Pose from Action: Unsupervised Learning of Pose Features based on Motion

Senthil Purushwalkam
http://www.cs.cmu.edu/~spurushw/
Robotics Institute,
Carnegie Mellon University

Abhinav Gupta
http://www.cs.cmu.edu/~abhinavg/
Robotics Institute,
Carnegie Mellon University

Abstract

Human actions are comprised of a sequence of poses. This makes videos of humans a rich and dense source of human poses. We propose an unsupervised method to learn pose features from videos that exploits a signal which is complementary to appearance and can be used as supervision: motion. The key idea is that given two poses, it should be possible to predict what motion caused the change between them. We represent each of the poses as a feature in a CNN (Appearance ConvNet) and generate a motion encoding from optical flow maps using a separate CNN (Motion ConvNet). The data for this task is automatically generated allowing us to train without human supervision. We demonstrate the strength of the learned representation by finetuning for Pose Estimation on the FLIC dataset, for static image action recognition on PASCAL and for action recognition in videos on UCF101 and HMDB51.

1 Introduction

Humans learn visual representations by observing and actively exploring the dynamic world around us. The manual labeling of images remains a significant bottleneck in exploiting a large number of images to learn visual representations. As a consequence, there has been rising interest in the area of unsupervised feature learning. A popular approach to unsupervised learning is training a network using standard back-propagation on an auxiliary task for which ground truth can be easily mined in an automated fashion. The hope is that the visual representation learned for this auxiliary task is generalizable and can be used for other tasks with simple fine-tuning. Owing to the rise of interest in unsupervised learning, many such auxiliary tasks have been proposed in the recent past. [3] proposes to predict the relative location of pairs of patches, which seems to generalize to object detection as shown in the paper. [1, 5] propose an approach to take pair of patches and predict the camera motion that caused the change. The ground-truth for this task is obtained via other sensors which measure ego-motion. Finally, [15] proposed an approach to sample a pair of patches separated in time via tracking and learn a representation which embeds these patches close by in the visual representation space (since they are different views of same object).

In this work, we argue that there is a complementary and strong signal in videos to supervise the training of these networks: motion patterns. We hypothesize that a representation in which visual concepts demonstrating similar motion patterns cluster together in the appearance space can prove useful for many computer vision tasks. We demonstrate how motion patterns in the videos can act as strong supervision to train the appearance network itself. For instance, when this method is applied to human action videos, we would expect the appearance representation to encode pose features. A key observation is that pairs of human poses are often associated with similar motions in between. For example, poses before and after swinging a bat are associated with the swinging motion. The proposed approach could possibly be used to learn different kinds of appearance representations based on different kinds of videos. Specifically, in this paper, we choose to work with human action videos since the learnt representations can be semantically associated to poses. We believe that this idea can provide the missing link in unsupervised learning of visual representations for human actions and human poses.

However, there is still one missing link: how do you define similarity of motion patterns. One way is to use distance metric on hand designed features (e.g., 3DHOG, HOF[13]) or the optical flows maps directly. Instead, inspired by the success of the two-stream network[10], we try to jointly learn convolutional features for both the RGB and optical flow at the same time. Our key idea is to have a triplet network where two streams with shared parameters correspond to the first and n^{th} frame in the video; and the third stream looks at $n-1$ optical flow maps. All the convolutional streams run in a feedforward manner to produce 4096 dimensional vectors. The three streams are then combined to classify if the RGB frames and optical flow channels correspond to each other *i.e.* does the transformation cause the change in appearance? Intuitively, solving this task requires the Appearance ConvNet to identify the visual structures in the frame and encode their poses. The Motion ConvNet is expected to efficiently encode the change in pose that the optical flow block represents. We evaluate our trained appearance network by finetuning on the task of pose estimation on the FLIC dataset[9], static image action recognition on PASCAL VOC[4], and action recognition on UCF101[11] and HMDB51[7]. We show that these models perform significantly better than training from random initialisation.

2 Approach

The goal of this work is to learn an appearance representation that captures pose properties without the use of any human supervision. We achieve this by formulating a surrogate task for which the ground truth labels are readily available or can be mined automatically. In simple terms, given a change in appearance, the task we formulate involves predicting what transformation causes it. For example, in Figure 1, given the appearance of Frame 1 and Frame13, we can predict that the transformation of 'swinging the bat' caused the change in appearance. In this section, we first develop an intuitive motivation for the surrogate task and then concretely explain how it can be implemented.

Suppose we want to train a model to predict if a Transformation $\mathbf{T}$ causes the change in Appearance $\mathbf{A} \rightarrow \mathbf{A}'$. We would need to have an robust way to encode $\mathbf{A}, \mathbf{A}'$ and $\mathbf{T}$ such that they capture all the information required to solve this task. More specifically, given an image, the appearance representation $\mathbf{A}$ needs to localise the object(s) that could undergo a transformation and encode its properties such as shape, size and more importantly, ***pose***. On the other hand, given a motion signal (like optical flow, dense trajectories [14], etc), the transformation representation $\mathbf{T}$ needs to express a robust encoding that is discriminative in the space of transformations.

We propose to learn a appearance representation $\mathbf{A}$ using a convolutional neural network. Instead of using hand-crafted approaches to encode a motion representation from optical flow, we propose to jointly learn it as a Transformation $\mathbf{T}$ using a separate convolutional neural network (overview in Figure 1). This framework allows us to use an unsupervised approach to jointly train the Appearance and Motion ConvNets. The key idea of our approach is that given two appearance features $\mathbf{A}$ and $\mathbf{A}'$, it should be possible to predict whether a Transformation $\mathbf{T}$ causes the change $\mathbf{A} \rightarrow \mathbf{A}'$. This idea is synchronous with [5], where the notion of ego-motions producing predictable transformations is used to learn an unsupervised model.

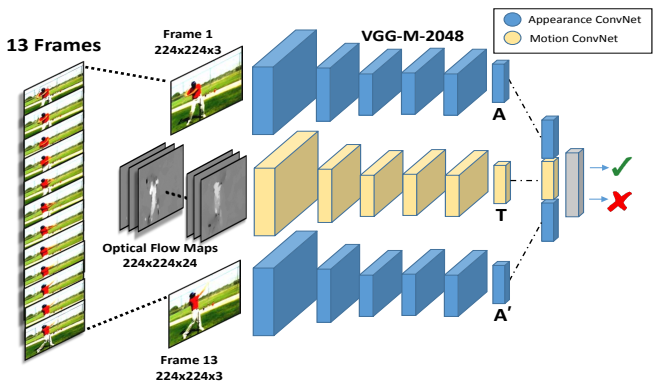

Figure 1: An overview of our approach. Predicting whether a transformation encoding T causes the change in appearance A→A' requires capturing pose properties.

Data Extraction

For training the binary classification model, we require a large collection of pairs of frames, the correct block of optical flow maps between them and multiple negative samples of optical flow blocks. As the training set, we use a large collection of video which contain humans performing actions. This set is formed by combining the training set videos from the UCF101 [11](split1), HMDB51 [7] (split1) and the ACT[16] datasets. For every pair of consecutive frames we precompute the horizontal and vertical directional optical flow maps using the OpenCV GPU implementation of the TVL1 algorithm[8].

As inputs to the Appearance ConvNet we randomly sample a spatial location and crop 224x224 patches at that location from two frames separated by $\Delta n(=12)$ frames. For the Motion ConvNet, we sample the 224x224 patches from each of the 12 horizontal and 12 vertical flow maps in between the two sampled frames at the same location, as the positive (label= 1) which gives us a 224x224x24 dimensional array. As the negative examples (label= 0), we randomly sample another 224x224x24 block from a random spatial location in a randomly picked video.

In summary, the joint unsupervised learning pipeline consists of one Motion ConvNet, two instances of the Appearance ConvNet and a two-layer fully connected neural network on top. The parameters of the two Appearance ConvNets are shared since we expect both networks to encode similar properties. This architecture allows us to use standard back propagation to train all the components simultaneously.

3 Experimental results

The efficacy of unsupervised feature representations are generally tested by finetuning for tasks for which the representation might prove useful. We follow a similar strategy and perform an extensive evaluation of our unsupervised model to investigate the transferability of the learned features. Since our unsupervised representation is expected to capture pose properties, we perform evaluation for the Pose Estimation and Action Recognition tasks.

Pose Estimation

We design a simple deep learning based pose estimation architecture to allow us the freedom to accommodate other unsupervised models. We copy the VGG-M[2] architecture till the fifth convolution layer (Conv5). This is followed by a deconvolution layer to upscale the feature maps. Then 1x1 convolutions are used to predict heat maps for each body point to be estimated. This network architecture is partly inspired from [12]. We use the centroid of the 20 highest scoring pixels as the prediction and evaluate on the FLIC dataset using the strict PCP and the PDJ evaluation metrics. In Table 1 & 2, we compare models initialised randomly, with our Appearance ConvNet, with a model pretrained for Action Classification on UCF101, the unsupervised model from [15] and an ImageNet classification pretrained model.

Table 1: Results for the Strict PCP Evaluation for Pose Estimation on the FLIC Dataset

	Body Part	
Initialisation	**Upper Arms**	**Lower Arms**
Random	51.9	19.3
Wang et. al Unsupervised[15]	52.8	19.7
UCF101 Action Classification Pretrained	46.7	17.8
Ours	**57.1**	**24.4**
ImageNet Classification Pretrained	65.6	34.3

Table 2: Results for the PDJ Evaluation for Pose Estimation on the FLIC Dataset

		Elbow				Wrist			
Initialisation	Precision→	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
Random		20.0	47.3	63.9	74.8	17.2	36.1	49.6	60.8
UCF101 Pretrained		18.5	44.8	61.0	71.1	16.5	34.8	45.2	53.2
Wang et. al Unsupervised[15]		23.0	48.3	66.5	**77.6**	19.1	36.6	46.7	55.1
Ours		**28.0**	**54.6**	**68.8**	**77.6**	**20.1**	**40.0**	**51.6**	**60.8**
ImageNet Pretrained		34.8	62.0	74.7	82.1	29.0	48.5	59.3	66.7

Action Recognition in Videos

For the task of action recognition, we use the first split of videos from the UCF101 and HMDB51 datasets. We use the same architecture as the Appearance ConvNet(VGG-M till FC6) followed by two randomly initialised fully-connected layers at the end to perform classification. For each video, we uniformly sample 25 frames and sample 224x224 crops from the corners and the center. These samples are used as input to compute the predictions for each of the samples and average them across all samples for a video to get the final prediction. Similar to before, we compare models intiailised randomly (from [10]), with the unsupervised model from [15], our Appearance Network and a ImageNet classification pretrained model (from [10]).

For action recognition in static images on PASCAL VOC, we use the same architecture. We follow [6] and use 50 randomly sampled images to finetune our models. The Appearance ConvNet initialisation shows an improvement of 7.4% over random initialisation and performs about 2.5% better than [6].

Table 3: Results for the Appearance Based action recognition on UCF101 and HMDB51

		Dataset	
Initialisation	**Finetuning/Training**	**UCF101**	**HMDB51**
Random	Full Network	42.5%	15.1%
Wang et. al Unsupervised[15]	Full Network	41.5%	16.9%
Ours	Full Network	**55.4%**	**23.6%**
Ours	Last 2 layers	41.4%	19.1%
ImageNet	Full Network	70.8%	40.5%

4 Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DoI/IBC) contract number D16PC00007. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *ICCV*, pages 37–45, 2015.

[2] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[3] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88 (2):303–338, June 2010.

[5] Dinesh Jayaraman and Kristen Grauman. Learning image representations equivariant to ego-motion. *arXiv preprint arXiv:1505.02206*, 2015.

[6] Dinesh Jayaraman and Kristen Grauman. Slow and steady feature analysis: Higher order temporal coherence in video. *arXiv preprint arXiv:1506.04714*, 2015.

[7] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*,, 2011.

[8] Julien Marzat, Yann Dumortier, and Andre Ducrot. Real-time dense and accurate parallel optical flow using cuda. 2009.

[9] Benjamin Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *In Proc. CVPR*, 2013.

[10] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[11] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[12] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *ICCV*, 2015.

[13] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*. BMVA Press, 2009.

[14] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*. IEEE, 2011.

[15] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015.

[16] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions ~ transformations. *CoRR*, abs/1512.00795, 2015.