一种基于动态贝叶斯网的视觉注意力识别方法

董力赓,邸慧军,陶霖密,徐光祐
(清华大学计算机科学与技术系清华信息科学与技术国家实验室(筹),北京 100084)

摘 要: 头部姿态估计是识别用户视觉注意力目标的主要依据.但在实际应用场合下,大范围头部姿态、低分辨率图像以及光照变化等因素使得可靠、准确的头部姿态估计难以实现.针对这些困难,提出一种基于动态贝叶斯网模型的视觉注意力目标识别方法.通过人脸图像与多个人脸姿态类别的相似度向量对头部姿态进行度量而不是显式的计算具体姿态值.模型融合多注意力目标、多用户位置、多摄像机图像等因素间的概率依赖关系并进行联合推理.智能厨房原型环境下的实验结果表明提出的模型是有效的.

关键词: 视觉注意力目标识别; 动态贝叶斯网; 智能厨房
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2011) 3A-140-07

Dynamic Bayesian Network Based Visual Focus of Attention Recognition

DONG Li-geng, DI Hui-jun, TAO Lin-mi, XU Guang-you
(*Department of Computer Science and Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Visual focus of attention recognition is usually based on head pose estimation. However, in a real application, it is difficult to accurately estimate the head pose due to large pose variations, low resolution images and varying illuminations. To handle the problem, we propose a dynamic Bayesian network model to infer the visual focus of attention. The head pose is not explicitly computed but measured by a similarity vector which represents the likelihoods of multiple face pose clusters. The model encodes the probabilistic relations among multiple foci of attention, multiple user locations and faces captured by multiple cameras. Data are collected in a prototype ambient kitchen and results show that the model is effective.

Key words: visual focus of attention recognition; dynamic Bayesian network; the ambient kitchen

1 引言

基于用户注意力的系统[1]在很多人机交互应用中起到了重要作用,其中视觉注意力目标(Visual Focus of Attention, VFOA) 特指用户眼睛所注意的目标.在人机交互中,通过理解用户的视觉注意力目标,可以理解用户的兴趣或者意图,从而可能给用户提供相应的主动服务.视觉注意力与头部姿态和人眼视线方向(eye gaze)有关.研究[2]表明,在很多情况下通过头部姿态足以识别用户的注意力目标.因为通常用户并不习惯于长时间斜着眼睛盯着某个目标,而会将头转过去正视该目标.此外,在实际应用场合下采集到的图像分辨率较低,使得精确的估计人眼视线方向非常困难.本文以智能厨房环境为例,研究在实际应用场合中如何根据头部姿态来识别用户的视觉注意力目标.智能厨房是一个在厨房中安装了摄像头、投影显示屏等设备并能够提供普适计算服务的智能空间.其工作台后的墙壁上装有多个投影显示屏来显示菜谱等相关信息,同时装有多个摄像机拍摄不同位置的用户.我们的目的是当用户在工作台前准备饭菜的时候根据拍摄的图像识别用户的注意力目标,即所注视的显示屏,从而提供更多的主动服务.

在智能厨房这种实际应用场景下识别视觉注意力目标的方法需要考虑以下因素.第一,用户可能站在厨房工作台前的多个地方准备饭菜并观看多个投影显示屏,从多个摄像机拍摄的图像包括大范围变化的头部姿态而不仅仅是接近正面的姿态.而实际应用场合下,大范围头部姿态、低分辨率图像以及光照变化等因素使得可靠、准确的头部姿态估计难以实现.所以需要对头部姿态做某种度量而不是显式的计算准确的姿态值.第二,根据用户相对摄像机的位置、摄像机获得的头部姿态以及多注意力目标的空间位置,才可以识别用户的视觉注意力.因此需要考虑如何融合多注意力目标、多用户位置和多摄像机之间的相互关系.第三,环境中的多个摄像机可以拍摄到用户在不同位置不同角度的图像,

收稿日期:2010-01-15;修回日期:2010-09-27
基金项目:国家自然科学基金(No.60673189, No.60873266, No.90820304).通信作者:陶霖密,副教授.

需要考虑如何融合多摄像机的信息,而不是仅仅使用单摄像机的信息.

近年来,越来越多的研究者开始研究视觉注意力目标的识别问题.Stiefelhagen 等[3]研究了在小型圆桌会议环境下参与者的注意力目标识别问题,会议桌上放置了一个全方向的摄像机.他们后来研究了在多个远距离摄像机环境下识别注意力目标的方法[4].Ba 和 Odobez[5]研究了在小型会议环境下如何利用会议中的上下文信息,包括用户的座位,对话过程中发言权的交替模式以及投影仪的变化情况等,并结合姿态信息,通过动态模型推理出会议中的群体交互行为.Otsuka 等[6]也研究了在小型会议环境下根据头部姿态和语音信息推理出群体交互行为.然而,会议环境下的用户主要坐在固定的座位上,其身体并没有多少移动.Smith 等[7]研究了户外环境下的注意力目标识别问题,他们主要分析路过的行人是否观看了墙上的海报,然而该文中的注意力目标只有一个.Zhang 等[8]也研究了在户外环境下监控用户的注意力目标,但是该文的头部姿态范围仅局限在接近正面的人脸姿态.这些研究者的工作主要处理固定位置下多注意力目标的识别,或者多用户位置下单个注意力目标的识别.然而,本文的应用环境包括多用户位置和多注意力目标.其次,先前的这些工作通常需要采用某些额外的方法来估计头部姿态值,包括基于 3D 模型或者基于 2D 表观模型的方法.本文应用环境下的图像分辨率较低使得 3D 方法不太可行.基于 2D 表观模型的方法通常需要额外的数据库训练人脸姿态模型.然而由于数据和模型的推广性问题,额外的数据库训练得到的人脸姿态模型在本文的应用环境中并不适用.考虑到我们的目的是识别用户的视觉注意力目标而不是准确的头部姿态值,本文主要关注是否有可能不通过显式的计算头部姿态值而根据人脸图像就可以识别出用户的视觉注意力目标.另外,先前的工作大部分仅采用单摄像机信息,并且用户注视不同目标时的姿态差异较大.仅文[4]采用了多摄像机信息计算头部姿态,并根据姿态值识别注意力目标.而本文融合多摄像机信息进行识别.

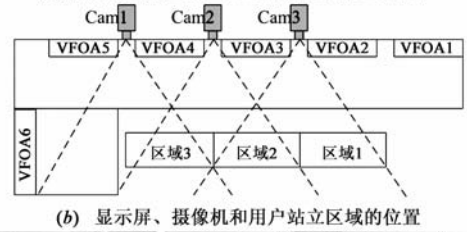
本文提出一种基于动态贝叶斯网模型的视觉注意力目标识别方法.模型最高层隐变量就是视觉注意力目标.通过人脸图像与多个人脸姿态类别的相似度向量对头部姿态进行度量而不是显式的计算具体的头部姿态值,该相似度向量是模型中间层的隐变量.模型的观测包括多摄像机拍摄的人脸图像和人脸位置.模型融合了多注意力目标、多用户位置、多摄像机拍摄的人脸图像、多个人脸姿态类别等因素之间的相互概率依赖关系,通过计算隐变量的最大后验概率来联合推理用户的视觉注意力目标.我们在智能厨房原型环境下采集了数据,实验结果表明提出的模型是有效的.

2　应用环境

本文以智能厨房为例,研究视觉注意力目标的识别方法.智能厨房[9]是一个在厨房中安装了摄像头、投影显示屏等设备并能够提供普适计算服务的智能空间.如图 1(*a*)所示,其工作台后的墙壁上装有 4 个投影显示屏,上面显示菜谱、做菜指南等相关信息.当用户准备做饭时,不同的显示屏会显示不同的菜谱.用户浏览不同的菜谱后会仔细阅读其最感兴趣的菜谱.通过分析用户长时间注视的目标,智能环境就可以知道其对哪个显示屏上的内容感兴趣,从而可以提供更多的主动服务,比如进一步在多个显示屏上显示该菜谱的详细制作步骤,帮助人们制作新的菜品.

由于单个摄像机视场有限,我们在墙上的不同位置安装了 3 个摄像机,这样当用户站在任意显示屏前的时候,都可以被摄像机拍摄到.我们将摄像机安装在两个显示屏边界中间的地方,这样摄像机就不会影响显示屏的内容.图 1(*b*)显示了显示屏和摄像机的位置关系,图 1(*c*)显示用户在观看某个显示屏时不同摄像机拍摄得到的图像.我们在每个显示屏的中心放一个纸片,并让用户注视不同的纸片来模拟观看显示屏的过程.除了在 4 个显示屏中心外(VFOA2,3,4,5),我们还在另外两个地方(VFOA1,6)放置纸片以获得更多的数据.我们让用户观看某个纸片一小会儿,然后转头观看另外的纸片.3 个摄像机同步记录视频.

(*a*) 智能厨房,墙上的4个显示屏,显示屏上的菜单



(*b*) 显示屏、摄像机和用户站立区域的位置

(*c*) 当用户站在区域2观看某个显示屏时3个摄像机拍摄得到的图像

图1

严格来讲,用户可能站在柜台前的任意位置.然而,根据我们的观察和经验,在真实的厨房中,用户倾向于在某个他比较习惯的位置准备饭菜.而且,由于墙上的显示屏可以提供饭菜制作信息,所以用户通常会

选择站在某个显示屏前.基于这些观察,为了简化数据的采集,我们将柜台前的空间根据显示屏的位置划分成 3 个区域,并让用户分别站在这 3 个区域注视不同的显示屏进行数据采集.

3 动态贝叶斯网模型

3.1 模型概述

本文提出的用于识别视觉注意力目标的动态贝叶斯网模型如图 2 所示.在模型中,隐变量 F_t 代表用户在 t 时刻的注意力目标,它的值就是 M 个可能的注意力目标(下面用 VFOA 表示).隐变量 C_t^i 表示在 t 时刻由摄像机 i 拍摄的人脸图像所属的姿态类别,它的值就是 K 个人脸姿态类别.观测变量 Z_t^i 表示在 t 时刻由摄像机 i 拍摄的人脸图像.观测变量 L_t^i 表示在 t 时刻由摄像机 i 拍摄的人脸在图像中的水平位置.在我们的模型中,我们忽略了人体身高的影响.该动态贝叶斯网模型描述了这些变量之间的概率依赖关系.人脸姿态类别 C_t^i 受当前的注意力目标 F_t 和人脸位置 L_t^i 约束.也就是说,当用户站在某个固定的位置并关注某个特定的目标时,通过摄像机 i 所拍摄的人脸姿态朝向将是确定的.人脸观测变量 Z_t^i 依赖于人脸姿态类别 C_t^i,即某个姿态下的人脸图像应该具有某些相同的特征.

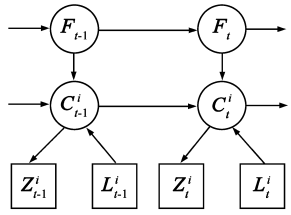


图2 动态贝叶斯网模型,下标 t:帧号,上标 i:摄像机编号

模型融合了多摄像机信息,摄像机编号通过 i 表示.除了视场的原因外,多摄像机在我们的应用场景中可以帮助更准确的识别 VFOA.比如,当用户站在区域 2 和区域 3 时,两个摄像机都可以拍摄到用户.在有些情况下,仅靠单摄像机获得的图像可能无法正确的识别用户的注意力目标.此时,通过另外一个摄像机的信息,也许就可以较容易的识别用户的注意力目标.如图 3 所示,用户站在区域 2 并分别观看 VFOA1 和 VFOA2.在这两个情况下摄像机 2 获得的图像非常相似并且难以区分.而从摄像机 3 获得的图像则有较大的差异,从而相对容易区分.因此,我们在模型中需要融合多摄像机信息.

图3 用户站在区域2并分别观看VFOA1和VFOA2时,摄像机2和3拍摄的人脸

3.2 模型各部分说明

根据该概率模型,可以写出所有变量之间的联合概率密度分布,公式如下.

$$P(F_{1:T}, C_{1:T}^{1:R}, L_{1:T}^{1:R}, Z_{1:T}^{1:R}) = \prod_{t=1}^{T} P(F_t \mid F_{t-1}) \prod_{i=1}^{R} P(C_t^i \mid C_{t-1}^i, L_t^i, F_t) P(Z_t^i \mid C_t^i) \tag{1}$$

下面描述模型的各个部分. $P(F_t \mid F_{t-1})$ 表示相邻时刻不同 VFOA 之间的转移概率矩阵.当用户注视某个目标的时候,一般情况下将有较大的概率继续注视该目标,有较小的概率转移到另外一个目标.令矩阵的对角元素值较大,而非对角元素值较小,可以增强时间上的平滑性.

$P(C_t^i \mid C_{t-1}^i, L_t^i, F_t)$ 表示人脸姿态类别 C_t^i 对 VFOA F_t,人脸位置 L_t^i 和上一时刻人脸姿态类别 C_{t-1}^i 的概率依赖关系.这是本模型的核心,它刻画了多个 VFOA,多个用户位置和多个摄像机拍摄的人脸之间的概率依赖关系.假设用户站在某个位置 L_t^i,并关注某个特定的目标 F_t 时,通过摄像机 i 所拍摄的人脸姿态类别 C_t^i 将是确定的.通过训练数据可以得到这个概率矩阵.

$P(Z_t^i \mid C_t^i)$ 表示人脸观测似然度.已知人脸姿态类别 C_t^i,该似然度表示人脸观测 Z_t^i 由该人脸姿态类别产生的概率,如下公式所示.

$$P(Z_t^i \mid C_t^i = k) = \frac{1}{\Lambda} \exp\left(\frac{-d^2(Z_t^i, M_k)}{\sigma^2}\right) \tag{2}$$

其中 Λ 是归一化因子,M_k 表示人脸姿态类别的图像子空间 $C_t^i = k$,而 $d(Z_t^i, M_k)$ 表示人脸图像到该图像子空间的距离,比如当图像投影到该子空间时的重构误差.假设一共有 K 个人脸姿态类别,则对某个观测而言,我们可以获得一个相似度向量 $\boldsymbol{S}$,其长度是 K,其中第 k 项为 $S_k = P(Z_t^i \mid C_t^i = k)$.

3.3 观测模型

首先用多姿态人脸检测器检测人脸进行初始化,然后结合人脸检测和在线模板更新方法进行人脸跟踪.多姿态人脸检测器采用类似文[10]的方法,并且检测范围局限在上帧人脸的附近.当无法检测到人脸时,采用在线模板更新方法[11]继续跟踪人脸.

当获得人脸之后,计算该人脸由多个人脸姿态类别产生的似然度作为相似度向量. $P(Z_t^i \mid C_t^i)$ 表示人脸观测 Z_t^i 由人脸姿态类别 C_t^i 产生的概率,见式(2).由于额外多姿态人脸数据库的扩展性问题,我们直接根据采集的数据训练人脸姿态类别.当用户站在某个固定的位置 l 关注某个特定的目标 f 时,通过摄像机 i 所拍摄的人脸姿态朝向将是确定的,不同人的人脸姿态也相似.我们用 $C(l, f, i)$ 表示这种情况下拍摄的人脸集合.假设将用户站立的位置离散化成 N 个区域,有 M 个注意力目标和 R 个摄像机,则总共将有 $M \times N \times R$ 种

人脸姿态类别.我们采用主分量分析(PCA)的方法对人脸姿态类别进行建模.

3.4 模型参数学习和设定

对 VFOA 之间的转移概率矩阵,当相邻时刻为相同目标时,我们设 $P(F_t = i \mid F_t = i) = P_f$,同时令跳转到其他目标的概率为相等的较小值,即 $P(F_t = j, j \neq i \mid F_t = i) = (1 - P_f)/(M - 1)$.通常设 P_f 的值接近 1,我们实验中取为 0.8.

$P(C_t^i \mid C_{t-1}^i, L_t^i, F_t)$通过计数法统计在某个固定的 L 和 F 下观测变量属于某个姿态类别 C 的概率.其中我们手工标注视频中用户所注视的 VFOA,而用户的位置 L 和当前属于某姿态类别 C 的概率则通过训练数据计算得到.首先我们假设 C_t 和 C_{t-1}之间互相独立,然后进行计数统计.统计完成后,通过令 $P(C_t^i = j \mid C_{t-1}^i = j, L_t^i, F_t)$增加 β 的方法增加不同姿态类别之间跳转的时间平滑性,然后再归一化.实验中取 β 为 0.2.人脸位置 L 和划分的用户区域数目对应,设置为 3.

$P(Z_t^i \mid C_t^i)$表示人脸似然度,由相似度向量 $\boldsymbol{S}$ 表示观测图像和每个人脸姿态类别的相似概率.这里我们并不确定性的决定人脸属于哪个姿态类别,而通过概率的方式进行模糊度量.因此我们需要调整式(2)中的归一化参数使得相似度向量 $\boldsymbol{S}$ 的最大值落在某个范围内.在实验中该范围定为[0.4 - 0.5].如果当前时刻没有人脸检测到,则令相似度向量为等概率,即人脸可能属于任何一个姿态类别.

4 模型推理

VFOA 的识别问题被看做是概率模型的推理问题.已知观测 Z 和 L,需要推理出隐变量 F 和 C.也就是说,目标函数是最大化以下联合概率密度分布.

$$(\hat{F}, \hat{C}) = \arg\max_{F,C} P(F, C, Z, L) \tag{3}$$

我们采用文[12]提出的对代价函数"自由能量"进行最小化的近似推理算法.自由能量使用较简单的概率密度分布 $Q(h)$来近似真实的后验概率密度分布 $P(h \mid v)$.其中,h 和 v 分别代表隐变量(F, C)和观测变量(Z, L).获得的近似 $Q(h)$然后用来计算目标函数.概率密度分布 $Q(h)$的公式如下.

$$Q(h) = \prod_{t=1}^{T} Q(F_t \mid F_{t-1}) \prod_{i=1}^{R} Q(C_t^i \mid C_{t-1}^i, F_t) \tag{4}$$

参照文献[12]的方法,自由能量可以写成:

$$E = \int_h Q(h) \ln \frac{Q(h)}{P(h, v)} \tag{5}$$

已知隐变量 F, C 在 $t-1$ 时刻的状态,通过最小化 E,可以得到:

$$\frac{\partial E_t}{\partial Q(C_t^i \mid C_{t-1}^i, F_t)} = 0 \Rightarrow Q(C_t^i \mid C_{t-1}^i, F_t)$$

$$\propto P(C_t^i \mid C_{t-1}^i, L_t^i, F_t) P(Z_t^i \mid C_t^i) \tag{6}$$

$$\frac{\partial E_t}{\partial Q(F_t \mid F_{t-1})} = 0 \Rightarrow Q(F_t \mid F_{t-1}) \propto P(F_t \mid F_{t-1}) \cdot \prod_{i=1}^{R} \prod_{C_{t-1}^i = 1}^{K} \sum_{C_t^i = 1}^{K} (P(C_t^i \mid C_{t-1}^i, L_t^i, F_t) P(Z_t^i \mid C_t^i))^{Q(C_{t-1}^i \mid F_{t-1})} \tag{7}$$

式(7)中,可以如下计算:

$$Q(C_t^i \mid F_t) = \int_{F_{t-1}, Q_{t-1}^i} Q(C_t^i \mid C_{t-1}^i, F_t) Q(C_{t-1}^i \mid F_{t-1}) Q(F_{t-1}) \tag{8}$$

然后就可以通过迭代的方法计算概率密度分布:

$$Q(F_t) = \int_{F_{t-1}} Q(F_t \mid F_{t-1}) Q(F_{t-1}) \tag{9}$$

$$Q(C_t^i) = \int_{F_t} Q(C_t^i \mid F_t) Q(F_t) \tag{10}$$

最后,选择概率最大的那个值 $\hat{F}_t$ 作为模型的最终推理结果.

$$\hat{F}_t = \arg\max_{F_t} Q(F_t) \tag{11}$$

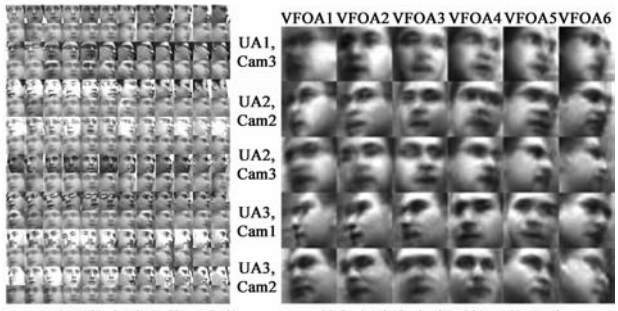
5 实验

我们共采集了 8 个用户的数据.用户分别站在 3 个不同的区域并注视 6 个不同的目标.3 个摄像机同时拍摄视频.总共得到 24 组视频,每组有来自 3 个摄像机的 3 段视频,采样率为 25fps,视频长度为 25 到 30 秒.图像大小为 360 × 288,其中人脸大小大约 40 × 40 像素左右.所有人脸图像通过直方图均衡化进行预处理以减弱光照的影响,然后人脸被缩放到 32 × 32 大小,转成向量,并令向量均值为 0,方差为 1.我们手工标注了用户注视某个 VFOA 的开始和结束时间.标注过 VFOA 真值的视频片段用来进行人脸姿态类别和模型参数的训练,以及最后的实验评估.模型推理是在整个视频上进行的.

5.1 人脸姿态类别

图 4(*a*)显示跟踪得到的人脸.每两行表示一个用户的数据,其中第一行是原始的人脸,第二行是与其最相似的人脸姿态类别的平均人脸.可以看出,大部分人脸的姿态与平均人脸的姿态是很相似的,显示姿态观测模型是合理的.另外,跟踪得到的人脸框并没有非常精确的对准,可能存在某些偏差.在我们的模型中,我们并不计算精确的姿态值,而是利用相似度向量作为模糊度量,通过这种方式推理 VFOA.

当用户站在区域 1 时,仅 3 号摄像机可以拍摄到用户人脸.当用户站在区域 2 和 3 时,分别有两个摄像机可以拍摄到用户人脸.因此,我们总共可以得到 5 × 6 = 30 个人脸姿态类别.对每个姿态类别而言,我们随机从每个用户的视频中选择 10 张人脸图像组成训练集,并通过主分量分析对训练集样本建模.图 4(*b*)显示每个姿态类别的平均人脸.



(*a*) 跟踪到的人脸和其对应的最相似的人脸姿态类别 (*b*) 所有人脸姿态类别的平均人脸

图4

5.2 注意力目标识别

我们通过两组实验来评估模型的有效性和扩展性.第一组实验中,我们使用所有 8 个人的数据进行训练和测试.第二组实验中,采用排一法交叉验证的方法,即用 7 个人的数据训练,用另外一个人的数据测试,分别将每个人的数据都做为测试集,总共运行 8 轮.

表 1 和表 2 分别是两组实验的结果,表中的百分比是将正确识别的帧数除以总帧数得到的识别正确率.正确率评估仅在有人工标注的视频片段上进行.第二组实验中,我们仅考虑测试集视频的结果,并给出 8 轮实验的平均结果.

表 1 训练集和测试集相同时的 VFOA 识别结果(%)

区域	VFOA1	VFOA2	VFOA3	VFOA4	VFOA5	VFOA6
1	100.00	99.78	100.00	97.87	98.47	100.00
2	100.00	89.00	94.33	97.25	95.89	98.32
3	100.00	92.21	86.61	93.72	97.71	98.50

表 2 排一法交叉验证时的 VFOA 识别结果(%)

区域	VFOA1	VFOA2	VFOA3	VFOA4	VFOA5	VFOA6
1	97.22	98.01	64.86	53.52	45.70	58.46
2	91.70	80.45	95.57	81.50	46.47	87.31
3	81.16	46.12	70.77	99.78	64.86	74.31

从表 1 可以看出,结果非常好.这是因为训练集和测试集的用户是相同的.第二组实验的结果不如第一组.这是可以理解的,因为测试集用户并未在训练集中出现过,而且不同用户的外貌和光照情况都有较大差别.可以看出,当用户站在区域 1 并且观看 VFOA4,5 和 6 时,精度不是很高.他们常常会被错误的识别成相邻的 VFOA.这是因为 VFOA4,5 和 6 之间的距离较近,而他们整体离区域 1 的距离较远,所以当用户注视这些不同的目标时,常常仅非常轻微的转动一下头部,甚至有时候头部都不转动,而仅转动眼球.另外一个原因是,当用户注视这些目标时,由 3 号摄像机拍摄得到的人脸姿态为接近全侧面,而全侧面附近的姿态变化从图像上来看非常微小并且难以区分.同样类似的情况出现在当用户站在区域 2 并注视 VFOA5 和 6,或者站在区域 3 并注视 VFOA1 和 2.如果排除这些情况,考虑到真实场景下拍摄的数据的困难性,实验结果还是可以接受的.

为评估使用多摄像机信息融合的必要性,我们测试仅使用单摄像机信息识别 VFOA 的结果.当用户站在区域 2 和 3 时,分别有两个摄像机可以拍摄到用户.图 5 列出了在这两个区域里使用单摄像机和双摄像机进行 VFOA 识别的结果比较.可以看出,大部分时间里,多摄像机的结果要好于单摄像机.尽管有些情况下单摄像机的结果要略好于双摄像机,但这只是在看个别 VFOA 时的结果,而看其他 VFOA 时则不是如此.表 3 列出了分别使用单摄像机和双摄像机进行 VFOA 识别的平均正确率.可以看出,多摄像机的平均正确率要比单摄像机好.这说明在模型中融合多摄像机信息是非常必要的.

表 3 用户站在区域 2 和 3 并关注不同 VFOA 的平均正确率

用户区域 2	正确率%	用户区域 3	正确率%
摄像机 2 + 3	80.50	摄像机 1 + 2	72.83
摄像机 2	68.29	摄像机 1	52.84
摄像机 3	74.91	摄像机 2	63.49

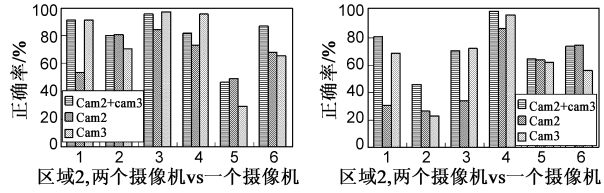


图5 单摄像机和多摄像机结果比较.左:区域2,右:区域3.X轴:VFOA下标,Y轴:正确率

为评估模型中隐状态的时间连续性对结果的影响,我们忽略注意力目标 F 和姿态类别 C 的时间连续性.令状态转移概率矩阵 $P(F_t|F_{t-1})$的值都为 $1/M$,这样当前时刻的注意力目标与前一时刻完全无关.令 $P(C_t^i|C_{t-1}^i, L_t^i = l, F_t = f)$的值都为 $1/K$,这样当前时刻的姿态类别与前一时刻完全无关.使用这些参数进行注意力目标识别的结果见表 4,识别的平均正确率比较见表 5.可以看出,考虑到时间连续性的模型参数下,结果的正确率要比不考虑时间连续性的结果好.这说明在模型中考虑时间连续性是合理的.

在前面的实验中,为了分析和验证算法的性能,我们采集了用户观看 6 个不同目标时的数据进行评估.从结果中可以发现,当用户看较远的目标时,容易发生混淆.在我们的实际场景中,厨房墙上只有 4 个显示屏.因此,我们排除的 1 号和 6 号 VFOA,只保留用户观看 4 个显示屏的数据(假设这 4 个显示屏的编号为 1 到 4),实验结果见表 6.为了对比,我们从表 2 中选择对应的注视 4 个显示屏的识别结果(即 VFOA2-5),见表 7.比较表 6 和表 7 的平均值,可以发现,如果只保留 4 个注视目

标进行识别,结果要明显好于 6 个注视目标的结果.这是因为较远的目标很容易混淆成相邻姿态,因此排除了最远的 1 号和 6 号 VFOA 后,识别的正确率就会明显提高.以上实验结果可以指导应用场景的设计.在真实的应用环境中,最好确定有限数目的注视目标,并尽可能的让目标之间的位置间隔较大.同时,用户最好站在多个目标的中间,且有多个摄像机可以采集到用户,这样才能较准确的识别用户的注视目标.

表 4　VFOA 识别结果:忽略隐状态的时间连续性

区域	VFOA1	VFOA2	VFOA3	VFOA4	VFOA5	VFOA6
1	94.21	92.05	64.67	52.86	47.04	56.02
2	89.21	73.32	92.20	77.11	47.95	87.13
3	76.29	40.71	62.02	98.48	67.15	73.82

表 5　是否考虑隐状态的时间连续性的平均正确率(%)比较

方法	区域 1	区域 2	区域 3
考虑时间连续性	69.63	80.50	72.83
不考虑时间连续性	67.81	77.82	69.75

表 6　目标为 4 个显示屏时的识别正确率(%)

区域	VFOA1	VFOA2	VFOA3	VFOA4	平均值
1	100	65.06	57.61	67.30	72.49
2	98.87	94.86	79.67	66.17	84.89
3	62.34	65.03	100	89.40	79.19

表 7　有 6 个目标时,4 个显示屏的识别正确率(%)

区域	VFOA1	VFOA2	VFOA3	VFOA4	平均值
1	98.01	64.86	53.52	45.70	65.52
2	80.45	95.57	81.50	46.47	76.00
3	46.12	70.77	99.78	64.86	70.38

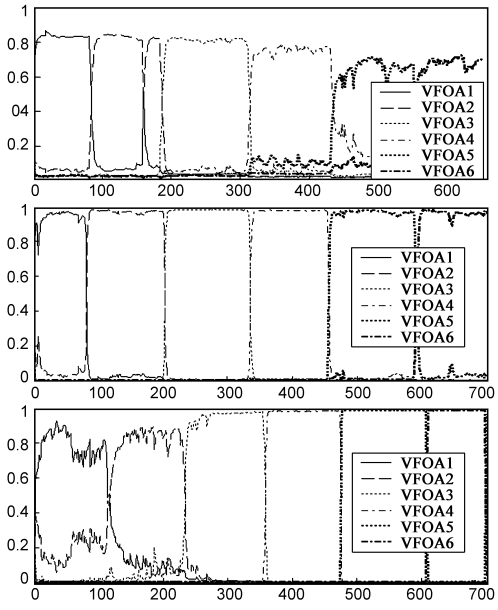


图6　概率推理过程.*X*轴是帧号,*Y*轴是概率值

图 6 显示了一些推理过程的概率分布图.从上到下的 3 张图分别代表当用户站在区域 1,2 和 3 时并关注不同 VFOA 时的推理结果.不同的线代表不同的 VFOA.在某个时刻,具有最高概率的线即表示当时最可能的 VFOA.可以看出,当用户站在区域 1 并关注 VFOA4,5 和 6 时,其最大的概率值要小于当关注 VFOA1,2 和 3 时的最大概率值.相似的情况发生在,当用户站在区域 3 时并关注 VFOA1,2 和 3 时,其最大的概率值要小于当关注 VFOA4,5 和 6 时的最大概率值.这是因为当用户关注距离其较远的不同目标时,他通常不会非常明显的转动头部,因此不同姿态的人脸差异不大,使得推理得到的最大概率值降低.

6　总结

本文提出了一种基于动态贝叶斯网模型的视觉注意力目标识别方法.通过人脸图像与多个人脸姿态类别的相似度向量对头部姿态进行度量而不是显式的计算具体姿态值.模型融合多注意力目标、多用户位置、多摄像机图像等因素间的概率依赖关系并进行联合推理.智能厨房原型环境下的实验结果表明提出的模型是有效的.目前的实验中,当用户注视较远距离的 VFOA 时,由于图像差异不大导致姿态度量不准确.未来我们将考虑调整 VFOA 的位置间隔增加姿态的区分性,并融合运动信息检测 VFOA 变化.同时,未来我们考虑将 VFOA 扩展到更多的地方,比如工作台上的不同区域等.

参考文献

[1] Roda C, Thomas J. Attention aware systems: Theories, applications, and research agenda[J]. Computers in Human Behavior, 2006, 22(4): 557 – 587.

[2] Langton S, Watt R, Bruce V. Do the eyes have it? Cues to the direction of social attention[J]. Trends in Cognitive Sciences, 2000, 4(2): 50 – 58.

[3] Stiefelhagen R. Tracking focus of attention in meetings[A]. Proceedings 4th IEEE International Conference on Multimodal Interfaces[C]. Washington, DC: IEEE Computer Society, 2002. 273 – 280.

[4] Voit M, Stiefelhagen R. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios[A]. Proceedings 10th International Conference on Multimodal Interfaces[C]. New York: ACM, 2008. 173 – 180.

[5] Ba S, Odobez J. Multi-person visual focus of attention from head pose and meeting contextual cues[J]. IEEE Trans on Patt Anal Mach Intelli, 2010, 32(3): 1 – 16.

[6] Otsuka K, Sawada H, Yamato J. Automatic inference of cross-modal nonverbal interactions in multiparty conversations[A].

Proceedings 9th International Conference on Multimodal Interfaces[C]. New York: ACM, 2007. 255 – 262.

[7] Smith K, Ba S, Gatica-Perez D, et al. Tracking the multi person wandering visual focus of attention[A]. Proceedings 8th International Conference on Multimodal Interfaces[C]. New York: ACM, 2006. 265 – 272.

[8] Zhang H, Toth L, Deng W, et al. Monitoring visual focus of attention via local discriminant projection[A]. Proceedings 1st ACM International Conference on Multimedia Information Retrieval[C]. New York: ACM, 2008. 18 – 23.

[9] Olivier P, Monk A, Xu, G, Hoey J. Ambient Kitchen: designing situated services using a high fidelity prototyping environment [A]. Proceedings 2nd International Conference on Pervasive Technologies Related to Assistive Environments [C]. New York: ACM, 2009. 47.

[10] Jones M, Viola P. Fast Multi-view Face Detection[R]. Cambridge, Massachusetts: Mitsubishi Electric Research Laboratories, Inc., 2003.

[11] Ross D, Lim J, Lin R, et al. Incremental learning for robust visual tracking[J]. Int Journal of Computer Vision, 2008, 77(1 – 3): 125 – 141.

[12] Frey B, Jojic N. A comparison of algorithms for inference and learning in probabilistic graphical models[J]. IEEE Trans on Patt Anal Mach Intelli, 2005, 27(9): 1 – 25.

作者简介

董力赓 男, 1981 年 1 月生于山西省河津市, 2010 年 1 月获得清华大学计算机系工学博士学位. 主要研究领域为计算机视觉、数字图像处理、模式识别等.

E-mail: dongligeng99@mails.thu.edu.cn

邸慧军 男, 1980 年 3 月生于内蒙古包头市, 2009 年 7 月获得清华大学计算机系工学博士学位. 现为清华大学计算机系博士后. 主要研究领域为计算机视觉、模式识别等.

E-mail: ajon@tsinghua.edu.cn