# Reading Notes

## ORIGINAL

Subhransu Maji, Lubomir Bourdev, and Jitendra Malik1. University of California, at Berkeley. Adobe Systems, Inc., San Jose, CA.

## CATEGORIES

[**Computer Vision**]: Action Recognization, Distributed Representation.

## KEYWORDS

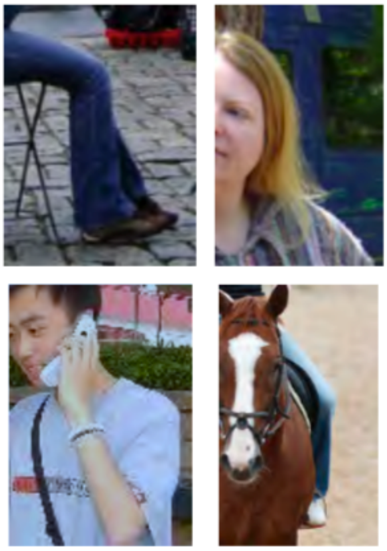
Pose Estimation, Action Recognization.

## AUTHOR

Nino Lau, School of Data and Computer Science, Sun-Yet-Sen University.

# 1. BACKGROUND

## Distributed Representation



Humans have a remarkable ability to infer actions from a still image. Sometimes we can identify the motion of at the people without video necessarily. We can draw the conclusion such as the orientations of their heads, torsos and other body parts with respect to the camera, whether they are sitting, standing, running or riding horses, their interactions with particular objects, etc. And clearly, we can do it from single image.

## Existing Problems

A classical way to approach the problem of action recognition in still images is to recover the underlying stick figure. This could be parameterized by the positions of various joints, or equivalently various body parts. By placing appropriate markers on joints, and using multiple cameras or range sensing devices, the entire kinematic structure of the human body can be detected, localized and tracked over time.

However, considering the single picture, incomplete figure and wide variety of clothing, this turn out to be a hard task. In this paper the authors take the position that recovering the precise geometric locations of various body parts is trying to solve a harder intermediate problem than necessary for our purposes. We advocate instead the use of a representation, the “poselet activation vector”, which implicitly represents the configuration of the underlying stick figure, and inferences such as head and torso pose, action classification, can be made directly from the poselet activation vector.

# 2. APPROACH

## Poselet Activation Vector

The framework is built on top of poselets trained from various images. The annotations are used to find patches similar in pose space to a given configuration of joints. Along with the appearance model one can also obtain the distributions of these joints and person bounding boxes conditioned on each poselet from the annotations.

Given the bounding box of a person in an image, our representation, called the poselet activation vector, consists of poselets that are consistent with the bounding box. Entry for each poselet type which reflects the degree of poselet type in that person. As the pose and appearance information is encoded at multiple scales they use this representation for both action recognition and 3D pose estimation from still images.

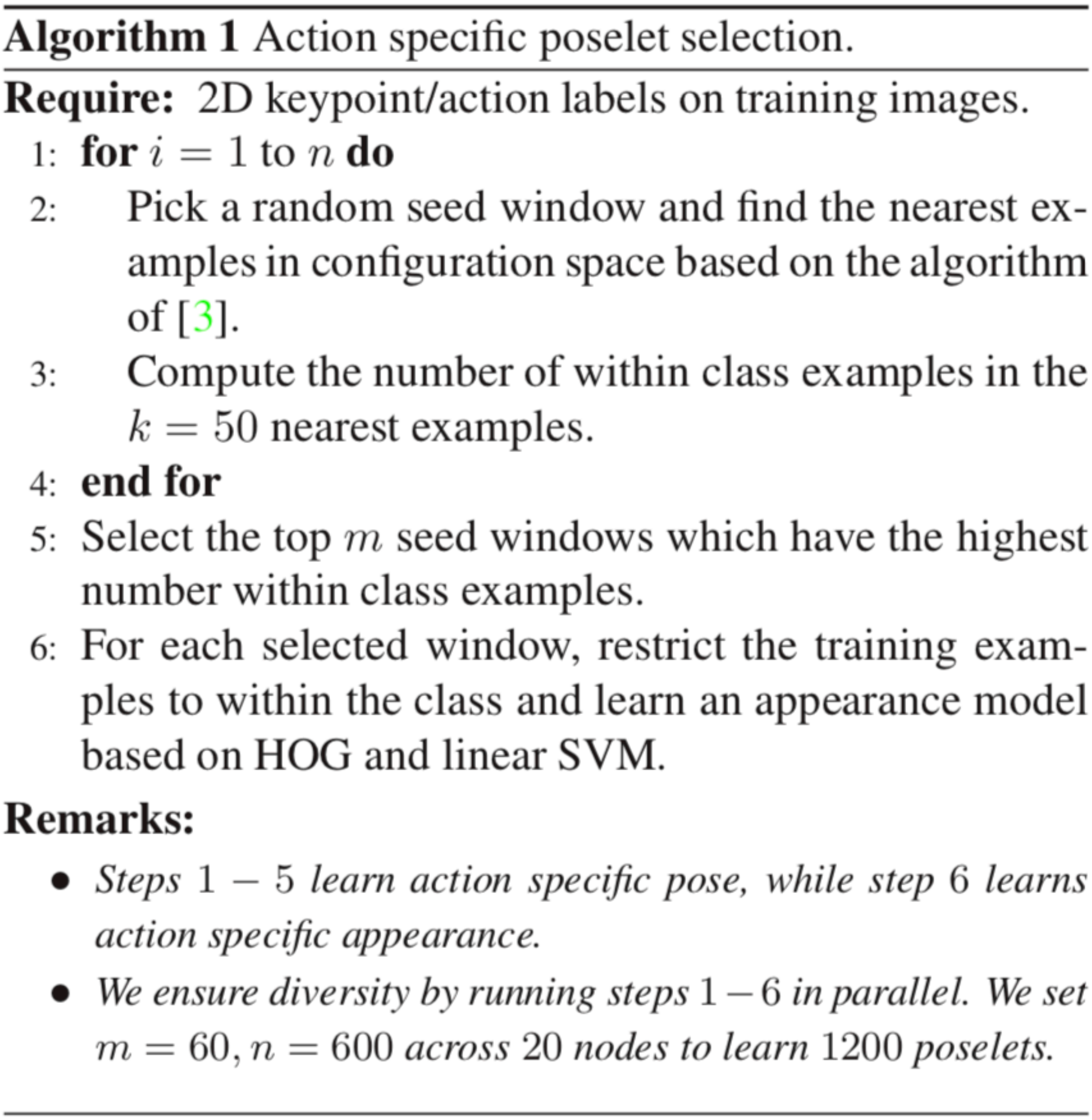
## 3D Pose Estimation from Still Images

Given the bounding box, our task is to estimate the 3D pose of the head and torso. Current approaches for pose estimation based on variants of pictorial structures are quite ill suited for this task as they do bad in distinguishing between a front facing and back facing person.

The pose of the person in encoded at multiple scales and often one can roughly guess the 3D pose of the person from various parts of the person and our representation based on poselets are an effective way to use this information. Our results show that we are able to estimate the pose quite well for both profile and back facing persons.

### Static Action Classification

To present the method for action classification, the authors develop an algorithm as shown.



**Algorithm 1** Action specific poselet selection.

**Require:** 2D keypoint/action labels on training images.

```
1: for $i = 1$ to $n$ do
2:     Pick a random seed window and find the nearest examples in configuration space based on the algorithm of [3].
3:     Compute the number of within class examples in the $k = 50$ nearest examples.
4: end for
5: Select the top $m$ seed windows which have the highest number within class examples.
6: For each selected window, restrict the training examples to within the class and learn an appearance model based on HOG and linear SVM.
```

**Remarks:**

- *Steps* $1 - 5$ *learn action specific pose, while step* $6$ *learns action specific appearance.*
- *We ensure diversity by running steps* $1-6$ *in parallel. We set* $m = 60, n = 600$ *across* $20$ *nodes to learn* $1200$ *poselets.*

## 3. EVALUATION

The authors train 1200 poselets on the PASCAL train 2010 + H3D trainval dataset. Instead of all poselets having the same aspect ratio, we used four aspect-ratios: 96×64,64×64,64×96and128×64and trained 300 poselets of each. In addition they fit a model of bounding box prediction for each poselet and construct the poselet activation vector by considering all poselet detections whose predicted bounding box overlaps the bounding box of the person, defined by the intersection over union > 0.20 and adding up the detection scores for each poselet type. They use this 1200 dimensional vector to estimate the pose of the person.

Estimating the pose of the head and torso separately, the authors find that average confusion matrix over 10-fold cross validation, for predicting four views left, right, front and back. The mean diagonal accuracy is 62.10% and 61.71% for predicting the head and the torso respectively and get the error in predicting the yaw averaged over 8 discrete views using 10-fold cross validation. At last, the authors develop a confusion matrix for the action classifier.

## 4. COMMENTS

The poselet activation vector for action recognition is significantly effective, which is well suited for estimating the 3D pose of persons as well as actions from static images.

After this, the author plan to deal with more disordered images with higher speed. And the future work includes exploration of representation for localizing body parts by combining with bottom-up priors and exploit pose-to-pose relations between people and objects to estimate better.