CS 331 Class Notes - October 11, 2012 (alahi@stanford.edu)

Aman (Delaitre, Learning Person-Object Interactions for Action Recognition in Still Images)

Main Contributions of Paper

Use stronger discriminatively trained body part and object detectors (poselets)

Objects: Use LSVM detectors

Body parts: Use 160 pretrained detectors returning x, y and sigma (scale)

Developed body part/object interaction representation

Representing pairwise interactions (manual annotation - no location, just label):

- One root (body part), one leaf (body part/object)

- *NOTE:* "Person" treated as object, "left leg" treated as body part

- Offset "v" (x, y, scale space)

- Fixed displacement cost "C" of leaf (determined by clusters) - covariance matrix penalizing distance between actual object location and expected object location

- Determine maximum "response" of object (pi) and body part (pj) - how strongly is this interaction represented?

Learning

1.) Determine set of positions of all positive detector responses

2.) NMS (Non-maximum suppression) - choose top detections

3.) Use clustering of object location / body parts

4.) Train SVM classifier (one vs. all)

How did you like the paper?

- Some interaction-object pairs were sketchy (for instance, accidentally correct classification of using a computer with leg -> sometimes leg is mapped to keyboard). Sometimes root->leaf, root = leaf

Vignesh (Weakly Supervised Learning of Interactions Between Humans and Objects)

What is meant by "weakly supervised"?

- Use some form of annotation in training or testing, action labels for individual images are the only labels available during training

Method Overview

Human Detection (Smart Engineering)

- Extension of Felzenszwalb DPM (Deformable Parts Model)
    - 2 models from mixture components of DPM, 1 face & 1 upper-body
- Must map all body parts on to common reference frame - using regressor
- Upper body is labelled in training set (manual annotation)
- Must cluster all 4 detectors into one human cluster
    - Used by weighted mean-shift algorithm
    - Then bad clusters filtered out using NMS

Weakly Supervised Action Detection (Most novel part of paper)

Input: training images with action labels, human annotations in detector

Output: Object bounding box, distribution of object position/scale relative to human

Method:

Detect Objects

Run "Objectness detector" - gives score of bounding box, object or not, choose top 500 regions/bounding boxes (cheaper processing)

Then, use MRF model to score human-object pairs

- MRF: create fully connected K-partite graph (every point of K clusters is connected to every point of every other K-1 clusters)
    - Edge weight: pairwise energy, Unary weight: unary energy
- Energy (to be minimized) cues:
    Unary: objectness measure (rewarded), overlap (penalized)

Human-Object pairwise: small object/human scale, human-object distances, object similarly oriented to human, scale covered must be same across images

Object pairwise (between objects across samples): Color Histogram similarity, BoW similarity

- Then, choose one box from each image that minimizes energy

HOI Model for Action Recognition

Given bounding box corresponding to object in each image, learn spatial and scale distribution of object boxes w.r.t human

Learn appearance model of object

Does fairly similarly in respect to "Koniusz" et a. [2]

How did you like the paper?

Too much engineering - overengineering