

基于音频事件检测和分类的音频监控系统背景 模型自适应方法研究

张爱英 倪崇嘉

(山东财经大学系统科学与信息处理研究所 济南 250014)

摘 要 随着监控系统中音频传感器应用的与日俱增,音频事件检测与分类已成为一个重要的研究课题。音频系统所处的音频环境(不同场所、不同噪声)非常复杂,以致检测与分类音频事件异常困难。因此,进行背景模型自适应从而适应不断变化的音频环境变得十分重要。提出了利用受限的最大似然线性回归方法对背景模型进行自适应。采用实际监控场景中的音频数据和模拟录制数据,研究了背景模型自适应方法以及如何有效地进行背景模型自适应。实验结果表明背景模型自适应可以提高目标声音事件的检测性能,减少系统误报。

关键词 音频事件检测与分类,背景模型自适应,受限的最大似然线性回归,监控系统

中图分类号 TP391.42 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.062

Research on Background Model Adaptation for Acoustic Event Detection and Classification Based on Acoustic Surveillance System

ZHANG Ai-ying NI Chong-jia

(Institute of System Science and Information Processing, Shandong University of Finance and Economics, Jinan 250014, China)

Abstract Acoustic event detection and classification have become an important research problem as the increasing use of audio sensors in surveillance system. In these systems, audio circumstance is very complicated, that is, different locations, different noises, which cause the acoustic event detection and classification to be very difficult. Therefore, it is important to implement the background model adaptation in order to adapt these variations of background. In this paper, we proposed to use the constrained maximum likelihood linear regression (CMLLR) to adapt background model. Using the real world data and simulated data, we investigate the background model adaptation approaches and strategies for background model adaptation. Experimental results show that background model adaptation can improve the performance of target acoustic event detection and classification, and also can greatly reduce the false alarm of target acoustic event detection and classification.

Keywords Acoustic event detection and classification, Background model adaptation, Constrained maximum likelihood linear regression (CMLLR), Surveillance system

1 引言

随着数字信号处理技术的飞速发展和互联网的迅速普及,数字音频处理技术的研究已进入一个快速发展的阶段,在音频信息检索、分类、内容理解等方面已经取得了长足的进步。近年来,随着音视频监控的发展以及其在应用中表现的不足,人们对音频监控也提出了较大的需求,主要表现在:1)虽然音频监控系统实现过程中的困难相当多,但人们对音影同步监控的需求越来越强烈;2)对于一些公共治安事件频发区域,如学校、歌厅、大型广场等,大部分都安装了视频监控设备,但很多地方声音监控还是空白。这些视频监控系统都没有办法对正在发生的紧急事件进行实时报警,只能事后提供监控信息。因此对这些地方进行音视频监控可以成为及时处理突发事件的最佳手段。事实上,作为计算听觉场景分析

(Computational Auditory Scene Analysis, CASA)^[1]研究领域的一个分支,音频检测和分类已成为一个非常活跃的研究方向。Espinoza M 等分析了基于深度神经网络的音频事件检测的特征抽取策略^[2];Plinge A 等提出利用 Bag-of-Features 方法对音频事件进行检测和分类^[3];Phan H 等提出了利用随机回归森林(Random Regression Forests)对音频事件进行检测和分类^[4];Parascandolo G 等利用长短时记忆递归神经网络对声音事件进行分类^[5];为了克服训练数据的不足,Lim H 等提出了利用迁移学习的方法对声音事件进行分类^[6]等。

随着基于音频事件检测和分类的音频监控系统的进一步应用,一些新的问题也随之出现^[7-10]。其中一个重要的问题就是:由于实际应用中场景的复杂多变,基于音频事件检测与分类的音频监控系统的音频事件模型的训练环境与测试环境越来越不匹配,使得音频监控系统的性能下降。为了减少训

到稿日期:2016-05-11 返修日期:2016-07-20 本文受国家自然科学基金(61305027),山东省自然科学基金(ZR2011FQ024)资助。

张爱英 女,讲师,主要研究方向为模式识别、数字信号处理等,E-mail: ayzhang@sdufe.edu.cn;倪崇嘉 男,博士,主要研究方向为模式识别、语音识别、音频分类等。

练和测试之间的不匹配,可以从两个方面来减少二者之间的差异:1)在特征空间通过将训练数据和测试数据进行同一类型的转换(如直方图均衡化),来减少二者之间的差异^[6];2)在模型空间进行模型自适应。模型自适应最初用于语音识别中的说话人自适应,其目的在于减小由于说话人的不同而对语音识别系统性能造成的影响^[12-15]。用于说话人自适应的方法有多种,一种是最大后验概率(Maximum A Posterior, MAP)说话人自适应^[12],另外两种是最大似然线性回归(Maximum Likelihood Liner Regression, MLLR)^[14]和受约束的最大似然线性回归(Constrained Maximum Likelihood Liner Regression, CMLLR)^[15]说话人自适应。MAP 说话人自适应方法通常需要较多的自适应数据才能获得较好的效果,而 MLLR 和 CMLLR 说话人自适应仅利用较少的数据就能获得较好的效果,且自适应的速度很快。对于音频事件检测和分类,尽管深度神经网络和递归神经网络等深度学习方法比常用的产生式学习方法(如高斯混合模型)能获得更好的识别效果,但它们也有一些不足,如模型训练需要较多的训练数据才能获得较好的效果、计算复杂度高、模型自适应的效果不明显等限制了它们在基于音频事件检测和分类的音频监控中的应用。

对于音频事件检测和分类的音频监控应用, Ntalampiras 等提出了基于最大后验概率(Maximum A Posteriori, MAP)的自适应框架来减少训练集和测试集之间的不匹配^[16]。在自适应的开始阶段,监督的半自动化方法用于确认真正属于音频事件的数据,这些数据可以用于自适应该音频事件的模型。之后,无监督的方法用于模型的自适应。在实际的监控系统获得的数据上评测了没有进行自适应、有监督的自适应和同时进行有监督和无监督的自适应这 3 种情况下系统的性能。与没有进行自适应的系统相比,同时进行有监督和无监督自适应的系统在 3 种不同的应用场景下的平均错误率减小了 70.48%。Ntalampiras 等提出的基于 MAP 方法的模型自适应方法有两点不足:1)对于基于音频事件检测和分类的音频监控系统来说,那些异常的声音事件在通常的情况下很少发生,通常系统检测出的这些事件基本上属于误报。如果在无监督的自适应阶段,这些误报作为某一异常的声音事件的数据用于该异常声音事件模型的自适应,其结果是自适应后反而导致系统性能的下降。2)由于 MAP 方法进行模型自适应时,通常需要大量的自适应数据,而且计算量很大,不利于实时音频监控系统。针对上述情况,本文提出了利用受约束的最大似然线性回归的方法,仅对背景模型进行自适应。其不仅有利于提高监控系统的实时性,减少系统响应时间,而且还能够避免由于利用产生误报的监控数据对目标声音事件进行模型自适应而造成系统检出目标声音事件的性能的下降。受约束的最大似然线性回归自适应方法仅应用较少的数据就能获得较好的自适应效果,同时该方法的计算复杂度低,有利于提高监控系统的实时性,减少系统响应时间。基于不同场景下的音频监控数据的实验结果表明,该方法能够减少系统的误报,并能够有效地检测异常声音事件。另外,本文还对基于受约束的最大似然线性回归的方法进行背景模型自适应的一些具体问题及策略进行了实验对比。

本文第 2 节介绍了受约束的最大似然线性回归方法;第

3 节描述了基于异常声音事件检测的音频监控系统框架;第 4 节给出了实验的结果及分析比较;最后总结全文。

2 受约束的最大似然线性回归 CMLLR

最大似然线性回归利用线性变换对高斯模型参数进行变换以实现对话人进行自适应。利用 MLLR 对话人进行自适应后,新的均值 $\hat{\mu}$ 和方差 $\hat{\Sigma}$ 计算如下:

$$\hat{\mu} = W\mu + b = A\xi \quad (1)$$

$$\hat{\Sigma} = LBL^T \quad (2)$$

其中, L 是 Σ 的 Choleski 因子。如果仅对均值进行自适应,则称为只有均值的 MLLR 自适应;如果对均值 μ 和方差 Σ 同时进行自适应,则称为标准 MLLR 自适应方法。

在标准的 MLLR 自适应框架下,进一步变化可以表示为式(3)和式(4):

$$\hat{\mu} = A_c \mu - b_c \quad (3)$$

$$\hat{\Sigma} = A_c \Sigma A_c^T \quad (4)$$

式(3)和式(4)中,由于采用同一变换矩阵 A_c 对均值和方差进行自适应变换,因此称其为受限的最大似然线性回归 CMLLR。

变换矩阵 A_c 可以利用最大期望算法进行估计,其辅助函数 $Q(\lambda, \hat{\lambda})$ 为:

$$Q(\lambda, \hat{\lambda}) = c - \frac{1}{2} \sum_{t=1}^T \sum_{m=1}^M \gamma_m(t) (c_m + \log(|\Sigma_m|) + (\hat{o}_t - \mu_m)^T \Sigma_m^{-1} (\hat{o}_t - \mu_m)) \quad (5)$$

其中, $\gamma_m(t)$ 为混合组件 m 在时间 t 的状态的概率, c, c_m 为常数, μ_m 和 Σ_m 是混合组件 m 的均值和方差, $\hat{o}_t = A_c^{-1} o_t + A_c^{-1} b_c = A o_t + b = W \zeta_t$, 对 $\hat{o}_t, A = A_c^{-1}, b = A b_c, W = [b^T A^T]^T, \zeta_t = [1 \ o_t]^T$ 。

按照逐行的方式计算 W , 其一次迭代计算可以表示为:

$$w_i = (a p_i + k_i) G_i^{-1} \quad (6)$$

其中, $p_i = [0 \ c_{i1} \ c_{i2} \ \dots \ c_{im}]$, $c_{ij} = \text{cof}(A_{ij})$ 为辅助因子的行向量, $G_i = \sum_{m=1}^M \frac{1}{\sigma_m^2} \sum_{t=1}^T \gamma_m(t) \zeta_i \zeta_i^T, k_i = \sum_{m=1}^M \frac{1}{\sigma_m^2} \sum_{t=1}^T \gamma_m(t) \zeta_i^T$ 。对于 MLLR 的公式细节和推导过程可以参见文献^[9, 10]。

当利用 CMLLR 对背景模型进行自适应时,利用 $\gamma_m(t)$ 计算出 G_i , 获得 G_i^{-1} 后,计算出 w_i , 进一步计算获得变换矩阵 A_c , 以更新背景模型的均值和方差。

3 系统框架

图 1 给出了基于音频事件检测和分类的音频监控系统的框架,分为特征抽取模块、音频事件检测和分类模块、模型训练模块等。特征抽取模块主要抽取声学特征。为了提高系统的鲁棒性,减少训练和测试之间的不匹配,采用特征空间的直方图均衡化方法对输入的特征进行变换。模型训练模块主要用于训练不同的声音事件模型。音频事件检测和分类模块利用训练好的模型,在检测和分类的过程中输入声音特征流中的不同类型的声音事件。该模块还包含了背景模型自适应的部分以及对检测和分类的结果进行后处理的部分。背景模型自适应部分就是利用 CMLLR 对背景模型进行自适应。图 2

给出了背景模型自适应的流程图。分类结果后处理部分利用置信度测量和最大最小滤波对模型的检测和分类结果进行处理,以提高系统性能。

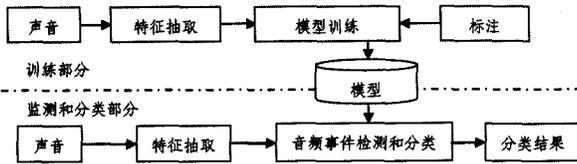


图1 基于音频事件检测和分类的音频监控系统的框架

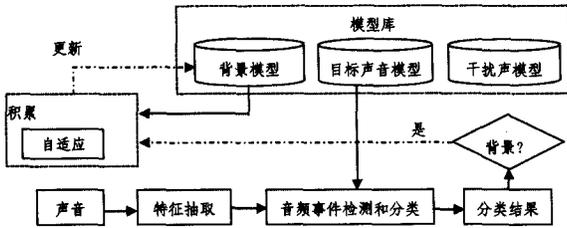


图2 背景模型自适应流程图

4 实验

4.1 数据库及实验建立

表1列出了不同的声音类型以及其声音片段的数目,共有3617段,每一种声音片段的长度为1~3s。根据系统的应用,文中把尖叫声、玻璃破碎声以及撞车声视为目标声音事件,因为这些声音事件的存在一般意味着在系统的应用场景中有不安全的存在。音频监控系统主要为当检测出目标声音事件时做出响应。

表1 不同的声音片段及数目

声音类型	声音片段数目
目标声音	一尖叫声(2460) 一玻璃破碎声(47) 一撞车声(50)
背景声音	一城市街道背景声音(100) 一购物中心背景声音(36) 一城市地铁站背景声音(51)
干扰声	一咳嗽声(17),哭声(45),笑声(44),脚步声(21) 一狗叫声(33),猫叫声(36),乌鸦叫声(46),麻雀叫声(58) 一风声(20),打雷声(26),雨声(20),河流声(12) 一汽车经过的声音(32),汽车发动声(31),汽车停车声(44), 汽车喇叭声(83),自行车铃声(23),摩托车声(28),火车经过的声音(15),撞击声(66),汽车故障声(43) 一铃声(13),电钻声(24),电锯声(23),爆炸声(9),枪声(28),警车声(33)

使用两类不同的数据集来评测系统性能。一类数据是模拟产生的数据用于测试系统检测目标声音事件能力。即在一个相对开阔安静的场地,在两个不同位置分别放置扬声器和录音设备,其中一个扬声器中随机地播放不同干扰声音数据和目标声音数据,另一个扬声器播放不同的背景噪声。录音设备放置在另外一个位置,用于收集扬声器发出的声音。同时为了测试声源位置对系统性能的影响,在不同距离的录音点采集这些声音,这一类数据共有17段,每段大约11min。另一类声音数据是在不同真实场景下的声音,这些声音数据经检查没有目标声音数据,这一类数据共27段,每段大约35min~24h,用这些数据测试系统的误报性能。

对于每一声音片段,以25ms的Hamming窗、以10ms为窗移或帧移抽取13维子带特征,连同其一阶、二阶差分共39

维特征用于表示声音片段的每一帧^[17]。3个状态的无跳转隐马尔科夫模型(Hidden Markov Model, HMM)用于建模除背景事件以外的其他声音事件,1个状态的HMM用于建模背景事件。不同声音事件模型的状态输出概率分布用不同数目的高斯混合模型进行建模。在本文中,用32个高斯混合模型对非目标声音事件的状态输出概率分布进行建模,用16个高斯混合模型对目标声音事件的状态输出概率分布进行建模,而用256个高斯混合模型对背景模型的状态输出概率分布进行建模。多条件训练(Multiple Condition Training)方法用于训练不同的声音事件模型。取自不同场景下的5种噪声数据在5种不同的信噪比(-5,0,5,10,15)与不同的声音数据进行混合获得的数据用于训练声音事件模型。HTK工具用于训练模型^[18]。本文中所有的音频数据的采样率为8000Hz。

4.2 实验结果与分析

4.2.1 背景模型自适应对系统检测目标声音事件影响的对比实验

表2列出了在没有进行背景模型自适应和进行背景模型自适应时,系统检测目标声音事件的实验结果。其目的是验证背景模型自适应对系统检出目标声音事件的能力是否会产生负面影响。

表2 目标声音事件的检测结果

文件	距离(m)	没有自适应			背景模型自适应		
		准确率(%)	召回率(%)	F-值	准确率(%)	召回率(%)	F-值
1	4.2	78.13	89.29	0.8334	78.13	89.29	0.8334
2		76.19	57.14	0.6530	76.19	57.14	0.6530
3		87.50	50.00	0.6364	87.50	50.00	0.6364
4	6.0	81.82	64.29	0.7200	81.82	64.29	0.7200
5		60.00	21.43	0.3158	60.00	21.43	0.3158
6		100.00	17.86	0.3031	100.00	17.86	0.3031
7	6.1	82.76	85.71	0.8421	82.76	85.71	0.8421
8		78.95	53.57	0.6383	80.95	60.71	0.6938
9		73.33	39.29	0.5117	76.93	35.71	0.4878
10	8.3	75.00	10.71	0.1874	75.00	10.71	0.1874
11		0.00	0.00	0.0000	0.00	0.00	0.0000
12		83.33	17.86	0.2942	100.00	21.43	0.3530
13	10.8	50.00	3.57	0.0666	50.00	3.57	0.0666
14		75.00	10.71	0.1874	75.00	10.71	0.1874
15		0.00	0.00	0.0000	0.00	0.00	0.0000
16	20.3	0.00	0.00	0.0000	0.00	0.00	0.0000
17		0.00	0.00	0.0000	0.00	0.00	0.0000
平均		—	58.94	30.67	0.3641	60.25	31.09

从两组实验的对比可以看到:1)背景模型自适应对系统检测目标声音事件的能力没有损害,并且能够提高系统检出目标声音事件的性能。经过背景模型自适应,系统检测目标声音事件的准确率有1.31%的提高,召回率有0.42%的提高。2)声源位置和录音位置之间的距离对系统的性能有很大的影响。当声源位置和录音位置之间的距离较小时,系统检测目标声音事件的准确率和召回率较高,系统的性能较好;随着声源位置和录音位置之间的距离逐步增大,系统的性能逐步下降。特别地,当两者之间的距离大于10m时,系统基本上不能检测出目标声音事件。

4.2.2 背景模型自适应减少系统误报的对比实验

表3列出了利用第二类评测数据,系统在没有进行背景模型自适应以及进行背景模型自适应的对比实验结果。其目的在于评测背景模型自适应对减少系统误报的效果。

表3 背景模型自适应对比实验结果

文件	场所	时长	没有进行背景模型自适应	进行背景模型自适应
1	街道	7小时09分54秒	7	10
2	街道	4小时10分52秒	10	11
3	街道	4小时42分33秒	8	5
4	街道	5小时22分2秒	14	11
5	街道	4小时56分3秒	2	2
6	街道	3小时37分48秒	15	16
7	街道	5小时0分0秒	37	1
8	街道	4小时41分20秒	350	3
9	街道	4小时23分5秒	1757	41
10	街道	4小时34分1秒	1015	78
11	街道	5小时33分11秒	2	2
12	街道	4小时42分4秒	10	10
13	街道	6小时20分7秒	3	2
14	商店	0小时36分40秒	3	2
15	商店	0小时39分0秒	40	34
16	商店	0小时38分30秒	3	3
17	商店	0小时35分0秒	4	4
18	商店	0小时36分0秒	68	51
19	商店	0小时35分0秒	4	5
20	商店	0小时36分30秒	5	4
21	地铁站	0小时32分52秒	16	16
22	地铁站	0小时35分0秒	29	29
23	地铁站	0小时32分45秒	22	21
24	地铁站	0小时25分45秒	23	21
25	地铁站	0小时37分50秒	23	23
26	地铁站	0小时17分30秒	18	19
27	地铁站	0小时16分20秒	20	16

从表3的实验结果可以看到,背景模型自适应可减少系统的误报。特别地,当训练背景模型的数据不能够很好地反映当前系统所处的音频场景时,自适应后背景模型可以很好地适应音频场景的变化。且从上面的实验结果可以看出,当音频文件的时长越长时,背景模型自适应的效果越显著。

4.2.3 背景模型自适应策略

表2和表3列出了进行背景模型自适应时的实验结果。事实上,在进行背景模型自适应时,还有一些具体的问题需要进行探讨,如利用哪些数据进行自适应?何时进行自适应?采用离线背景模型自适应还是采用在线背景模型自适应?这些涉及到具体的背景模型自适应的问题,我们将根据实验结果对这些问题作出回答。

首先,随机抽取一段2h的声音片段,利用该声音片段,比较了没有进行背景模型自适应和随机选择5min的数据进行离线自适应对系统性能的影响。表4列出了系统误报统计结果。同时还发现,利用多于5min的数据,如10min的数据进行自适应与利用5min数据进行自适应对系统性能类似。这里只列出了利用5min的数据进行自适应的结果。从该实验结果可以看到,利用5min的数据进行背景模型自适应可以很好地减少系统的误报,同时从该实验结果得到启发:如果在某一个时间段内,系统检测出较多的目标声音事件,可能是系统所处的环境发生了较大的变化,需要立刻进行背景模型自适应以适应背景的这种变化。

表4 利用5min数据进行背景模型自适应的对比实验结果

随机选择的数据开始	没有自适应	利用5min数据进行自适应
0h:30m:25s	108	4
0h:45m:00s	108	1
1h:12m:20s	108	1
1h:46m:50s	108	1
0h:16m:31s	108	1

其次,考虑到在公共场所下,白天和晚上的场景可能完全不同。因此,在考虑背景模型自适应的同时也考虑每隔固定的时间间隔对背景模型进行自适应。表5列出了实验结果。

表5 不同的时间间隔进行背景模型自适应时的对比实验结果

数据	时长(h)	没有进行自适应	固定时间间隔进行自适应				
			0.5h	1h	2h	5h	10h
1	20	32	27	26	27	27	29
2	21	3144	2	2	2	3	4
3	16	19	17	17	17	17	17
4	24	6	5	6	6	4	4
5	24	43	44	44	44	44	44

从上面的实验结果可以看到:1)背景模型自适应能够减少系统的误报,特别是在某场景中某背景音频事件持续发生,而该背景音频事件没有出现在之前的背景模型建模的数据中。数据2的实验结果就说明了这一点。2)每隔5h对背景模型进行自适应可以获得不错的效果。考虑到系统的实时性要求,不能过于频繁地进行背景模型自适应和长时间不进行背景模型自适应。因此,在进行背景自适应时,每隔5h对背景模型进行自适应。

最后,针对离线背景模型自适应和在线背景模型自适应对系统性能的影响,利用实验进行对比。表6列出了实验结果。

表6 离线背景模型自适应和在线背景模型自适应对比实验结果

数据	时长(h)	离线	在线
1	20	29	27
2	21	6	3
3	16	19	17
4	24	6	4
5	24	40	44

进行离线背景模型自适应时,随机选择5min的数据进行背景模型自适应。进行在线背景模型自适应时,当开启背景自适应后,其后积累的5min背景事件数据用于对背景模型进行自适应。从该实验结果可以看到,采用在线方式的背景模型自适应可以有效地减少系统的误报,获得不错的效果。

基于上述一系列背景自适应策略的实验,积累自适应、固定的时间间隔以及利用5min背景数据进行在线自适应的策略用于背景模型的自适应是非常有效的方法,能取得较好的效果。表2和表3的实验结果就利用了该背景模型自适应方法和策略。

结束语 本文考虑到基于音频事件检测和分类的音频监控系统的实时性和无人值守的自动化要求,提出了利用受限的最大似然线性回归(CMLLR)方法对背景模型进行自适应。实验结果表明,背景模型自适应可以大大减少基于音频事件检测和分类的音频监控系统的误报,同时背景模型自适应不但没有对系统检测目标声音事件的性能产生负面影响,反而能够提高系统检出目标声音事件的性能。通过对背景模型自适应策略的实验,可以看到:1)仅利用5min的背景数据对背景模型进行自适应就可以获得较好的自适应效果;2)固定时间间隔的背景模型自适应不仅可以减少系统的误报,而且还可以提高系统的运行效率;3)在线背景模型自适应较离线背景模型自适应能够更好地适应系统所处环境的变化。本文的研究不仅可以直接应用于基于音频事件检测和分类的音频监控系统中,而且对以后基于音频事件检测和分类的音频监控系统的发展具有重要的借鉴和参考价值。在今后的研究中,

将进一步优化背景模型自适应的方法,以减少计算的复杂度,提高系统的实时性及准确性。

参 考 文 献

- [1] Wang D L, Brown G J. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications [M]. Wiley-IEEE Press, 2006
 - [2] Espi M, Fujimoto M, Kinoshita K, et al. Feature Extraction Strategies in Deep Learning Based Acoustic Event Detection [C] // INTERSPEECH. 2015:2922-2926
 - [3] Plinge A, Grzeszick R, Fink G A. A Bag-of-Features Approach to Acoustic Event Detection [C] // 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014:3704-3708
 - [4] Phan H, Maab M, Mazur R, et al. Random Regression Forests for Acoustic Event Detection and Classification [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1):20-31
 - [5] Parascandolo G, Huttunen H, Virtanen T. Recurrent Neural Networks for Polyphonic Sound Event Detection in Real Life Recordings [C] // 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016:6440-6444
 - [6] Lim H, Kim M J, Kim H. Cross-Acoustic Transfer Learning for Sound Event Classification [M] // 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016:2504-2508
 - [7] Atrey P K, Maddage N C, Kankanhalli M S. Audio based Event Detection for Multimedia Surveillance [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2006:813-816
 - [8] Zhuang X, Zhou X, Hasegawa-Hohnson M, et al. Real-world Acoustic Event Detection [J]. Pattern Recognition Letter, 2010, 31(12):1543-1551
 - [9] Zhang A Y. Using Hierarchical Method to Improve Real Time for Audio-based Surveillance System [C] // International Symposium on Chinese Spoken Language Processing (ISCSLP). 2014:570-573
 - [10] Rabaoui A, Davy M, Rossignol S, et al. Using One-Class SVM and Wavelets for Audio Surveillance [J]. IEEE Trans. on Information Forensics and Security, 2008, 3(4):763-775
 - [11] Angel D L T, Peinado A M, Segura J C, et al. Histogram Equalization of Speech Representation for Robust Speech Recognition [J]. IEEE Trans. on Speech and Audio Processing, 2005, 13(3):355-366
 - [12] Lee C H, Lin C H, Juang B H. A Study on Speaker Adaptation of Parameters of Continuous Density Hidden Markov Models [J]. IEEE Trans. on Signal Processing, 1991, 39(4):806-814
 - [13] Kaltenmeier A, Regel P, Trotter K. Fast Speaker Adaptation for Speech Recognition Systems [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1990:133-136
 - [14] Legetter C, Woodland P. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models [J]. Computer Speech and Language, 1995, 9(2):171-185
 - [15] Povey D, Yao K. A Basis Method of Robust Estimation of Constrained MLLR [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011:4460-4463
 - [16] Ntalampiras S, Potamitis I, Fakotakis N. An Adaptive Framework for Acoustic Monitoring of Potential Hazards [J]. EURASIP Journal on Audio, Speech, and Music Processing, 2009, 10:1-15
 - [17] Kryze D, Rigazio L, Junqua J C. A New Noise-Robust Subband Front-End and Its Comparison To PLP [J]. J. Chem. educ, 2000, 31(15):269
 - [18] Young S, Evermann G, Gales M, et al. The HTK Book [OL]. <http://htk.eng.cam.ac.uk/docs/docs.shtml>
-
- (上接第 279 页)
- [2] Lee S, Yoo C D. Robust video fingerprinting for content-based video identification [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(7):983-988
 - [3] Indyk P, Iyengar G, Shivakumar N. Finding pirated video sequences on the internet; Technical report [R]. Stanford University, 1999
 - [4] Chen L, Stentiford F W M. Video sequence matching based on temporal ordinal measurement [J]. Pattern Recognition Letters, 2008, 29(13):1824-1831
 - [5] Coskun B, Sankur B, Memon N. Spatio-temporal transform based video hashing [J]. IEEE Transactions on Multimedia, 2006, 8(6):1190-1208
 - [6] Leon G, Kalva H, Furht B. Video identification using video tomography [C] // IEEE International Conference on Multimedia and Expo. IEEE, 2009:1030-1033
 - [7] Ji Qing-ge, Tan Zhi-feng, Lu Zhe-ming, et al. An Improved Video Identification Scheme Based on Video Tomography [J]. IEICE Transactions on Information and Systems, 2014, 97(4):919-927
 - [8] Oostveen J, Kalker T, Haitsma J. Feature extraction and a database strategy for video fingerprinting [M] // Recent Advances in Visual Information Systems. Springer Berlin Heidelberg, 2002:117-128
 - [9] Zhao Wan-Lei, Ngo Chong-Wah, Tan Hung-Khoon, et al. Near-duplicate keyframe identification with interest point matching and pattern learning [J]. IEEE Transactions on Multimedia, 2007, 9(5):1037-1048
 - [10] Barrios J M, Bustos B. Competitive content-based video copy detection using global descriptors [J]. Multimedia tools and applications, 2013, 62(1):75-110
 - [11] Esmaeili M M, Ward R K. Robust video hashing based on temporally informative representative images [C] // International Conference on Consumer Electronics (ICCE). 2010:179-180
 - [12] Malek Esmaeili M, Ward R K, Fatourech M. Fast matching for video/audio fingerprinting algorithms [C] // IEEE International Workshop on Information Forensics and Security (WIFS). 2011:1-6
 - [13] MUSCLE-VCD-2007 [OL]. <http://www.wrocq.inria.fr/imedia/civr-bench/index.html>
 - [14] Awad G, Over P, Kraaij W. Content-based video copy detection benchmarking at TRECVID [J]. ACM Transactions on Information Systems, 2014, 32(3):1-40