基于动态贝叶斯网络的音视频联合说话人跟踪

金乃高 [1] 殷福亮 [1] 陈 喆 [1]

摘 要 将多传感器信息融合技术用于说话人跟踪问题, 提出了一种基于动态贝叶斯网络的音视频联合说话人跟踪方法. 在动态贝叶斯网络中, 该方法分别采用麦克风阵列声源定位、人脸肤色检测以及音视频互信息最大化三种感知方式获取与说话人位置相关的量测信息; 然后采用粒子滤波对这些信息进行融合, 通过贝叶斯推理实现说话人的有效跟踪; 并运用信息熵理论对三种感知方式进行动态管理, 以提高跟踪系统的整体性能. 实验结果验证了本文方法的有效性.

关键词 说话人跟踪, 动态贝叶斯网络, 粒子滤波, 麦克风阵列
中图分类号 TP391

Audio-visual Speaker Tracking Based on Dynamic Bayesian Network

JIN Nai-Gao[1] YIN Fu-Liang[1] CHEN Zhe[1]

Abstract Multi-sensor data fusion technique is applied to speaker tracking problem, and a novel audio-visual speaker tracking approach based on dynamic Bayesian network is proposed. Based on the complementarity and redundancy between speech and image of a speaker, three kinds of perception methods, including sound source localization based on microphone array, face detection based on skin color information, and maximization mutual information based on audio-visual synchronization, are proposed to acquire the tracking information. In the framework of dynamic Bayesian network, particle filtering is used to fuse the tracking information, and perception management is achieved to improve the tracking efficiency by information entropy theory. Experiments using real-world data demonstrate that the proposed method can robustly track the speaker even in the presence of perturbing factors such as high room reverberation and video occlusions.

Key words Speaker tracking, dynamic Bayesian network, particle filter, microphone array

说话人跟踪是人机交互研究中的重要课题, 在视频会议系统、多媒体系统、机器人等领域有着广泛的应用. 例如, 在视频会议系统中, 说话人跟踪可为摄像机转向控制与语音拾取提供位置信息. 另外, 移动机器人也需要根据说话人的当前位置进行路径规划. 鉴于视频会议场景与机器人工作环境的复杂性, 如何提高说话人跟踪系统的精度与鲁棒性成为当前迫切需要解决的问题.

传统的说话人跟踪方法可分为基于计算机视觉的人脸或人体跟踪方法[1] 与基于计算机听觉的声源定位方法[2]. 这些方法仅利用单一的媒体信息, 无法获得目标的全部特征, 只有在特定的条件下才能获得较好的跟踪效果, 难以适应复杂的动态环境. 例如, 人脸跟踪方法容易受到视频遮挡以及光照、姿态变化等因素的影响, 而背景噪声与房间混响则制约着声源定位方法的性能. 众所周知, 即使在复杂的环境中人类的感知系统也能够准确地定位说话人, 这是大脑对听觉和视觉信息进行融合的结果. 因此, 说话人跟踪系统应该充分利用说话人语音信息与图像信息之间的相关性与互补性, 以进一步增强跟踪系统对复杂环境的适应能力. 文献 [3] 通过声源定位确定说话人的大致位置, 然后使用人脸跟踪的精确定位结果引导摄像机的指向. 文献 [4] 则首先检测出视角中的多个人脸区域, 将其作为说话人的备选位置, 然后使用音频信息进行声源定位, 确定说话人的实际位置. 近年来, 采用信息融合方法的音视频联合说话人跟踪技术成为研究的主要趋势, 基于神经网络[5]、粒子滤波[6]、贝叶斯网络[7] 的信息融合方法均已成功应用于音视频联合说话人跟踪问题, 且取得了较好的跟踪效果.

本文提出了一种基于动态贝叶斯网络的音视频联合说话人跟踪方法. 该方法在动态贝叶斯网络框架下, 根据说话人音频与视频信息之间的互补性与相关性, 通过说话人的语音、图像以及音视频互信息获取与说话人位置相关的跟踪信息, 通过粒子滤波对这些信息加以融合, 进而确定说话人的空间位置. 同时, 本文运用信息熵理论对三种感知方式进行动态管理, 以提高跟踪系统的实时性. 实验结果表明, 本文方法能够在复杂的环境下实现说话人的有效跟踪.

收稿日期 2007-07-09 收修改稿日期 2007-11-26
Received July 9, 2007; in revised form November 26, 2007
国家自然科学基金 (60772161, 60372082) 资助
Supported by National Natural Science Foundation of China (60772161, 60372082)
1. 大连理工大学电子与信息工程学院 大连 116023
1. School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023
DOI: 10.3724/SP.J.1004.2008.01083

1 音视频联合说话人跟踪系统基本框架

说话人的语音与图像之间具有互补性与相关性, 如图 1 所示. 两者之间的互补性体现在音频信息具有全方位特性, 但其定位精度较差; 视频信息的获取虽然受到摄像机视角的限制, 却可以提供精确的定位信息. 另外, 视频信息不受背景噪声以及房间混响等声学环境的影响, 音频信息则与视觉场景的复杂性无关. 说话人的语音与图像之间的相关性体现在说话人语音与唇动信息之间具有相关性; 两个麦克风之间的时延与图像中人脸是分别通过麦克风阵列与摄像机对说话人位置进行观测的结果, 二者之间也存在内在的相关性.

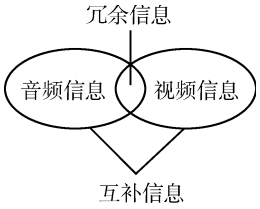


图 1 音视频信息之间的互补性与相关性

Fig. 1 The complementarity and redundancy between audio and visual information

人类的视觉感知系统与听觉感知系统既可以独立工作, 又可以通过互相协作来共同感知说话人的空间位置. 与其对应, 工作在复杂环境下的说话人跟踪系统既可以利用音频或视频信息独立地实现说话人跟踪, 同时又可以联合音频与视频信息, 采用信息融合技术求解说话人跟踪问题. 为此, 本文提出了一种基于动态贝叶斯网络的音视频联合说话人跟踪方法. 该方法利用说话人双模态信息 (语音与图像) 的互补性, 分别将基于麦克风阵列的声源定位方法与基于肤色的人脸跟踪方法作为听觉和视觉感知手段, 这样就可以提高定位精度, 增强跟踪系统对复杂环境的适应能力; 同时考虑到说话人双模态信息之间具有内在的相关性, 跟踪系统可以利用语音与图像特征之间的互信息来增强系统的可靠性.

为了提高跟踪系统对复杂环境的适应能力, 本文在跟踪过程中引入反馈机制, 在感知环节与融合环节之间通过双向信息传递来提高系统的跟踪性能, 从而使说话人跟踪系统具有完整的观测、融合、决策和协调功能, 如图 2 所示. 在图 2 中, 感知环节从说话人的语音与图像中获取说话人的位置信息, 并以似然函数的形式向高层的融合环节传递证据信息. 在融合环节中, 融合中心利用粒子滤波算法进行贝叶斯推理, 根据系统的动态特性以及感知环节提供的证据信息更新后验概率密度函数, 从而确定说话人的空间位置; 同时, 运用信息熵理论对三种感知方式进行动态管理, 并以先验知识的形式向感知环节传递规划信息. 跟踪系统根据环境的不同, 合理分配有限的计算资源, 从而有效地提高跟踪系统的性能.

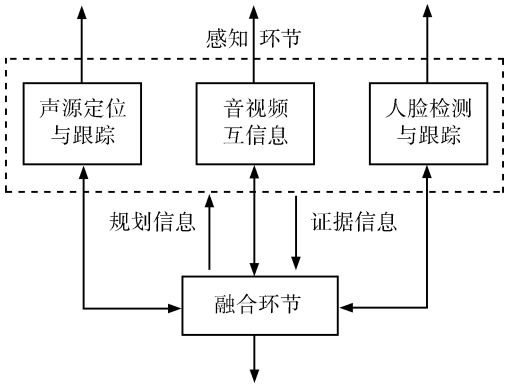


图 2 音视频联合说话人跟踪系统

Fig. 2 Audio-visual speaker tracking system

本文下面分别介绍说话人跟踪系统中三种感知方法的具体实现过程以及基于粒子滤波的信息融合方法与基于信息熵的感知方式管理方法.

2 说话人跟踪系统中感知方法的实现

2.1 基于麦克风阵列的声源定位方法

本文采用基于麦克风阵列的声源定位技术, 根据声源到麦克风阵列的时延来获取说话人的位置信息. 在声源刚刚发出信号时, 混响信号总是比直达信号延迟一段时间到达, 语音建立信号便是指这段先于混响信号到达的语音信号. 人耳优先效应 (Precedence effect) 实验表明, 在房间混响较强的情况下, 人耳可以利用未被混响信号污染的语音建立信号 (Onset signal) 准确地判断出声源方向. 本文借鉴人耳的定位机制, 从麦克风阵列获取的多路语音中提取出无混响影响的建立信号, 以增强定位系统的抗混响能力.

设直达语音信号通过窄带滤波器后输出的包络为 $s(t)$, 混响信号为 $e_{\rm echo}(t)$. 在语音包络信号的初始段, 信号的幅度明显高于混响的幅度. 若 $|s(t)|/|e_{\rm echo}(t)|$ 超过指定阈值, 便可将 $s(t)$ 视为无混响影响的建立信号[8]. 每经过 τ_{fe} 时间, 对麦克风接收到的第 l 帧语音信号 $m^l(t)$ 进行加窗短时傅里叶变换, 得到频域信号 $M^l(k)$. 设混响衰减系数为 λ, 根据混响的指数衰减特性, 第 l 帧、第 k 频带处的最大混响幅度 $M^l_{\rm echo}(k)$ 可估计为

$$M^l_{\rm echo}(k)=\max\{\lambda^n|M^{l-n}(k)|\}\quad n=1,2,\cdots,l-1,\ \ 0<\lambda<1 \tag{1}$$

$M^l_{\text{echo}}(k)$ 的递推计算公式为

$$M^l_{\text{echo}}(k)=\max\{\lambda M^{l-1}_{\text{echo}}(k),\ \ \lambda|M^{l-1}(k)|\} \quad (2)$$

将 $M^l(k)$ 的幅度与最大混响幅度 $M^l_{\text{echo}}(k)$ 进行比较, 当两者的比值大于预定阈值 T 时, 便可将 $M^l(k)$ 视为无混响影响的建立信号.

提取出语音的建立信号后, 本文采用相位变换 (Phase transform, PHAT) 方法进行时延估计[9], 然后通过几何方法确定声源的方位角与俯仰角. 麦克风阵列与摄像机的摆放位置如图 3 所示, 其中摄像机位于麦克风阵列的中心, 垂直方向上两个麦克风之间的距离为 H, 水平方向上两个麦克风之间的距离为 D, 声源的俯仰角为 ϕ, 方位角为 θ. 设声速为 c, 垂直麦克风之间的时延估计为 $\hat{\tau}_{01}$, 水平麦克风之间的时延估计为 $\hat{\tau}_{23}$, 则声源的俯仰角 ϕ 与方位角 θ 分别为

$$\phi=\arcsin\left(\frac{\hat{\tau}_{01}c}{H}\right) \quad (3)$$

$$\theta=\arcsin\left(\frac{\hat{\tau}_{23}c}{D\cos\phi}\right) \quad (4)$$

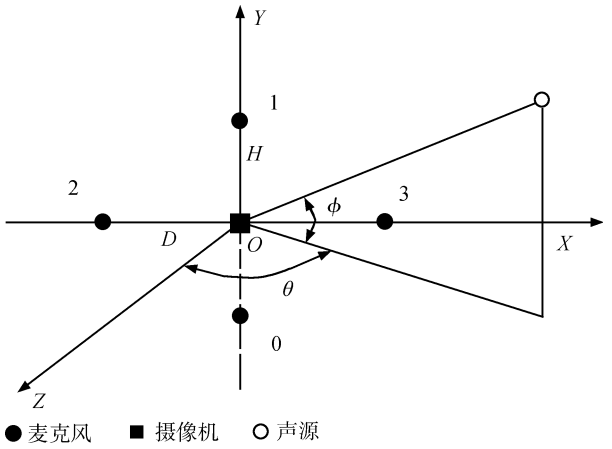


图 3　麦克风阵列与摄像机摆放示意图

Fig. 3　The placement of microphone array and camera

假设摄像机的光轴与 Z 轴重合, 焦距为 f, 摄像机成像平面大小为 $S_x\times S_y$, 图像水平方向与垂直方向上的像素点数分别为 N_x 与 N_y, 图像中心坐标为 (x_c,y_c). 为了便于音视频信息融合, 将声源定位的角度 (θ,ϕ) 映射为图像坐标点 $\boldsymbol{X}_a=(X^x_a,X^y_a)$, 即

$$X^x_a=f\frac{N_x}{S_x}\tan\theta+x_c \quad (5)$$

$$X^y_a=f\frac{N_y}{S_y}\frac{\tan\phi}{\cos\theta}+y_c \quad (6)$$

2.2　基于肤色的人脸跟踪方法

在通常的光照条件下, 人脸肤色会聚集在色彩空间中某个特定的区域内, 视频图像中的肤色区域可以通过建立合适的肤色模型进行提取. 基于肤色的人脸跟踪方法不受人脸姿态、尺寸改变和部分遮挡的影响, 因此本文将其作为视觉感知手段, 进行说话人跟踪.

假设人脸目标是以点 $\boldsymbol{X}$ 为中心, 长短轴分别为 h_y 与 h_x 的椭圆. 本文采用 HSV 与 YCrCb 混合空间肤色模型, 将 HCrCb 三个颜色分量分别量化为 N_H、N_{Cr} 与 N_{Cb} 级, 目标模板为 H 空间的 N_H 级一维直方图和 $CrCb$ 空间的 $N_{Cr}\times N_{Cb}$ 级二维直方图.

设函数 $b^n(\boldsymbol{X}^{(i)})$ 为像素点 $\boldsymbol{X}^{(i)}$ 至相应直方图中颜色索引值 u^n 的映射, 函数 $k(\cdot)$ 为高斯核函数 $K(\cdot)$ 的轮廓函数, 则考虑空间位置信息的加权直方图为

$$\begin{aligned}&p^n_u(\boldsymbol{X})=\\&C_h\sum_{i=1}^{N}k\left(\left\|h^{-1}(\boldsymbol{X}-\boldsymbol{X}^{(i)})\right\|^2\right)\delta(b^n(\boldsymbol{X}^{(i)})-u^n)\\&u^1=1,\cdots,N_{Cr}\times N_{Cb},\ u^2=1,\cdots,N_H\end{aligned} \quad (7)$$

其中 C_h 为归一化系数, $\delta(\cdot)$ 为 Kronecker delta 函数, 核函数的窗宽 $h=\text{diag}\{h_x,h_y\}$.

本文采用均值漂移 (Mean shift) 算法[10], 通过最大化候选区域直方图 p_u 与参考目标直方图 q_u 之间的 Bhattacharyya 系数来跟踪人脸的位置 $\boldsymbol{X}_v$, 即

$$\begin{aligned}\rho(p_u(\boldsymbol{X}),q_u)=&\sum_{u=1}^{N_{cr}\times N_{cb}}\sqrt{p^1_u(\boldsymbol{X})q_u}\cdot\sum_{u=1}^{N_H}\sqrt{p^2_u(\boldsymbol{X})q_u}\\\boldsymbol{X}_v=&\arg\max_{\boldsymbol{X}}\rho(p_u(\boldsymbol{X}),q_u)\end{aligned} \quad (8)$$

2.3　基于音视频互信息最大化的说话人跟踪方法

基于音视频互信息最大化的定位方法利用说话人语音与唇动可视语音之间的相关性, 通过最大化两者之间的互信息确定说话人的位置。该方法可以消除由于混响产生的虚假声源[11], 适用于视角中同时存在多个人脸时的情形. 语音的音频与视频特征之间的复杂关系需要采用统计的方法进行描述, 信息论中的互信息为定量计算两个随机变量间的关联程度提供了一种有效的工具. 设特征矢量的熵为 H, 音频特征为 a_i、视频特征为 v_i, 则音视频互信息可以描述为[12]

$$\begin{aligned}&I(\boldsymbol{A},\boldsymbol{V})=H(\boldsymbol{A})+H(\boldsymbol{V})-H(\boldsymbol{A},\boldsymbol{V})=\\&\sum_i p(a_i)\log p(a_i)-\sum_j p(v_j)\log p(v_j)+\\&\sum_{i,j}p(a_i,v_j)\log p(a_i,v_j)\end{aligned} \quad (9)$$

本文分别采用语音能量与唇部椭圆内像素点的个数作为音频与视频特征进行互信息计算. 对于视频特征的提取, 本文首先使用帧间差分法获得感兴趣区域, 然后根据唇色信息精确估计嘴唇位置.

假设特征矢量服从高斯分布, 则其熵值 H 取决于随机变量的方差. 设服从高斯分布的音频特征、视频特征与音视频联合特征矢量的协方差矩阵分别为 Σ_A、Σ_V 与 Σ_{AV}, 则音视频互信息可以描述为

$$I(\boldsymbol{A},\boldsymbol{V}) = \frac{1}{2}\log\frac{|\det(\Sigma_A)|\cdot|\det(\Sigma_V)|}{|\det(\Sigma_{AV})|} \tag{10}$$

当音频与视频特征空间的维数都是一维时, 音视频互信息可以通过音频特征与视频特征之间的 Pearson 相关系数 ρ_k 进行计算[12], 即

$$I(\boldsymbol{A},\boldsymbol{V}) = -\frac{1}{2}\log(1-\rho_k) \tag{11}$$

为了提高互信息计算的鲁棒性, 本文对互信息进行多帧平滑处理. 设描述唇型的椭圆模板的中心坐标为 (x,y), 平滑后的音视频互信息为 $I(x,y)$, 将最大音视频互信息所对应的椭圆模板中心点作为说话人的位置估计 $\boldsymbol{X}_m$, 即

$$\boldsymbol{X}_m = \arg\max_{(x,y)} I(x,y) \tag{12}$$

3 说话人跟踪系统中融合方法的实现

动态贝叶斯网络作为贝叶斯网络随时间变化的动态扩展, 适合于对动态不确定性问题进行建模. 动态贝叶斯网络可以描述具有多个通道的复杂随机过程, 它为基于多传感器信息融合的目标跟踪问题提供了一种可行的解决方案[13]. 本文采用动态贝叶斯网络来融合说话人的语音与图像信息, 通过粒子滤波进行动态贝叶斯网络推理, 进而求解复杂环境下的说话人跟踪问题; 同时, 应用信息熵原理对第 2 节给出的三种感知方式进行动态管理, 根据环境的不同动态地改变感知手段, 从而提高说话人跟踪系统的整体性能.

3.1 基于动态贝叶斯网络的信息融合方法

说话人跟踪系统的动态贝叶斯网络描述如图 4 所示. 在动态贝叶斯网络中, 对说话人的音频与视频信息进行处理, 将得到的人脸图像 $\boldsymbol{y}_t^V$、麦克风之间的时延估计 $\boldsymbol{y}_t^A$ 以及音视频互信息 $\boldsymbol{y}_t^I$ 作为观测变量 $\boldsymbol{Y}_t = (\boldsymbol{y}_t^A, \boldsymbol{y}_t^V, \boldsymbol{y}_t^I)$, 将说话人的位置 $\boldsymbol{X}_t$ 作为状态变量. 说话人跟踪系统将观测变量 $\boldsymbol{Y}_t$ 作为证据信息, 通过贝叶斯推理完成证据在网络中的传播, 以确定说话人的位置 $\boldsymbol{X}_t$.

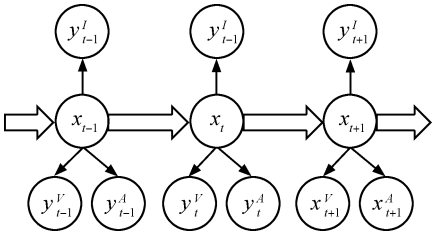


图 4 基于动态贝叶斯网络的说话人跟踪

Fig. 4 Audio-visual speaker tracking based on dynamic Bayesian network

近年来, 粒子滤波已经成为解决非线性、非高斯动态系统最优估计问题的有效方法. 粒子滤波利用一组随机粒子及其对应的重要性权值来描述贝叶斯网络中变量的概率密度函数, 通过证据在贝叶斯网络中的传播来计算后验概率密度或滤波概率密度的近似值. 设从重要性采样函数 $\pi(\boldsymbol{X}_t|\boldsymbol{Y}_{1:t})$ 中抽样获取的粒子集为 $\{\boldsymbol{X}_t^{(i)}, w_t^{(i)}, i=1,\cdots,N\}$, 其中 $w_t^{(i)}$ 为第 i 个粒子 $\boldsymbol{X}_t^{(i)}$ 的权值, 则滤波概率概率密度 $p(\boldsymbol{X}_t|\boldsymbol{Y}_{1:t})$ 可以表示为

$$p(\boldsymbol{X}_t|\boldsymbol{Y}_{1:t}) = \sum_{i=1}^{N} w_t^{(i)}\delta(\boldsymbol{X}_t - \boldsymbol{X}_t^{(i)}) \tag{13}$$

其中

$$w_t^{(i)} \propto \frac{p(\boldsymbol{X}_t^{(i)}|\boldsymbol{Y}_{1:t})}{\pi(\boldsymbol{X}_t^{(i)}|\boldsymbol{Y}_{1:t})} \tag{14}$$

下面给出采用粒子滤波实现贝叶斯推理的具体过程. 本文使用 Langevin 过程建立说话人的运动模型, 结合状态转移概率密度函数 $p(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})$ 与第 2 节中基于自底而上数据驱动的跟踪结果, 共同构建重要性概率密度函数, 进而生成采样粒子.

给定图像坐标 $\boldsymbol{X}_t^{(i)}$, 该点所对应的时间延迟理论值为 τ_{01} 与 τ_{23}. 声源定位方法的似然函数 $p(\boldsymbol{y}_t^A|\boldsymbol{X}_t^{(i)})$ 为

$$D_\tau = \sqrt{(\tau_{01}-\hat{\tau}_{01})^2 + (\tau_{23}-\hat{\tau}_{23})^2}$$

$$p(\boldsymbol{y}_t^A|\boldsymbol{X}_t^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma_a}\exp\left(-\frac{D_\tau^2}{2\sigma_a^2}\right) \tag{15}$$

给定候选目标区域中心的图像坐标 $\boldsymbol{X}_t^{(i)}$, 由 Bhattacharyya 系数确定的 Bhattacharyya 距离为 $D_b(\boldsymbol{X}_t^{(i)}) = \sqrt{1-\rho(p_u(\boldsymbol{X}_t^{(i)}), q_u)}$, 则人脸跟踪方法的似然函数 $p(\boldsymbol{y}_t^V|\boldsymbol{X}_t^{(i)})$ 为

$$p(\boldsymbol{y}_t^V|\boldsymbol{X}_t^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\frac{D_b^2(\boldsymbol{X}_t^{(i)})}{2\sigma_v^2}\right) \tag{16}$$

给定图像坐标 $\boldsymbol{X}_t^{(i)}$, 最大互信息定位方法的似

然函数 $p\left(\boldsymbol{y}_t^I|\boldsymbol{X}_t^{(i)}\right)$ 可以通过计算该点所对应的音视频互信息得到.

在图 4 中, t 时刻的说话人位置 $\boldsymbol{X}_t$ 的后验滤波概率密度可以通过 $t-1$ 时刻的跟踪结果以及 t 时刻的似然函数来确定, 即

$$p(\boldsymbol{X}_t|\boldsymbol{Y}_t) \propto \sum\nolimits_{\boldsymbol{X}_{t-1}} p(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1}) \times \prod_{j=A,V,I} p(\boldsymbol{y}_t^j|\boldsymbol{X}_t) \tag{17}$$

粒子权值 $w_t^{(i)}$ 的递推形式为

$$w_t^{(i)} \propto w_{t-1}^{(i)} \prod_{j=A,V,I} p(\boldsymbol{y}_t^j|\boldsymbol{X}_t) \tag{18}$$

故, 基于最小均方误差 (Minimum mean square error, MMSE) 准则的跟踪结果为 $\hat{\boldsymbol{X}}_t = \sum_{i=1}^N \boldsymbol{X}_t^{(i)} w_t^{(i)}$.

3.2　基于信息熵的感知方式管理

在说话人跟踪系统中, 为了充分利用有限的计算资源, 本文针对不同的场景选择合理的感知方式, 以增强说话人跟踪系统的整体性能.

多传感器信息融合的目的之一是减少目标状态的不确定性. 本文根据信息熵理论, 采用说话人空间位置的后验条件熵 $H(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ 来表示目标位置 $\boldsymbol{X}_t$ 的平均不确定度. 设目标位置 $\boldsymbol{X}_t$ 与量测 $\boldsymbol{Y}_t$ 之间的互信息量为 $I(\boldsymbol{X}_t;\boldsymbol{Y}_t)$, 量测获取的信息增量 ΔI_t 可用互信息之差表示. 例如获取视频量测 $\boldsymbol{y}_t^V$ 后, 信息增量 ΔI_t^V 为

$$\Delta I_t^V = I(\boldsymbol{X}_t;\boldsymbol{y}_t^A,\boldsymbol{y}_t^V) - I(\boldsymbol{X}_t;\boldsymbol{y}_t^A) = H(\boldsymbol{X}_t|\boldsymbol{y}_t^A) - H(\boldsymbol{X}_t|\boldsymbol{y}_t^A,\boldsymbol{y}_t^V) \tag{19}$$

下面是信息融合中关于条件熵的两个定理[14].

定理 1. 设状态变量 $\boldsymbol{X}_t$ 与观测变量 $\boldsymbol{Y}_t$ 之间不独立, 则融合系统的条件熵满足

$$H(\boldsymbol{X}_t|\boldsymbol{y}_t^j) \geq H(\boldsymbol{X}_t|\boldsymbol{Y}_t), \quad j = A,V,I \tag{20}$$

定理 2. 当量测的各分量之间相互独立时, 融合系统输出的不确定性最小.

上述两个定理表明, 在跟踪系统中, 多个传感器的融合输出可以减少位置估计的不确定性. 另外, 为了获取精确的位置估计, 应充分利用目标的不同特征, 尽量采用不同类型的传感器, 以减小量测的相关性. 音视频联合说话人跟踪系统便是充分利用说话人语音与图像之间的互补性与相关性来减少说话人位置估计的不确定性.

在说话人跟踪系统的中, 说话人动态模型描述的状态转移过程使得目标位置的不确定性增大, 跟踪的目的就是通过量测获取信息增量, 以减小状态估计的不确定性[15]. 状态转移产生的不确定性增量 ΔI_t^d 为

$$\Delta I_t^d = H(\boldsymbol{X}_{t+1}|\boldsymbol{Y}_t) - H(\boldsymbol{X}_t|\boldsymbol{Y}_t) \tag{21}$$

$t+1$ 时刻量测所获得的信息增量 ΔI_t^m 为

$$\Delta I_t^m = H(\boldsymbol{X}_{t+1}|\boldsymbol{Y}_t) - H(\boldsymbol{X}_{t+1}|\boldsymbol{Y}_{t+1}) \tag{22}$$

说话人的语音、图像与音视频互信息所提供的证据信息对跟踪的贡献可用后验滤波概率密度函数 $p\left(\boldsymbol{X}_t|\boldsymbol{Y}_t\right)$ 与似然函数 $p\left(\boldsymbol{y}_t|\boldsymbol{X}_t\right)$ 之间的 Kullback-Leibler 距离来评价. Kullback-Leibler 距离的定义为

$$KL_t = \sum_{\boldsymbol{X}_t} p(\boldsymbol{X}_t|\boldsymbol{Y}_t) \ln \frac{p(\boldsymbol{X}_t|\boldsymbol{Y}_t)}{p(\boldsymbol{y}_t|\boldsymbol{X}_t)} \tag{23}$$

KL_t 的值表明了后验滤波概率密度与似然函数之间的逼近程度. KL_t 越小, 表明该证据信息对减少位置估计不确定性的贡献越大. 基于信息熵的感知方式管理就是合理选择感知方式, 以有效利用有限的计算资源. 当 ΔI_t^d 较大时, 表明说话人位置的不确定性较强, 此时需要激活所有的感知方式来获取足够证据信息. 若增加一种感知方式后 ΔI_t^m 的变化不明显, 则表明该感知方式对减少位置的不确定性贡献较小, 此时可以停止这种感知方式的运行, 这样便可以有效地分配有限的计算资源, 在保证系统跟踪性能的同时, 提高跟踪系统的实时性.

4　实验结果与分析

本文建立的说话人跟踪系统由 PC 机、音频信号采集板、麦克风阵列以及摄像头组成. 麦克风阵列包括 4 个全指向麦克风, 水平麦克风之间的距离为 0.30 m, 垂直麦克风之间的距离为 0.24 m. 实验房间大小为 7.4 m × 4.0 m × 3.3 m. 摄像头采集图像的帧率为 20 帧/秒; 语音采样率为 44.1 kHz, 2 205 点 (50 ms) 组成一帧, 窗函数为汉明 (Hamming) 窗. 在语音建立信号的提取过程中, 选取 $\tau_{fe} = 0.008$ s, 混响衰减系数 λ 为 0.9, 阈值 $T = 1.7$. 摄像机的焦距为 $f = 6$ mm, 图像尺寸 $N_x = 320$ 与 $N_y = 240$, 图像中心的像素坐标为 $(x_c, y_c) = (160, 120)$, 摄像机成像平面 $S_x = 6.4$ mm, $S_y = 4.8$ mm. 肤色跟踪中采用的椭圆模板的长宽轴分别为 $h_x = 24$ 与 $h_y = 16$, 在 HCrCb 混合空间肤色模型中, 三个颜色分量量化等级为 16. 粒子数为 200, 跟踪系统每 0.5 s 给出一次跟踪结果.

在较理想环境下, 麦克风阵列声源定位结果 (转换至图像坐标系中)、人脸肤色检测结果以及最大化音视频互信息得到的说话人位置估计结果如图 5 (见下页) 所示. 从实验结果可以看到, 三种感知方法均可为说话人跟踪提供有效信息.

为了验证本文方法的有效性, 本文首先对声源定位方法、人脸跟踪方法以及本文提出的双模态说话人跟踪方法的跟踪性能加以比较. 三种跟踪方法在水平方向上的跟踪误差如表 1 所示. 在实验的 0～2 s 内, 由于存在其他说话人干扰且背景噪声较强, 导致声学环境比较恶劣, 此时声源定位方法的性能较差, 有时甚至给出错误的位置信息; 在 2～4 s 内, 由于存在视频遮挡与多个人脸干扰, 基于肤色的人脸跟踪方法无法判定人脸与说话人之间的对应关系, 导致跟踪失效. 由此可见, 仅利用语音或图像信息的单模态说话人跟踪系统缺乏足够的鲁棒性, 难以适应复杂多变的外界环境. 由于麦克风阵列声源定位方法不受光照变化与遮挡的影响, 当光照变化、视频遮挡以及出现其他人脸干扰时, 本文方法通过利用说话人的音频信息仍然能够继续有效跟踪说话人; 由于人脸跟踪方法与房间噪声以及混响等声学环境无关, 本文方法在恶劣的声学环境下仍然能够准确地跟踪说话人.

下面将本文方法与文献 [6] 提出的音视频联合说话人跟踪方法进行比较. 在复杂场景下, 当说话人的人脸与其他人脸干扰比较接近以及房间混响较强时, 文献 [6] 方法有时会得到错误的跟踪结果; 本文方法利用不受混响影响的语音建立信号进行声源定位, 并通过语音信号与说话人唇动信息之间的相关性来排除其他人脸干扰的影响, 仍然能够继续跟踪说话人, 如图 6 所示.

由以上实验结果可以看出, 本文方法正是充分利用说话人的音频与视频信息之间的互补性与相关性, 有效地提高了复杂环境下说话人跟踪系统的性能.

5 结论

本文提出了一种适用于复杂环境的音视频联合

(a) 声源定位 (b) 肤色检测 (c) 最大化互信息

(a) Sound source localization (b) Skin color detection (c) Maximization mutual information

图 5 三种感知方法的跟踪结果

Fig. 5 The tracking results of three kinds of perception methods

表 1 三种方法在水平方向上的跟踪误差比较

Table 1 Comparison of tracking errors of three kinds of methods in horizontal direction

时间 (s)	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
音频跟踪的绝对误差 (度)	8.5	14.3	10.8	16.5	1.7	1.5	1.4	1.6	1.5
视频跟踪的绝对误差 (度)	0.5	0.6	0.4	0.7	0.5	2.8	4.2	6.8	7.4
本文方法的绝对误差 (度)	0.6	0.4	0.5	0.5	0.6	1.4	1.2	1.4	1.6

图 6 复杂环境下本文方法的跟踪结果

Fig. 6 The tracking results of the proposed method in complex environments

说话人跟踪方法. 该方法综合利用说话人音视频信息之间的互补性与相关性, 采用动态贝叶斯网络融合说话人的语音和图像信息, 通过粒子滤波进行贝叶斯推理, 实现说话人的有效跟踪, 提高了说话人跟踪系统的精度与鲁棒性. 另外, 本文采用信息熵方法对感知方式进行管理, 减少了计算复杂度, 提高了跟踪系统的实时性. 今后的研究工作将充分利用视频图像中的肤色、运动、轮廓信息以及多路说话人语音所提供的定位信息, 来进一步提高说话人跟踪系统的性能.

References

1 Cheng C, Ansari R. Kernel particle filter for visual tracking. *IEEE Signal Processing Letters*, 2005, **12**(3): 242−245

2 Smaragdis P, Boufounos P. Position and trajectory learning for microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(1): 358−368

3 Wang C, Griebel S, Brandstein M, Hsu B. Real-time automated video and audio capture with multiple cameras and microphones. *Journal of VL SI Signal Processing Systems*, 2001, **29**(1-2): 81−99

4 Wilson K, Rangarajan V, Checka N, Darrell T. Audiovisual arrays for untethered spoken interfaces. In: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces. Pittsburg, USA: IEEE, 2002. 389−394

5 Wrigley S N, Brown G J. Physiologically motivated audiovisual localization and tracking. In: Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal: Interspeech, 2005. 773−776

6 Vermaak J, Gangnet M, Blake A, Perez P. Sequential Monte Carlo fusion of sound and vision for speaker tracking. In: Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver, Canada: IEEE, 2001. 741−746

7 Lo D, Goubran R A, Dansereau R M. Robust joint audio-video talker localization in video conferencing using reliability information-II: Bayesian network fusion. *IEEE Transactions on Instrumentation and Measurement*, 2005, **54**(4): 1541−1547

8 Huang J, Ohnishi N, Sugie N. Sound localization in reverberant environment based on the model of the precedence effect. *IEEE Transactions on Instrumentation and Measurement*, 1997, **46**(4): 842−846

9 Chen J D, Benesty J, Huang Y T. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on Applied Signal Processing*, 2006, **2006**(12): 1−19

10 Fashing M, Tomasi C. Mean shift is a bound optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(3): 471−474

11 Hershey J, Movellan J. *Audio-Vision: Using Audio-Visual Synchrony to Locate Sounds*. Cambridge, MA, USA: MIT Press, 2000. 813−819

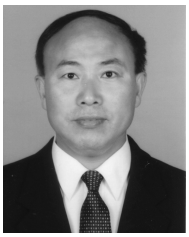
12 Cover T M, Thomas J A. *Elements of Information Theory*. New York, USA: Wiley, 1991

13 Singhal A, Brown C. Dynamic Bayes net approach to multimodal sensor fusion. In: Proceedings of SPIE Conference on Sensor Fusion and Decentralized Control in Autonomous Robotic Systems. Pittsburgh, PA, USA: SPIE, 1997. 2−10

14 Sun Ji-Xiang, Shi Hui-Min, Wang Hong-Qiang. The theory relative to entropy in information fusion. *Chinese Journal of Computers*, 2003, **26**(7): 796−801
(孙即祥, 史慧敏, 王宏强. 信息融合中的有关熵理论. 计算机学报, 2003, **26**(7): 796−801)

15 Liu Xian-Xing, Shen Shi-Lei, Pan Quan, Zhang Hong-Cai. An algorithm of sensor management based on information entropy. *Acta Electronica Sinica*, 2000, **28**(9): 39−41
(刘先省, 申石磊, 潘泉, 张洪才. 基于信息熵的一种传感器管理算法. 电子学报, 2000, **28**(9): 39−41)

金乃高 大连理工大学电子与信息工程学院博士. 主要研究方向为信息融合, 语音信号处理. 本文通信作者.
E-mail: naigao.jin@gmail.com
(**JIN Nai-Gao** Ph. D. at the School of Electronic and Information Engineering, Dalian University of Technology. His research interest covers data fusion and speech processing. Corresponding author of this paper.)



殷福亮 大连理工大学电子与信息工程学院教授. 主要研究方向为语音信号处理, 阵列处理与宽带无线通信技术.
E-mail: flyin@dlut.edu.cn
(**YIN Fu-Liang** Professor at the School of Electronic and Information Engineering, Dalian University of Technology. His research interest covers speech signal processing, array signal processing, and broadband wireless communication.)

陈 喆 大连理工大学电子与信息工程学院副教授. 主要研究方向为语音信号处理, 阵列处理与宽带无线通信技术.
E-mail: eeyin@dlut.edu.cn
(**CHEN Zhe** Associate professor at the School of Electronic and Information Engineering, Dalian University of Technology. His research interest covers speech signal processing, array signal processing, and broadband wireless communication.)