Human Action Recognition Using Deep Networks

I. Introduction

Human action recognition from images has broad applications. It has immense use in areas such as tracking, semantic web-search, video annotation, surveillance, interactive game design, and many others involving semantic understanding of the human environment. A technique which can recognize / identify human actions in unconstrained settings would be all the more useful - being more widely, commonly applicable. We thus propose a solution for human action recognition from informally captured solitary images in unstructured natural environments.

The problem of action recognition is one of assigning semantic labels to the action(s) being performed by person(s) in some environment. Solutions for human action recognition have quite often operated on image sequences (videos), depth data streams and/or inertial sensing data, exploiting the temporal relationships between the series of detected body poses. Hidden Markov Models (HMMs) have been popularly and successfully deployed in such settings ([2], [7]).

More recently, the vision community has shown active interest in human action classification from single still images ([12], [9] for instance). A still snapshot, although devoid of any temporal information, quite often contains several useful cues which when considered holistically may suffice to identify an action/activity. These indicative cues not only pertain to humans' pose but also involve the objects they are interacting with. For instance, the actions classes indicated in Figure 1, would be easily evident to a human.

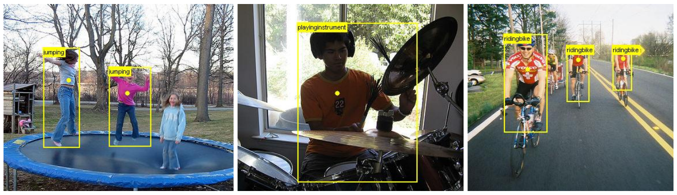

Fig. 1. Human action examples from Pascal VOC dataset, [3], with the classified actions indicated over bounding boxes. Note that the environment context, together with pose information, makes the actions rather obvious to a human.

In the aforementioned setting, the problem of action recognition essentially focuses on inference over interactions between parsed human pose and salient objects (in contrast to focusing on temporal analysis when processing data streams). Such a problem setting has much broader applicability, and in fact augments temporal analysis methodologies for action recognition. Image snapshots, today, are pervasive and abundant. They are easily accessible and acquirable - with the web serving as a vast, diverse and ever increasing pool. This is in contrast to data from other sensing modalities such as inertial or 3D/depth which requires specialized hardware, and is not nearly as pervasive.

Deep learning based approaches in recent years have shown much promise in image classification tasks, being able to learn rich feature hierarchies from unlabeled data and capable of capturing complex invariance in visual patterns - when abundant training data is available, as is the case here. A very popular and successful deep learning architecture is based on Convolutional Neural Networks (CNNs) utilizing a high number of layers (*deep*) and large breadth ([13] presents a good overview). Of particular note is the recently introduced R-CNN algorithm ([15]), which is based on CNNs applied on to regions in image and was highly successful in object detection tasks.

This paper investigates, explores a similar deep Convolution Neural Network based approach for the complex problem of classifying human actions from 2D images collected from the internet ([3]). Our strategy, illustrated in *Figure 2*, is based on extracting rich feature hierarchies through CNNs operating over salient image regions (poselets, *Sec.* III-A). The image level features thus learnt are then assigned action labels using a straight-forward (One vs One) Support Vector Machine (SVM) based classification framework. As can be seen in *Sec.* IV, the employed methodology has yielded promising results in our experiments till now - validating the proposed approach and warranting ensuing research.

II. Related Work

While there has been a larger focus on action recognition in unstructured and unconstrained videos ([4], [6]), there is little work involving action recognition in still images [7]. Most of the past work builds action recognition algorithm on top of the output of pose estimation algorithm. Different researchers use different representation for pose. Maji et. al. [11] represent pose as 3D orientation of head and torso. Ikizler and Pinar (cite) represent pose using a histogram of oriented rectangle which basically histogram of human body part locations and orientations. The feature vector for the image is crafted from the pose information and other aspects such as interaction with objects and people and use it for action recognition. Later a classifier such as SVM is trained over the feature vectors.

We believe that pose information is crucial for action recognition. For example, cycling or horse-riding usually have bent pose, and clicking a photograph usually has hands held near face. However, there might be other subtle clues important for action recognition that may be ignored in the hand-crafted feature vectors.

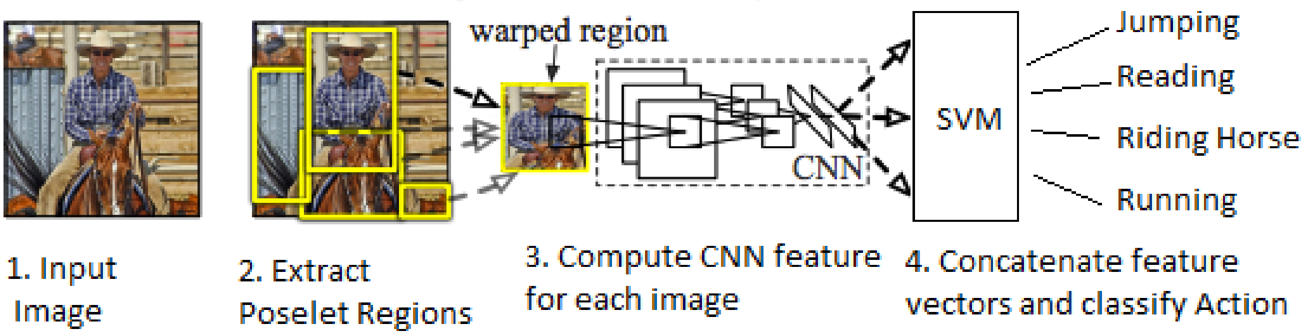
warped region
CNN
SVM
Jumping
Reading
Riding Horse
Running
1. Input Image
2. Extract Poselet Regions
3. Compute CNN feature for each image
4. Concatenate feature vectors and classify Action

Fig. 2. (Best viewed in color) **System Overview** (1) takes an input image and bounding box of person (2) we calcuate top 5 poselet regions (2) extract 4096 length CNN feature for each region and complete image (4) conctenate 6 features vectors and run SVM classifer to get action label.

Fukushimas Neocognitron [14] suggests that human brain uses a heirarchical, multi-stage process, to extract features for recogntion tasks. Building on this, LeCun et. al. proposed a Convolutional Neural Network (CNN) method for deep learning and successfully applied it to Optical Character Recognition [5] and then to generic object recognition tasks. In a usual setting, CNN is used to extract features, over which classifier such as SVM or softmax is trained. However, most recent recognition tasks used hand-engineered features instead of CNNs. Recently Krizhevsky et al. [18] showed a big gain over existing methods in image classification on the ImageNet challenge using CNNs [19]. This rekindled interest in CNNs. Donahue et al. [20] found that Krizhevsky et al.s model can be used to extract features that are broadly useful for a large variety of tasks. Basically, it can be concluded that the same trained CNN can be used to extract features for muliple recognition tasks, for example object recognition, action recognition, attribute classfication, etc.

III. Approach

As mentioned earlier, our recognition methodology constitutes of of three stages - poselet detection, CNN based feature extraction, and subsequent classification using SVM. The input to our system is a single 2D image with bounding boxes around each person in the image and the output is the estimated action of the people in the image. Each stage of the algorithm is described in reasonable detail below.

A. Poselet Detection

We first preprocess all images (rather the bounding boxes around interacting humans) using a model detection scheme based on salient parts (poselets, [1]). Poselet detectors use clustering to ascertain regions of similarities between people across images (Figure 3 indicates three identified poselets around the face area). Modeling humans using such a part based technique is robust to variations in human poses and occlusions, so long as there are clusters of trained poselets with enough similarity to the current image.

B. Feature Extraction using Convolution Neural Network

Parts based models perform very well for pose detection because they incorporate pose variation in the model, but the drawback is that they are limited by shallow low-level features. To solve this problem, features are extracted from every single poselet using CNNs.

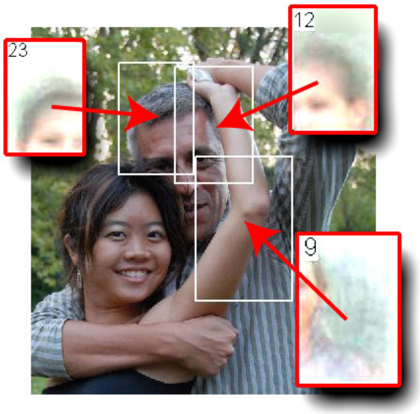

Fig. 3. The regions corresponding to three poselets [11]

We use a pretrained model of CNN to extract features from poselets. The pretrained model is provided by Girshick et al. [17]. The net is first trained over the ImageNet dataset and then fine tuned over the VOC 2012 dataset of 20 action object categories. Selective Search is used to extract regions and each region is then passed through the CNN. The architecture of the CNN is 5 convolutional layers and 2 fully connected layers. The networks input is 150,528 (227x227x3)-dimensional, and the number of neurons in the networks remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096. Rectified nonlinear units are used for activation in convolution layers and a dropout of 0.5 is used in the fully connected layers. The extracted feature has 4096 dimensions.

C. Classification using SVM

Features are extracted from each poselet and concatenated. We use the top 5 poselets in their respective order and the full image of the person from ground truth bounding box. A one-against-one SVM model with polynomial kernel of order four was then trained to classify different actions. The SVM is trained for the 4 classes. We do not use a softmax layer for final classification as we leverage multiple poselets for each action.

We try two approaches for action feature representation:

Approach 1: We concatenate the feature vectors as 150*4096 feature vector where the unactivated poselets are treated as zero.

Approach 2: We concatenate the feature vectors as 6*4096 feature vector using only the poselets and the full image.

We show in our evaluation that the second approach performs significantly better than the first approach.

IV. Evaluation

We utlize the popular PASCAL VOC 2010-2012 for testing out our methodology. It constitutes of a large collection of labelled images colected over the web, with labels identifying 10 distinct action categories (like Reading, Riding horse and so forth).

For evaluation purposes, considering the scope of this project and time/resource constraints, only a subset (about half) of the expansive dataset was considered and classification was performed over 4 action categories. We considered varied activities of jumping, reading, riding a horse and running.

The data has 119 instances of reading, 117 instances of riding horse, 66 instances of jumping and 246 instances of running. 5-fold cross validation over SVM outputs was performed to evaluate the performance of our approach.

The final results have been summarized in the I. Thesee

Actions	Reading	Riding house	Jumping	Running	mAP
Approach 1	0.670	0.349	0.20	0.559	0.446
Approach 2	0.812	0.686	0.633	0.756	0.723

TABLE I

Average Precision for each of the action category

V. Discussion

In this paper, we describe our approach for action recognition from a single still image. We show our results for 4 classes. We are unable to compare our results to approach by other researchers, because most of the results are available for 10-class action recognition [15] and their code are publicly unavailable to run on our classes. However, our results are promising and validate our hypothesis that deep networks based feature learning is more discriminative than handmade features like HOG.

We use a pretrained model and we show the top 5 activations in the image after feature extraction. We also show the top 5 poselets obtained for each person and show that these are quite similar while poselets have lesser redundancy.

This shows that we do not need to train a model on the poselets separately for the action classification and shows the generalizibility of the features learned through deep neural networks. Gkioxari et al. [15] have trained on the actual ground truth boxes added with boxes having an overlap with ground truth ≤ 0.7 to increase the traing data, while we use the same pretrained model from RCNN and thus do not need any extra training making us much more efficient. We also use top 5 poselets rather than choosing the poselets unlike Gkioxari et al. [8]. We would like to extend our approach to all the 10 classes and would like to also test whether increasing the number of poselets or choosing them more intelligently improves our performance.

References

[1] Bourdev, Lubomir, and Jitendra Malik. ”Poselets: Body part detectors trained using 3d human pose annotations.” Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009.

[2] Brand, Matthew and Oliver, Nuria and Pentland, Alex. ”Coupled hidden Markov models for complex action recognition.”Computer Vision and Pattern Recognition, 1997.

[3] Everingham, Mark, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. ”The Pascal Visual Object Classes Challengea Retrospective.”

[4] Laptev, Ivan, et al. ”Learning realistic human actions from movies.” Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

[5] LeCun, Yann, et al. ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[6] Liu, Jingen, Jiebo Luo, and Mubarak Shah. ”Recognizing realistic actions from videos in the wild.” Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[7] Lv, Fengjun, and Ramakant Nevatia. ”Recognition and segmentation of 3-D human action using hmm and multi-class adaboost.” Computer VisionECCV 2006. Springer Berlin Heidelberg, 2006. 359-372.

[8] Gkioxari, Georgia, Bharath Hariharan, Ross Girshick, and Jitendra Malik. ”Using k-poselets for detecting people and localizing their keypoints.” In CVPR, 2014.

[9] Maji, Subhransu, Lubomir Bourdev, and Jitendra Malik. ”Action recognition from a distributed representation of pose and appearance.” Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[10] Zhang, Ning, et al. ”PANDA: Pose Aligned Networks for Deep Attribute Modeling.” arXiv preprint arXiv:1311.5591 (2013).

[11] Maji, Subhransu, Lubomir Bourdev, and Jitendra Malik. ”Action recognition from a distributed representation of pose and appearance.” Computer Vision and Pattern Recognition (CVPR), 2011.

[12] Yao, Bangpeng, and Li Fei-Fei. ”Modeling mutual context of object and human pose in human-object interaction activities.” Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[13] Yu, K. and Ng, A. ”http://ufldl.stanford.edu/eccv10-tutorial/”. Tutorial in ECCV 2010.

[14] Fukushima, Kunihiko. ”Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.” Biological cybernetics 36.4 (1980): 193-202.

[15] Gkioxari, Georgia, et al. ”R-CNNs for Pose Estimation and Action Detection.” arXiv preprint arXiv:1406.5212 (2014).

[16] Ikizler, Nazl, and Pnar Duygulu. ”Histogram of oriented rectangles: A new pose descriptor for human action recognition.” Image and Vision Computing 27.10 (2009): 1515-1526.

[17] Girshick, Ross, et al. ”Rich feature hierarchies for accurate object detection and semantic segmentation.” arXiv preprint arXiv:1311.2524 (2013).

[18] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ”Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems. 2012.

[19] Deng, Jia, et al. ”ImageNet Large Scale Visual Recognition Competition 2012 (ILSVRC2012).” (2012).

[20] Donahue, Jeff, et al. ”Decaf: A deep convolutional activation feature for generic visual recognition.” arXiv preprint arXiv:1310.1531 (2013).