<< Social Interactions : A First-Person Perspective >>

Kevin Presents

Use Egocentric wearable camera

Why Egocentric?

1) Natural videos of family and friends

2) Occlusion is less likely

3) It is not practical to track all the individuals and record all the interactions with static cameras

Overview

two types of features

1) faces : location, attention, social role
2) first person motion

faces and attention: estimate location + orientation of faces in space

why faces?

1. faces and their attention play a important role in social interactions
2. robust detection compared to other object detection

Finding faces

1. track faces - Pitpatt
2. estimate 3D location of every face
3. estimate orientation(yaw, pitch, roll) in 2D

only subset of individuals is visible in each frame : build local map of faces around first-person

Attention : Face location+ orientation gives approximate line of sight

Goal: Find 3D location of where each person is looking at

If someone is looking at somewhere it is likely someone else is looking at the same place

MRF

Unary potential

First term : a face is looking at some location

Second term: doesn't allow the person to look at themselves

Third term : to look at another face

Pairwise Potential

For two people it is likely to look at the place where other face is looking at i.e., the if there is a person looking at some location that the probability that another person is looking at the same place is higher

Approximate MRF

Similar to Alpha-expansion

Attention Results

Even if we cannot detect some face we can infer where she/he looks at

Even if some face is facing other direction at looking at you

Manually label who each person is looking at in 1000 frames

Split half for train/test

71.4% accuracy

Features

1. location of faces around first-person

2. First person head movement

3. attention and roles

    for each individual x :

    -number of faces looking at x

    - whether first-person looks at x

    - If there is mutual attention between x and first-person

    - number of faces looking at where x is attending

Temporal Model

HCRF

Labels y are binary for detection, multiple values for recognition

each x is a frame in HCRF

each term in HCRF

first : the hidden variable and feature vector

second : hidden variable and label y

third : pairwise between hidden variables

all the w's are learned from training

Dataset

8 subjects, >42 hours of video

each day subset of individuals with GoPro camera (1280*720, 30 fps, >2million images)

Experiments

6 labels : dialogue, discussion, monologue, walk dialogue, walk discussion, background

walk dialogue and walk discussion seems to be confusing

social network

cluster faces into multiple bins and manually assign each bin to an individuals

weight connection based on the frequency of appearance of each face

::Discussion::

They use the person's height to estimate the distance from the camera

Big face is close and small face is far

So video features are not used? They do, they use temporal model

They assume the faces sizes are similar

And use tripod to fix the camera and ask a person to stand a certain position and do calibration

They are not using audio

Probably we can retrieve this kind of interaction information from only audio analysis

The problem seems to be intereting

Very new problem space, early explore

<< Social roles in Hierarchical Models for Human Activity Recognition >>

CVPR 2012

Vignesh presents

Three levels of abstraction : Overall event in the video, action, social role

Social role: attacker, defender, other

Model formulation

Hierarchical

Every frame has an attacker

Every role is connected to attacker

Every role is connected to another role within spatial vicinity

Graph structure dependent on choice of attacker

Action potential: feature

Unary role potential : dependencies between role of the person and action label, position of the player with respect to the role

Pair-wise role potential : dependencies between a pair of social roles under an event

Structured SVM formulation:

Max margin learning

Needs inference at each iteration

Inference:

Exact inference is NP-hard

Coordinate ascent keeping one layer fixed at a time

Easy to infer action labels since there is no structure

Inferring Role Labels

Fix the role of an attacker(k), and infer the other player roles using Loopy BP

Choose the best choice of attacker role which maximized your score

Hockey dataset:

5 roles(attacker, first defenders, man-marking, defenders defend against space, other)

11actions(pass, dribble, shot, receive, tackle, prepare, stand, jog, run, walk, save)

58 videos

3 events(attack pay, free hit, penalty corner)

Nursing dataset

4 roles(helping, visit, reside, falling in a hospital)

2 events(fall & non-falling)

so falling corresponds to attacker role

Sample/numerical results (Hockey)

There is a huge drop when they ignore the interaction between action

Using full model results better confusion matrix than using only unary potential model

Sample results (Nursing)

A lot of falling scene… ha ha ha… ☺

Visualization of pairwise weights

They are not symmetrical

Conclusions

Novel hierarchical model relating three layers of abstraction

Showed that knowledge of social roles helps activity recognition

Social role recognition decided by interaction between roles in a video

::Discussion::

How the trajectories are used? Extract HoG features and use pool it from every 3 frames and use SVM

Don't track ball

Seems to be similar to Bangpeng's work

Any further comments about these two papers?

Two difference social role papers : pioneering the problem

It also connects computer vision to other fields