Unsupervised Learning of Long-Term Motion Dynamics for Videos

Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, Li Fei-Fei
Stanford University
{zelunluo,boya,dahuang,alahi,feifeili}@cs.stanford.edu

Abstract

We present an unsupervised representation learning approach that compactly encodes the motion dependencies in videos. Given a pair of images from a video clip, our framework learns to predict the long-term 3D motions. To reduce the complexity of the learning framework, we propose to describe the motion as a sequence of atomic 3D flows computed with RGB-D modality. We use a Recurrent Neural Network based Encoder-Decoder framework to predict these sequences of flows. We argue that in order for the decoder to reconstruct these sequences, the encoder must learn a robust video representation that captures long-term motion dependencies and spatial-temporal relations. We demonstrate the effectiveness of our learned temporal representations on activity classification across multiple modalities and datasets such as NTU RGB+D and MSR Daily Activity 3D. Our framework is generic to any input modality, i.e., RGB, depth, and RGB-D videos.

1. Introduction

Human activities can often be described as a sequence of basic motions. For instance, common activities like brushing hair or waving a hand can be described as a sequence of successive raising and lowering of the hand. Over the past years, researchers have studied multiple strategies to effectively represent motion dynamics and classify activities in videos [38, 20, 42]. However, the existing methods suffer from the inability to compactly encode long-term motion dependencies. In this work, we propose to learn a representation that can describe the sequence of motions by learning to predict it. In other words, we are interested in learning a representation that, given a pair of video frames, can predict the sequence of basic motions (see in Figure 1). We believe that if the learned representation has encoded enough information to predict the motion, it is discriminative enough to classify activities in videos. Hence, our final goal is to use our learned representation to classify activities in videos.

To classify activities, we argue that a video representation needs to capture not only the semantics, but also the motion dependencies in a long temporal sequence. Since robust representations exist to extract semantic information [29], we focus our effort on learning a representation that encodes the sequence of basic motions in consecutive frames. We define basic motions as atomic 3D flows. The atomic 3D flows are computed by quantizing the estimated dense 3D flows in space and time using RGB-D modality. Given a pair of images from a video clip, our framework learns a representation that can predict the sequence of atomic 3D flows.

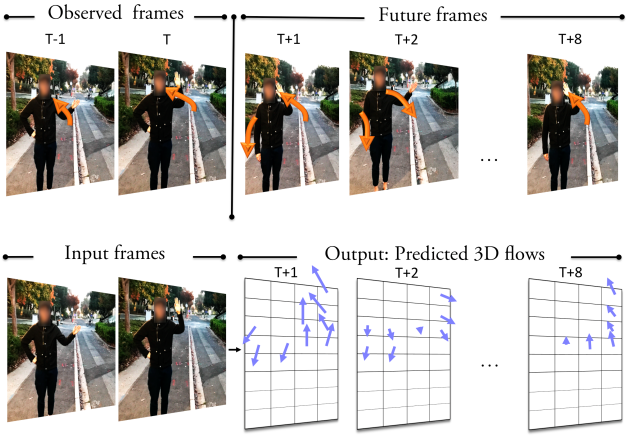

Figure 1. We propose a method that learns a video representation by predicting a sequence of basic motions described as atomic 3D flows. The learned representation is then extracted from this model to recognize activities.

Our learning framework is unsupervised, *i.e.*, it does not require human-labeled data. Not relying on labels has the following benefits. It is not clear how many labels are needed to understand activities in videos. For a single image, millions of labels have been used to surpass human-level accuracy in extracting semantic information [29]. Consequently, we would expect that videos will require several orders of magnitude more labels to learn a representation in a supervised setting. It will be unrealistic to collect all these labels.

Recently, a stream of unsupervised methods have been proposed to learn temporal structures from videos. These

methods are formulated with various objectives - supervision. Some focus on constructing future frames [32, 23], or enforcing the learned representations to be temporally smooth [53], while others make use of the sequential order of frames sampled from a video [17, 44]. Although they show promising results, most of the learned representations still focus heavily on either capturing semantic features [17], or are not discriminative enough for classifying activities as the output supervision is too large and coarse (*e.g.*, frame reconstruction).

When learning a representation that predicts motions, the following properties are needed: the output supervision needs to be of i) low dimensionality, ii) easy to parameterize, and iii) discriminative enough for other tasks. We address the first two properties by reducing the dimensionality of the flows through clustering. Then, we address the third property by augmenting the RGB videos with depth modality to reason on 3D motions. By inferring 3D motion as opposed to view-specific 2D optical flow, our model is able to learn an intermediate representation that captures less view-specific spatial-temporal interactions. Compared to 2D dense trajectories [38], our 3D motions are of much lower dimensionality. Moreover, we focus on inferring the sequence of basic motions that describes an activity as opposed to tracking keypoints over space and time. We claim that our proposed description of the motion enables our learning framework to predict longer motion dependencies since the complexity of the output space is reduced. In Section 5.2, we show quantitatively that our proposed method outperforms previous methods on activity recognition.

The contributions of our work are as follows:

(i) We propose to use a Recurrent Neural Network based Encoder-Decoder framework to effectively learn a representation that predicts the sequence of basic motions. Whereas existing unsupervised methods describe motion as either a single optical flow [37] or 2D dense trajectories [38], we propose to describe it as a sequence of atomic 3D flows over a long period of time (Section 3).

(ii) We are the first to explore and generalize unsupervised learning methods across different modalities. We study the performance of our unsupervised task - predicting the sequence of basic motions - using various input modalities: RGB $\rightarrow$ motion, depth $\rightarrow$ motion, and RGB-D $\rightarrow$ motion (Section 5.1).

(iii) We show the effectiveness of our learned representations on activity recognition tasks across multiple modalities and datasets (Section 5.2). At the time of its introduction, our model outperforms state-of-the-art unsupervised methods [17, 32] across modalities (RGB and depth).

2. Related Work

We first present previous works on unsupervised representation learning for images and videos. Then, we give a brief overview on existing methods that classify activities in multi-modal videos.

Unsupervised Representation Learning. In the RGB domain, unsupervised learning of visual representations has shown usefulness for various supervised tasks such as pedestrian detection and object detection [1, 26]. To exploit temporal structures, researchers have started focusing on learning visual representations using RGB videos. Early works such as [53] focused on inclusion of constraints via video to autoencoder framework. The most common constraint is enforcing learned representations to be temporally smooth [53]. More recently, a stream of reconstruction-based models has been proposed. Ranzato et al. [23] proposed a generative model that uses a recurrent neural network to predict the next frame or interpolate between frames. This was extended by Srivastava et al. [32] where they utilized a LSTM Encoder-Decoder framework to reconstruct current frame or predict future frames. Another line of work [44] uses video data to mine patches which belong to the same object to learn representations useful for distinguishing objects. Misra et al. [17] presented an approach to learn visual representation with an unsupervised sequential verification task, and showed performance gain for supervised tasks like activity recognition and pose estimation. One common problem for the learned representations is that they capture mostly semantic features that we can get from ImageNet or short-range activities, neglecting the temporal features.

RGB-D / depth-Based Activity Recognition. Techniques for activity recognition in this domain use appearance and motion information in order to reason about non-rigid human deformations activities. Feature-based approaches such as HON4D [20], HOPC [21], and DCSF [46] capture spatio-temporal features in a temporal grid-like structure. Skeleton-based approaches such as [5, 22, 35, 39, 50] move beyond such sparse grid-like pooling and focus on how to propose good skeletal representations. Haque *et al*. [4] proposed an alternative to skeleton representation by using a Recurrent Attention model (RAM). Another stream of work uses probabilistic graphical models such as Hidden Markov Models (HMM) [49], Conditional Random Fields (CRF) [12] or Latent Dialect Allocation (LDA) [45] to capture spatial-temporal structures and learn the relations in activities from RGB-D videos. However, most of these works require a lot of feature engineering and can only model short-range action relations. State-of-the-art methods [15, 16] for RGB-D/depth-based activity recognition report human level performance on well-established datasets like MSR-DailyActivity3D [14] and CAD-120 [33]. However, these datasets were often constructed under various constraints,

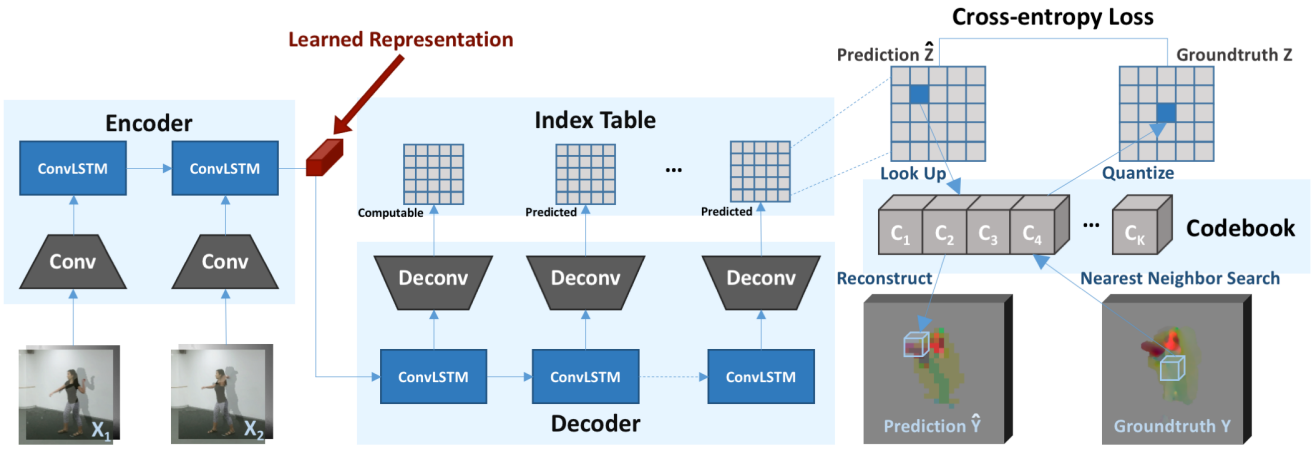

Figure 2. Our proposed learning framework based on the LSTM Encoder-Decoder method. During the encoding step, a downsampling network (referred to as "Conv") extracts a low-dimensionality feature from the input frames. Note that we use a pair of frames as the input to reduce temporal ambiguity. Then, the LTSM learns a temporal representation. This representation is then decoded with the upsampling network (referred to as "Deconv") to output the atomic 3D flows.

including single-view, single background, or with very few subjects. On the other hand, [27] shows that there is a big performance gap between human and existing methods on a more challenging dataset [27], which contains significantly more subjects, viewpoints, and background information.

RGB-Based Activity Recognition. The past few years have seen great progress on activity recognition on short clips [13, 51, 28, 38, 40]. These works can be roughly divided into two categories. The first category focuses on handcrafted local features and Bag of Visual Words (BoVWs) representation. The most successful example is to extract improved trajectory features [38] and employ Fisher vector representation [25]. The second category utilizes deep convolutional neural networks (ConvNets) to learn video representations from raw data (*e.g.*, RGB images or optical flow fields) and train a recognition system in an end-to-end manner. The most competitive deep learning model is the deep two-stream ConvNets [42] and its successors [43, 41], which combine both semantic features extracted by ConvNets and traditional optical flow that captures motion. However, unlike image classification, the benefit of using deep neural networks over traditional handcrafted features is not very evident. This is potentially because supervised training of deep networks requires a lot of data, whilst the current RGB activity recognition datasets are still too small.

3. Method

The goal of our method is to learn a representation that predicts the sequence of basic motions, which are defined as atomic 3D flows (described in details in Section 3.1). The problem is formulated as follows: given a pair of images $\langle \mathbf{X_1}, \mathbf{X_2} \rangle$, our objective is to predict the sequence of atomic 3D flows over T temporal steps: $\langle \widehat{\mathbf{Y}}_\mathbf{1}, \widehat{\mathbf{Y}}_\mathbf{2}, ..., \widehat{\mathbf{Y}}_\mathbf{T} \rangle$, where $\widehat{\mathbf{Y}}_\mathbf{t}$ is the atomic 3D flow at time t (see Figure 2). Note that $\mathbf{X_i} \in \mathbb{R}^{H \times W \times \mathcal{D}}$ and $\widehat{\mathbf{Y}}_\mathbf{t} \in \mathbb{R}^{H \times W \times 3}$, where $\mathcal{D}$ is the number of input channels, and H, W are the height and width of the video frames respectively. In Section 5, we experiment with inputs from three different modalities: RGB only ($\mathcal{D} = 3$), depth only ($\mathcal{D} = 1$), and RGB-D ($\mathcal{D} = 4$).

The learned representation – the red cuboid in Figure 2 – can then be used as a motion feature for activity recognition (as described in Section 4). In the remaining of this section, we first present details on how we describe basic motions. Then, we present the learning framework .

3.1. Sequence of Atomic 3D Flows

To effectively predict the sequence of basic motions, we need to describe the motion as a low-dimensional signal such that it is easy to parameterize and is discriminative enough for other tasks such as activity recognition. Inspired by the vector quantization algorithms for image compression [9], we propose to address the first goal by quantizing the estimated 3D flows in space and time, referred to as atomic 3D flows. We address the discriminative property by inferring a long-term sequence of 3D flows instead of a single 3D flow. With these properties, our learned representation has the ability to capture longer term motion dependencies.

Reasoning in 3D. Whereas previous unsupervised learning methods model 2D motions in the RGB space [37], we propose to predict motions in 3D. The benefit of using depth information along with RGB input is to overcome difficulties such as variations of texture, illumination, shape, viewpoint, self occlusion, clutter and occlusion. We augment the RGB videos with depth modality and estimate the 3D flows

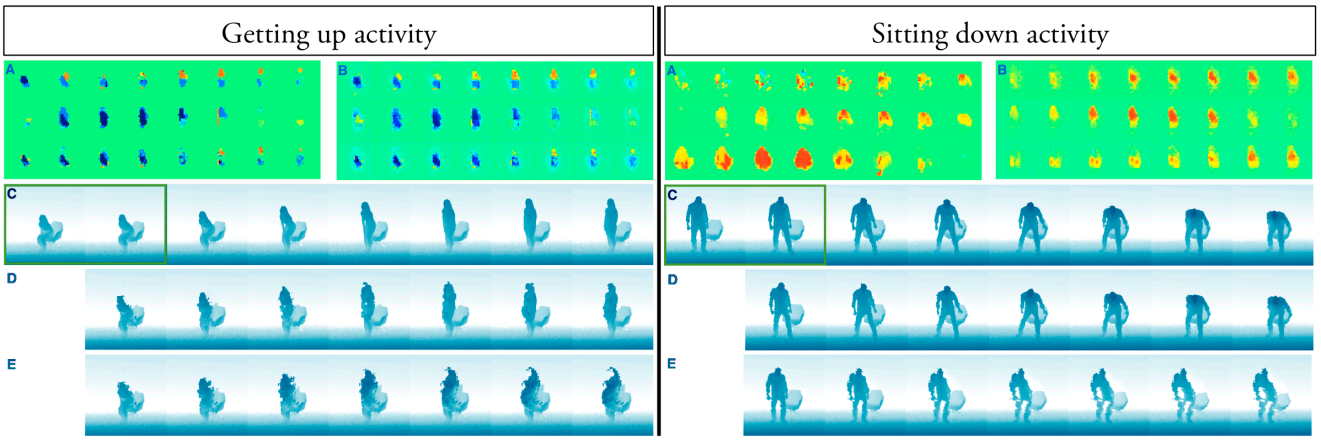

Figure 3. Qualitative results on predicting motion: two examples of long-term flow prediction (8 timesteps, 0.8s). The right hand side illustrates the "Getting up" activity whereas the right side presents the "Sitting down" activity. A: Ground truth 3D flow. Each row corresponds to flow along x, y, z direction respectively. B: Predicted 3D flows. C: Ground truth depths. The two frames in green boxes are the input. D: Depth reconstructed by adding ground truth depth and predicted flow. E: Depth reconstructed by adding the previous reconstructed depth and predicted flow, except for the first frame, in which case the ground truth depth is used.

[8] in order to reduce the level of ambiguities that exist in each independent modality.

Reasoning with sequences. Previous unsupervised learning methods have modeled motion as either a single optical flow [37] or a dense trajectories over multiple frames [38]. The first approach has the advantage of representing motion with a single fixed size image. However, it only encodes a short range motion. The second approach addresses the long-term motion dependencies but is difficult to efficiently model each keypoint. We propose a third alternative: model the motion as a sequence of flows. Motivated by the recent success of RNN to predict sequence of images [34], we propose to learn to predict the sequence of flows over a long period of time. To ease the prediction of the sequence, we can further transform the flow into a lower dimensionality signal (referred to as atomic flows).

Reasoning with atomic flows. Flow prediction can be posed as a regression problem where the loss is squared Euclidean distance between the ground truth flow and predicted flow. Unfortunately, the squared Euclidean distance in pixel space is not a good metric, since it is not stable to small image deformations, and the output space tends to smoothen results to the mean [23]. Instead, we formulate the flow prediction task as a classification task using $\mathbf{Z} = \mathcal{F}(\mathbf{Y})$, where $\mathbf{Y} \in \mathbb{R}^{H \times W \times 3}$, $\mathbf{Z} \in \mathbb{R}^{h \times w \times K}$, and $\mathcal{F}$ maps each non-overlapping $M \times M$ 3D flow patch in $\mathbf{Y}$ to a probability distribution over K quantized classes (*i.e.*, atomic flows). More specifically, we assign a soft class label over K quantized codewords for each $M \times M$ flow patch, where $M = H/h = W/w$. After mapping each patch to a probability distribution, we get a probability distribution $\mathbf{Z} \in \mathbb{R}^{h \times w \times K}$ over all patches. We investigated three quantization methods: k-means codebook (similar to [37]), uniform codebook, and learnable codebook (initialized with k-means or uniform codebook, and trained end-to-end). We got the best result using uniform codebook and training the codebook end-to-end only leads to minor performance gain. K-means codebook results in inferior performance because the lack of balance causes k-means to produce a poor clustering.

Our uniform quantization is performed as follows: we construct a codebook $\mathbf{C} \in \mathbb{R}^{K \times 3}$ by quantizing bounded 3D flow into equal-sized bins, where we have $\sqrt[3]{K}$ distinct classes along each axes. For each $M \times M$ 3D flow patch, we compute its mean and retrieve its k nearest neighbors (each represents one flow class) from the codebook. Empirically, we find having the number of nearest neighbors $k > 1$ (soft label) yields better performance. To reconstruct the predicted flow $\widehat{\mathbf{Y}}$ from predicted distribution $\widehat{\mathbf{Z}}$, we replace each codebook distribution as a linear combination of codewords. The parameters are determined empirically such that $K = 125$ (5 quantized bins across each dimension) and $M = 8$.

3.2. Learning framework

To learn a representation that encodes the long-term motion dependencies in videos, we cast the learning framework as a sequence-to-sequence problem. We propose to use a Recurrent Neural Network (RNN) based Encoder-Decoder framework to effectively learn these motion dependencies. Given two frames, our proposed RNN predicts the sequence of atomic 3D flows.

Figure 2 presents an overview of our learning framework, which can be divided into an encoding and decoding steps. During encoding, a downsampling network (referred to as "Conv") extracts a low-dimensionality feature from the

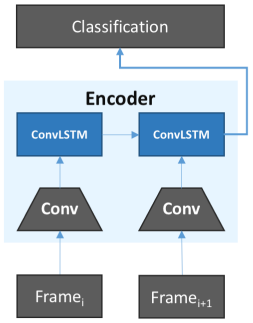

Figure 4. Our proposed network architecture for activity recognition. Each pair of video frames is encoded with our learned temporal representation (fixing the weights). Then, a classification layer is trained to infer the activities.

input frames. Then, the LTSM runs through the sequence of extracted features to learn a temporal representation. This representation is then decoded with the upsampling network ("Deconv") to output the atomic 3D flows.

The LSTM Encoder-Decoder framework [34] provides a general framework for sequence-to-sequence learning problems, and its ability to capture long-term temporal dependencies makes it a natural choice for this application. However, vanilla LSTMs do not take spatial correlations into consideration. In fact, putting them between the upsampling and downsampling networks leads to much slower convergence speed and significantly worse performance, compared to a single-step flow prediction without LSTMs. To preserve the spatial information in intermediate representations, we use the convolutional LSTM unit [47] that has convolutional structures in both the input-to-state and state-to-state transitions. Here are more details on the downsampling and upsampling networks:

Downsampling Network ("Conv"). We train a Convolutional Neural Network (CNN) to extract high-level features from each input frame. The architecture of our network is similar to the standard VGG-16 network [29] with the following modifications. Our network is fully convolutional, with the first two fully connected layers converted to convolution with the same number of parameters to preserve spatial information. The last softmax layer is replaced by a convolutional layer with a filter of size $1 \times 1 \times 32$, resulting in a downsampled output of shape $7 \times 7 \times 32$. A batch normalization layer [7] is added to the output of every convolutional layer. In addition, the number of input channels in the first convolutional layer is adapted according to the modality.

Upsampling Network ("Deconv"). We use an upsampling CNN with fractionally-strided convolution [31] to perform spatial upsampling and atomic 3D flow prediction. A stack of five fractionally-strided convolutions upsamples each input to the predicted distribution $\widehat{\mathbf{Z}} \in \mathbb{R}^{h \times w \times K}$, where $\widehat{\mathbf{Z}}_{\mathbf{ij}}$ represents the unscaled log probabilities over the $(i,j)^{th}$ flow patch.

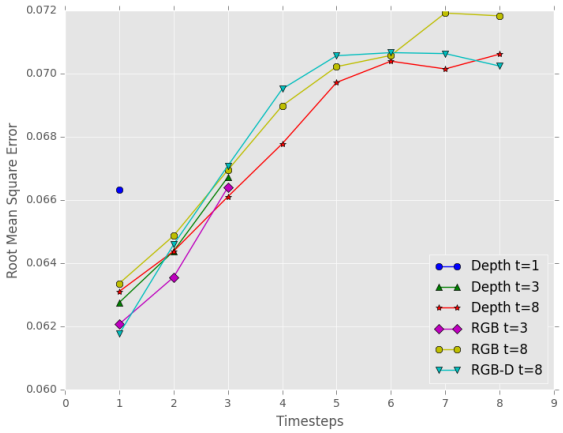

Figure 5. Motion prediction error on NTU-RGB+D. We plot the per-pixel root mean square error of estimating the atomic 3D flows with respect to time across different input modalities.

3.3. Loss Function

Finally, we define a loss function that is stable and easy to optimize for motion prediction. As described in section 3.1, we define the cross-entropy loss between the ground truth distribution $\mathbf{Z}$ over the atomic 3D flow space $\mathbf{C}$ and the predicted distribution $\widehat{\mathbf{Z}}$:

$$\mathcal{L}_{ce}(\mathbf{Z}, \widehat{\mathbf{Z}}) = -\sum_{i=1}^{H'} \sum_{j=1}^{W'} \sum_{k=1}^{K} \mathbf{w_k} \mathbf{Z_{ijk}} \log \widehat{\mathbf{Z}}_{\mathbf{ijk}} \qquad (1)$$

where $\mathbf{w} \in \mathbb{R}^K$ is a weighting vector for rebalancing the loss based on the frequency of each atomic flow vectors.

The distribution of atomic 3D flows is strongly biased towards classes with small flow magnitude, as there is little to no motion in the background. Without accounting for this, the loss function is dominated by classes with very small flow magnitudes, causing the model to predict only class 0 which represents no motion. Following the approach in [52], we define the class weight $\mathbf{w}$ as follow:

$$\mathbf{w} \propto \left((1-\lambda)\widetilde{\mathbf{p}} + \frac{\lambda}{K}\right)^{-1} \qquad \text{and} \qquad \sum_{k=1}^{K} \widetilde{\mathbf{p}}_k \mathbf{w_k} = 1 \quad (2)$$

where $\widetilde{\mathbf{p}}$ is the empirical distribution of the codewords in codebook $\mathbf{C}$, and λ is the smoothing weight.

4. Activity recognition

The final goal of our learned representation is to classify activities in videos. We use our encoder architecture

Methods	Depth	RGB
Our architecture only	37.5	34.1
Our method (with 2D motion)	58.8	–
Our method (3-step prediction)	62.5	54.7
Our method (8-step prediction)	**66.2**	56

Table 1. Quantitative results on activity recognition using the **NTU-RGB+D** dataset [27] with the following input modalities: depth and RGB. We report the mean AP in percentage on our ablation study as well as our complete model (in bold). We report the meanAP in percentage.

from unsupervised learning for activity recognition. A final classification layer is added on top of the encoder output to classify activities.

To study the effectiveness of our learned representation, we consider the following three scenarios:

1. Initialize the weight of our architecture randomly and learn them with the labels available for the supervision task (referred to as "architecture only" in Table 1);

2. Initialize the weights with our learned representation and fine-tune on activity recognition datasets;

3. Keep the pre-trained encoder fixed and only fine-tune the last classification layer.

Note that we don't combine our learned representation with any pre-trained semantic representation (such as the fc7 representation learned on ImageNet [24]). We argue that for our model to learn to predict the basic motions, it needs to understand the semantic content.

We follow the same data sampling strategy described in [28]. During training, a mini-batch of 8 samples is constructed by sampling from 8 training videos, from each of which a pair of consecutive frames is randomly selected. For scenario (i) and (iii), the learning rate is initially set to 10^{-4} with a decay rate of 0.96 every 2000 steps. For scenario (ii), the initial learning rates of encoder and the final classification layer are set to 10^{-5} and 10^{-4} respectively, with the same decay rate. At test time, we uniformly sample 25 frames from each video and average the scores across the sampled frames to get the class score for the video.

Our presented classification method is intentionally simple to show the strength of our learned representation. Moreover, our method is computationally effective. It runs in real-time since it consists of a forward pass through our encoder. Finally, our learned representation is compact $(7 \times 7 \times 32)$ enabling implementation on embedded devices.

Methods	Depth
HOG [19]	32.24
Super Normal Vector [48]	31.82
HON4D [20]	30.56
Lie Group [35]	50.08
Skeletal Quads [3]	38.62
FTP Dynamic Skeletons [6]	60.23
HBRNN-L [2]	59.07
2 Layer P-LSTM [27]	62.93
Shuffle and Learn [17]	47.5
Our method (Unsupervised training)	**66.2**

Table 2. Quantitative results on depth-based activity recognition using the **NTU-RGB+D** dataset [27]. The first group (row) presents the state-of-the-art supervised depth-map based method; the second group reports the supervised skeleton-based methods; The third one includes skeleton-based deep learning methods; The fourth is a recently proposed unsupervised method we implemented; The final row presents our complete model. We report the mean AP in percentage.

5. Experiments

We first present the performance of our unsupervised learning task, *i.e.*, predicting the sequence of motion, using various input modalities including RGB, depth, and RGB-D. Then, we study the effectiveness of our learned representations on classifying activities across multiple modalities and datasets.

5.1. Unsupervised Learning of Long-term Motion

Dataset. We use the publicly available NTU RGB+D dataset [27] to train our unsupervised framework. The dataset contains 57K videos for 60 action classes, 40 subjects and 80 viewpoints. We split the 40 subjects into training and testing groups as described in [27]. Each group consists of 20 subjects where the training and testing sets have 40,320 and 16,560 samples, respectively.

Training details. We use a mini-batch of size 16. The model is trained for 50 epochs with an initial learning rate of $1e^{-4}$ using the Adam optimizer [10]. We divide the learning rate by 10 whenever validation accuracy stops going up. The network is L_2 regularized with a weight decay of $5e^{-4}$. For classification, we use a smoothing $\lambda = 0.5$.

Evaluation. We measure the root mean square error (RMSE) between the ground truth flow $\mathbf{Y}$ and the predicted flow $\widehat{\mathbf{Y}}$. F1 score is used to measure the classification error between the ground truth index table $\mathbf{Z}$ and the predicted index table $\widehat{\mathbf{Z}}$.

Results. In Figure 5, we plot the prediction error with respect to different input modalities (RGB, depth, RGB-D) and prediction time (3 and 8 timesteps). We also report the

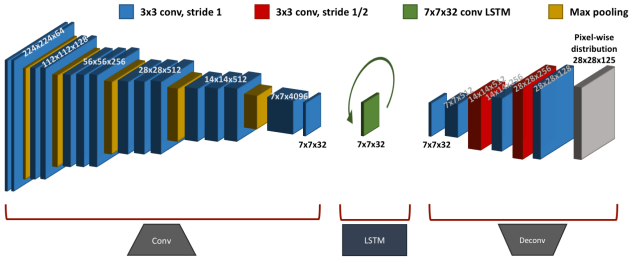

Figure 6. The detailed architecture of our model. The Conv shows the architecture of our downsampling network; LSTM represents the encoder-decoder framework; Deconv shows the architecture of our upsampling network.

prediction error of using a single input frame to predict the next frame similar to [37] (blue dot). The error is intuitively the highest since there are ambiguities when reasoning with a single input image. Interestingly, all input modalities perform very similarly when predicting 8 timesteps. The RGB modality is quite competitive to the other two modalites although the 3D information is not measured. When all 4 channels are used, *i.e.*, RGB-D input, the performance is still similar to using the other modality. The overall error linearly increases with the first 4 frames and stabilizes for the final 4 frames. All methods that predict only the next 3 frames have similar prediction errors compared to the ones that predict a longer sequence. Consequently, our model has enough capacity to learn a harder problem, *i.e.*, predicting long sequences. In Figure 3, we qualitatively show the prediction output using depth modality. We illustrate the results by reconstructing the input frame (depth image) from the predicted flows. Our method has not been trained to accurately reconstruct the signal. Nevertheless, the reconstructed signals convey the accuracy of the prediction.

5.2. Activity Recognition

We compare our activity recognition performance with state-of-the-art supervised methods for each modality. In addition, we perform the ablation studies for our unsupervised methods and compare with the a recently-proposed unsupervised method.

Our method with 2D motion. Instead of predicting 3D motion, we predict 2D motion in the form of quantized 2D optical flow.

Our method with 3-step prediction. We predict motions for the next three frames. Note that our proposed method uses 8-step prediction.

Shuffle and Learn [17]. Given a tuple of three frames extracted from a video, the model predicts whether the three frames are in the correct temporal order or not. We implemented the above model using TensorFlow and trained on the NTU RGB-D dataset for the sequential verification task, following the same data sampling techniques and unsupervised training strategies as specified in [17].

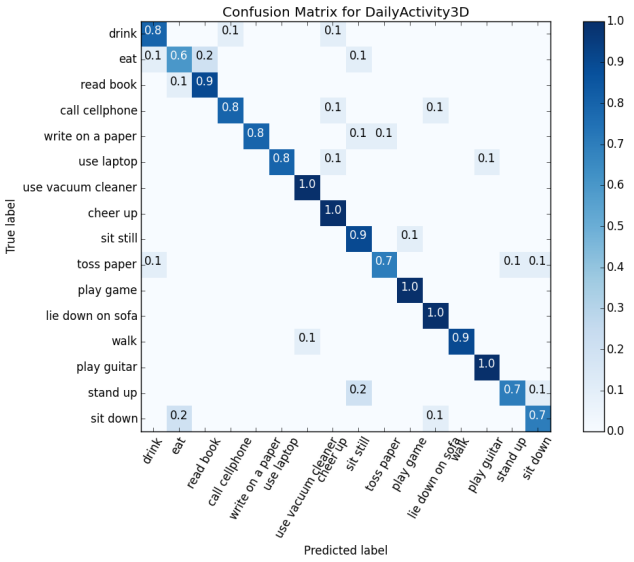

Figure 7. The confusion matrix for action recognition on **MSR-DailyActivity3D** dataset [14]. Activities with large motions are better classified than the ones with fine-grained motion.

5.2.1 Depth-based Activity Recognition

Dataset. We train and test our depth-based activity recognition model on two datasets: NTU-RGB+D and MSRDailyActivity3D [14]. For NTU-RGB+D, we follow the cross-subject split as described in [27]. The MSRDailyActivity3D dataset contains 16 activities performed by 10 subjects. We follow the same leave-one-out training-testing split as in [11]. We intentionally use this extra MSRDailyActivity3D dataset that is different from the one we use for unsupervised training to show the effectiveness of our learned representation in new domains (different viewpoints and activities).

Results on NTU-RGB+D. Table 2 shows classification accuracy on the NTU-RGB+D dataset. The first group of methods use depth maps as inputs, while the second and the third use skeleton features. Methods in the third group are deep-learning based models. Our proposed method outperforms the state-of-the-art supervised methods. We use our learned representation that predicts the next 8 frames without fine-tuning it on the classification task. Interestingly, fine-tuning the weights of our encoder did not give a boost in performance.

Ablation study on NTU-RGB+D. In Table 1, we present more insights on our design choices. We first show that by using our encoder architecture without pre-training it to predict the motion (referred to as "our architecture only"), the classification accuracy (mean AP) is the lowest. We then show that modeling 3D motion instead of 2D motion positively impacts the performance. Finally, we report the results when shorter sequences (3-step prediction) are en-

Methods	Depth
Dynamic Temporal Warping [18]	54.0
Actionlet Ensemble [39]	85.8
HON4D [20]	85
3D Trajectories [11]	72
Our method (Unsupervised training)	**86.9**

Table 3. Quantitative results on activity recognition using the **MSRDailyActivity3D** dataset [14]. Methods in italic require full skeleton detection. Our method has learned a video representation from a different dataset and has not fine-tuned on this dataset. We report the meanAP in percentage.

coded during our unsupervised training. Increasing the sequence length to 8 time-steps increases the classification accuracy. The discrimination power of our representation is increased by encoding longer-term dependencies. For the sake of completeness, we also fine-tune our activity recognition model using RGB videos from the NTU RGB-D dataset. We notice that the results are comparable to depth-based activity recognition and follow the same trend for ablation studies (*i.e.*, predicting longer motion in 3D yields better performance).

Results on MSRDailyActivity3D. Table 3 presents classification accuracy on the MSRDailyActivity3D dataset [14] and Figure 7 its confusion matrix. Methods in italic require skeleton detection, while the fourth one makes use of dense 3D trajectories. Note that our unsupervised learning task – predicting the basic motions – has not been trained on these activities and viewpoints. Nevertheless, we outperform previous work specially the method based on the 3D trajectories by a large margin (+15%). Our compact representation of the 3D motion is more discriminative than the existing representation for 3D trajectories [38].

5.2.2 RGB-based Activity Recognition

Dataset. We train and test our RGB-based activity recognition model on the UCF-101 dataset [30] to compare with state-of-the-art unsupervised methods [17, 36] in this domain. The dataset contains 13,320 videos with an average length of 6.2 seconds and 101 different activity categories. We follow the same training and testing protocol as suggested in [28]. However, note that we are not training the unsupervised task on the UCF-101 dataset. Instead, the model is pretrained on the RGB videos from NTU-RGB+D dataset. We want to study the capacity of our learned representation to be used across domains and activities.

Results on UCF-101. Table 4 shows classification accuracy for RGB-based activity recognition methods on the UCF-101 dataset. By initializing the weights of our supervised model with the learned representation, our model (*i.e.*, our method w/o semantics) outperforms two recent unsupervised video representation learning approaches [17, 36]. Note that although the unsupervised LSTM [32] method outperforms all other methods, it uses a ConvNet pretrained on ImageNet for semantic feature extraction, whilst the other methods do not make use of extra semantic information. To compare with [32], we use a VGG-16 network pretrained on ImageNet to extract semantic features (*i.e.*, fc7 feature) from input images, and add a softmax layer on top of it. We combine the softmax score from our model with the semantic softmax score by late fusion.

Methods	RGB
S: Deep two stream [42]	91.4
U: Shuffle and Learn [17]	50.2
U: VGAN [36]	52.1
U: Our method (w/o semantics)	**53.0**
U: Unsupervised LSTMs [32]	75.8
U: Our method (w/ semantics)	**79.3**

Table 4. Quantitative results on activity recognition using the **UCF-101** dataset [30]. The first group presents the state-of-the-art supervised (S) method; the second group reports unsupervised (U) methods without using ImageNet semantics; the third shows unsupervised (U) methods with ImageNet semantics. We report the meanAP in percentage.

6. Conclusions

We have presented a general framework to learn long-term temporal representations for videos across different modalities. By using our proposed sequence of atomic 3D flows as supervision, we can train our model on a large number of unlabeled videos. We show that our learned representation is effective and discriminative enough for classifying actions as we achieve state-of-the-art activity recognition performance on two well-established RGB-D datasets. For future work, we aim to explore the performance of our method on RGB based datasets such as ActivityNet or other supervised tasks beyond activity recognition. We want to use other free labels from videos such as predicting 3D scenes interactions from RGB frames. We also want to come up with a compact representation for dense trajectory, which can effectively reduce background motions in many existing datasets.

Acknowledgement. We would like to start by thanking our sponsors: Stanford Computer Science Department, Intel, ONR MURI, and Stanford Program in AI-assisted Care (PAC). Next, we specially thank Juan Carlos Niebles, Serena Yeung, Kenji Hata, Yuliang Zou, and Lyne Tchapmi for their helpful feedback. Finally, we thank all members of the Stanford Vision Lab and Stanford Computational Vision and Geometry Lab for their useful comments and discussions.

References

[1] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[2] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[3] G. Evangelidis, G. Singh, and R. Horaud. Skeletal quads: Human action recognition using joint quadruples. In *International Conference on Pattern Recognition*, pages 4513–4518, 2014.

[4] A. Haque, A. Alahi, and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1229–1238, 2016.

[5] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *European Conference on Computer Vision (ECCV)*, October 2016.

[6] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015.

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[8] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers. A primal-dual framework for real-time dense rgb-d scene flow. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 98–104. IEEE, 2015.

[9] T. Kim. Side match and overlap match vector quantizers for images. pages 170–185, 1992.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[11] M. Koperski, P. Bilinski, and F. Bremond. 3d trajectories for action recognition. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4176–4180. IEEE, 2014.

[12] H. S. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *ICML (3)*, pages 792–800, 2013.

[13] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 204–212, 2015.

[14] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–14. IEEE, 2010.

[15] C. Lu, J. Jia, and C.-K. Tang. Range-sample depth feature for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 772–779, 2014.

[16] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1809–1816, 2013.

[17] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[18] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 137–146. Eurographics Association, 2006.

[19] E. Ohn-Bar and M. Trivedi. Joint angles similarities and hog2 for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 465–470, 2013.

[20] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723, 2013.

[21] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European Conference on Computer Vision*, pages 742–757. Springer, 2014.

[22] H. Rahmani and A. Mian. 3d action recognition from novel viewpoints. In *CVPR, June*, 2016.

[23] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[25] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[26] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013.

[27] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

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

[30] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[31] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[32] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *CoRR, abs/1502.04681*, 2, 2015.

[33] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011.

[34] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[35] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014.

[36] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. *arXiv preprint arXiv:1609.02612*, 2016.

[37] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2443–2451, 2015.

[38] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[39] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012.

[40] L. Wang, Y. Qiao, and X. Tang. Mining motion atoms and phrases for complex action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2680–2687, 2013.

[41] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. *CoRR*, abs/1505.04868, 2015.

[42] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.

[43] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *CoRR*, abs/1507.02159, 2015.

[44] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[45] C. Wu, J. Zhang, S. Savarese, and A. Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4370, 2015.

[46] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2834–2841, 2013.

[47] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810, 2015.

[48] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 804–811, 2014.

[49] Y. Yang, I. Saleemi, and M. Shah. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1635–1648, 2013.

[50] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*, pages 50–65. Springer, 2014.

[51] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[52] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *CoRR*, abs/1603.08511, 2016.

[53] W. Zou, S. Zhu, K. Yu, and A. Y. Ng. Deep learning of invariant features via simulated fixations in video. In *Advances in neural information processing systems*, pages 3212–3220, 2012.