# 音视频信息融合的说话人跟踪算法研究

曹 洁,郑景润
CAO Jie, ZHENG Jingrun
兰州理工大学 电气工程与信息工程学院,兰州 730050
College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China

**CAO Jie, ZHENG Jingrun. Speaker tracking based on audio-video information fusion. Computer Engineering and Applications, 2012, 48(13):118-124.**

**Abstract:** In order to solve the defects of tracking using only audio and video information, a novel speaker tracking algorithm based on audio-video information fusion using importance particle filter is proposed. The proposed algorithm performs in a closed-loop tracking system where five modules that are bottom tracking, fusion center, importance particle filtering, tracking results output and results feedback work together to make the system best. At the bottom tracking module, based on the complementarity between speech and image of a speaker, both mean shift tracking based on face color information and sound source localization using time delay of arrival from microphone array are adopted to acquire tracking information, and they are integrated in the fusion center to obtain audio-video fused importance function and fused likelihood model. Then the fused data are processed by importance particle filter to output the tracking results, and the results are returned dynamically to the skin color tracking module and sound source localization module. Such a closed-loop system ensures the proposed algorithm performs in real-time. Experiments using AMI Meeting Corpus data demonstrate that the proposed approach is more better than those trackers utilizing only audio or video information at robustness and accuracy, and reaches an average tracking error of 9.32%.

**Key words:** object tracking; sound source localization; skin color tracking; mean shift; importance particle filter

**摘 要:** 针对单独的音频和视频信息跟踪的缺陷,提出了一种音视频信息融合的粒子滤波跟踪算法。采用闭环跟踪框架,分为底层跟踪、融合、重要性粒子滤波、跟踪输出和反馈五个环节。底层跟踪环节利用说话人脸部肤色信息进行均值漂移跟踪的同时,利用说话人声音信号到达麦克风阵列的时间延迟进行跟踪定位;融合环节对这两者得到的跟踪信息进行整合,得出基于音视频信息融合的重要性函数和融合似然模型;滤波环节利用重要性粒子滤波算法对融合的数据进行滤波处理;跟踪环节根据滤波结果对说话人进行跟踪;反馈环节将跟踪结果动态反馈给人脸肤色跟踪和声源定位跟踪模块。流程化的闭环处理过程保证了算法的实时性。最后,采用AMI会议语料库对该算法进行测试,结果表明该算法平均误跟率仅为9.32%,比使用单一音频或视频信息的跟踪算法稳定性好、准确性高。

**关键词:** 对象跟踪;声源定位;肤色跟踪;均值漂移;重要性粒子滤波
**文章编号:** 1002-8331(2012)13-0118-07 **文献标识码:** A **中图分类号:** TP391

## 1 引言

说话人跟踪是当前多传感器信息融合领域的重要研究课题,广泛应用在视频会议、智能环境、人机交互、机器人、安防监控等需要对对象进行实时跟踪

**基金项目:** 甘肃省自然科学基金(No.1010RJZA046);甘肃省基本科研业务费项目(No.0914ZTB148)。
**作者简介:** 曹洁(1966—),女,教授,博导,主要研究领域为智能交通系统、多传感器信息融合等;郑景润(1984—),男,硕士研究生,主要研究方向为信息融合理论及应用、目标跟踪等。E-mail:zhengjingrun@163.com
**收稿日期:** 2011-03-03 **修回日期:** 2011-04-30 **CNKI出版日期:** 2011-08-04
DOI:10.3778/j.issn.1002-8331.2012.13.026 http://www.cnki.net/kcms/detail/11.2127.TP.20110804.1611.151.html

的领域。基于麦克风阵列的声源定位方法[1]与基于计算机视觉的人脸或人体跟踪方法[2]分别利用说话人的音频信息与视频信息估计说话人的空间位置，已成为解决说话人跟踪问题的基本手段。但这些方法仅利用了单一的音频信息或视频信息，无法有效识别和排除跟踪过程中的干扰。例如，摄像机采集的图像容易受到视频遮挡、姿态变化和人物交错等因素的影响，而在强背景噪音或房间混响的情况下，利用声源定位进行说话人跟踪的准确性也大为下降。

人类大脑能对所获取的音频信息和视频信息进行融合处理，帮助人们在非常复杂的环境中也能够准确跟踪和识别说话人。借鉴这种融合机理，研究人员采用音视频信息融合的方法来跟踪说话人。文献[3]将音频信息映射到像平面上，然后运用粒子滤波方法将说话人形状信息和音频信息进行融合，进而跟踪说话人。文献[4]采用SMC（Sequential Monte Carlo）方法来融合头部轮廓线和声音时延两方面的信息，得到说话人的位置。文献[5]采用隐变量概率图来分别描述音频信息和视频信息，通过把音频信号之间的时间延迟对应到目标图像的空间定位而实现两者的融合，最后采用EM（Expectation-Maximization）算法得到人物的位置估计。如今，国际上一些学者开始尝试将音频视频信息融合的方法应用于更复杂的场景[6-8]。

国内此方面的研究尚处于起步阶段，文献[9]以Importance Condensation算法融合声源信息和人物图像信息，在智能教室环境中可以跟踪一个前台讲演者。文献[10]利用摄像机、麦克风等提取说话人的音频与视频信息，通过动态贝叶斯网络进行融合，实现了复杂背景下的说话人定位与跟踪。

本文结合同源目标音频信息和视频信息互补的特性，在重要性粒子滤波框架下，提出了一种基于音频信息和视频信息融合的说话人跟踪算法。实验结果表明，该算法可以较好地跟踪小型会议场景中的主要说话人，较单一的颜色跟踪或声源定位算法在可靠性、准确性方面有明显提高，并能满足实时跟踪的要求。

## 2　重要性粒子滤波IPF

粒子滤波（Particle Filter，PF）利用随即样本的加权和表示所需的后验概率密度，得到状态估计值。由于非参数化的特点，它摆脱了解决非线性滤波问题时随机量必须满足高斯分布的制约，已成为一种解决非线性、非高斯动态系统最优估计的有效方法[11-12]，能很好地处理说话人跟踪问题。在实际计算中，后验概率分布被用来产生粒子样本，但通常真正的后验概率分布的解析式很难得到，因而从后验分布中抽取样本就显得更加困难。IPF（Importance Particle Filter）是PF的一种改进算法，它选取一个已知的、容易采样的重要性函数 $q(X_t|Z_t)$，从中抽取粒子样本，以粒子加权和的形式逼近后验概率密度 $p(X_t|Z_t)$，并能有效克服粒子匮乏问题。

**定义**1　在时刻 $t=0,1,2,\cdots$，以 $X_{0:t}$ 代表目标状态（说话人位置），$Z_{0:t}$ 代表观察值（音频信号或视频信号），$t=0$ 代表初始时刻，初始先验概率密度为 $p(X_0)$。

重要性粒子滤波跟踪算法分以下三个步骤：

（1）采样

从构建的重要性函数 $q(X_t|Z_t)$ 中抽取一组粒子集，用 $\{X_t^{(i)}, w_t^{(i)}, i=1,2,\cdots,N\}$ 表示，其中 $X_t^{(i)}$ 为 $t$ 时刻第 $i$ 个粒子的状态，其相应的权值为 $w_t^{(i)}$，则后验概率密度可以用式（1）表示：

$$p(X_t|Z_t)=\sum_{i=1}^{N} w_t^{(i)}\delta(X_t-X_t^{(i)}) \tag{1}$$

$$w_t^{(i)}\propto\frac{p(Z_{1:t}|X_{0:t}^{(i)})p(X_{0:t}^{(i)})}{q(X_{0:t}^{(i)}|Z_{1:t})} \tag{2}$$

式（1）中，$\delta(\cdot)$ 为单位冲击函数（狄拉克函数），即 $\delta(x-x_t)=0(x\neq x_t)$，且 $\int\delta(x)\mathrm{d}x=1$。当采样粒子的数目很大时，式（1）便可近似逼近真实的后验概率密度函数。权值 $w_t^{(i)}$ 由式（2）计算得出。

任意函数 $f(X_t)$ 的期望估计用式（3）表示：

$$E(f(X_t)|Z_t)=\frac{1}{N}\sum_{i=1}^{N} f(X_t^{(i)})\frac{p(X_t^{(i)}|Z_t)}{q(X_t^{(i)}|Z_t)}=\frac{1}{N}\sum_{i=1}^{N} f(X_t^{(i)})w_t^{(i)} \tag{3}$$

（2）权值更新

假定系统服从Markov过程，且对于给定状态，观察值是独立的，则有：

$$q(X_{0:t}|Z_{1:t})=q(X_{0:t-1}|Z_{1:t-1})\cdot q(X_t|X_{0:t-1},Z_{1:t}) \tag{4}$$

$$p(X_{0:t})=p(X_0)\prod_{k=1}^{t} p(X_k|X_{k-1}) \tag{5}$$

$$p(Z_{1:t}|X_{1:t})=\prod_{k=1}^{t} p(Z_k|X_k) \tag{6}$$

将式（4）、（5）、（6）带入式（2）中，得权值的递归估计为：

$$w_t^{(i)}\propto w_{t-1}^{(i)}\frac{p(Z_t|X_t^{(i)})p(X_t^{(i)}|X_{t-1}^{(i)})}{q(X_t^{(i)}|X_{0:t-1}^{(i)},Z_{1:t})},\ \sum_{i=1}^{N} w_t^{(i)}=1 \tag{7}$$

根据马尔科夫性,即当前状态只与前一时刻状态有关,则有:

$$q\left(X_t^{(i)}|X_{0:t-1}^{(i)},Z_{1:t}\right)=q\left(X_t^{(i)}|X_{t-1}^{(i)},Z_t\right) \tag{8}$$

则式(7)中粒子权值可简化为:

$$w_t^{(i)}\propto w_{t-1}^{(i)}\frac{p(Z_t|X_t^{(i)})p(X_t^{(i)}|X_{t-1}^{(i)})}{q(X_t^{(i)}|X_{t-1}^{(i)},Z_t)},\sum_{i=1}^{N}w_t^{(i)}=1 \tag{9}$$

(3)跟踪输出

以加权后的粒子作为最终的跟踪结果。利用式(3),以 $f(\boldsymbol{X}_t)=\boldsymbol{X}_t$ 可计算得出 $\boldsymbol{X}_t$ 的条件均值;以 $f(\boldsymbol{X}_t)=\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}$ 可计算得出 $\boldsymbol{X}_t$ 的条件协方差。

## 3　音频视频融合说话人跟踪框架

说话人的图像和语音之间具有互补性,如音频信息定位精度低但具有全方位性,视频信息定位精度高但受到摄像机视角的限制;音频信息连续性好且不受视觉场景复杂度的影响,视频信息则不受房间混响和背景噪声等的影响。因此,在小型会议环境下,为提高跟踪的有效性和准确性,可以将音频信息和视频信息融合起来实现说话人的跟踪。

如图1所示,算法采用闭环结构,引入反馈机制以提高系统对复杂环境的适应能力。音频跟踪模块和视频跟踪模块为底层模块,分别进行声源定位和人脸肤色跟踪;融合中心将这些信息进行融合,IPF模块进行重要性粒子滤波;跟踪模块给出融合后的跟踪结果,并动态地反馈给底层跟踪模块,从而提高系统的整体性能。

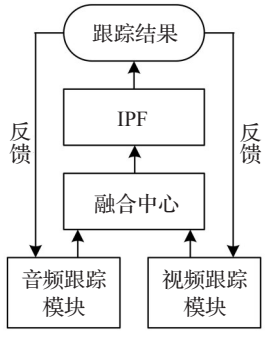


**图1　音视频信息融合的跟踪算法框图**

## 4　跟踪算法

### 4.1　基于肤色的人脸跟踪

在一般条件下,人脸肤色会聚集在色彩空间某个特定的区域内,它不易受人脸姿态、尺寸改变及部分遮挡的影响。因此通过对视频图像中的肤色区域进行处理,能有效地对说话人进行跟踪。

均值漂移(Mean-Shift)算法是一种非参数密度估计算法和高效的模式匹配算法,已成功应用于目标跟踪领域[13-15]。设视频帧图像中人脸目标是以点 $X_t^v$ 为中心,高和宽分别为 $h_y$ 和 $h_x$ 的矩形。

以 $\{X_i^*\}_{i=1}^n$ 表示参考目标区域像素的归一化坐标,以坐标原点为中心。设函数 $b(X_i^*)$ 为 $\{X_i^*\}_{i=1}^n$ 至相应直方图中颜色索引值 $u\,(u=1,2,\cdots,m)$ 的映射,即 $b:\rightarrow\{1,2,\cdots,m\}$。函数 $k(x)$ 为核函数,它对距离中心较远的像素点赋予较小的权重。则参考目标模型 $h_{\mathrm{obj}}$ 的加权直方图 $h_u^{\mathrm{obj}}$ 用式(10)表示:

$$h_u^{\mathrm{obj}}=C\sum_{i=1}^{n}k\left(\|X_i^*\|\right)^2\delta\left[b\left(X_i^*\right)-u\right] \tag{10}$$

其中,$C$ 为归一化常数,使得 $\sum_{u=1}^{m}h_u^{\mathrm{obj}}=1,C=\frac{1}{\sum_{i=1}^{n}k\left(\|X_i^*\|\right)^2}$,

$\delta[\cdot]$ 为Kronecker Delta函数。

以 $\{X_i\}_{i=1}^{n_h}$ 表示候选目标区域像素的归一化坐标,若 $X$ 表示当前帧的中心坐标点,采用同样的核函数 $k(x)$,核函数窗宽 $h=\mathrm{diag}\{h_x,h_y\}$,则候选目标直方图 $h_u(X)$ 用式(11)表示:

$$h_u(X)=C_h\sum_{i=1}^{n_h}k\left(\left\|\frac{X-X_i}{h}\right\|\right)^2\delta\left[b\left(X_i\right)-u\right] \tag{11}$$

其中,$C_h$ 为归一化常数,使得 $\sum_{u=1}^{m}h_u(X)=1$,$C_h=\frac{1}{\sum_{i=1}^{n_h}k\left(\left\|\frac{X-X_i}{h}\right\|\right)^2}$。核函数:

$$k(x)=\begin{cases}\frac{1}{2}c_d^{-1}(d+2)(1-x) & \text{if } x\le 1\\ 0 & \text{others}\end{cases} \tag{12}$$

其中,$d$ 为空间维数,$c_d$ 为 $d$ 维空间的单位量。

假定参考目标 $h_{\mathrm{obj}}$ 的颜色直方图是稳定的,利用递归梯度搜索策略来找寻与 $h_{\mathrm{obj}}$ 最相似的区域。定义 $h_u^{\mathrm{obj}}$ 与 $h_u(X)$ 的相似性即Bhattacharyya系数 $\rho[h_u^{\mathrm{obj}},h_u(X)]=\sum_{u=1}^{m}\sqrt{h_u^{\mathrm{obj}}\cdot h_u(X)}$,则 $\rho[h_u^{\mathrm{obj}},h_u(X)]$ 最大化时即可得到新的图像帧中目标的位置。为了跟踪当前帧的目标 $X_t^v$,先前帧被用作初始估计,也就是 $\hat{X}_0=X_{t-1}^v$,$X_{t-1}^v$ 中的上标"$v$"表示基于颜色的跟踪。Mean-Shift算法具体跟踪步骤如下:

(1)设 $l$ 代表均值偏移循环次数,设 $l=0$ 表示均值漂移初始化。

(2)计算在初始化位置 $\hat{X}_l$ 处的候选目标颜色直

方图 $h_{\hat{X}_l}$，并计算 $h_{\hat{X}_l}$ 与 $h_{\text{obj}}$ 之间Bhattacharyya系数 $\rho[h_{\text{obj}}, h_{\hat{X}_l}]=\sum_{u=1}^{m}\sqrt{h_u^{\text{obj}}\cdot h_u^{\hat{X}_l}}$。

(3)计算含有梯度信息的权重 $\pi_i$，并使用梯度信息得到下一位置 $\hat{X}_N$。

$$\hat{X}_N=\frac{\sum_{i=1}^{nh}X_i\pi_i}{\sum_{i=1}^{nh}\pi_i},\ \pi_i=\sum_{u=1}^{m}\delta\left[b(X_i)-u\right]\sqrt{\frac{h_u^{\text{obj}}}{h_u^{\hat{X}_l}}}$$

(4)计算 $\hat{X}_N$ 处的候选目标的颜色直方图 $h_{\hat{X}_N}$，并计算 $h_{\hat{X}_N}$ 与 $h_{\text{obj}}$ 之间的Bhattacharyya系数 $\rho[h_{\text{obj}}, h_{\hat{X}_N}]=\sum_{u=1}^{m}\sqrt{h_u^{\text{obj}}\cdot h_u^{\hat{X}_N}}$。

(5)若 $\rho[h_{\text{obj}}, h_{\hat{X}_N}]>\rho[h_{\text{obj}}, h_{\hat{X}_l}]$，转第(6)步；否则，令 $\hat{X}_N=\frac{1}{2}(\hat{X}_l+\hat{X}_N)$，并转第(4)步。

(6)若 $\|\hat{X}_N-\hat{X}_l\|<\varepsilon$（$\varepsilon$ 小于一个像素值），算法停止；否则令 $l=l+1$ 且 $\hat{X}_l=\hat{X}_N$，转第(2)步。

当算法停止时时，$\rho[h_u^{\text{obj}}, h_u(X)]$ 达到最大化，此时，$\hat{X}_l$ 就是目标位置 $\hat{X}_t^v$ 的新估计。

由此，建立基于肤色跟踪的重要性函数，如式(13)所示：

$$q_v(X_t^v|X_{t-1}^v, Z_t^v)=N(\hat{X}_t^v, \hat{\boldsymbol{\Sigma}}_t^v) \tag{13}$$

式(13)中，$N$ 表示二维正态分布，$\hat{X}_t^v$ 为期望，即由肤色跟踪得到的目标位置；$\hat{\boldsymbol{\Sigma}}_t^v$ 为协方差矩阵，表示Mean-Shift肤色跟踪的不确定性。

在实际应用中，为了将目标与背景区分开来，使用如下判别模型：对于一个给定的目标假设 $X_t^{(i)}$，以 $h_{X_t^{(i)}}^f$ 为其颜色直方图，$h_{X_t^{(i)}}^b$ 为其背景颜色直方图，计算两个直方图之间的相似性即Bhattacharyya系数，用式(14)表示：

$$\rho(h_{X_t^{(i)}}^f, h_{X_t^{(i)}}^b)=\sum_{u=1}^{m}\sqrt{h_{X_t^{(i)}}^f(u)\cdot h_{X_t^{(i)}}^b(u)} \tag{14}$$

式中，$u$ 表示直方图颜色索引，而 $h_{X_t^{(i)}}^f$ 与 $h_{X_t^{(i)}}^b$ 之间的差异性给出了目标的似然性即 $X_t^{(i)}$ 的颜色似然模型，用式(15)表示：

$$p(Z_t^v|X_t^{(i)})=1-\rho(h_{X_t^{(i)}}^f, h_{X_t^{(i)}}^b) \tag{15}$$

式(15)中，差值越大即差异性越大则表示是目标的似然性越大，其相应采样粒子的权值也越大。

### 4.2　基于声源定位的跟踪

基于视频的跟踪只能提供人物的位置，而基于音频的跟踪能确定正在说话的是哪个人。本文采用声源定位SSL(Sound Source Localization)跟踪，利用GCC-PHAT(Generalized Cross Correlation-Phase Transform)估计出声源信号到达麦克风对之间的时延TDOA(Time Delay of Arrival)[16-17]，以此确定说话人的位置。设 $s(t)$ 为说话者声源信号，$x_1(t)$ 和 $x_2(t)$ 分别为麦克风对中两个麦克风收到的信号，则 $x_1(t)$ 和 $x_2(t)$ 可表示为：

$$\begin{cases}x_1(t)=s(t-D)+h_1(t)*s(t)+n_1(t)\\x_2(t)=s(t)+h_2(t)*s(t)+n_2(t)\end{cases} \tag{16}$$

式(16)中，$D$ 为信号到达两个麦克风之间的时延，$h_1(t)$ 和 $h_2(t)$ 表示回响，$n_1(t)$ 和 $n_2(t)$ 表示加性噪声。

设信号与噪声不相关，则 $D$ 可以通过GCC-PHAT估计出来，如式(17)所示：

$$D=\arg\max\hat{R}_{x_1x_2}(\tau) \tag{17}$$

$$\hat{R}_{x_1x_2}(\tau)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{x_1(\mathrm{e}^{\mathrm{j}\omega\tau})x_2^*(\mathrm{e}^{\mathrm{j}\omega\tau})}{\left|x_1(\mathrm{e}^{\mathrm{j}\omega\tau})x_2^*(\mathrm{e}^{\mathrm{j}\omega\tau})\right|}\mathrm{e}^{\mathrm{j}\omega\tau}\mathrm{d}\omega \tag{18}$$

式(18)中，$\hat{R}_{x_1x_2}(\tau)$ 为 $x_1(t)$ 和 $x_2(t)$ 的互相关函数，$x_1(\mathrm{e}^{\mathrm{j}\omega\tau})$ 和 $x_2(\mathrm{e}^{\mathrm{j}\omega\tau})$ 分别为 $x_1(t)$ 和 $x_2(t)$ 的傅里叶变换，$(\cdot)^*$ 表示共轭，$|\cdot|$ 表示模。

得到了时延 $D$，声源所在方向 $\theta$ 就可以计算出来。设 $A$、$B$ 为两个麦克风的位置，其中点为 $O$，$S$ 为声源所在位置，如图2所示。

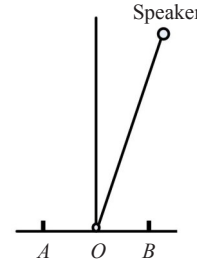


**图2　声源定位模型图**

在实际场景中，距离 $|OS|>>|AB|$，则 $\theta$ 可以由式(19)估计出来：

$$\theta=\arccos\frac{D\cdot V}{|AB|} \tag{19}$$

式(19)中，$V=342$ m/s，为声音在空气中的传播速度。

通常情况下，人们很少做垂直运动，而多以水平运动为主，则需要更多地关注说话者的水平位置 $X_t^a$。因此，要将 $\theta$ 转换为物体状态即水平位置 $X_t^a$。将麦克风和摄像头放置在使它们的中心在水平面上重合的位置，以 $O$ 为摄像头的光学中心点，$X_F$ 为摄像机的水平视野区域，$X_R$ 为摄像机的水平分辨

率,则 $X_t^a$ 可由式(20)计算得到:

$$\hat{X}_t^a=\frac{X_R/2}{\tan(X_F/2)\cdot\tan\theta} \tag{20}$$

由此,建立基于SSL跟踪的重要性函数,如式(21)所示:

$$q_a(X_t^a|X_{t-1}^a,Z_t^a)=N(\hat{X}_t^a,\hat{\boldsymbol{\Sigma}}_t^a) \tag{21}$$

式(21)中,$N$ 表示二维正态分布,$\hat{X}_t^a$ 为期望,即由SSL跟踪得到的目标位置;$\hat{\boldsymbol{\Sigma}}_t^a$ 为协方差矩阵,表示SSL跟踪的不确定性。

在真实的室内环境中,通常存在轻微的噪声(如计算机风扇声等),以及房间混响等,使得声音信号的互相关曲线 $\hat{R}_{x_1x_2}(\tau)$ 有多个峰值,为此需要建立更加精确的音频似然模型,以达到快速准确地跟踪。对于一个给定目标假设 $X_t^{(i)}$,其音频似然模型定义为在互相关曲线 $\hat{R}_{x_1x_2}(\tau)$ 上,它自身的高度 $\hat{R}_{x_1x_2}(D^{(i)})$ 与最高峰值 $\hat{R}_{x_1x_2}(D)$ 之间的比率,用式(22)表示:

$$p(Z_t^a|X_t^{(i)})=\frac{\hat{R}_{x_1x_2}(D^{(i)})}{\hat{R}_{x_1x_2}(D)} \tag{22}$$

式(22)比值越大,则表示真正声源的可能性越大,从而采样粒子的权值也会越高。

### 4.3 音视频信息融合与结果反馈

假定基于颜色的视频似然模型与基于SSL的音频似然模型之间相互独立,根据概率乘法原理,得到一个音频视频融合似然模型,用式(23)表示:

$$p(Z_t|X_t^{(i)})=p(Z_t^v|X_t^{(i)})\cdot p(Z_t^a|X_t^{(i)}) \tag{23}$$

式(23)在式(9)中用来计算并更新粒子权值。

在分别完成音频跟踪和视频跟踪之后,通过融合中心融合音频和视频跟踪模块的结果,得到一个音视频融合重要性函数,用如式(24)所示的混合高斯分布表示:

$$q\left(X_t|X_{t-1},Z_t\right)=\lambda_a q_a\left(X_t^a|X_{t-1}^a,Z_t^a\right)+\lambda_v q_v\left(X_t^v|X_{t-1}^v,Z_t^v\right) \tag{24}$$

式(24)中,$\lambda_a$ 和 $\lambda_v$ 分别为音频跟踪和视频跟踪的可靠因子,由式(25)给出。这样,音视频融合重要性函数就由音频和视频重要性函数组合而成,使得单一音频跟踪或视频跟踪失败时,融合算法仍是鲁棒的。

在一个非静止的环境中,物体外观会发生改变,背景杂波也会进一度使跟踪复杂化。因此,要为音频跟踪和视频跟踪定义一个可靠度。当前粒子集 $\{X_t^{(i)},w_t^{(i)},i=1,2,\cdots,N\}$ 表示目标状态的后验分布,通过重要性函数与后验分布相似或不相似的对的对比,计算出可靠因子 $\lambda_a$ 和 $\lambda_v$,用式(25)表示:

$$\begin{cases}\lambda_a=\sum\limits_{X_t^{(i)}}\sqrt{w_t^{(i)}\cdot q_a\left(X_t^{(i)}|X_{0:\ t-1},Z_t^a\right)}\\ \lambda_v=\sum\limits_{X_t^{(i)}}\sqrt{w_t^{(i)}\cdot q_v\left(X_t^{(i)}|X_{0:\ t-1},Z_t^v\right)}\end{cases} \tag{25}$$

这种性能评价准则类似于Bhattacharyya系数,但前者是基于加权粒子的,其直观解释为:如果一个单一的重要性函数(音频或视频)与后验分布有很大重叠,则它是一个好的重要性函数,应该更多地信任相应的跟踪。

融合后的跟踪结果可进一步反馈给音频跟踪和视频跟踪,如式(26)和(27)所示:

$$X_t^a=\lambda_a\hat{X}_t^a+(1-\lambda_a)\sum_{X_t^{(i)}}w_t^{(i)}\cdot X_t^{(i)} \tag{26}$$

$$X_t^v=\lambda_v\hat{X}_t^v+(1-\lambda_v)\sum_{X_t^{(i)}}w_t^{(i)}\cdot X_t^{(i)} \tag{27}$$

式(26)和式(27)中,$\hat{X}_t^a$ 和 $\hat{X}_t^v$ 分别为音频跟踪和视频跟踪对 $X_t$ 的估计,$\sum\limits_{X_t^{(i)}}w_t^{(i)}\cdot X_t^{(i)}$ 为融合中心对 $X_t$ 的估计。可靠因子 $\lambda_a$ 和 $\lambda_v$ 发挥着自动校准的角色,若单一跟踪(音频或视频)是可靠的,则该跟踪的当前状态更多地依赖于其自身的估计;否则,它更多地依赖于融合中心的估计。

## 5 仿真实验及分析

采用AMI Meeting Corpus[18]的会议中IS1003a.C进行仿真实验。会议长度为15分钟,有4人参加,其中的女性为主持人(主要发言人),其他3人在她发言完毕后轮流提出自己的观点。会议房间为一小型智能会议环境,房间中设置的麦克风阵列(桌上绿色物体)、耳麦、摄像机等完成音频和视频数据的实时采集。实验采用Intel Pentium4 3.00 GHz CPU,768 MB内存的服务器;图像尺寸为320×240,帧率15帧/秒,先验分布协方差为30像素,后验分布协方差为20像素;人脸矩形模板的高宽分别为 $h_y=24$ 和 $h_x=16$,采样粒子为60个。

对同一序列分别作了音频、视频及融合的跟踪仿真。如图3,主持人突然移动时,出现大量噪声,使声源定位出现很大误差。图4中围巾颜色与肤色相近,对肤色跟踪造成很强干扰;或者当说话者背对摄像机时,发生视频遮挡,致使肤色跟踪误差较大甚至失效。而音视频融合的跟踪,鲁棒性、准确性均非常好,如图5所示。

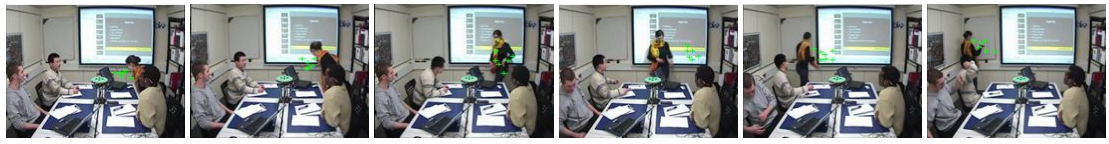

图3　音频跟踪结果图

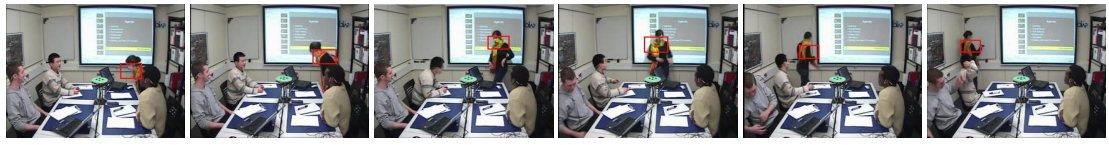

图4　视频跟踪结果图

图5　音视频融合跟踪结果图

图6是对选取的一段长3分钟的视听会议资料进行的均方误差统计，在0~30 s内，参会者都没有移动，也没有遮挡现象，只有主持人一个人讲话，此时3种跟踪方法都能跟踪到说话者的位置信息；在30~80 s内和130~180 s内，分别出现说话者移动和相互交谈的情况，声源定位误差明显增大，很难跟踪到说话者的位置；在80~130 s内，发生视频遮挡现象，此时肤色跟踪难以判断说话者脸部区域，甚至导致跟踪目标丢失。表1是对3种跟踪算法的性能比较，表中误跟率是基于跟踪效果统计得出的，运行时间指算法首次跟踪到说话人所需要的时间。

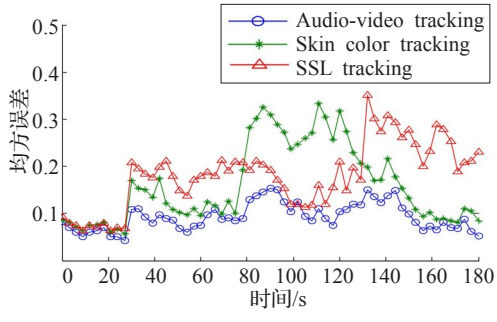


图6　3种跟踪方法的均方误差图

表1　3种跟踪方法的性能比较

| 算法 | 误跟率/(%) | 运行时间/ms |
|---|---|---|
| 声源定位 | 18.17 | 1.217 |
| 视频跟踪 | 16.03 | 0.921 |
| 音视频融合 | 9.32 | 0.785 |

本文方法融合了音频视频两方面的信息，在说话者移动和房间混响时通过人脸肤色跟踪说话者位置，在视频遮挡时通过声源定位仍能有效跟踪说话者，鲁棒性较好。如表1所示，音视频融合的跟踪方法在水平方向的平均误跟率为9.32%，远低于人脸肤色跟踪的16.03%和声源定位的18.17%。在同样条件下，音视频融合算法首次跟踪到说话人的时间也是三者中最快的，说明其实时性最好。

## 6　结论

提出了一种智能会议环境下的音视频信息融合的说话人跟踪算法。首先分别利用人脸肤色跟踪和声源定位跟踪提出各自的重要性函数和似然模型，然后结合说话人视频图像信息和语音信息之间的互补性，以重要性粒子滤波为工具建立融合重要性函数和融合似然模型，最后进行粒子滤波，实现了小型智能会议环境下的说话人跟踪。文中采用流程化的闭环处理框架，在跟踪过程中引入了反馈环节，提高了跟踪的实时性和完整性，平均误跟率仅为9.32%。

## 参考文献：

[1] Mungamuru B，Aarabi P.Enhanced sound localization[J]. IEEE Transactions on Systems，Man and Cybernetics：Part B，2004，34(3)：1526-1540.

[2] Verma R C，Schmid C，Mikolajczyk K.Face detection and tracking in a video by propagating detection probabilities[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，2003，25(10)：1215-1228.

[3] Gatica-Perez D，Lathoud G，McCowan I，et al. Audio-visual speaker tracking with importance particle filters[C]// Proceedings of IEEE International Conference on Image Processing(ICIP2003).[S.l.]：IEEE Press，2003：25-28.

[4] Vermaak J，Blake A，Gangnet M，et al.Sequential Monte Carlo fusion of sound and vision for speaker tracking[C]// Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV2001).Vancouver，Canada：IEEE Press，2001：741-746.

[5] Beal M，Attias H，Jojic N.Audio-video sensor fusion

with probabilistic graphical models[C]//Proceedings of the 7th European Conference on Computer Vision(ECCV 2002).Berlin,Heidelberg:Springer,2002,2350:736-750.

[6] Checka N,Wilson K W,Siracusa M R,et al.Multiple person and speaker activity tracking with a particle filter[C]//Proceedings of IEEE International Conference on Acoustics,Speech,and Signal Processing(ICASSP 2004).[S.l.]:IEEE Press,2004:881-884.

[7] Chen Y,Rui Y.Real-time speaker tracking using particle filter sensor fusion[J].Proceedings of the IEEE,2004,92(3):485-494.

[8] Bernardin K,Stiefelhagen R.Audio-visual multi-person tracking and identification for smart environments[C]//Proceedings of the 15th International Conference on Multimedia.New York,USA:ACM,2007:661-670.

[9] 李昕.基于音频视频信息融合的人物跟踪及其应用[D].北京:清华大学,2005.

[10] 金乃高,殷福亮,陈喆.基于动态贝叶斯网络的音视频联合说话人跟踪[J].自动化学报,2008,34(9):1083-1089.

[11] 胡士强,敬忠良.粒子滤波算法综述[J].控制与决策,2005,20(4):361-365.

[12] Gordon N J,Salmond D J,Smith A F M.Novel approach to nonlinear/non-Gaussian Bayesian state estimation[J].IEEE Proceedings on Radar,Sonar and Navigation,1993,140(2):107-113.

[13] Comaniciu D,Ramesh V,Meer P.Real-time tracking of non-rigid objects using mean shift[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR2000).Hilton Head Island,South Carolina:IEEE Press,2000:142-149.

[14] 何文媛,韩斌,徐之,等.基于粒子滤波和均值漂移的目标跟踪[J].计算机工程与应用,2008,44(11):61-64.

[15] 陈爱斌,蔡自兴,董德毅.一种基于目标和背景加权的目标跟踪算法[J].控制与决策,2010,25(8):1246-1250.

[16] Rui Y,Florencio D.Time delay estimation in the presence of correlated noise and reverberation[EB/OL].Microsoft Research(2003-01).http://research.microsoft.com/ven-us/um/people/yongrui/ps/icassp04.pdf.

[17] 何伟俊,周非.基于粒子滤波的TOA/TDOA融合定位算法研究[J].传感技术学报,2010,23(3):404-407.

[18] AMI:Augmented Multi-party Interaction.AMI meeting orpus[EB/OL].(2008)[2011-01].http://www.amiproject.org/ami-scientific-portal/meeting-corpus.

(上接113页)

率受到距离基站远近和移动速度的影响很大,用户之间的有效速率分配不是很均匀。从图10可以看出,在经过动态优化后,UE1至UE5这5个用户终端虽然信道状况差别很大,但是都可以享受到比较公平的带宽分配,在动态优化的前后对比中,各个终端速率之和,即$T_{max}$始终都维持在2.3 Mb/s左右。

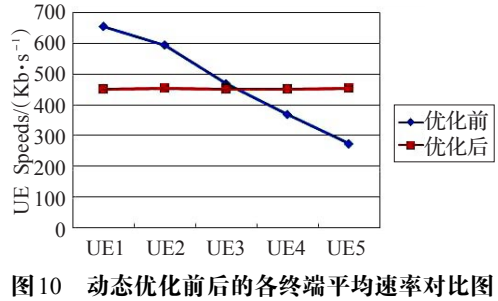


**图10 动态优化前后的各终端平均速率对比图**

## 5 结论

针对TFRC多用户无线链路视频传输,提出了基于动态优化的速率控制方法。本文方法充分利用了自适应调制编码特性,根据每个用户的信道状况,动态地调节编码调制方式,从而优化每个用户的速率。通过分析和仿真实验,可以看出本文方法在总的最大有效速率限制下,一定程度上保证了整体用户群速率的公平分配和优化。

## 参考文献:

[1] Chiu D,Jain R.Analysis of the increase and decrease algorithms for congestion avoidance in computer networks[J].Journal of Computer Networks and ISDN Systems,1989,17:1-14.

[2] Denda W R,Mauve M.A survey on TCP-friendly congestion control[J].IEEE Network,2001,15:28-37.

[3] Floyd S,Handley M,Padhye J,et al.RFC 3448 TCP rriendly rate control(TFRC):protocol specification[S].Network Working Group,2003-01.

[4] Chen Minhua.Rate control for streaming video over wireless[C]//Proceedings of IEEE INFOCOM 2004,2004.

[5] 3GPP TSG-RAN WG1#17,Adaptive Modulation and Coding(AMC),Stockholm,Sweden,20-24 Oct,2000.

[6] Kolding T.Performance aspects of WCDMA systems with high speed downlink packet access(HSDPA)[C]//Proceedings of VTC 2002,Vanconver BC,Canada,2002.

[7] Parekh A,Gallager R.A generalized processor sharing approach to flow control in integrated services networks:the single node case[J].IEEE/ACM Transactions on Networking,1993,1(3):344-357.

[8] Shao H R,Shen C,Gu D,et al.Dynamic resource for high-speed downlink packet access wireless channel[C]//Proceedings of the 23rd International Conference on Distributed Covnputing Systems,2003.