahi, H.; Collobert, R.; and Weston, J. 2009. Deep learning from temporal coherence in video. In *ICML*.

Noroozi, M., and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*.

Pintea, S. L.; van Gemert, J. C.; and Smeulders, A. W. M. 2014. Déjà vu: - motion prediction in static images. In *ECCV*, 172–187.

Purushwalkam, S., and Gupta, A. 2016. Pose from action: Unsupervised learning of pose features based on motion. *CoRR* abs/1609.05420.

Schmid, M. C.; Schmiedt, J. T.; Peters, A. J.; Saunders, R. C.; Maier, A.; and Leopold, D. A. 2013. Motion-sensitive responses in visual area v4 in the absence of primary visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33(48):18740–5.

Sharma, G.; Jurie, F.; and Schmid, C. 2012. Discriminative spatial saliency for image classification. In *CVPR*.

Simonyan, K., and Zisserman, A. 2014a. Two-stream convolutional networks for action recognition in videos. In *NIPS*.

Simonyan, K., and Zisserman, A. 2014b. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR* abs/1212.0402.

Srivastava, N.; Mansimov, E.; and Salakhutdinov, R. 2015. Unsupervised learning of video representations using lstms. In *ICML*.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2016. Anticipating visual representations from unlabeled video. In *CVPR*.

Walker, J.; Gupta, A.; and Hebert, M. 2015. Dense optical flow prediction from a static image. In *ICCV*.

Wang, X., and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *ICCV*, 2794–2802.

Wiskott, L., and Sejnowski, T. J. 2002. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation* 14(4):715–770.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L. J.; and Li, F. 2011. Human action recognition by learning bases of action attributes and parts. In *ICCV*.

Zhang, Y.; Cheng, L.; Wu, J.; Cai, J.; Do, M. N.; and Lu, J. 2016. Action recognition in still images with minimum annotation efforts. *T-IP* 25(11):5479–5490.

Zhang, R.; Isola, P.; and Efros, A. A. 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*.

Zheng, Y.; Zhang, Y.; Li, X.; and Liu, B. 2012. Action recognition in still images using a combination of human pose and context information. In *ICIP*.