基于DSP的说话人定位跟踪系统的设计

曹 洁[1],何裔玺[2]
CAO Jie[1], HE Yixi[2]

1.兰州理工大学 计算机与通信学院,兰州 730050
2.兰州理工大学 计算机与通信学院,通信工程系,兰州 730050
1.School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
2.Department of Tele-communication, School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

CAO Jie, HE Yixi. Design of speaker location and tracking system based on DSP. Computer Engineering and Applications, 2013, 49(1):163-166.

Abstract: Aiming at the problem of inaccurate location and tracking for speaker in a meeting room, a method of location and tracking of audio visual fusion based on Digital Signal Processing(DSP) is proposed. Kalman filter and Mean-shift algorithm are used to seek optimal situation of speaker for visual location and tracking. Meanwhile, it uses Time Difference of Arrival to locate the target. Then Kalman information centre made audio and visual fused in order to advance stability of audio and visual system. The experimental results show that the processing for 320 pixels×240 pixels image achieves 20 frame/s, and the proposed method can rise target's location and tracking precision of 17%, compared with single mode system, and improve the stability.
Key words: information fusion; audio location; target tracking; Kalman filter

摘 要: 针对室内说话人实时定位跟踪不准确的问题,提出了一种基于TMS320DM6437硬件平台的音视频融合定位跟踪方法。该方法利用Kalman滤波器和Mean-shift算法搜寻说话人最优位置进行视频定位跟踪。同时,采用到达时间差的音频方法进行目标位置估计。由Kalman信息整合中心进行音视频融合,以提高视听系统定位跟踪的稳定性。实验结果表明,与单模态定位跟踪系统相比,该方法对320×240的图像可实现平均20 frame/s的跟踪速度,能提高目标定位跟踪准确度17%,改进效果明显且稳定。
关键词: 信息整合;声源定位;目标跟踪;Kalman滤波器
文献标志码: A **中图分类号:** TP391.4 doi:10.3778/j.issn.1002-8331.1105-0521

1 引言

基于DSP的音视频技术的运动目标定位跟踪是目前一个很有价值的研究方向。DSP数字信号处理器以高速度、高精度、高实时性的数据采集与处理技术,在诸如工业、交通运输业、医学、军事、航空航天等领域发挥着极其重要的作用。近些年来,人们提出了许多方法,并结合硬件平台,设计出体积小的嵌入式跟踪系统,用于运动目标的跟踪中,取得了一定的效果。如刘晓辉等人采用的扩展Kalman滤波的主动跟踪技术[1],可以根据目标的连续运动状态参数预测摄像头的运动,但该方法不能有效地控制摄像机对非连续运动状态的目标跟踪;赵鹏等人分析了基于粒子滤波的跟踪算法[2]中的退化现象,设计了使用每个权值的确切值来指导重采样的新算法,基于DSP平台实现了对远距离弱小目标进行10 frame/s的跟踪,但随着目标在图像上尺寸的增大,跟踪的实时性会严重下降;文献[3]结合神经网络给出了视听整合的方法,但没有给出音频到视频映射的模型;文献[4]结合基于语音合成的隐性马尔可夫模型,设计了一种增强的语音理解算法,通过参数对每一帧进行跟踪,但该方法在语音合成时操作冗繁,实时性受到了较大影响。

针对上述问题,本文采用TI公司的DaVinci系列TMS320DM6437数字视频硬件开发平台作为硬件处理平台,采用音视频融合双模态的方法,对目标说话人进行定位与跟踪。通过Mean-shift算法与Kalman滤波器相结合

基金项目: 甘肃省自然科学基金(No.1014ZSB064);甘肃省财政厅项目(No.0914ZTB148)。
作者简介: 曹洁(1966—),教授,博士生导师,现任兰州理工大学计算机与通信学院院长,主要研究方向为信息融合、智能交通系统、嵌入式系统和计算机控制系统等;何裔玺(1984—),男,硕士研究生,研究方向:智能信号处理。E-mail:caoj@lut.cn
收稿日期: 2011-05-26 **修回日期:** 2011-07-25 **文章编号:** 1002-8331(2013)01-0163-04
CNKI**出版日期:** 2011-10-13 http://www.cnki.net/kcms/detail/11.2127.TP.20111013.0954.029.html

对说话人目标进行视频跟踪。同时,运用到达时间差方法进行音频定位。以此消除或降低跟踪实时性差,不能跟踪非连续运动目标及使音视频不能较好配准的问题。通过上述方法得到音视频信息,被传送到Kalman信息整合中心,进行信息整合,从而估计出最终的系统状态,得到准确的目标位置。

2　系统的硬件设计

系统主要由CCD摄像机、视频处理子系统、显示器、音频处理模块、麦克风和DSP处理控制器部分构成。硬件系统框图如图1。

2.1　TMS320DM6437硬件平台的特点

TMS320DM6437是TI公司针对数字视频领域开发的高性能、32位定点DSP达芬奇(DaVinci(TM))技术的处理器[5]。采用TI第3代超长指令集结构(VelociTI.3)的TMS320C64x+DSP内核,主频可达600 MHz,支持8个8位或4个16位并行MAC运算,峰值处理能力高达4 800 MIPS,可实时处理8路CIF或3路D1格式的H.264编码算法。含有的视频处理子系统(VPSS)极大地支持了视频前端(VPFE)预处理与视频后端(VPBE)显示。一个专用的单通道视频输入接口,既可以方便地与各种数字视频输入标准接口,还具有常用的视频预处理功能;一个专用的单通道视频输出接口,既可以提供多种模拟视频输出标准,还可以提供各种数字视频输出标准接口。I2C总线,可无缝接口视频解码器/编码器和音频Codec的控制口,方便实现音/视频编解码器的控制。

视频处理子系统(VPSS)的视频前端预处理(VPFE)部分,由CCD控制器(CCDC)、预览引擎、图像大小调整器、硬件3A(H3A)统计发生器以及柱状图模块组成。此部分可从传感器接收原始图像/视频数据,或从视频解码器接收多种形式的YUV视频数据。当CCDC的输出,要求额外的图像处理,可将原始输入图像转换成最终处理过的图像送往DSP芯片进行视频处理。视频后端处理(VPBE)部分,由屏幕显示模块、视频编码器(VENC)和数字LCD控制器(DLCD)组成。这部分主要功能是将视频数据和显示数据整合后,以YC_bC_r形式提供给视频编码器(VENC)。视频和数据从DDR2等外部存储器读取。音频信号通过多通道缓冲串行口(McBSP)、模拟接口芯片(AICs)和串行A/D和D/A设备连接并进行处理,并通过DSP处理器进行处理。

2.2　信号获取

视频图像通过CCD摄像机获取图像,经过A/D转换器,输入到视频前端预处理模块。此模块可将原始输入图像转换成最终处理过的图像送往DSP芯片进行视频处理。然后信号被输入到DM6437的视频捕捉口,最后将图像数据经DM6437的视频显示口输出到D/A转换器,按一定的标准转化为模拟图像数据显示在显示器上。

音频信号由麦克风进行采集,将模拟信号送入音频编解码芯片TLV320AIC33,然后模数转换,信号成为数字信号。该CODEC芯片与DSP有两个通道,其中一个通道用于传输经 TLV320AIC33处理的音频数字信号。信号在DM6437中,由到达时间差算法进行处理,所得信号以坐标的形式来表示当前目标所在的位置,以此进行定位。

3　系统的软件算法设计

系统中,分别对音、视频部分进行设计。通过Mean-shift算法和Kalman滤波器来对视频目标进行跟踪,能够使目标运动过快之时,不会因为Mean-shift算法此时跟踪的不准确而造成跟踪准确度下降,甚至丢失目标的情况出现[6]。

现有的声源定位方法大致可以分为三类:(1)基于最大输出功率的可控波束形成技术。该方法对麦克风所接收到的语音信号进行滤波求和,在空间形成特定指向的波束,波束输出功率最大的点就是声源的位置。波束形成本质上是一种最大似然估计方法,需要声源和环境噪声的先验知识。在实际使用中,这种先验知识往往很难获得。(2)基于高分辨率谱估计的定向技术。其来源于一些现代高分辨率谱估计技术,虽然该方法成功的应用于一些阵列信号处理的应用,但在语音定位中的效果不佳。(3)基于到达时间差(TDOA)技术[7]。其在导航系统、声纳系统等领域有广泛的应用,却在运算量上远远小于可控波束形成的谱估计法,可在实际系统中实时实现。于是采用到达时间差

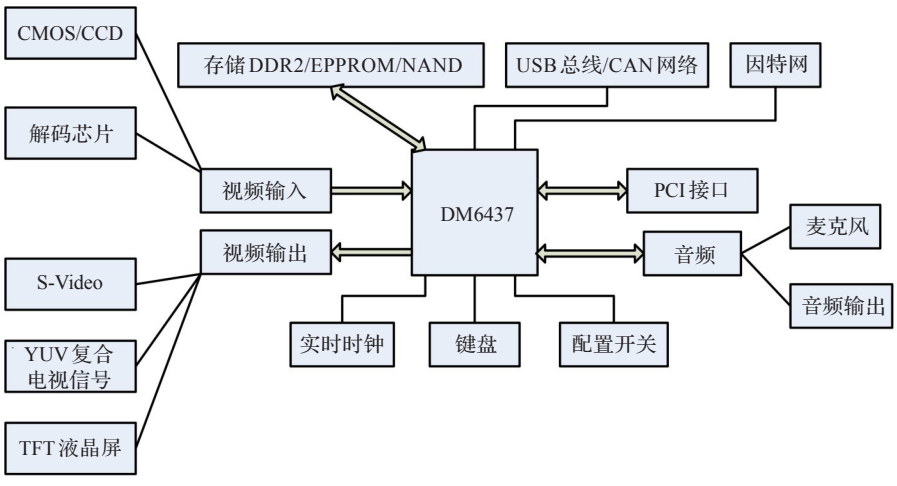

图1　基于DSP的运动目标音视频定位跟踪硬件系统框图

方法进行声源的定位。

3.1　视频跟踪方法

目标跟踪的实质就是在视频连续帧中确定目标的位置，它可以通过目标的特征匹配来实现。目前比较流行的特征匹配算法是Mean-shift算法[8]，这种算法虽然对目标位置跟踪可靠，计算速度快，但是当目标运动过快的时候，算法的跟踪效果会有所下降，在一些情况下可能造成目标丢失。于是，本文采用了一种Mean-shift与Kalman滤波器相结合的方法，进行目标跟踪，以提高跟踪的准确性和稳定性。

(1)目标模型进行建立[9]

设目标区域中心为 X_0，假设有 n 个像素由 $\{x_i\}, i=1,2,\cdots,n$ 来表示。特征值个数为 m。则目标模型 $(u=1,2,\cdots,m)$ 的特征值估计概率密度为：

$$\hat{q}_u = C\sum_{i=1}^{n} k\left(\left\|\frac{x_0 - x_i}{h}\right\|^2\right)\delta[b(x_i)-u] \tag{1}$$

其中，$k(x)$ 是核函数的轮廓函数，h 为轮廓窗半径。由于遮挡和背景的影响，靠近目标模型中心的像素比离边缘较近的模型更可靠。$k(x)$ 给每一个像素点赋一个权值，靠近中心的像素点会被赋予一个较大的权值，离中心越远，则赋予的权值也越小。

(2)候选目标模型

在当前帧中，候选目标采用同样的核函数 $k(x)$ 以及半径为 h 的窗。接着，估计基于核函数的候选目标 $(u=1,2,\cdots,m)$ 的特征值的密度，则目标位置即可由 $k(x)$ 表示的轮廓函数所表示的矩形跟踪窗的中心位置来描述。

(3)相似度函数

相似度函数描述的是目标模型和候选目标模型在初始帧时的相似度。此处用Bhattacharyya系数作为相似性函数，则定义相似度函数为：

$$\hat{\rho}(y) = \rho(\hat{p}(y), \hat{q}) = \sum_{u=1}^{m}\sqrt{\hat{p}_u(y), \hat{q}_u} \tag{2}$$

上式的值在0和1中间。若 $\hat{\rho}(y)$ 值越大，那么两个模型的相似度就越高。候选模型通过计算当前帧中不同的候选区域得到，其中使候选区域最大的 $\hat{\rho}(y)$ 即为目标在这一帧的位置。

(4)Kalman滤波建模

当目标运动过快的时候，Mean-shift算法的跟踪效果会有所下降。此时，采用Kalman滤波器确定初始时搜索窗口的中心点[10]。在连续的 $(k-1)$ 帧中，用目标中心位置作为它的运动轨迹。文中，这些中心位置将作为Kalman滤波器的观测值。先用Kalman滤波预测第 m 帧时刻目标的位置，然后用Mean-shift算法在该目标位置邻域内搜寻目标的最优位置；再以这个搜寻到的最优位置作为Kalman滤波进行下一帧运算时的观测值，并如此迭代。经过这样的几次迭代，目标区域就逐渐地从起始位置移动到真实目标位置，完成对于目标的跟踪[11]。

3.2　音频定位方法

图2显示了三个麦克风及相关声源的位置[12]。因为要用音频信息来确定声源的位置，所以，首先估计目标的音频信息，然后预测说话状态。计算到达时间差(Time Difference of Arrival，TDOA)的值，并用互相关的方法来估计声源信号。运用观测时间差值，对音频信息进行估算。图3所示为处理音频信息的流程图。

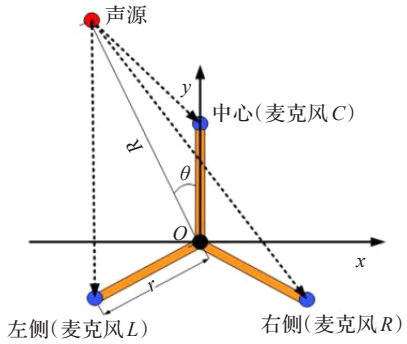

图2　麦克风声源定位图

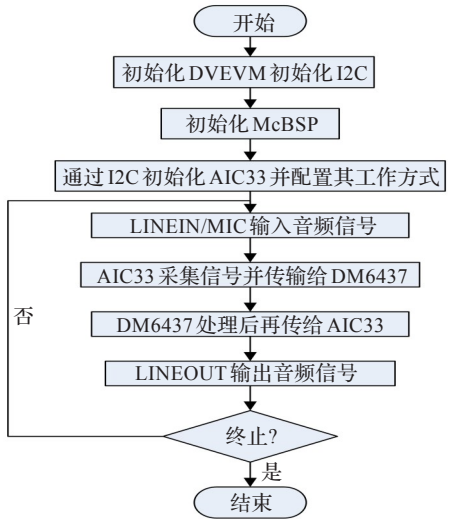

图3　音频处理流程图

3.3　音频、视频信号配准

将音频信号映射到图像时，传感器的不同导致获得的信号到达融合中心的时间也不尽相同，所以对视听信号进行时间配准是视听信息处理重要的一环。时间配准的核心就是将各个传感器数据统一到扫描周期较长的一个传感器数据上。这里采用数据丢弃法，即通过对异类信号的对比找出周期较长的信号作为对准信号，对小周期信号进行“滤波”，在对准信号周期内只允许一帧信号(被击中信号)被融合，其余信号被丢弃。如图4所示，a 为大周期信号，b 为小周期信号。黄色区域代表时间窗口，在时间窗口内到达的第一个帧将被击中。时间窗口必须大于 b 信号的周期，以保证至少有一帧数据被击中。

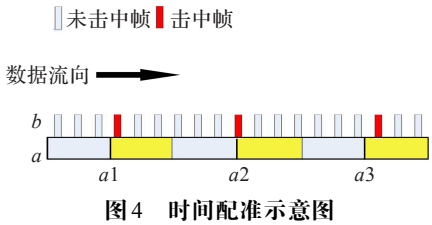

图4　时间配准示意图

3.4　音频与视频融合

设定说话人的说话与否的状态不会影响其视频信息。音视频信息联合定位跟踪涉及到两个传感器,包括麦克风阵列传感器和摄像机传感器。Kalman滤波器对音视频信息进行融合的框架图如图5所示。图5中,音频信号通过Kalman滤波器能够较好地处理噪声干扰下引起的滤波不稳定因素,使之能具有环境适应能力和定位效果,能更准确地对系统的状态进行评估。通过Kalman滤波,从而得到更加准确的定位的结果,$x_1(k)$是音频信息声源定位的结果;同时,采用Kalman滤波器确定初始时搜索窗口的中心点。在连续的$(k-1)$帧中,用目标的中心点作为目标运动的轨迹。这些中心位置将作为滤波器的观测值,以此预测第m帧时刻目标的位置,$x_2(k)$是视频跟踪信息的结果,分别对$z_1[k]$和$z_2[k]$进行Kalman滤波获得局部状态$\hat{x}_1[k|k]$和$\hat{x}_2[k|k]$。因为不同传感器会使得所以获得的信号到达融合中心的时间不同,所以,在此时需要将音视频信号进行配准,使之能够同步。于是,需要为异类信号找出找到一个周期较长的信号作为校准信号,这样使得在对准信号周期内的音视频信号能够被融合,将这两个局部状态带入到信息融合中心进行融合,从而估计出最终的系统状态$\hat{x}[k|k]$。

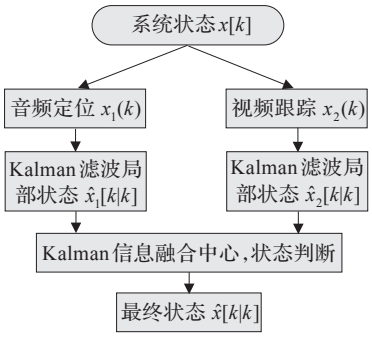

图5　音视频融合框图

4　实验结果与分析

考虑到实际运用系统的需要,进行了针对性的实验。实验设备:PC机(主频2.8 GHz,1 GB内存),摄像头,显示器,TI公司新出品的DaVinci系列TMS320DM6437数字视频硬件开发平台。开始实验需将上述设备按要求连接到DM6437实验硬件平台上,视频解码器的图像采样率为20 frame/s,分辨率为320×240。并进行调试。待调试完毕后,在PC机上,使用Code Composer Studio v3.3(CCS v3.3)软件,将事先编写调试好的系统程序烧写到硬件开发平台上。此时硬件即可独立进行设定的定位跟踪任务,并在显示器上显示出来。本实验设计为一个说话人在会议室环境中一边说话一边移动。首先,用音频和视频信息分别对声源进行定位跟踪;然后,再利用本文方法所介绍的音视频融合方法对声源进行定位跟踪。由此可得到三种跟踪定位方法的结果,以及真实位置在图像x、y、z方向上的误差均值与标准差单位(因为实验过程中,z方向上表示高度的值基本保持不变,变化的主要是x和y方向上的量),如表1所示。

表1　三种定位跟踪方法与真值的误差均值、标准差的比较

方法	误差均值(x,y)	误差标准差(x,y)
音频	(7.69,8.16)	(14.21,12.26)
视频	(7.11,4.89)	(8.18,6.54)
音视频联合	(5.13,4.61)	(6.32,6.41)

从表1中可以看出,采用音视频整合技术对目标声源进行定位及用视频进行跟踪的方法结果要好于只采用音频或视频单模态的定位跟踪结果。

图6是两组跟踪实验的对比图。其中,a组表示的是运用视听融合的方法对目标进行的跟踪效果图;b组表示的是运用视频方法对目标进行跟踪的效果图。通过对比,可以看到b组图能够较好地对目标进行跟踪,但有时会出现不准确的现象,比如在b1、b4、b6处。在改进之后,通过视听融合的方法进行的实验效果要比b组的效果更好、更准确。a组中的矩形框较之b组,能够更准确地对目标进行跟踪,不会再出现b组中的上述跟踪不准确的问题。用视听融合的方法,可将对目标定位跟踪准确度提高17%。实时性方面,测得在上述配置的条件下,跟踪中每帧需要大约50 ms的计算时间。实验中还计算了跟踪、定位的误差均值和误差标准差。由上述对图6的分析可以看出,单模态视频跟踪在目标移动开始不久会偏离目标的真实位置;而双模态下,平均只偏离1个像素左右,最大偏差不超过4个像素。这说明在绝大数情况下,运用音、视频双模态方法对目标进行处理,能够紧跟目标,提高目标定位跟踪的可靠性和准确性。

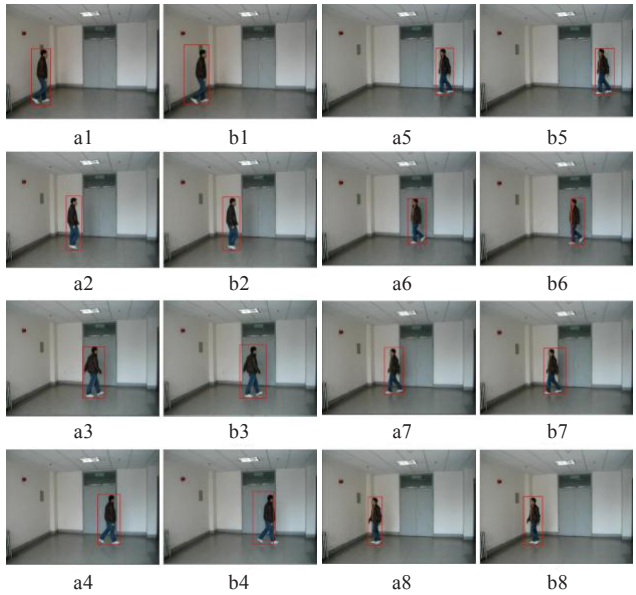

图6　视频序列跟踪比较

5　结论

本文从硬件与软件着手,较好地将Mean-shift算法和Kalman滤波进行视频跟踪和到达时间差方法对说话人进行定位综合在一起,提出了基于音视频融合技术对说话人

(下转190页)

具有较强的鲁棒性。同时无论是相对于传统的边缘提取方法,还是相对于对边缘检测具有很好的信噪比和检测精度的Canny算子,本文提出的方法都具有更好的检测效果。

通过实验发现本文方法仍有一些可以改进的地方。例如,如何改进能量函数来消除曲率变化较大的点以得到更好雨滴的边界图像;如何能够在保证准确找到雨滴图像轮廓的基础上,增加轮廓线的收敛速度,更加快速地找到雨滴边界,以利于大量信息的快速自动化处理;如何能够进一步改进轮廓线的初始化控制点,找到一个更加合理的初始化轮廓线等等。

5 结束语

本文将主动轮廓法用于对数字相机拍摄的图片进行雨滴检测,改进了主动轮廓线初始控制点的获取方法,实现了自动寻找检测区域和标定目标形心。实验结果表明,采用的轮廓线初始控制点获取方法和检测区域自动标定,无须任何人为干预就能够很好地接近目标的边缘;最终结果也能够很好地逼近目标的真实边界。如何能够使计算机在承担较少的计算量的情况下,快速地逼近雨滴的真实边界,确定最佳的控制点数,以及如何为主动轮廓线法找到一个更好的初始化轮廓线使结果更加精确,是本文下一步的研究方向。

参考文献:

[1] 冯圆.河南春季一次冷锋降水过程的云物理结构分析[J].解放军理工大学学报:自然科学版,2005(6):591-597.

[2] 安英玉,金凤岭.地面雨滴谱观测的图像自动识别方法[J].应用气象学报,2008,19(2):188-192.

[3] Kass M,Witkin A,Terzopoulos D.Snakes:active contour models[J].International Journal of Computer Vision,1988(4):321-331.

[4] Williams D,Shah M.A fast algorithm for active contours and curvature estimation[J].Computer Vision,Graphics and Image Processing:Image Understanding,1992,55(1):14-26.

[5] Amini A A,Tehrani S,Weymouth T E.Using dynamic programming for minimizing the energy of active contours in the presence of hard constraints[C]//Proc of the 2nd International Conference on Computer Vision,1988:95-99.

[6] 侯立华.改进Snake模型在医学超声图像分割中的应用[J].计算机仿真,2008,25(8):183-185.

[7] 谢凤英,赵丹培.Visual C++数字图像处理[M].北京:电子工业出版社,2008.

[8] 李天庆,张毅.Snake模型综述[J].计算机工程,2005,31(9):1-3.

[9] 吕伟涛,马颖,张阳,等.降水粒子图像采集装置:中国,实用新型专利ZL201120073006.3[P].2011.

(上接166页)

进行定位跟踪的新方法。基于DM6437的硬件平台能够有效协调视频图像和音频信号的数据。在硬件平台的基础上,用C语言完成算法的设计编写工作。本文算法进行实验操作的实际情况,较之单模态的算法及其仿真性质,其计算时间虽有所增大,但对目标定位跟踪的实时性和鲁棒性均有大幅提高。今后的研究工作将通过进一步对算法及代码优化,来满足硬件平台处理程序的运算量和实时性的更高要求。

参考文献:

[1] 刘晓辉,陈小平.基于扩展Kalman滤波的主动视觉跟踪技术[J].计算机辅助工程,2007,16(2):32-37.

[2] 郝志成,朱明.智能目标检测与跟踪系统的设计与实现[J].光电工程,2007,34(1):27-31.

[3] 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.

[4] Coelho L,Braga D,Garcia-Mateo C.Kalman tracking linear predictor for vowel intelligibility enhancement on European Portuguese HMM based speech synthesis[C]//Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing(ICASSP),2010:4734-4737.

[5] 俞一彪.DSP技术与应用基础[M].北京:北京大学出版社,2009.3.

[6] Ali A,Terada K.A framework for human tracking using Kalman filter and fast mean shift algorithms[C]//Proceedings of IEEE 12th International Conference on Computer Vision Workshops(ICCV Workshops),2009:1028-1033.

[7] 王大中,李晓妮.基于麦克风阵列的语音信号实时时延估计[J].吉林大学学报:信息科学版,2009,27(2):133-138.

[8] 马加庆,韩崇昭.一类基于信息融合的粒子滤波跟踪算法[J].光电工程,2007,34(4):22-25.

[9] Chu Hongxia,Wang Kejun.Target tracking based on mean shift and improved Kalman filtering algorithm[C]//Proceedings of IEEE International Conference on Automation and Logistics,2009:808-812.

[10] 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,Hilton Head Island,SC,USA,June 13-15,2000,2:142-149.

[11] Feng Shimin,Guan Qing,Xu Sheng.Human tracking based on mean shift and Kalman filter[C]//Proceedings of IEEE International Conference on Artificial Intelligence and Computerational Intelligence,2009:518-522.

[12] Lim Y,Choi J.Speaker selection and tracking in a cluttered environment with audio and visual information[J].IEEE Transactions on Consumer Electronics,2009,55:1581-1589.