Actom Sequence Models for Efficient Action Detection[Adrien Gaidon Zaid Harchaoui Cordelia Schmid, CVPR 2011]

Presenter: Evan Shieh
Time: 9:05 - 9:50AM Nov 1, 2012

Agenda

- Context learning temporal structure actoms,
- Actom sequence model(ASM) and its feature representation

Ideology

-Detection(if and when) is important than classification
-Temporal structure vs bag of features
-Introduction of actom, action unit which represents the characteristic for an action

[DISCUSSION]

Is the concept of actom really innovative?
-A similar concept is moveme which is atomic motions
-Generative approach for multi-scale detection

Method

-Actoms are specific to each action class and central frames of a clip
-Need to manually annotate actoms
-Training data: Each clip to annotate has a specific action with boundary
-Each actom is associated with a radius(timespan of an action). An adaptive radius determination method is used: $r_i = 1/(2-\rho)$ and ρ is the overlap ratio between two actoms $\rho \in [0,1]$ and ρ is set as 0.75
- Weight is applied to how far away from the center of actoms
-ASM vector is L_1 -normalized

Feature

3 Features are used as building blocks

[DISCUSSION]

-spatio-temporal interest points(STIPS)-> Harris in 3D
-HOG 3D spatial gradient in time without direction of motion
-HOF 3D temporal magnitude of the actions with directions of motion

ASM classifier

-Binary SVM action classifiers are trained using intersection kernel like the intersection of two histogram which is similarity measure for salient features.

[DISCUSSION] about negative examples

-Weakly and randomly initialized examples which may include false positive examples
-There are 3x-6x more negative examples compared to positive examples
-Negative examples are important in the discriminative model while negative examples are not important in the generative model
-It is good to have negative examples similar to positive examples

[Discussion]
-The guest raised a question: negative and positive examples (classification) are at action level(feifei) or actom level(evan)? Kevin convinced us that they are at action level.
-How to find the center of actoms?(bangpeng) randomly placed/ learn a distrubution of the center
- scott's factor is not clear in the presentation yet

Result
-Baseline is bag of feature+ temporal grid
-ASM uses sliding windows

Strengths[DISCUSSION]
-The idea of actoms is intuitive and one question is how many actoms are ideal?
- Probabilistic and generative (but sensitive to actom annotation)
- Good experimental control (partially rigid temporal structure)
- Flexible utilization of temporal structure

Final comment
-hard to scale up

Searching Video for Complex Activities with Finite State Models[Nazlı ˙Ikizler and David Forsyth, CVPR 2007]

Presenter: Cameron Schaeffer
Time: 9:53-10:30AM Oct 31, 2012

Introduction

-Analogy between HMM node and syllables
one phoneme is about 3 sound bytes
one word is about 3 phoneme
one action is about 3-10 frames of jump
one activity is roughly 3 actions
-The goal is to be able to search video for general activities without labeling every possible activity

[DISCUSSION]

-Action is small and punctual in time

-Activity is a composition of several actions

-Currently the definition of action and activity is messy in human recognition and this paper try to define these terms in a formal way

Joint level HMM

-A HMM for each action for each 3D part independently

-9 actions for arms and 6 actions for legs

-Features are vectors quantized value of 3D joint angles of arms/legs

-40 quantized clusters for each of our 4 limb 4d vector

-Each frame's feature vector : average feature vectors

-3D model is used (viewing angle invariant)

-Grouping hidden state and grouping action models to (arm/leg) activity model

-Dynamic programming is used to match projections of 3D motion capture data to 2D arm/leg bounding boxes over snippets of multiple frames

-Legs and arms are trained independently: segment the leg and arm separately

Regular Expression Machine

-Query language

-Very simple equation P(string|frames)

Strengths

-Cloth color invariant: body segment tracker-part based model

-Viewing angle invariant: synthesize smoothly between frames("linking step")

[DISCUSSION]

What is to be learned

-Learn actions in HMM for arms and legs

-Learn joint feature vectors

-Action label only appears in the training examples

-No semantic meaning to high level (no classificiation)

-Sliding window for HMM is used

-If an action lasts too long, then penalty is applied

-The duration can be determine (the probability to stay in the same state)

-HMM(2D data) VS SVM(2D data): both are 2D in testing. But HMM uses the lifted 3D representation

Presenter's Comment

Many concepts are not explained clearly in the paper