# mmASL: Environment-Independent ASL Gesture Recognition Using 60 GHz Millimeter-wave Signals

Panneer Selvam Santhalingam, Alamin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, Raja Kushalnagar

Computer Science Department, George Mason University

Department of Science, Technology and Mathematics, Gallaudet University

1. **概述**

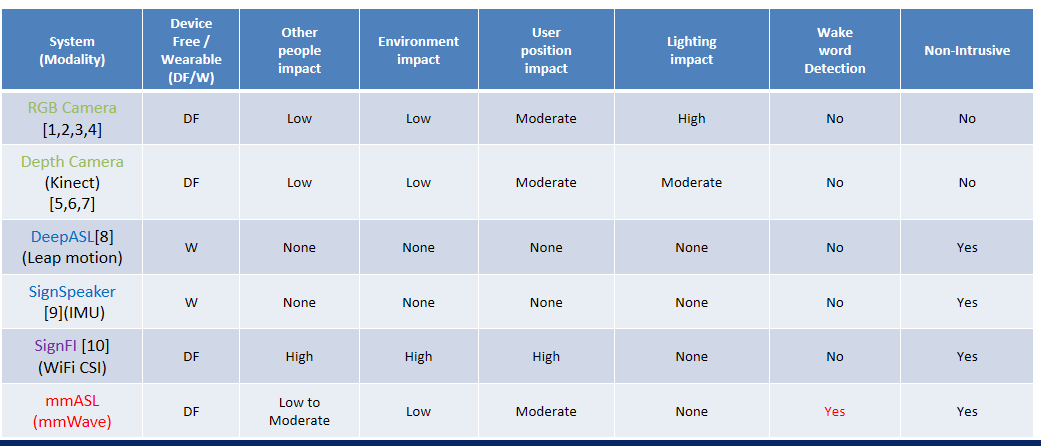
在过去的几年里，家庭助手开始流行（Google home，Amazon Echo），然而语音控制的方式导致这些设备不适用于有严重听力障碍的用户。美国有超过50万的听力障碍人士使用ASL（American Sign Language）进行交流。因此，一个能够识别ASL的家庭助手有很大需求。

毫米波是一项新兴的技术，前景很广阔，它具有波的一般特性，同时还有很多优点：高带宽，高频率的特点使得它能容纳更多的信息，也就能更好的用于细粒度的识别工作；方向性的特点使得发送和返回的信号具有更高的能量，能够传输更远；信道的稀疏性使得相对于WIFI CSI它能够更好地抵抗多径效应和环境带来的影响。这些特点让毫米波成为了较优的手势识别信号，本文实现的系统，就选用了毫米波来进行ASL手势识别。

mmASL系统有两大主要功能，一是ASL唤醒词识别，二是ASL识别。两个功能都用到了频谱图+机器学习的方式，但在细节上略有不同。唤醒词识别使用的是CNN模型，ASL识别使用的是MTL（Multi-Task Learning）模型。系统的平均识别精度达到87%。

主要贡献：将毫米波应用于有很好现实意义的手势识别并实现了完整的系统，展现了毫米波的广阔前景和丰富的应用领域。

1. **相关工作**

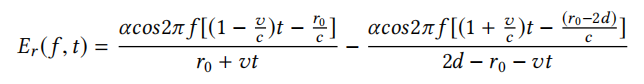


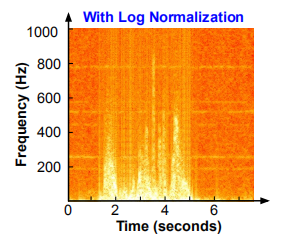
1. **挑战**

* 唤醒词检测（Wake-word detection）：要做到可靠，低时延，在房间内的任何位置都可以实现，同时能够忽略其他无关人员带来的影响。
* ASL手势识别（ASL sign recognition）：以识别ASL符号为目标的手势识别。作为一个家庭助手，应当能够识别大量的手势。
* 鲁棒性（Robust）：不同甚至相同用户执行的标志之间可能存在很大差异，同时还有其他干扰人员的存在和环境变化。需要克服这些干扰。

1. **主要方法**

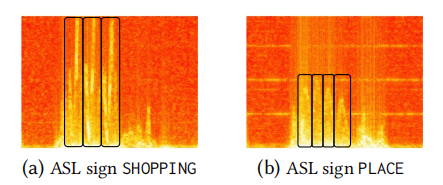
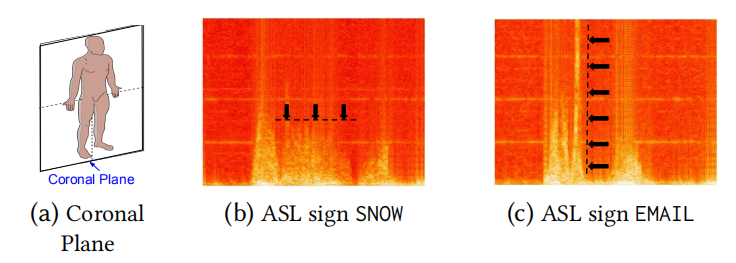
针对以上三个挑战，研究人员首先证明了用毫米波作为信号的可行性。为了使mmASL能可靠地识别ASL符号，它需要从接收到的信号中提取出能够特征表示所执行符号的分量。mmASL使用了1Mhz的正弦波作为基带信号，将手势反射轮廓作为Tx和Rx之间无线信道。用户做出一个动作，会对波产生特定的反射，同时墙壁也会对波进行反射，最后接收端收到的信号就是这两种波叠加产生的复合波。利用了多普勒效应和多普勒频移的知识求解收到的信号强度：

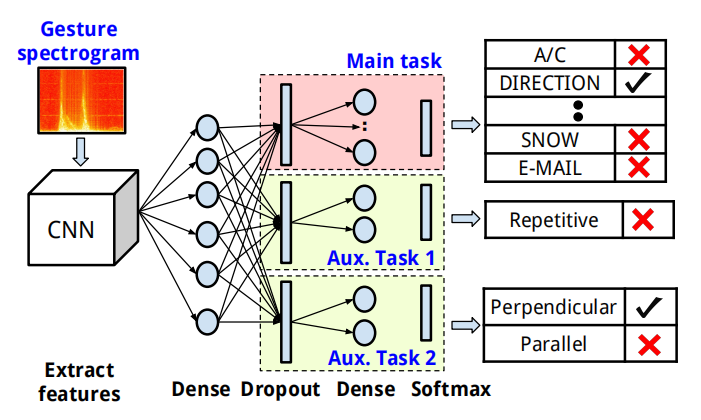
这里，Tx发射一个余弦波cos2πft，α代表收到的波，c是光速，d是发送端到墙壁的距离，Rx在t=0时刻处于坐标r0，在时间t到达r=r0+vt，最后通过这个式子的到了电场Er，也就是信号强度。

他们先是在实验中观察到最大多普勒频移不超过1khz，为了避免在16mhz信号上使用1khz低通滤波器导致很差的滚降和阻带衰减，首先使用平滑的方式将信号降到8khz（也就是分成2000个样本的窗口大小）。然后使用了1khz低通滤波器和10hz高通滤波器进行滤波。10赫兹高通滤波器消除了诸如呼吸和姿势变化等低频人类活动的影响。将得到的滤波信号用于使用短时傅里叶变换（STFT）绘制光谱图，窗口大小为800个样本（100ms），同时每1ms滑动一次窗口（8个样本）。最后，使用幅度值的对数变换进行对数归一化，并加强低强度分量，就得到了如图所示的结果，这种图就是频谱图。

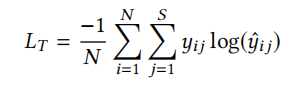
接着再对特定的ASL符号进行测试，发现得到的空间频谱图具有特异性，这就证明理论上利用毫米波进行ASL手势识别是可行的，接下来就需要解决如何识别的问题。

他们与ASL相关的专家进行了合作，提取了ASL具有的两大特性：一个是重复，一个是motion direction。左侧两幅图指出了重复性。当手垂直于冠状面移动时，在频谱图中可以观察到更高频率的高强度分量，因为与Rx相比，垂直移动可能导致更高的相对速度。相反，与冠状面平行运动的迹象会导致相对较低的强度和低频成分。右侧的图表示这一点，两个词的手势分别是平行和垂直运动。

提取出这两个特性之后利用了一个多任务的机器学习模型进行学习。

所有任务共享初始隐藏层，然后共享特定于任务的输出层。它有两个辅助任务（重复运动和运动方向），通过注意力集中和表征偏差来提高符号识别的主要任务的性能。与传统的多任务学习相比，只关心主要任务的执行情况。辅助任务有助于主要任务集中注意力于相关特征以及学习有效表征。由于mmASL输入为光谱图像，选择了基于已建立的图像识别模型的网络设计[11]。Alex Net利用多个卷积层，增加滤波器计数（支持分层特征表示学习）和maxpool层（用于下采样），然后是密集层和softmax层（用于最终预测）。该模型采用了相似的设计方法，通过卷积层学习分层特征表示，然后是所有任务共享的密集层和任务特定的丢失（以避免过度拟合），以及用于单个任务分类的softmax层。

对于每个任务T，使用预测分布和目标分布的交叉熵作为损失函数（LT）。其中N是样本数，S是ASL符号的数目，y和yˆ分别是基本真实值和预测概率。

三个任务的组合损失函数为L = Lm + λ1La1 + λ2La2。λ1和λ2分别为辅助任务1和辅助任务2的重要性系数。将λ1设为0.75，将λ2设为0.5（观察到，λ1和λ2的设定值大于0.5且小于0.75会产生类似的结果）。仅仅调整重要性系数的设定是根据每个任务的性能采用了特定于任务的早期停止方法[12]得到的。

系统的特色功能WAKE WORD RECOGNITION唤醒词检测，主要的两个挑战，第一是在扫描的扇形区数量和每个扇形区扫描的时间之间做出权衡，第二就是如何消除其他人带来的影响。使用的系统提供25个扇形区域扫描，每个相隔5°。选取的标准是选出的扇形区域能够覆盖这25个扇形覆盖的地方，同时具有最小的重叠部分。标号为2，5，9，12，16，20的扇形区域，每个扇形区域是由主瓣和旁瓣组成的，这样使得唤醒词在相邻的区域内也可以识别出来，尽管强度比较低。

唤醒必然伴随着突然的动作变化，这会导致能量的突增，可以作为一个开始识别的特征，仅仅依靠能量作为标准会受到他人运动的干扰，所以采用了能量检测＋空间图谱的方式，简单概括就是在收到突增能量时对该扇形区域进行ASL检测，判断是否为事先设定的唤醒词手势。因为设定的唤醒词特定且数量少，所以使用了CNN的学习方式，没有用前面提到的多任务模型。

1. **实验记录**

|  |  |
| --- | --- |
| **唤醒词检测实验设置** | |
| 实验设备 | * Phased antenna array（发送毫米波） * SiBeam 60 Ghz RF front（变频） * Reflective surface（分离Rx和Tx信号） * I/Q and beam control channels * Ni multi-FPGA mmWave platform（核心） * Clock |
| 对比设备 | * Kinect |
| 测试情景 | * 单人随机选取测试点和训练点 * 其他用户站在同一扇区 * 其他用户站在不同扇区 * 其他用户从背后走过 * 其他用户从面前走过 |
| 样本数量 | * 3700 |
| 测试频率 | * 3s/样本 |
| 测试人数 | * 3 |
| 测试环境 | * R1-Classroom * R2-Lab * R3-Conference room |
| **ASL手势检测实验设置** | |
| 实验设备 | * Phased antenna array（发送毫米波） * SiBeam 60 Ghz RF front（变频） * Reflective surface（分离Rx和Tx信号） * I/Q and beam control channels * Ni multi-FPGA mmWave platform（核心） * Clock |
| 对比设备 | * Kinect |
| 测试情景 | * 单人随机选取测试点和训练点 * 其他用户站在同一扇区 * 其他用户站在不同扇区 * 其他用户从背后走过 * 其他用户从面前走过 |
| 样本数量 | * 12236 |
| 测试词数 | * 50 |
| 测试人数 | * 15 |
| 测试环境 | * R1-Classroom * R2-Lab * R3-Conference room |

|  |  |
| --- | --- |
| **唤醒词检测评估内容** | |
| 唤醒词内容的影响 | * 对比了不同唤醒词对精度的影响 * 以柱状图的形式记录结果 |
| 扫描时间的影响 | * 对比了3种扫描时间对精度的影响 * 1s，2s，3s * 以柱状图的形式记录结果 |
| 学习模型的影响 | * 对比了3种机器学习模型对精度的影响 * 对比了3种机器学习模型对误报率的影响 * SVM,RF,CNN * 以柱状图的形式记录结果 |
| 干扰人员的影响 | * 对比了有无干扰人员对精度的影响 * 以柱状图的形式记录结果 |

|  |  |
| --- | --- |
| **ASL识别评估内容** | |
| 平均识别精度 | * 展示了多个房间下不同学习模型达到的平均识别精度 * 以柱状图的形式记录结果 |
| 最高识别精度 | * 展示了多个房间下不同学习模型达到的最高识别精度 * 以柱状图的形式记录结果 |
| 有无受过训练的影响 | * 展示了未受过训练人员测试的平均识别精度 * 展示了未受过训练人员测试的最高识别精度 * 以柱状图的形式记录结果 |
| 干扰人员的影响 | * 对比了有无干扰人员对精度的影响 * 以柱状图的形式记录结果 |
| 信号处理的影响 | * 对比了有无信号处理对精度的影响 * 以折线图的形式记录结果 |

1. **总结及启发**

* 该文章利用毫米波对用户执行的ASL进行识别，并开发出了一个基础识别功能完整同时具有手势唤醒功能的家庭助手系统。
* 美中不足的地方是，整个系统使用的设备非常复杂，整体显得臃肿，不能直接供大众使用。另外，从实验设置可以看出，mmASL目前只能对面对着设备的用户动作进行识别，其他的方向并未作考虑。
* 系统各个功能的实现都使用到了机器学习这一工具，同时应对不同的场景使用了不同的模型，我认为这非常值得学习，在面对一个需要使用机器学习的问题时先分析针对这个问题使用各种模型的优缺点，进行取舍。
* 结合目前正在进行的ContextLoc的实验工作，学习到将未经过训练的人员纳入测试可以提高结果的可信度。

1. **参考文献**

[1] Oya Aran, Thomas Burger, Alice Caplier, and Lale Akarun. 2009. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition* 42, 5 (2009), 812 – 822.

[2] Nicholas Michael, Dimitris Metaxas, and Carol Neidle. 2009. Spatial and Temporal Pyramids for Grammatical Expression Recognition of American Sign Language. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, Pennsylvania, USA) *(Assets ’09)*. ACM, New York, NY, USA, 75–82.

[3] T. Starner, J. Weaver, and A. Pentland. 1998. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 12 (Dec 1998), 1371–1375.

[4] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu. 2018. Recognizing American Sign Language Gestures from Within Continuous Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2145–214509.

[5] Cao Dong, M. C. Leu, and Z. Yin. 2015. American Sign Language alphabet recognition using Microsoft Kinect. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 44–52.

[6] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2015. Sign Language Recognition using 3D convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.

[7] N. Pugeault and R. Bowden. 2011. Spelling it out: Real-time ASL fingerspelling recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 1114–1119.

[8] Biyi Fang, Jillian Co, and Mi Zhang. 2017. DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems* (Delft, Netherlands) *(SenSys ’17)*. ACM, New York, NY, USA, Article 5, 13 pages.

[9] Jiahui Hou, Xiang-Yang Li, Peide Zhu, Zefan Wang, Yu Wang, Jianwei Qian, and Panlong Yang. 2019. SignSpeaker: A Real-time, High-Precision SmartWatch-based Sign Language Translator. In *To appear in Mobicom 2019* (Los Cabos, Mexico).

[10] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki. 2012. Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques. In *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*. 248–251.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105.

[12] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial Landmark Detection by Deep Multi-task Learning. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 94–108.