应用光学
Journal of Applied Optics
SCOPUS,CSCD等来源期刊
中文核心期刊、中国科技核心期刊

基于音视频信息融合的目标检测与跟踪算法

黄战华 陈智林 张晗笑 曹雨生 申苜弘

Object detection and tracking algorithm based on audio-visual information fusion

HUANG Zhanhua, CHEN Zhilin, ZHANG Hanxiao, CAO Yusheng, SHEN Muhong

引用本文:
黄战华, 陈智林, 张晗笑, 等. 基于音视频信息融合的目标检测与跟踪算法[J]. 应用光学, 2021, 42(5): 867-876. DOI: 10.5768/JAO202142.0502007
HUANG Zhanhua, CHEN Zhilin, ZHANG Hanxiao, et al. Object detection and tracking algorithm based on audio-visual information fusion[J]. Journal of Applied Optics, 2021, 42(5): 867-876. DOI: 10.5768/JAO202142.0502007

在线阅读 View online: https://doi.org/10.5768/JAO202142.0502007

文章编号：1002-2082 (2021) 05-0867-10

基于音视频信息融合的目标检测与跟踪算法

黄战华，陈智林，张晗笑，曹雨生，申苜弘

（天津大学 精密仪器与光电子工程学院 光电信息技术教育部重点实验室，天津 300072）

摘　要：针对单一视觉跟踪算法易受遮挡影响的缺陷，提出一种基于音视频信息融合的目标检测与跟踪算法。整个算法框架包括视频检测与跟踪、声源定位、音视频信息融合跟踪 3 个模块。视频检测与跟踪模块采用 YOLOv5m 算法作为视觉检测的框架，使用无迹卡尔曼滤波和匈牙利算法实现多目标的跟踪与匹配；声源定位模块采用十字型麦克风阵列获取音频信息，结合各麦克风接收信号的时延计算声源方位；音视频信息融合跟踪模块构建音视频似然函数和音视频重要性采样函数，采用重要性粒子滤波作为音视频融合跟踪的算法，实现对目标的跟踪。在室内复杂环境下对算法性能进行测试，结果表明该算法跟踪准确率达到 90.68%，相较于单一模态算法具有更好的性能。

关键词：目标跟踪算法；音视频融合；目标检测；声源定位

中图分类号：TN206　　**文献标志码：**A　　**DOI：**10.5768/JAO202142.0502007

Object detection and tracking algorithm based on audio-visual information fusion

HUANG Zhanhua, CHEN Zhilin, ZHANG Hanxiao, CAO Yusheng, SHEN Muhong

（Key Laboratory of Opto-electronics Information Technology (Ministry of Education), School of Precision Instruments and Opto-electronics Engineering, Tianjin University, Tianjin 300072, China）

Abstract: Aiming at the defect that the single vision tracking algorithm is easily affected by the occlusion, an object detection and tracking algorithm based on the audio-video information fusion was proposed. The whole algorithm framework included three modules: video detection and tracking, acoustic source localization, audio-video information fusion tracking. The YOLOv5m algorithm was adopted by the video detection and tracking module as the framework of visual inspection, and the unscented Kalman filter and Hungary algorithm were used to achieve multi-object tracking and matching. The cross microphone array was adopted by the acoustic source localization module to obtain the audio information, and according to the time delay of receiving signals of each microphone, the acoustic source orientation was calculated. The audio-video likelihood function and audio-video importance sampling function were constructed by the audio-video information fusion tracking module, and the importance particle filter was used as the audio-video information fusion tracking algorithm to achieve object tracking. The performance of the algorithm was tested in complex indoor environment. The experimental results show that the tracking accuracy of the proposed algorithm reaches 90.68%, which has better performance than single mode algorithm.

Key words: object tracking algorithm; audio-visual fusion; object detection; acoustic source localization

引言

基于视觉的目标跟踪技术广泛应用于视频会议、智能监控、智能机器人等领域。现今常用的目标跟踪算法包括粒子滤波 (particle filtering)、均值

收稿日期：2021-05-06；　**修回日期：**2021-05-28

基金项目：国防科技创新特区“无障碍语言交流系统总体设计与关键技术攻关”项目（19-H863-00-KX-001-002-01）

作者简介：黄战华（1965—），男，博士，教授，主要从事光电图像处理与模式识别、光电子信息技术及多媒体计算机应用与控制研究。E-mail: zhanhua@tju.edu.cn

漂移 (meanshift)、卡尔曼滤波 (Kalman filtering, KF) 等[1-3]。基于视觉的目标跟踪精确度较高，但容易受到遮挡、光照等因素影响，因此存在一定的误跟现象。声源定位技术[4]可以测得声源的位置信息，虽然声源定位的精度相对较低，但不受视觉场景的影响并且测量范围更宽。考虑到单一使用视频或音频跟踪定位的缺点，试图将音频信息和视频信息融合，综合两种模态的信息实现目标检测与跟踪，使系统具有更高的准确率和鲁棒性。

视频信息与音频信息是两种不同模态的信息。多种模态的信息既能实现互补，也能提高信息的可靠性。音视频信息的融合就是一种多模态融合的方向之一。通过音视频信息的融合，可以实现复杂环境下的目标检测与跟踪。文献 [5] 采用序列蒙特卡洛方法融合头部轮廓和声源定位信息，实现说话人的定位。文献 [6] 提出一种融合目标轮廓、颜色、声源位置的说话人跟踪算法，得到稳定的跟踪效果。国内相关研究起步较晚。文献 [7] 采用重要性粒子滤波实现在智能教室环境下对演讲者的跟踪。文献 [8] 将均值漂移算法嵌入到粒子滤波算法中，将音视频跟踪结果通过粒子滤波算法融合，得到融合跟踪的结果。

本文在重要性粒子滤波算法的基础上，提出一种基于同源音视频信息融合的目标检测与跟踪框架，并设计了实现相关功能的硬件系统。实验结果表明，该算法可以有效利用音视频信息进行检测与跟踪，相较单一模态算法具有更高的准确率。

1 算法总体框架

算法总体框架如图 1 所示，包括视频检测与跟踪、声源定位、音视频信息融合跟踪 3 个模块。本系统先分别从视频和音频 2 个底层层面实现目标的跟踪定位，最后在决策层将视频和音频跟踪的结果融合，得到基于音视频融合的跟踪定位结果。

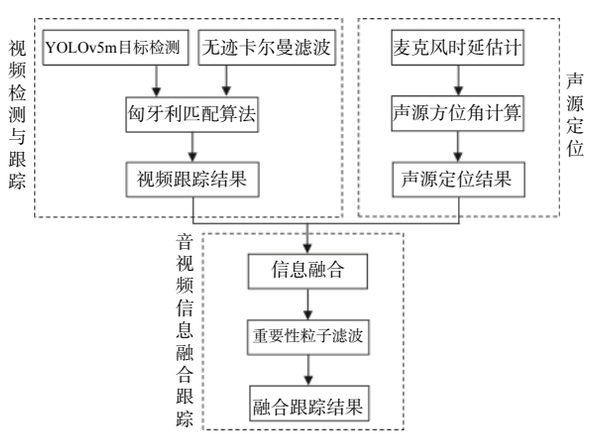


图 1 算法总体框图

Fig. 1 Block diagram of algorithm

视频检测与跟踪模块采用 YOLOv5m 算法对人等生活中常见的运动目标进行检测，通过无迹卡尔曼滤波 (unscented Kalman filtering, UKF) 跟踪算法对多目标进行跟踪预测，再通过匈牙利匹配算法将检测结果和跟踪结果匹配，得到视频跟踪结果。

声源定位模块采用基于时延估计（time difference of arrival, TDOA）的定位算法，用广义互相关函数 (generalized cross correlation, GCC) 和相位变化加权函数（phase transform, PHAT）估算出各个麦克风接收到声源信号的时间差，再结合麦克风阵列的空间拓扑结构计算出声源的方位角，最后将方位角投影至相机二维像面，得到声源定位结果。

音视频信息融合跟踪模块在决策层构建音视频信息的似然函数和重要性采样函数，将视频跟踪结果和声源定位结果融合，最后采用重要性粒子滤波算法对融合信息进行跟踪定位，实现对目标状态的最优估计。

1.1 视频目标检测与跟踪算法

基于视觉的目标检测与跟踪算法可以在没有遮挡的情况下，实现精确度较高的多目标检测与跟踪。研究人员在卷积神经网络的基础上提出一系列目标检测算法[9-12]。为了兼顾模型在复杂环境下的检测能力，同时需要满足系统的实时检测需求，采用 YOLOv5m 作为视频目标检测算法，模型结构如图 2 所示。图像统一缩放至 640×640×3 像素输入，经过特征提取网络和特征金字塔（FPN）后得到 80×80、40×40、20×20 像素 3 个尺度的高层特征。经过非极大值抑制（NMS）处理，得到最优的检测框。

YOLOv5m 卷积层特征提取的流程图如图 3 所示。

目标跟踪算法可以根据目标当前状态预测下一时刻的状态，从而实现目标跟踪任务。为兼顾对非线性目标的跟踪能力，同时考虑到实际应用场景中多目标跟踪任务的实时性需求，本文采用无迹卡尔曼滤波作为基于视频的目标跟踪算法。UKF 在 KF 的基础上，采用 UT 变换 (unscented transformation) 得到 Sigma 点集，计算概率分布的均值

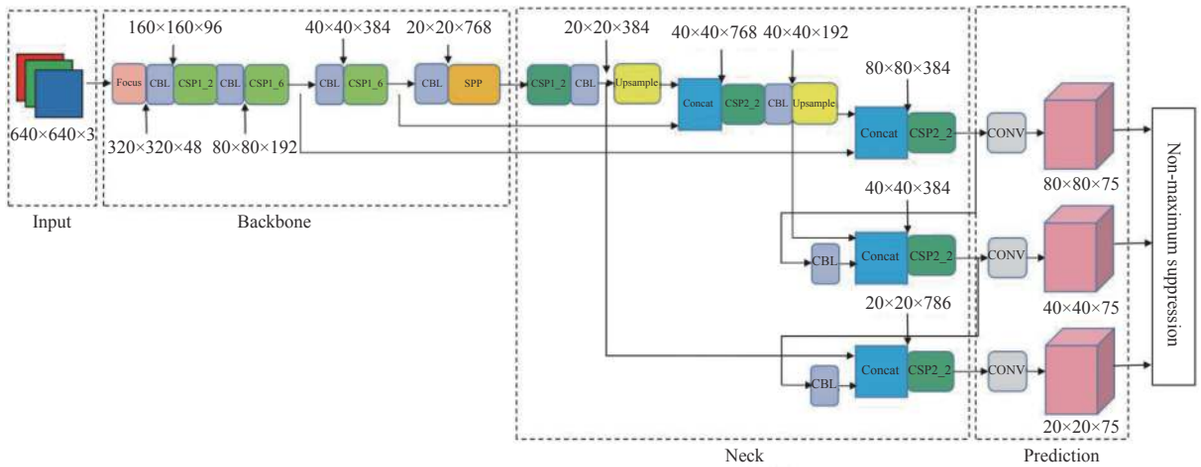


图 2　YOLOv5m 模型结构图

Fig. 2　Structure diagram of YOLOv5m model

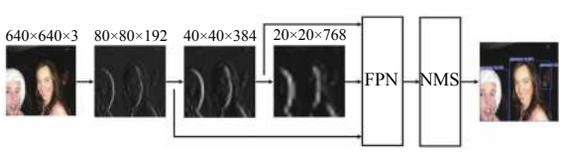


图 3　YOLOv5m 特征提取流程图

Fig. 3　Flow chart of YOLOv5m feature extraction

和协方差, 实现对非线性概率密度分布的近似, 具有精度高、计算量较小等优点[13]。

多目标跟踪任务中需要将经过 UKF 得到的跟踪结果和目标检测结果匹配。本文采用基于 GIoU (generalized intersection over union) 的匈牙利匹配算法进行匹配。GIoU 可用来衡量 2 个框的相交程度, 设检测框为 D, 跟踪框为 T, I 为能将 D 和 T 包含的最小封闭图形, 则 D 与 T 的 GIoU 可用 G 表示为

$$G=\frac{|D\cap T|}{|D\cup T|}-\frac{|I\backslash(D\cup T)|}{|I|} \tag{1}$$

根据 GIoU 可将检测框与跟踪框按如下关系匹配:

$$\begin{cases}(D_i,T_i)=\max\left(\sum_{i=1}^{Q}G_i\right)\\ G_i\geqslant t\end{cases} \tag{2}$$

式中: (D_i,T_i) 表示第 i 个配对的检测框和跟踪框; Q 为总配对数; t 为判定检测框与跟踪框可以配对的阈值。

1.2　声源定位算法

视频目标检测与跟踪可以从视觉层面确定目标的位置, 而基于音频的声源定位可以具体确定发出声音的目标, 并且可以在目标受视觉遮挡时辅助目标的跟踪定位。本文采用 TDOA 算法进行声源定位[14-15], 相比于其他声源定位算法, 基于 TDOA 的算法具有计算量小、实时性高、硬件易于实现等优点。

假设存在 2 个麦克风 M_1 和 M_2, 2 个麦克风接收到的音频信号为 $x_1(t)$ 与 $x_2(t)$, 由 GCC-PHAT 算法估算出 $x_1(t)$ 与 $x_2(t)$ 的时延 τ_{12} 为

$$R_{12}(\tau)=\int_0^{\pi}\frac{X_1(\omega)X_2^*(\omega)}{\left|X_1(\omega)X_2^*(\omega)\right|}\exp(-\mathrm{j}\omega\tau)\mathrm{d}\omega$$
$$\tau_{12}=\tau_1-\tau_2=\arg\max R_{12}(\tau) \tag{3}$$

式中: $X_1(\omega)$ 和 $X_2(\omega)$ 分别是 $x_1(t)$ 和 $x_2(t)$ 的傅里叶变换; $(.)^*$表示复共轭; $R_{12}(\tau)$ 为 $x_1(t)$ 与 $x_2(t)$ 的广义互相关函数。

算得阵列中各个麦克风之间时延后, 就可以结合阵列的空间结构计算声源的方位。本系统采用十字型阵列, 如图 4 所示。

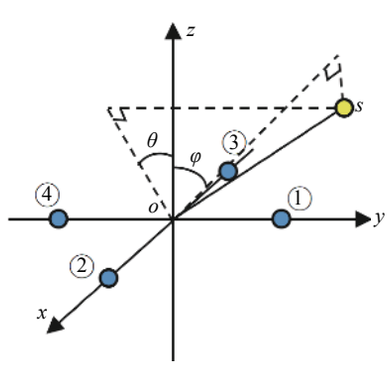


图 4　麦克风阵列空间结构

Fig. 4　Spatial structure of microphone array

图 4 中, ①、②、③、④分别表示阵列的 4 个麦克风, 坐标分别为 $(0,a,0)$、$(a,0,0)$、$(-a,0,0)$、$(0,-a,0)$,

4 个麦克风在空间内呈十字型分布，摄像头位于原点 o 处。声源为 s，其直角坐标表示为 (x,y,z)；os 在平面 xoz 的投影与 z 轴的夹角为 θ，取值范围为 (−90°,90°)；os 在平面 yoz 的投影与 z 轴的夹角为 φ，取值范围为 (−90°,90°)。设声音在空气中传播的速度为 v，且通过 GCC-PHAT 算法计算得到麦克风④与麦克风①的时延为 $\tau_{41}=\tau_4-\tau_1$，麦克风③与麦克风②的时延为 $\tau_{32}=\tau_3-\tau_2$。

通过 τ_{41} 和 τ_{32} 来计算声源的方位角 θ 和 φ：τ_{41} 和 τ_{32} 可以确定声源 s 分别在平面 xoz 和平面 yoz 的投影在 2 个双曲线上。当 $|os|\gg 2a$ 时，可认为声源 s 在平面 yoz 和平面 xoz 的投影位于对应双曲线的渐近线上，由此可计算出 θ 和 φ 为

$$\theta=\begin{cases}\operatorname{arccot}\left(\sqrt{\left(\dfrac{2a}{\tau_{32}v}\right)^2-\left(\dfrac{\tau_{41}}{\tau_{32}}\right)^2-1}\right) & \tau_{32}>0\\ 0 & \tau_{32}=0\\ -\operatorname{arccot}\left(\sqrt{\left(\dfrac{2a}{\tau_{32}v}\right)^2-\left(\dfrac{\tau_{41}}{\tau_{32}}\right)^2-1}\right) & \tau_{32}<0\end{cases}$$

$$\varphi=\begin{cases}\operatorname{arccot}\left(\sqrt{\left(\dfrac{2a}{\tau_{41}v}\right)^2-\left(\dfrac{\tau_{32}}{\tau_{41}}\right)^2-1}\right) & \tau_{41}>0\\ 0 & \tau_{41}=0\\ -\operatorname{arccot}\left(\sqrt{\left(\dfrac{2a}{\tau_{41}v}\right)^2-\left(\dfrac{\tau_{32}}{\tau_{41}}\right)^2-1}\right) & \tau_{41}<0\end{cases} \tag{4}$$

然后将声源方位映射到相机靶面上，设相机靶面的长宽分别为 l_x 和 l_y，像素尺寸为 $l_p\times l_p$，焦距为 f，则可得声源在最终所拍得图像的坐标 (X,Y) 为

$$(X,Y)=\left(\frac{l_x}{2l_p}+\frac{f\tan\varphi}{l_p},\frac{l_y}{2l_p}-\frac{f\tan\theta}{l_p}\right) \tag{5}$$

1.3 音视频信息融合跟踪算法

视频检测与跟踪的精度较高，但易受遮挡等因素影响；声源定位不受视觉场景的影响，但精度较低且易受到噪声的干扰。只采用单一模态对目标进行跟踪存在缺陷，如果同时捕捉被检测目标的视频信息和音频信息，将两种不同模态的信息互补，可以实现精确度更高且更可靠的目标跟踪定位。

由于重要性粒子滤波算法不局限于线性高斯系统，且具有优良的可扩展性和普适性，本文采取重要性粒子滤波[16]作为信息融合方法，基本思想是基于后验概率抽取状态粒子来表示目标概率密度分布，通过对粒子群的加权均值来近似跟踪目标的位置。实际应用中从后验概率抽取样本非常困难，因此引入重要性采样 (importance sampling)，通过重要性采样密度函数抽取样本。

为了将音视频信息融合，首先需要构建音视频信息的似然函数。假设视频似然函数和音频似然函数相互独立，则可通过概率相乘的方式构建音视频信息融合的似然函数为

$$p\left(z_k|x_k^{(i)}\right)=p\left(z_k^{(v)}|x_k^{(i)}\right)p\left(z_k^{(a)}|x_k^{(i)}\right) \tag{6}$$

式中：$x_k^{(i)}$ 为 k 时刻第 i 个粒子的状态；$z_k^{(v)}$、$z_k^{(a)}$ 和 z_k 分别是 k 时刻的视频观测值、音频观测值和融合观测值。考虑到不同应用场景中视频信息和音频信息的可靠程度不同，为了使系统能够在不同的场景中保持良好的鲁棒性，在计算重要性采样函数时需要对视频重要性函数和音频重要性函数加权处理[17]，如 (7) 式所示：

$$\begin{aligned}q\left(x_k^{(i)}|x_{k-1}^{(i)},z_k\right)=&\lambda_v q_v\left(x_k^{(i)}|x_{k-1}^{(i)},z_k^{(v)}\right)+\\&\lambda_a q_a\left(x_k^{(i)}|x_{k-1}^{(i)},z_k^{(a)}\right)\end{aligned} \tag{7}$$

式中：λ_v 和 λ_a 分别是视频重要性函数和音频重要性函数的权值，用来衡量 2 个重要性函数可靠程度。当单一模态失效时，重要性函数仍然可以基于另一模态计算，系统是稳定可靠的。k 时刻的可靠因子 λ_v 和 λ_a 可由 (8) 式计算：

$$\begin{cases}\lambda_v=\displaystyle\sum_{x_k^{(i)}}\sqrt{w_k^{(i)}q_v\left(x_k^{(i)}|x_{k-1}^{(i)},z_k^{(v)}\right)}\\ \lambda_a=\displaystyle\sum_{x_k^{(i)}}\sqrt{w_k^{(i)}q_a\left(x_k^{(i)}|x_{k-1}^{(i)},z_k^{(a)}\right)}\end{cases} \tag{8}$$

式中：$w_k^{(i)}$ 为在 k 时刻第 i 个粒子的权值。当视频或音频重要性函数与后验分布有更多重叠时，说明该模态的信息更加可靠，其可靠因子的值也会更大。

整个重要性粒子滤波算法流程如下：

1）k−1 时刻的粒子集为 $\left\{x_{k-1}^{(i)},w_{k-1}^{(i)}\right\}_{i=1}^{N}$，$x_{k-1}^{(i)}$ 和 $w_{k-1}^{(i)}$ 分别表示 k−1 时刻第 i 个粒子的状态和权值，N 表示粒子个数。

2）转移至 k 时刻，通过 k−1 时刻计算的可靠因子，利用 (7) 式计算重要性采样函数 $q(x_k^{(i)}|x_{k-1}^{(i)},z_k)$，视频重要性函数和音频重要性函数选用 UKF 重要性函数。由重要性采样函数采样得到 k 时刻粒子集 $\left\{x_k^{(i)}\right\}_{i=1}^{N}$。

3）系统观测，由视频和音频似然函数计算融合似然函数：

① 视频似然函数为

$$p\left(z_k^{(v)}|x_k^{(i)}\right)=\frac{1}{\sqrt{2\pi}\sigma_v^2}\exp\left\{\frac{-\left[\left(v_x-x_x^{(i)}\right)+\left(v_y-x_y^{(i)}\right)\right]^2}{2\sigma_v^2}\right\} \tag{9}$$

式中：(v_x,v_y) 是经过 UKF 算法得到视频跟踪目标的二维坐标；σ_v^2 是目标跟踪的观测方差；$\left(x_x^{(i)},x_y^{(i)}\right)$ 是第

i 个粒子的二维坐标。

② 音频似然函数为

$$p\left(z_k^{(a)}|x_k^{(i)}\right)=\frac{1}{\sqrt{2\pi}\sigma_a^2}\exp\left(\frac{-\left[\left(a_x-x_x^{(i)}\right)+\left(a_y-x_y^{(i)}\right)\right]^2}{2\sigma_a^2}\right) \tag{10}$$

式中: (a_x,a_y) 是声源定位结果映射到摄像头像面的二维坐标; σ_a^2是声源定位的观测方差。

③ 由 (6) 式计算音视频融合似然函数。

4) 更新粒子的权值

$$w_k^{(i)}\propto w_{k-1}^{(i)}\frac{p\left(z_k|x_k^{(i)}\right)p\left(x_k^{(i)}|x_{k-1}^{(i)}\right)}{q\left(x_k^{(i)}|x_{k-1}^{(i)},z_k\right)},\sum_{i=1}^{N}w_k^{(i)}=1 \tag{11}$$

5)状态估计为

$$x_k=\sum_{i=1}^{N}w_k^{(i)}x_k^{(i)} \tag{12}$$

6)随机线性重采样, 缓解粒子退化问题。

7)当既没有获得视频信息, 又没有获得音频信息时, 算法结束; 否则返回第 2 步。

2　系统硬件设计

系统硬件如图 5 所示, 硬件部分包括视频采集模块、音频采集模块、数据转接模块和上位机。硬件部分实物图如图 6 所示。

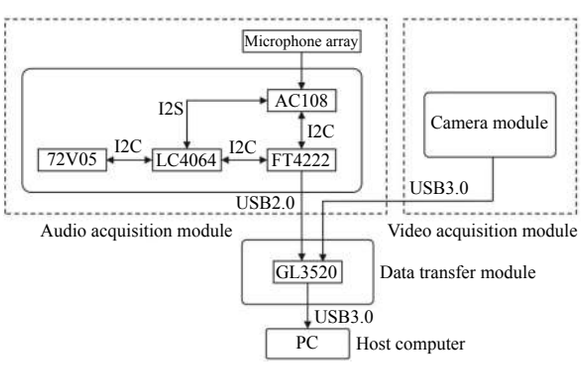


图 5　系统硬件设计框图

Fig. 5　Block diagram of system hardware design

图 6　硬件部分实物图

Fig. 6　Physical drawing of hardware part

视频采集模块采用 1080 P 摄像头, 感光芯片型号为 HM2131(1/2.7″), 像素尺寸为 3 μm×3 μm, 焦距为 4.262 mm, 视场角为 77°, 帧率为 30 fps, 输出 USB 2.0 信号。

音频采集模块采用 4 个硅麦克风作为拾音器件。麦克风阵列采集音频信号, 经由放大电路初步放大后, 再经过 AC108 芯片二次放大并将模拟信号转化为数字信号输出(I2S 信号)。整个音频电路板的逻辑控制、时序控制由 LC4064 芯片控制。通过 LC4064 芯片将 AC108 芯片输出音频数据存储于 72V05 芯片。当 72V05 半满时, LC4064 芯片读取 72V05 芯片中的数据, 再通过 I2C 总线的 SDA 线传输到 FT4222 芯片。FT4222 芯片将 I2C 信号转化为 USB2.0 信号输出。

数据转接模块采用 GL3520 芯片将视频采集模块输出信号和音频采集模块输出信号转化为 USB3.0 信号输出, 实现音视频信息的同步传输。最后将数据传输至上位机进行处理。

上位机硬件环境为 Intel(R) Core(TM) i5-8300H CPU, 主频 2.30 GHz, 内存 16 G, 显卡为 GTX 1060 6G。软件环境为 Windows10 操作系统, 算法在 Python3.7 环境下运行, 深度学习框架为 PyTorch1.5.1。

3　实验结果与分析

3.1　视频目标检测模型训练

考虑到应用场景中需要对生活中常见目标进行检测, 因此选用 VOC2007 数据集作为训练集和测试集。数据集包括人、狗等 20 个类, 随机选取 7000 张图片作为训练集, 剩下 2963 张图片作为测试集。

训练环境: 处理器为 Intel(R) Xeon(R) CPU E5-1620 v2, 主频 3.70 GHz, 内存 32 G, 显卡为 NVIDIA 1080 Ti。软件环境为 Windows10 操作系统, 算法在 Python3.7 环境下运行, 深度学习框架为 Pytorch1.5.1。

YOLOv5m 模型在训练中迭代次数 epoch 设置为 150 次, 模型初始学习率为 0.000 1, 采用 GIoU_Loss 作为损失函数。模型训练结果随迭代次数增加在测试集上的表现如图 7 所示。GIoU_Loss、mAP50 分别表示训练的模型在测试集上的 GIoU 损失和平均准确率。从图中可以看出, 在前 50 个 epoch, 模型的 mAP50 整体处于上升态势; 当训练至第 50 epoch 后, 模型整体趋于稳定。mAP50 达到了 86.9%, 虚警概率为 14.9%, 模型训练效果良好。

目标检测效果如图 8 所示.

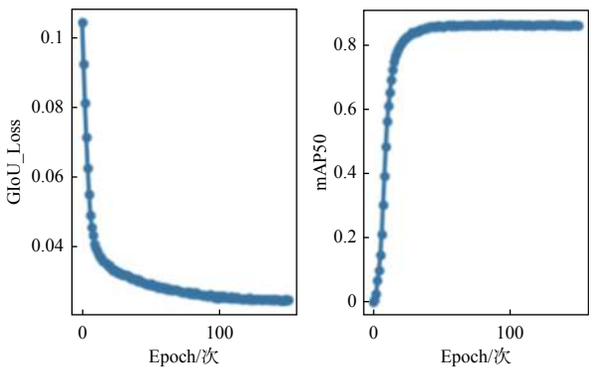


图 7 YOLOv5m 训练结果图

Fig. 7 Training result chart of YOLOv5m

图 8 目标检测效果图

Fig. 8 Renderings of object detection

将 YOLOv5m 检测结果作为系统量测值，对目标进行 UKF 跟踪。目标状态向量的初始化为 $\hat{x}_0$，误差协方差矩阵初始化为 $\boldsymbol{P}_0$，过程噪声矩阵为 $\boldsymbol{Q}_k$，量测噪声矩阵为 $\boldsymbol{R}_k$，有：

$$\hat{x}_0 = [x_0, v_{x,0}, y_0, v_{y,0}]^{\mathrm{T}} = [x_0, 0, y_0, 0]^{\mathrm{T}} \tag{13}$$

$$\boldsymbol{P}_0 = \mathrm{diag}\left(10^3, 10, 10^3, 10\right) \tag{14}$$

$$\boldsymbol{Q}_k = \mathrm{diag}\left(0.05, 0.05\right) \tag{15}$$

$$\boldsymbol{R}_k = \mathrm{diag}\left(20^2, 20^2\right) \tag{16}$$

(13) 式中 x_0 和 y_0 为量测坐标。状态转移矩阵 $\boldsymbol{F}$ 为

$$\boldsymbol{F} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{17}$$

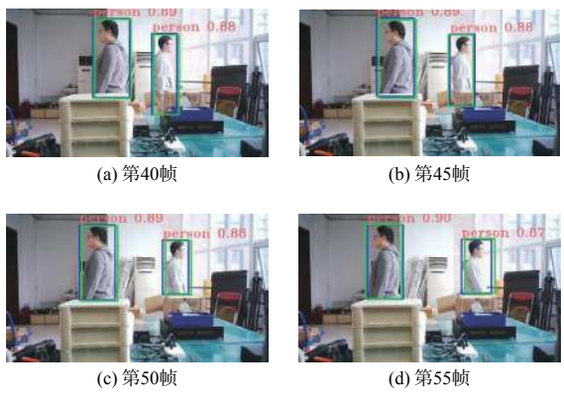
利用 UT 变化获得 Sigma 点集，通过 $\boldsymbol{F}$ 传播后得到新 Sigma 点集的预测与协方差矩阵，最后利用量测更新系统状态和误差协方差矩阵。

在室内环境对算法性能进行测试，如图 9 所示。图 9 中蓝色框为检测框，绿色框为跟踪框。采用基于 GIoU 的匈牙利算法将跟踪框与检测框匹配，设置 (2) 式的匹配阈值 $t = 0.3$，匹配结果如表 1。

在室内环境下录制 8 min 的音视频，对其中视频信息进行基于 YOLOv5m+UKF 的检测与跟踪实验，统计得检测框与跟踪框的匹配率为 98.7%。

3.2 声源定位实验

为了验证图 2 所设计麦克风阵列在声源定位任务中的性能，设计如下实验。为了兼顾定位精确度和整个阵列轻便性，设置 a = 28.28 mm，即 4 个麦克风呈正方形排布。由于麦克风阵列呈对称排布，只需在三维空间的第一象限设置声源验证系统性能即可。声源分别设置在 (0,0,500)、(0,0,800)、(0,200,400)、(0,500,500)、(200,0,500)、(200,500,500)、(150,400,200) 共 7 处，单位为 mm。将喇叭放置于声源点位播放语音对话，每组录制 10 段音频。采样率为 48 kHz，信噪比约为 15 dB。以声源位于 (0,0,500) 其中一次实验为例，给出麦克风①④所接收的音频信号波形以及二者的 GCC 函数波形，如图 10 所示.



(a) 第40帧 (b) 第45帧 (c) 第50帧 (d) 第55帧

图 9 YOLOv5m+UKF 实验效果

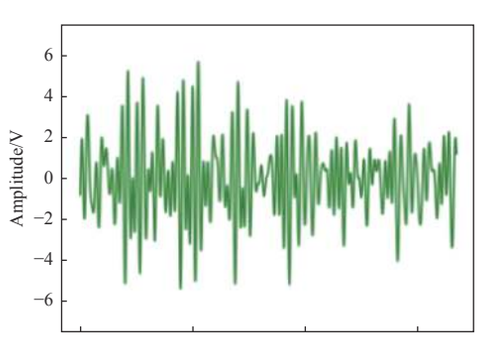
Fig. 9 Experimental effect of YOLOv5m + UKF

表 1 跟踪框与检测框匹配效果

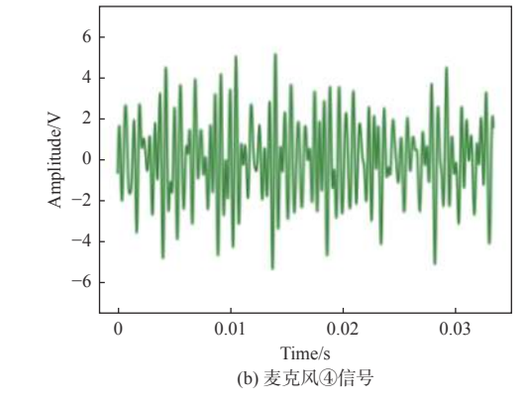
Table 1 Matching effect of tracking box and detection box

帧数序号	40	45	50	55
左侧目标GIoU值	0.918	0.847	0.859	0.856
右侧目标GIoU值	0.842	0.863	0.804	0.796

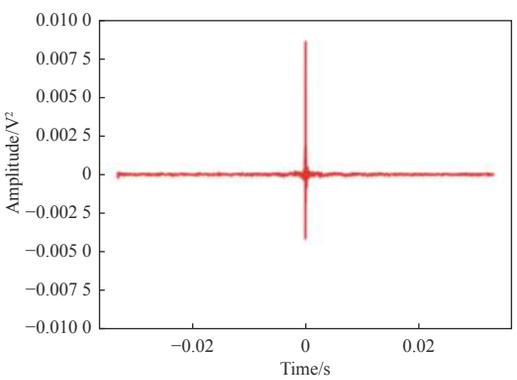
由图 10(c) 可得时延为 0.05 个采样点，也就是 1.042×10^{-6} s。采用 1.2 节声源定位算法计算得方位角 θ 和 φ，并统计 θ 和 φ 误差的绝对值均值，如表 2 所示。



(a) 麦克风①信号



(b) 麦克风④信号



(c) 麦克风①④信号的GCC函数

图 10　音频信号波形图

Fig. 10　Audio signal waveform

由表 2 所示, 2 个方位角中 φ 表示水平方位角, θ 表示俯仰角。从整体上看, 平均误差在 2°～3.5°之间, 定位精度整体上满足系统需求。

3.3　音视频信息融合的跟踪实验

为了验证音视频融合跟踪算法的效果, 使用

表 2　方位角 θ 和 φ 平均误差

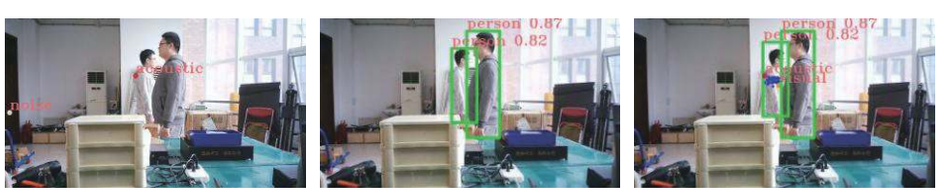
Table 2　Average error of azimuth θ and φ

声源坐标/ mm	θ平均误差/ (°)	φ平均误差/ (°)
(0,0,500)	3.418	2.891
(0,0,800)	3.254	2.724
(0,200,400)	2.921	2.230
(0,500,500)	3.396	2.486
(200,0,500)	2.855	2.492
(200,500,500)	2.423	2.362
(150,400,200)	2.827	2.069

图 4 所示的摄像头与十字型麦克风阵列 (a=28.28 mm), 在室内复杂环境下录制总时长为 8 min 的音视频。视频中包括诸如多人对话、两人重叠、杂物遮挡目标等场景; 音频中存在脚步声、碰撞声等噪声干扰。视频定位观测方差为 900 个像素, 音频定位观测方差为 2 500 个像素, 粒子个数为 50。

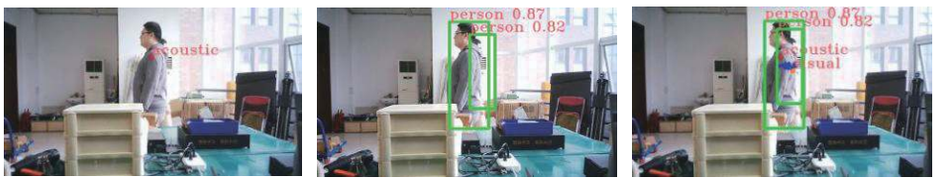
对录制的数据分别进行声源定位、基于视频的检测跟踪(YOLOv5m+UKF)、音视频融合的检测跟踪实验。跟踪结果对比如图 11 所示。

图 11 为 3 种算法在一段音视频的效果对比。视频中两人相对走过, 从左向右移动的人(称为 A)边走边说话, 从右向左走动的人(称为 B)不发出声音。图 11 左侧为声源定位结果, 用红点表示声源定位, 白点表示噪声大致位置; 中间为 YOLOv5m+UKF 跟踪结果, 用绿框表示; 右侧为音视频融合的检测跟踪结果, 红点和黄点分别表示 A 的声源定位和视觉跟踪结果, 蓝点为粒子滤波点集, A 的绿


(a) 第 67 帧


(b) 第 77 帧


(c) 第 80 帧



(d) 第 86 帧

(e) 第 96 帧

图 11　3 种算法跟踪效果对比

Fig. 11　Comparison of tracking effect of three algorithms

框为融合跟踪结果，由于 B 不发声，B 的绿框为视觉跟踪结果。图 11 中各帧的声源定位、A 的视觉跟踪、A 的融合跟踪(加权重要性函数)、A 的融合跟踪(不加权重要性函数)以及 A 人工标定的真实位置的具体坐标如表 3 所示。

表 3　3 种算法跟踪坐标数值

Table 3　Tracking coordinate values of three algorithms

帧数序号	声源定位	A视觉跟踪	A融合跟踪（加权）	A融合跟踪（不加权）	A标定位置
67	(833,366)	(923,427)	(903,387)	(869,378)	(911,425)
77	(925,353)	(1002, 426)	(972, 397)	(961,372)	(945,422)
80	(953,322)	(1029, 428)	(989, 374)	(982,361)	(961,422)
86	(950,320)	(1058, 429)	(1009, 378)	(998,369)	(982,422)
96	(982,501)	(1085, 476)	(1076, 483)	(1039,494)	(1042,484)

人工标注数据集中被遮挡目标的真实框，用 G 代表跟踪框与真实框的 GIoU，采用 1-G 衡量跟踪结果与真实结果的误差。由于声源定位结果只有坐标没有跟踪框，因此用该目标视觉跟踪框的尺寸作为声源定位框的尺寸。图 11 中的音视频 67 帧~96 帧的误差曲线如图 12 所示。

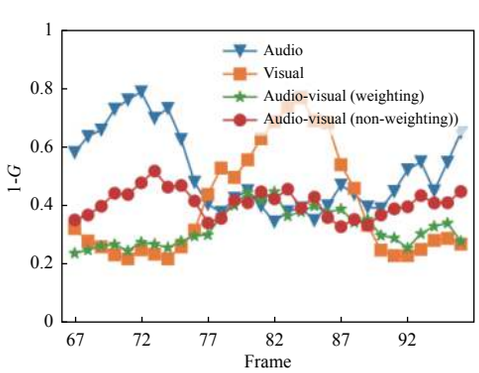


图 12　音视频序列误差曲线图

Fig. 12　Error curves of audio and video sequences

分析可知，声源定位精度相对视频跟踪较低，当噪声较大时会使声源定位结果偏离目标。视频跟踪在没有遮挡时具有较高的精度，但是当目标被遮挡时，跟踪定位的精度会下降。音视频融合跟踪算法可以有效抑制单一模态的失效，增强跟踪系统的鲁棒性。采用加权音视频重要性函数的融合跟踪可以自适应调节音频和视频的可靠度，具有更强的抑制噪声的能力。

当跟踪框与真实框满足：

$$1-G<0.5 \tag{18}$$

则认为该目标的跟踪结果准确，否则认为误跟。统计 3 种算法在所有音视频序列的跟踪准确率和平均每帧运行时间，如表 4 所示。融合检测跟踪的准确率 90.68% 远高于声源定位的 74.48% 和视频检测跟踪的 83.46%，平均每帧运行时间 29.2 ms 小于视频每帧间隔时间 33.3 ms。

表 4　3 种算法性能对比

Table 4　Performance comparison of three algorithms

算法	准确率/%	平均每帧运行时间/ms
声源定位	74.48	4.1
视频检测跟踪	83.46	23.0
融合检测跟踪	90.68	29.2

4　结论

本文提出一种音视频信息融合的检测与跟踪算法框架，并设计了音视频采集的硬件设备。采用 YOLOV5m 作为目标检测框架，使用 UKF 算法对多目标跟踪，使用匈牙利算法匹配检测与跟踪结果；采用 GCC-PHAT 作为时延估计算法，采用十

字形麦克风阵列实现声源定位; 在粒子滤波的基础上, 构造音视频似然函数和音视频重要性函数, 对音视频信息进行融合跟踪。经验证, 算法提高了跟踪的精确度与可靠性, 跟踪准确率为 90.68%, 高于声源定位和视频检测跟踪的准确率。

本算法所采用的声源定位算法只能定位一个声源。当所处环境同时存在多个发声目标时, 系统跟踪性能会下降, 后续将改进声源定位算法以实现多声源定位。

参考文献:

[1] 尹宏鹏, 陈波, 柴毅, 等. 基于视觉的目标检测与跟踪综述[J]. 自动化学报, 2016, 42(10): 1466-1489.
YIN Hongpeng, CHEN Bo, CHAI Yi, et al. Vision-based object detection and tracking: a review[J]. Acta Automatica Sinica, 2016, 42(10): 1466-1489.

[2] 许婉君, 侯志强, 余旺盛, 等. 基于颜色和空间信息的多特征融合目标跟踪算法[J]. 应用光学, 2015, 36(5): 755-761.
XU Wanjun, HOU Zhiqiang, YU Wangsheng, et al. Fusing multi-feature for object tracking algorithm based on color and space information[J]. Journal of Applied Optics, 2015, 36(5): 755-761.

[3] 邵辰琳, 杨卫平, 张志龙. 基于简单线性迭代聚类超像素的meanshift跟踪[J]. 应用光学, 2017, 38(2): 193-199.
SHAO Chenlin, YANG Weiping, ZHANG Zhilong. Meanshift tracking algorithm based on SLIC superpixel[J]. Journal of Applied Optics, 2017, 38(2): 193-199.

[4] 崔玮玮, 曹志刚, 魏建强. 声源定位中的时延估计技术[J]. 数据采集与处理, 2007, 22(1): 90-99.
CUN Weiwei, CAO Zhigang, WEI Jianqiang. Time delay estimation techniques in source location[J]. Journal of Data Acquisition and Processing, 2007, 22(1): 90-99.

[5] 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.

[6] CHEN Y, RUI Y. Real-time speaker tracking using particle filter sensor fusion[J]. IEEE, 2004, 92(3): 485-494.

[7] 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.
LI Xin. Human tracking based on audio visual information fusion and its application[D]. Beijing: Tsinghua University, 2005.

[8] 谢静. 基于音视频融合的定位跟踪算法[D]. 天津: 天津大学, 2009.
XIE Jing. Algorithm of localization and tracking based on audio and visual Fusion[D]. Tianjin: Tianjin University, 2009.

[9] GIRSHICK R, DONAHUE J, DARREL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE: 2014: 580-587.

[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems.USA:IEEE, 2015: 91-99.

[11] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2016: 21-37.

[12] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2017: 7263-7271.

[13] 石勇, 韩崇昭. 自适应UKF算法在目标跟踪中的应用[J]. 自动化学报, 2011, 37(6): 755-759.
SHI Yong, HAN Chongzhao. Adaptive UKF method with applications to target tracking[J]. Acta Automatica Sinica, 2011, 37(6): 755-759.

[14] 行鸿彦, 杨旭, 张金玉. 基于四元传声器阵列的声源全方位定位算法[J]. 仪器仪表学报, 2018, 39(11): 43-50.
XING Hongyan, YANG Xu, ZHANG Jinyu. Sound source omnidirectional location algorithm based on four-element microphone array[J]. Chinese Journal of Scientific Instrument, 2018, 39(11): 43-50.

[15] 孙建红, 张涛, 焦琛. 麦克风数量与阵型对声源定位性能的影响[J]. 电子测量与仪器学报, 2019, 33(11): 14-21.
SUN Jianhong, ZHANG Tao, JIAO Chen. Influence of array and the number of microphones on the localization performance of sound source[J]. Journal of Electronic Measurement and Instrumentation, 2019, 33(11): 14-21.

[16] 昝孟恩, 周航, 韩丹, 等. 粒子滤波目标跟踪算法综述[J]. 计算机工程与应用, 2019, 55(5): 14-23+65.

ZAN Meng'en, ZHOU Hang, HAN Dan, et al. Survey of particle filter target tracking algorithms[J]. Computer Engineering and Applications, 2019, 55(5): 14-23+65.

[17] 曹洁, 郑景润. 音视频信息融合的说话人跟踪算法研究[J]. 计算机工程与应用, 2012, 48(13): 118-124.
CAO Jie, ZHENG Jingrun. Speaker tracking based on audio-video information fusion[J]. Computer Engineering and Applications, 2012, 48(13): 118-124.