Modeling 3D Environments through

Hidden Human Context

要解决的问题：1.物体检测，2.物体放置

解决办法：Infinite Latent Condition Random Field(ILCRF)

主要思想:将human pose考虑在内

一般的CRF，把每一个物体作为一个node，物体之间的关系作为边

ILCRF；物体作为可见node，人作为潜在的node（latent variable），object affordance作为边，使用potential function表示，连接human和物体的label，由于每个场景中人的个数是不固定的，使用Dirichlet Process来获得，

然后用maximum a posteriori(MAP)来学习object affordances，Gibbs Sampling来学习样本label

Scene arrange,训练阶段，多学了object-object结构

Multi-Target Tracking by Discrete-Continuous Energy Minimization

问题：Muiti-target主要分为两个子问题：1.识别视频中的target（data association），2.得到target的运动轨迹

解决方法：提出一种方法能将data association 和trajectory estimation统一用一个模型来表示，提出了“minimization of a consistent discrete-continuous energy.”

使用连续的曲线来表示轨迹，使用离散的多labelCRF模型来表示data association

Social Grouping for Multi-Target Tracking

and Head Pose Estimation in Video

问题：multi-camera监控摄像头中的Muiti-target tracking 和头部姿势估计

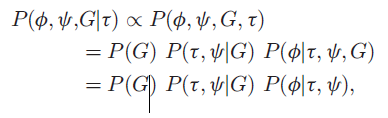
解决办法：social grouping 人们会成群结队的走，而且轨迹速度目的地相同，更有可能互相看或看同一个地方

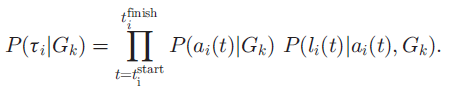
需要解决两个MAP问题：

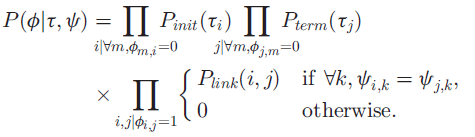


第一个MAP问题：

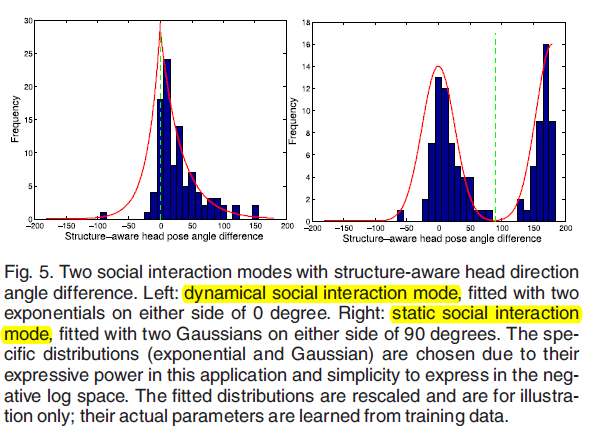
分块来看



表示的是Social Grouping算法得到的相似度，其中分别表示group k出现在camera ai(t)的可能性，后面表示t时间时ai中的group会出现在位置li(t)的概率,使用Gaussian表示，中心是uk,a(t)



算法命名为SGB(Social Grouping Behavior)



Analysing Domain Shift Factors between Videos and Images for Object Detection

目标：比较视频和图片在物体识别中效果的不同

评价标准：得到的结果和gt相比>50%算正确，一类的performance是有平均Precision（AP）来定的，整体performance由中值AP（mAP）来定的

detector：DPM和RCNN

步骤：

1. 提出了个公式来量化factor的影响
2. 找到方法使两个trainingset在公式上更加相似
3. 对比detector在处理前后的表现效果

能够量化由于在不同领域training而产生的performance gap，用以上方法对每一个factor进行处理

Spatial Location Accuracy

TrainingSet分为四种，对视频使用自动识别的算法PRE，FVS，以及手动人工标注的视频和图片

Appearance Diversity

手动将视频set中相同物体，不同viewpoint和遮挡的图片分为一组，appearance diversity用unique samples的数量来表示，修改方法，删掉重复的samples并且让两个domain大小相同（YTO因为重复会删掉很多，VOC因为要和YTO数量保持一致而删掉很多）

影响：VOCtest时，使用VOCtrain在减少了sample之后，mAP有了明显的减少，因为sample减少了一半左右，使用YTO则没有明显变化，因为只有重复的被删除了。YTOtest结果相似。说明相近的samples并没有带来过多的信息，这说明仅需要对视频每一个shot选择一部分很少的镜头就可以了。 同时，在testVOC差距缩小,testYTO差距增大，可以理解，因为对VOC来说效果变差了，而对YTO没有影响。

Image Quality:使用的数据集市Unique Samples

Measurement：计算一个bounding-box里HOG cells的梯度大小的总和，并用size来泛化.平均gradient energy:VOC 4.4 YTO 3.2，原因 视频压缩，运动模糊，低色彩对比度

Equalization：将VOC进行模糊，使得二者的模糊程度相同（Gaussian blur and motion blur）,motion blur的参数（sigema和K）针对每一类都不同，使得gradient energy两类相同

Impact：对两个dataset进行模糊，结果对trainVOC模糊的影响更大，trainYTO对结果影响较小，因为本来YTO已经有了motionblur，再模糊不会丢失更多的信息

两种blur都能减小test VOC 的gap。对testYTO来说motionBlur会增大gape

Aspect Distribution: 在两个领域中一样类型的比较少

Measurement： dKL

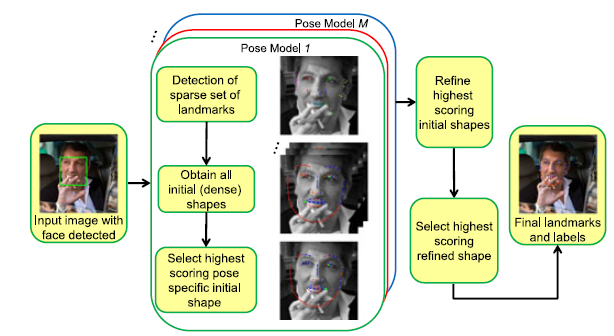
Equalization：建立子集，需要找到最大的自己使得满足不等式，使用greedy forward selection algorithm，同时不断向里加欧氏距离最小的一对sample

Towards a Unified Framework for Pose,

Expression, and Occlusion Tolerant

Automatic Facial Alignment

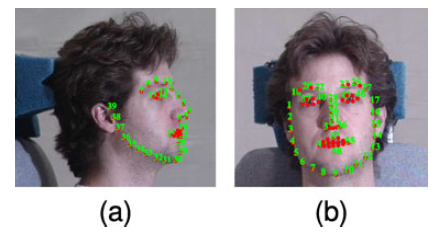
目标：对姿势表情遮挡容忍度较高的人脸对齐



1. Sparse Landmark Detection

首先找一些稀疏的key landmark作为seed landmark

使用的数据库CMU的MPIE，正面图像（45度角以下）使用68个点，侧面39个点



选择其中的8个点作为seed landmark:两眼中间的两个，鼻子，下巴尖，嘴角两个，和脸部边缘靠近耳朵的两个点，侧面的话只有5个点

训练阶段对每一个landmark的expression，pose的局部表现进行建模，【用了表现各方面的旋转角度（various yaw angle）使用了10个模型，对正面角度的张嘴表情使用了6个模型，但为了保证速度这个阶段没有用，只有在最后的refinement阶段用就可以了e（那为什么要提呢】

模型的构建过程是，要建立一个分类器，目标是将landmark周围的local texture和其他的landmark和occlusion区别开，方法是从正样本中gt的位置以及附近区域提取feature（HOG），负样本是其他location的descriptor，并用PCA降维，分类器使用Real AdaBoost训练，AdaBoost也有最高得分，可以进行比较。

1. Dense Landmark Alignment and Optimal Shape Initialization

解释了一大段说初始化很重要 我并没有看太明白

解释一下how shape models for CLMS work:脸的形状s使用N个坐标（x,y）来表示的，然后再使用GPA进行归一化，消除大小/旋转等影响，使其计入到参考坐标系内，然后根据s的中值和PCA子空间（来表示脸部的 空间变化？）来建立PDM（Point Distribution Models），脸部shape可以表示为：本文中并没有用！！！他妈的！其中b是形状系数(能够最好的表示目前已经根据局部特征顶下位置的landmarks)，T是旋转缩放等变换系数



本文中建立了16个PDM来表示不同的pose和表情，同时本文并没有用上述的公式，没有采用PCA!!!而是将所有的shapes都保留在了dictionary中，后面使用了l1-regularized least squares来确定shape系数，不过还是会保存sm

选取了10个正面的pose mean shape用来确定最佳初始化，这个脸需要和其他大部分脸大体上相近，所以会有一个评分系统，具体算法：没看懂 反正就是一些公式来计算subspace和现在这个landmark之间的距离，然后再用负样本重新算一遍，最后将正负样本的结果做比，得到一个比值，最后将所有landmark的比值求平均数R,最后的结果是样本点总数/R

1. shape refinement

迭代的fitting过程 基于ASM和CLM

每一个landmark现在的位置周围会有一个window，其内部每一个像素的localtexture都会scored并分类使用的是自己的local texture分类器（什么分类器来着想不起来了），然后每个landmark都会想邻域中分数最高的位置移动，直到收敛，但是在迭代的过程中生成的脸要进行regularized，为了保证生成的还是个脸…

实验：

使用CMU multi-PIE(MPIE)数据库，使用6495张图片进行对shape and texture models进行训练，正面68个landmark，侧面39，根据角度聚类成10个bins，bin中的每张图片landmark个数相同，这10个是微微张嘴或者完全闭嘴的，同时会有相似的数据是最完全张开的6个

对比：与其他好几个算法对比，其他的训练集都不大一样，为了fair，我们训练用了好几个训练集（？）

初始化问题也需要解决，选择了opencv自带的Viola-Jones脸检测

有些对box的要求很高，COFWtest set的box不是很好，所以做了一个square crop

大部分方法都使用了正面68个点的识别方法，但是有些也用了49个点，所以我们的结果reported for both these cases,使用最大可能common landmarks（？）。对于COFW，数据集中只有29个手工标定的点，所以我们用25个landmarks使得双方都有，同时在对PCPR的对比中选择了full set 29个landmark，在与Tree-DPM对比中使用了39个点，因为只有MPIE数据库包含了超过45的转角，

每张图片的fitting error 要根据两个眼角的距离进行归一化

MNFE：Mean normalized Fitting Error

表1：训练over the largest number 68点

表2：only interior facial landmarks 49点 –》第二个普遍表现更好

图8：累计误差分布，用来比较方法的准确程度。我们的方法最好

图9：体现了NFE在不同算法的结果中对pos的敏感程度，每个testset中的图片被根据角度进行分类，图中显示的是不同的数据库，然后计算这个bin中的所有的NFE

图10：体现了表情对结果的影响，具体方法如9

表3：在细节上对比了我们的方法和RCPR在遮挡上的不同，trainLFPW+COFW