MARCH 2021

EFFICIENT FACE DETECTION & TRACKING IN VIDEO SEQUENCES BASED ON DEEP LEARNING

Notes

Face tracking is Hard. deveation
I will always have some error rate. But it's effectiveness can Be ifluanced by a few things

1) target whole individual moves too fast screws up the update to the tracking

2) the face it's self rotates or some details about the face chunge again messing up the tracking

3) The enviornmet impact (change of lighting, etc.) cause difficulty

Deep ML has exploded in popularity It has shown promising results for face detection.

In this paper they will propose a efficient face detection & tracking Deep ML framework.

1. they first propose model for initial face extraction
2. they propose efficient model that can deal with Moment, lighting changes, etc
3. they that care of this Scale deviation problem

1. Face detetion

People Have tryed many things. most recently: YOLO.

Face Detection Falls into 2 categorys

a) Region Based

b) Sliding Window Based

Procides to explain relevent work in Both

Part 1 continued But Now on face tracking

3 types of face tracking

a) feature-Based

b) modul - Based

c) learning-Based

a) has poor effects Because of the difficulty to extract feature from target-in-motion

b) also not that good due to afformation Problems caused by the real world. & they have high latency

2. there face Detection model

so this is a Rather Deep convolutional net with residual connections. Inbetween the conv layers they note that they add something called a SEN

Layer. this layer is supposed to fuse the features of each conv layer. It "recalaborates" Learns the importance of each feature. it "enhances usefull info" & "suppresses useless info". It works by after the conv first do squeeze then Excitation.

Squeeze- why? in early layers conv has small receptive field so this leads to poor understanding of global feature.

So to fix this we compress the whole feature map for each channel. $H \times W \times C \Rightarrow 1 \times 1 \times C$

this is now the summary statistic for the feature globally & we will use it in the next step

Excitation - This is where we do the "recalibration", we will figure out which feature relate, which are important, which are redundent etc.

it does this via a gating mechanism where it assigns importance via learned weights. If you look at the equation in the paper it feels a little like an attention mechanism. & after these linear layers (2 of them) is passed through a sigmoid to get importance score.

Loss functions

so sigmoid is very popular in Image classification, But it can't meet the needs of inter-class distance increase & intra-class variance decrease.

They go through itterations on the soft max function improving it to fit the needs better till they reach Addative-Margin-SoftMax the best one

Now for the face tracking model

So they will be using a reggression based model where we will extract facial features from 2 adjacent frames then predict the next spot. the Network has 4 parts

P1 -
P2 - P4
P3 -

P1 & P2 are the earlyer RNET models. P4 is a FC layer & P3 is a correction Network.

the Correction Network helps with more Stable learning since there can be scale deviation this layer is supposed to adjust the size of the face detected.

Model Trainig:

if we know position of face in last 2 frames then it can't move much to the next frame. This is good prior knowledge. Because of that new center is

$C'_x = C_x + w \cdot \Delta x$ ← Random Var sampled from Laplace Distribution

$C'_y = C_y + h \cdot \Delta y$

(C_x, C_y) = center now

w, h = width, height

Input to the model is

1. video sequance
2. target face in the initialization window

we detect (C_x, C_y) from #2

For Evaluation

2 Data Sets

1. images of faces
2. videos from VTB

Resuts show their models very good.