

Human Localization Based on Audio-Video System on Mobile Robot

Kai Wen, Junchuan Liu

School of Mechanical and Automation, Beihang University, Beijing

Email: kwenbuaa@gmail.com

Received: Oct. 25th, 2012; revised: Nov. 11th, 2012; accepted: Nov. 26th, 2012

Abstract: In this paper, we introduce an audio-video system to help mobile robot to locate human. For the position of human, the auditory system of robot is used to locate the sound azimuth. Then, the survey uses video system to detect the abscissa of human face in the video for horizontal angle compensation; the robot is then guided to the user. The innovation of this paper is to put forward a kind of method to locate target people using a quaternary cross microphone array and monocular vision. The off-line results show that the audio-video system is feasible to locate human in the intelligent information household robot.

Keywords: Mobile Robot; Sound Source Localization; Face Detection

基于移动机器人听觉 - 视觉的目标人定位

温 凯, 刘军传

北京航空航天大学机械工程及自动化学院, 北京

Email: kwenbuaa@gmail.com

收稿日期: 2012 年 10 月 25 日; 修回日期: 2012 年 11 月 11 日; 录用日期: 2012 年 11 月 26 日

摘 要: 介绍了一种用于移动机器人搜索目标人方位的听觉和视觉一体的定位方法。针对目标人的位置, 利用机器人听觉定位系统进行声源水平方向初定位。在此基础上, 用视觉检测人脸中心在摄像头视野中的位置, 进行水平角度补偿, 引导机器人走向目标人。本文的创新点在于提出了一种四元十字麦克风阵和单目视觉联合对目标人进行定位的方法。离线实验表明, 这种声视一体的定位方法在家用机器人目标人定位上是可行的。

关键词: 移动机器人; 声源定位; 人脸检测

1. 引言

近年来, 融合平板电脑技术和移动机器人技术的智能信息家用机器人已经开始进入人们的生活。用户使用机器人的信息服务时, 经常需要走到机器人前去操作, 非常不便。为方便使用, 在用户召唤时让机器人自己走到用户前, 是很有意思也很有必要的。

对目标人的定位是实现机器人响应用户召唤走到用户面前的基础, 利用声源或视觉对目标人进行定位是一种便捷有效的方法, 得到了国内外众多研究者

的关注。

Huang 等人^[1]利用安装在机器人头部成正四面体结构的四个麦克风进行声源轴向角和仰角的定位, 互相关函数和互功率谱相位差分别被用于时延估计。

Valin 等人^[2]开发研制了避障机器人。他们在机器人上搭载了 8 个麦克风, 机器人具有鲁棒的定位性能和通过使用波束成形和粒子滤波能跟踪几个同时声源。

Ishi^[3]对机器人在办公室环境和室外环境下的声

源定位进行了研究。用到的麦克风阵列是十四通道的，声源定位方法是多重信号分类方法(MUSIC)。

国内邵怀宗等^[4]基于麦克风均匀圆阵，采用改进的 MUSIC 算法对近场声源进行定位。顿向明等人^[5]用五元麦克风阵列和改进的时延估计算法来定位储罐内爬壁机器人的位姿。15 m 内该系统定位距离误差在 12 cm 内，方向角误差在 15° 以内。

在上述声源定位方法中，五元麦克风阵列和十四通道阵用到的麦克风阵元数量较多，不便在家用机器人上安装。采用八通道麦克风阵列和正四面体结构可获得较高的定位精度，但对机器人上的安装空间和安装精度提出了更高的要求。

Okuno 和 Nakadai^[6]等融合听觉事件形成的听觉流与视觉事件形成的视觉流生成联合流，以控制 SIG 机器人注意力的转移，听觉用的是双耳听觉，视觉用了两个 CCD 摄像头。听觉事件为声源方向估计，视觉事件为多人人脸检测。

河北工业大学的陈涛、张明路等人^[7]，用双耳听觉系统确定声源的粗方位，然后用双目视觉系统进行水平方位角误差补偿，垂直标高角的评估和声源距离的评估。

使用双目视觉定位需要对摄像机进行精确标定，并且对图像处理和配准有很高的要求，在家庭环境中使用很容易受到环境因素的干扰。

本文在移动机器人上设计了一个四通道十字型的麦克风阵列来获取目标人的水平方位角，定位精度要在 15° 内，以便机器人声源定位，转过相应角度后，目标人出现在机器人视野中。然后用单目摄像头检测目标人脸在摄像头图像中的位置。视觉检测结果用来对听觉定位进行补偿，具体检测流程如图 1 示。

本文离线验证了声源定位和人脸检测算法，探索了声视一体定位算法的可行性。

2. 声视一体定位原理

2.1. 听觉系统对声源的定位

2.1.1. 定位原理

麦克风阵列定位的基本原理是将麦克风在空间布置成一定几何形状的阵列，通过检测和计算各麦克风所得信号的时延，然后根据时延估计值和阵元几何关系来确定目标的位置。

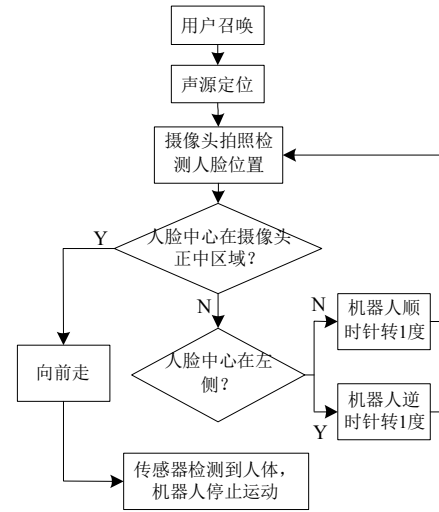


Figure 1. The process of localization based on audio-video system
图 1. 声视一体定位流程

2.1.2. 时延估计算法

本文采用广义互相关算法(GCC)^[8]作为时延估计算法。

麦克风 i 和 j 在 t 时刻接收到的信号幅值 $x_i(t)$ 和 $x_j(t)$ 为：

$$x_i(t) = \alpha_i s(t - \tau_i) + n_i(t) \quad (1)$$

$$x_j(t) = \alpha_j s(t - \tau_j) + n_j(t) \quad (2)$$

其中， $s(t)$ 为声源在 t 时刻接收到的信号幅值， α_i 、 n_i 和 τ_i 是声源到 i 麦克风的幅值衰减系数，加性噪声和时延值。 α_j 、 n_j 和 τ_j 是声源到 j 麦克风的幅值衰减系数，加性噪声和时延值。接收信号 $x_i(t)$ 和 $x_j(t)$ 的互相关函数 $R_{ij}(\tau)$ 为：

$$R_{ij}(\tau) = E(x_i(t)(x_j(t - \tau))) \quad (3)$$

由互相关函数与互功率谱的关系^[7]可得：

$$R_{ij}(\tau) = \int_0^\pi G_{ij}(\omega) e^{j\omega\tau} d\omega \quad (4)$$

两信号的广义互相关函数^[7]可表示为：

$$R_{ij}^g(\tau) = \int_0^\pi \psi_{ij}(\omega) G_{ij}(\omega) e^{j\omega\tau} d\omega \quad (5)$$

式 5 中 $\psi_{ij}(\omega)$ 为加权函数，本文加权函数为 $\psi_{ij}(\omega) = 1/|G_{ij}(\omega)|$ ，在频域乘以加权函数可以削弱噪声的影响，起到了锐化互相关函数和抗噪声效果。由信号的相关性可知，求得 $R_{ij}(\tau)$ 的峰值时刻对应的 τ 值，就是 i, j 两麦克风的时延值。

2.1.3. 四元十字位置估计算法

出于成本、定位精度和复杂度的考虑, 本文选择四元十字麦克风阵列, 如图 2 所示。

M1~M4 表示 1~4 号麦克风, 坐标系以 M1~M4 所在平面为 XOY 平面, 垂直麦克风阵列中心向上的轴为 Z 轴, 规定在麦克风阵列 XY 平面, 以 X 正方向为始边, 逆时间为正方向, 顺时针为负方向。声源的位置为 $S(x, y, z)$, 各阵元到中心间距 d 为 20 cm, $L_i (i=1 \sim 4)$ 表示声源到 i 麦克风的距离。 r 和 θ 分别表示声源到麦克风阵列中心的距离和水平角。在三维坐标系中, 利用几何关系, 可得下列方程组:

$$\begin{cases} x^2 + y^2 + z^2 = r^2 \\ (x-d)^2 + y^2 + z^2 = L_1^2 \\ x^2 + (y-d)^2 + z^2 = L_2^2 \\ (x+d)^2 + y^2 + z^2 = L_3^2 \\ x^2 + (y+d)^2 + z^2 = L_4^2 \\ L_2 - L_1 = \tau_{12}c \\ L_3 - L_1 = \tau_{13}c \\ L_4 - L_1 = \tau_{14}c \end{cases} \quad (5)$$

上式中, c 是常温下声速, τ_{ij} 表示声音到达麦克风 i 和 j 的时延, 当声源在不同的象限时, 解方程组(5)可得:

当 θ 在第一象限, 即 $\tau_{13} \leq 0$, $\tau_{24} = \tau_{14} - \tau_{12} < 0$ 时, 或 θ 在第四象限, 即 $\tau_{13} < 0$, $\tau_{24} = \tau_{14} - \tau_{12} \geq 0$ 时,

$$\theta = \arctan \frac{\tau_{14} - \tau_{12}}{\tau_{13}} \quad (6)$$

当 θ 在第二象限, 即 $\tau_{13} > 0$, $\tau_{24} = \tau_{14} - \tau_{12} \leq 0$ 时,

$$\theta = \arctan \frac{\tau_{14} - \tau_{12}}{\tau_{13}} + 180^\circ \quad (7)$$

当 θ 在第三象限, 即 $\tau_{13} \geq 0$, $\tau_{24} = \tau_{14} - \tau_{12} > 0$ 时,

$$\theta = \arctan \frac{\tau_{14} - \tau_{12}}{\tau_{13}} - 180^\circ \quad (8)$$

由式(6)~(8)可看出, 声源三维坐标 (x, y, z) 在水平方位角计算过程中被抵消了。水平方位角可用各麦克风间的时延值 $R_{ij}(\tau)$ 求得。此方程组还可以求出声源到阵元中心的距离和声源的俯仰角。但本系统主要应用的是水平角, 所以距离和俯仰角不做研究。

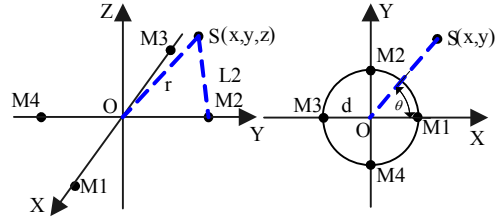


Figure 2. Microphone array
图 2. 麦克风阵列

2.2. 视觉人脸检测

人脸检测就是从各种不同图像中确定人脸的存在, 并且确定人脸的位置和数量。2001 年 Viola 和 Jones^[9]提出了 Adaboost 人脸检测方法。他们的方法是检测正确率最高, 速度最快的算法之一。

2.2.1. Adaboost 人脸检测原理

Adaboost 人脸检测方法是一种基于积分图、分类器和 Adaboost 算法的方法。

1) 特征的选取, 特征值和分类器

在人脸检测过程中, 需要对候选图像进行分析, 判断是否为人脸, 多数人脸检测系统都是使用特征对人脸模式进行建模, 这些特征都应有一定的人脸和非人脸区分性。Adaboost 用于人脸检测时, 需要从人脸中抽取大量的简单特征。本算法选择由 Rainer Lienhart^[10]等人提出的扩展的 Haar 型特征, 如图 3~5 所示。

Haar 型特征值的定义是黑色矩形和白色矩形在图像子窗口中对应的区域的灰度级总和之差, 而积分图法是一种快速计算特征值的方法。



Figure 3. Edge feature
图 3. 边缘特征



Figure 4. Surrounded feature
图 4. 环绕特征



Figure 5. Line feature
图 5. 线特征

分类器是 Viloa^[11]等人通过计算每个 Haar 特征的特征值是否达到人脸域值,从而判定一个图形区域是否为人脸区域的规则。

2) Adaboost 算法

Adaboost 算法的基本思想就是将大量分类能力一般的弱分类器通过一定方法叠加(boost)起来,构成一个分类能力很强的强分类器,再将若干个强分类器串联成为分级分类器完成图像搜索检测。每一层分类器经过 Adaboost 算法训练后都能让几乎全部正例样本通过,而拒绝很大一部分非人脸样本。最终通过所有层筛选的样本被接受为人脸窗口。检测过程如图 6 所示。

2.2.2. 实现方法

本文的人脸检测是在 Windows 下利用 VS2008,以 OpenCV2.3.1 为基础,利用库中的相关函数为基础进行处理。

OpenCV 的人脸检测函数采用了 Viola & Jones 人脸检测方法,主要是调用训练好的瀑布级联分类器 cascade 来进行模式匹配。

2.2.3. 脸部检测流程

OpenCV 人脸检测流程如图 7 所示。

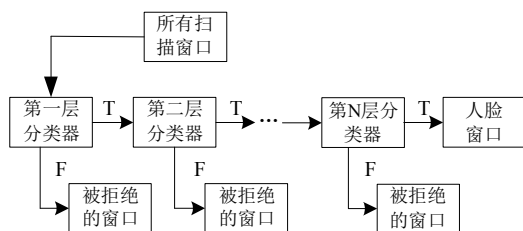


Figure 6. The detection cascade
图 6. 级联检测器

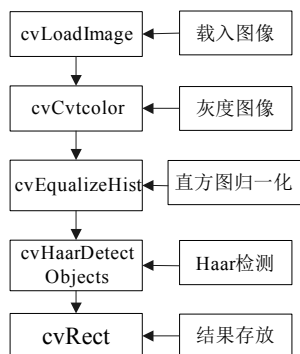


Figure 7. Process of OpenCV face detection
图 7. OpenCV 脸部检测流程图

首先输入图像,当检测到脸部后需利用 `cvCvtColor()`转化为灰度图像,利用 `cvEqualizeHist()`进行直方图归一化处理。再用 `cvLoad()`从文件中加载人脸检测模版 `CvHaarClassifierCascade` 变量,然后利用 `cvHaarDetectObjects()`来进行检测,函数使用针对某目标物体训练的级联分类器在图像中找到包含目标物体的矩形区域和位置,并且将这些区域和位置作为一序列的矩形框返回,最终检测结果保存在 `cvRect` 变量中。

3. 离线定位实验

3.1. 声源定位实验

麦克风实验数据采集是在长 10 m 宽 5 m 高 4 m 的走廊大厅里进行的。麦克风阵列在机器人上的布置如图 8 所示。人站在距离机器人 3 m 处,当采集卡上的指示灯亮时,对机器人说话,机器人上的声音采集卡采集 1 秒种人说话的声音,过程如图 9 所示。

采集硬件是置于机器人上自制的 STM32 单片机采集卡。软件程序是用定时器控制采集间隔,每 100 us 进入一次中断,在中断中用 DMA 将四路 AD 转化后的麦克风信号存储到 FLASH 中。

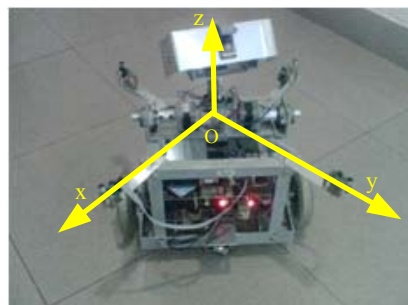


Figure 8. Robot prototype
图 8. 机器人样机

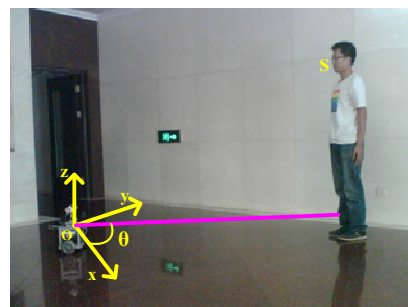


Figure 9. Sound acquisition
图 9. 声音采集过程

离线处理时，将 FLASH 中的数据通过串口传到 PC 机上，在 Matlab 上将采到的四路语音信号分别代入式(1)~(5)中求得时延值，再将时延值代入式(6)~(8)中即可求得目标人水平方位角。处理过程的图示如图 10 和 11。声源定位结果如表 1 所示。

图 10 显示的是目标人在阵元中心 135 度时，各麦克风采集到的声音信号幅值。图 11 是 1，4 麦克风互相关函数图，求出图中互相关函数峰值时刻与中心点的点数并除以采样率，即可得到时延值。

表 1 中人的位置 表示目标人相对于机器人坐标系的水平方位角，计算位置 是由麦克风阵列计算出的目标人的水平方位角。由表 1 对数据结果进行分析，可以看出，误差基本在 10 度以内，满足设计目标。经过分析，走廊里人走路脚步声以及单片机的 AD 转化精度等，均会对声源定位的结果造成影响。

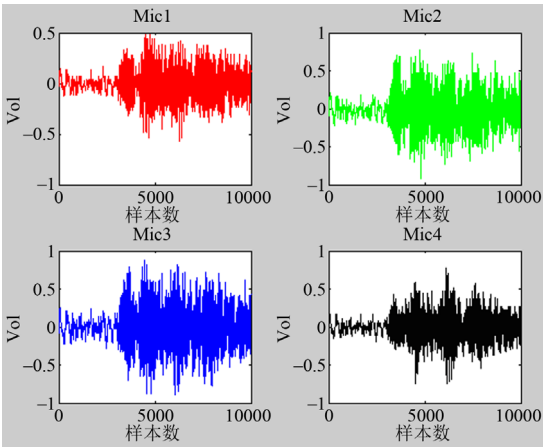


Figure 10. Voice signal collected by every microphone
图 10. 各麦克风收到的语音信号

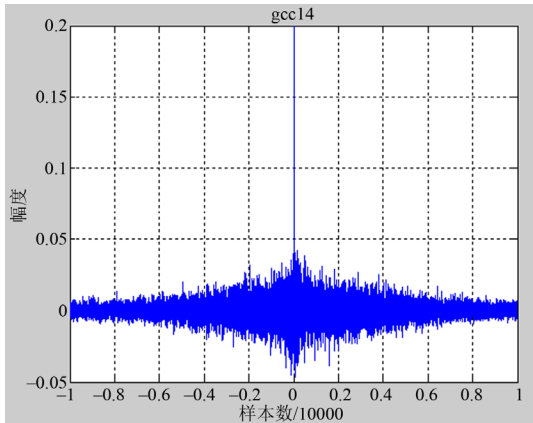


Figure 11. Cross-correlation function of microphone 1 and microphone 4
图 11. 麦克风 1 与 4 的互相关函数图

Table 1. Experiment result
表 1. 测试结果

人的位置 $\theta(^{\circ})$	0	30	45	60	90	120
计算位置 $\theta(^{\circ})$	0	23.2	45	60.3	90.0	126.9
人的位置 $\theta(^{\circ})$	135	150	180	-30	-45	
计算位置 $\theta(^{\circ})$	130.6	144.5	180	-26.6	-39.8	
人的位置 $\theta(^{\circ})$	-60	-90	-120	-135	-150	
计算位置 $\theta(^{\circ})$	-58.0	-90	-119.7	-135	-148.0	

3.2. 人脸检测结果

在 Matlab 上求出声源定位角后，将此角度输给机器人，控制机器人转过相应角度。下一步用机器人摄像头拍摄照片，将所拍照片在 PC 机上进行人脸检测，求出人脸位置和区域，补偿声源定位的水平角。

人脸检测用到的图片是用机器人操作系统自带的摄像头拍摄所得。摄像头是 OV9650 的 130 万像素 CMOS 摄像头，所拍图片尺寸是 1280 × 960。人脸检测的过程，是将采集到的照片放到 PC 机上，在 VS2008 下用 OpenCV2.3.1 处理。

几种不同背景和光照条件下的检测结果如图 12(a)~(d)所示，图像下面的数字表示图像在水平方向的坐标值，单位是像素点数。图 12(a)~(d)中人脸中心水平坐标值分别是 617、726、525 和 519，红圈中的区域表示检测到的人脸区域。

结果表明，视觉检测算法在不同光照、背景繁简和人脸正对摄像头并且没有遮挡的情况下，均可以检测到人脸区域和位置。

本文设定摄像头图像左边区域是横坐标在 0~600 像素点内的区域，中心区域为图像横坐标在 600~680 像素点内的区域，右边区域为图像横坐标在 680~1280 像素点内的区域。

按照图 1 中的流程，当检测到人脸中心横坐标后，就可以根据检测到的坐标值调整机器人姿态了。图 12(a)中人脸中心坐标为 617，在图像正中区域，所以机器人不用校正，向前走；图 12(b)中人脸中心坐标为 726，在图像右边区域，机器人需顺时针方向转动校正，图 12(c)和(d)中人脸中心分别为 525 和 519，人脸中心在图像左边区域，机器人需逆时针方向转动校正，待目标人出现在视野正中区域后，则向前走，当传感器检测到人体后，机器人停止运动。

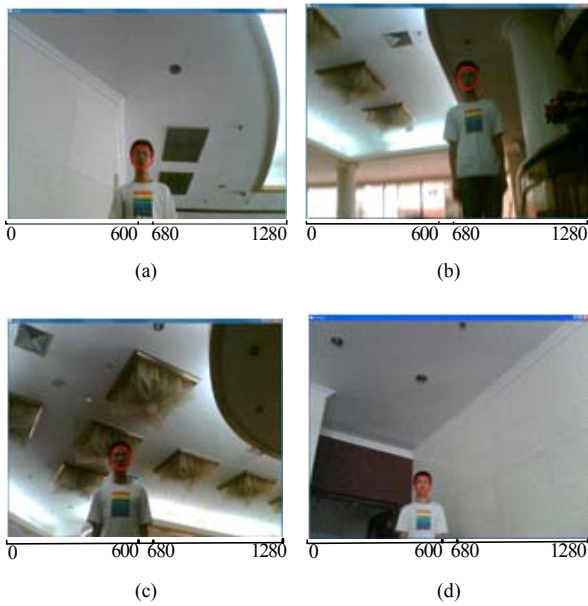


Figure 12. Results of face detection: (a) Bright light; (b) Dark light; (c) Complex background; (d) Simple background
图 12. 人脸检测结果: (a) 光线明亮; (b) 光线较暗; (c) 背景复杂; (d) 背景简单

4. 结论与展望

本文的实验是将机器人摄像头上拍摄的图像和机器人上的 STM32 单片机采集到的麦克风数据，离线在 PC 机上处理的。通过对麦克风阵列的离线定位结果分析，可以看出，四元十字的麦克风阵列，可以满足移动机器人对目标人在水平方位角的初步定位。在此基础上，基于 OpenCV 的人脸检测系统，可以较好地检测出人脸中心位置，以便对声源定位结果进行补偿。离线实验表明，此定位方法可以用于家用移动机器人目标人的定位。

本文用离线的方式探索了采用声视觉联合定位

方法对目标人进行定位的可行性，下一步工作是在机器人上位机 Tiny210 上采集麦克风的语音，在 Tiny210 的 Linux 系统上实时定位声源，并将人脸检测程序移植到 Tiny210 的 Linux 操作系统中，实时对人脸进行检测，实现机器人对目标人的在线定位。

参考文献 (References)

- [1] J. Huang, K. Kume, A. Saji, et al. Robotic spatial sound localization and its 3-D sound human interface. Tokyo: Proceedings of the 1st International Symposium on Cyber Worlds, 2002: 191-197.
- [2] J. M. Valin, F. Michaud, J. Rouat, et al. Robust localization and tracking of simultaneous moving sound source using beamforming and particle filtering. Robotics and Autonomous Systems Journal, 2007, 55(3): 216-228.
- [3] C. T. Ishi, O. Chatot, H. Ishiguro, et al. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. St Louis: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009: 2027-2032.
- [4] 邵怀宗, 林静然, 彭启琮等. 基于麦克风阵列的声源定位研究[J]. 云南民族大学学报(自然科学版), 2004, 13(4): 256-258.
- [5] 顿向明, 缪松华, 沈静等. 基于麦克风阵列的储罐内爬壁机器人定位技术[J]. 机器人, 2012, 34(4): 460-465.
- [6] K. Nakadai, K. Hidai, H. Mizoguchi, et al. Real-time auditory and visual multiple-object tracking for humanoids. Seattle: Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001: 1425-1436.
- [7] 陈涛, 张明路, 付灵丽. 基于机器人听觉-视觉系统的声源目标定位[J]. 计算机应用, 2009, 9(9): 2471-2480.
- [8] H. Charles, G. Knapp and C. Carter. The generalized correlation method for estimation of time delay. IEEE Transactions On Acoustics, Speech, and Signal Processing, 1976, 24(4): 320-327.
- [9] P. Viola, M. Jones. Robust real time object detection. Vancouver: 2nd International Workshop on Statistical and Computational Theories of Vision: Modeling, Learning, Computing, and Sampling, 13 July 2001: 151-155.
- [10] R. Lienhart, J. Maydt. An extended set of haar-like features for rapid object detection. IEEE ICIP, 2002, 1: 900-903.
- [11] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features. Kauai: IEEE Conference on Computer Vision and Pattern Recognition, 2001.