

# 基于视觉方法的手语翻译系统

应逸雯

（电子与电气工程系 指导教师：张宏）

# 目录

- 项目背景
- 效果展示
- 实施过程
- 结果分析
- 总结&展望

# 项目背景

- 手语是聋哑人的主要沟通方式，但大多数健听人不会手语。
- 设计一个手语翻译系统，辅助聋哑人与健听人交流。

# 项目背景

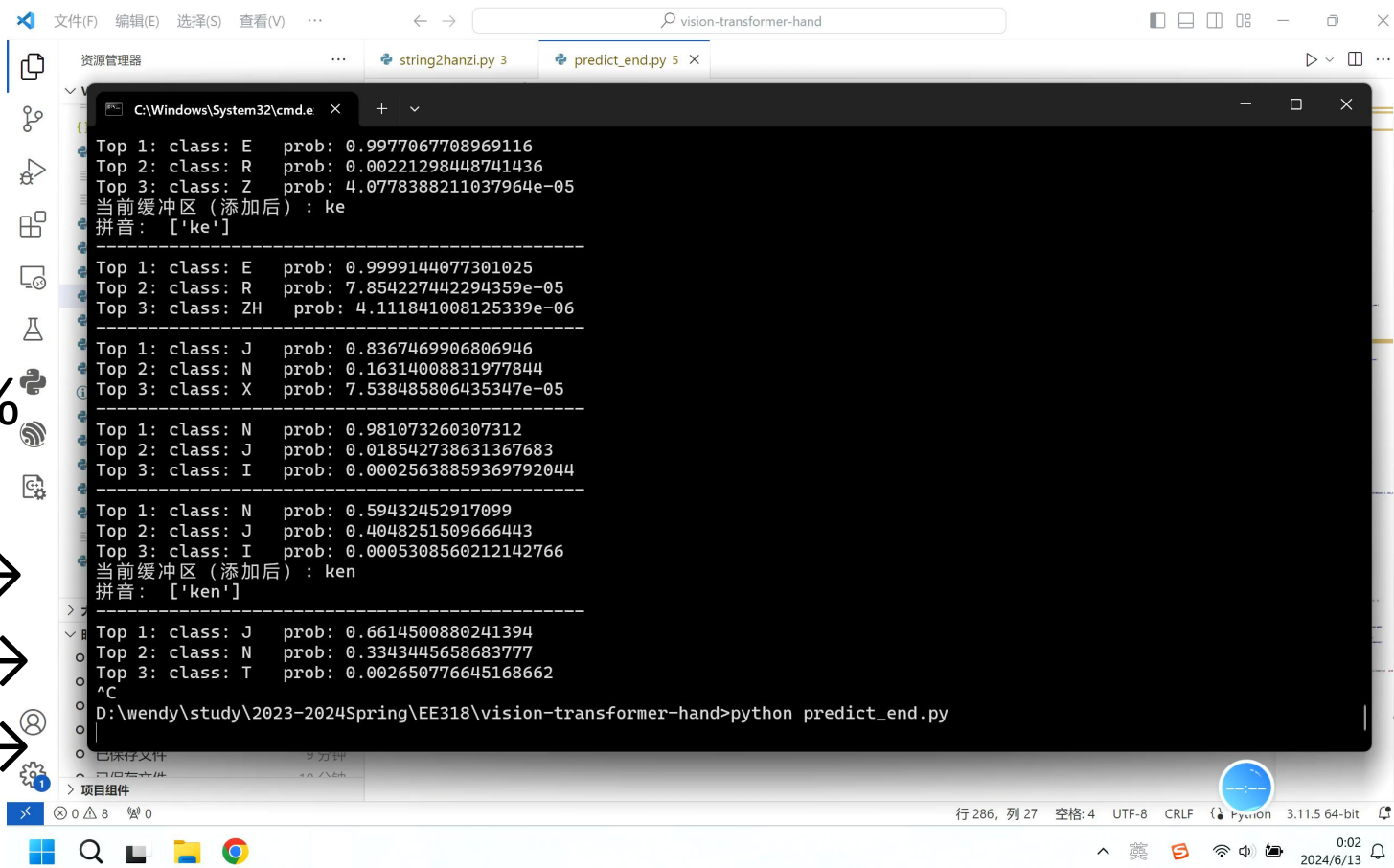
- 手语识别——视觉/传感器手套
  - 视觉手势识别——静态/动态
  - 深度学习模型——2DCNN/3DCNN/RNN/LSTM/ViT
  - 相机类型——RGB相机/深度相机/RGB-D相机/红外相机
- 
- 通过RGB图像和深度学习技术，识别汉语手语和数字的静态手势。

# 效果展示（拼音）

拼音  
——30分类

基于ViT  
准确率：75%

识别→检验→  
分割成拼音→  
转换为文本→  
输出语音



```
C:\Windows\System32\cmd.exe
Top 1: class: E prob: 0.9977067708969116
Top 2: class: R prob: 0.00221298448741436
Top 3: class: Z prob: 4.0778388211037964e-05
当前缓冲区（添加后）: ke
拼音: ['ke']

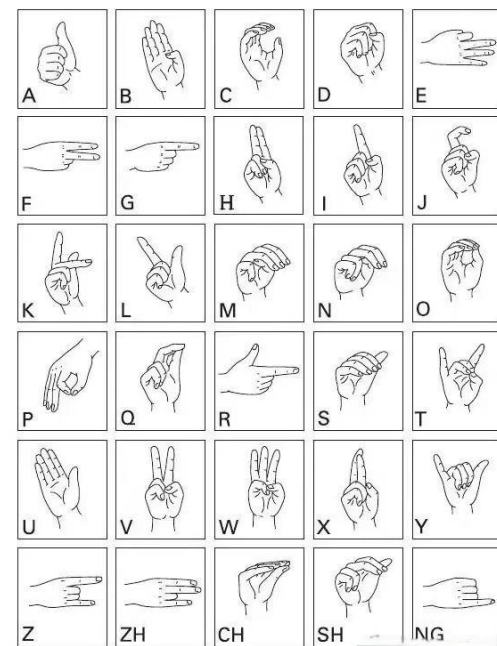
Top 1: class: E prob: 0.9999144077301025
Top 2: class: R prob: 7.854227442294359e-05
Top 3: class: ZH prob: 4.111841008125339e-06

Top 1: class: J prob: 0.8367469906806946
Top 2: class: N prob: 0.16314008831977844
Top 3: class: X prob: 7.538485806435347e-05

Top 1: class: N prob: 0.981073260307312
Top 2: class: J prob: 0.018542738631367683
Top 3: class: I prob: 0.00025638859369792044

Top 1: class: N prob: 0.59432452917099
Top 2: class: J prob: 0.4048251509666443
Top 3: class: I prob: 0.0005308560212142766
当前缓冲区（添加后）: ken
拼音: ['ken']

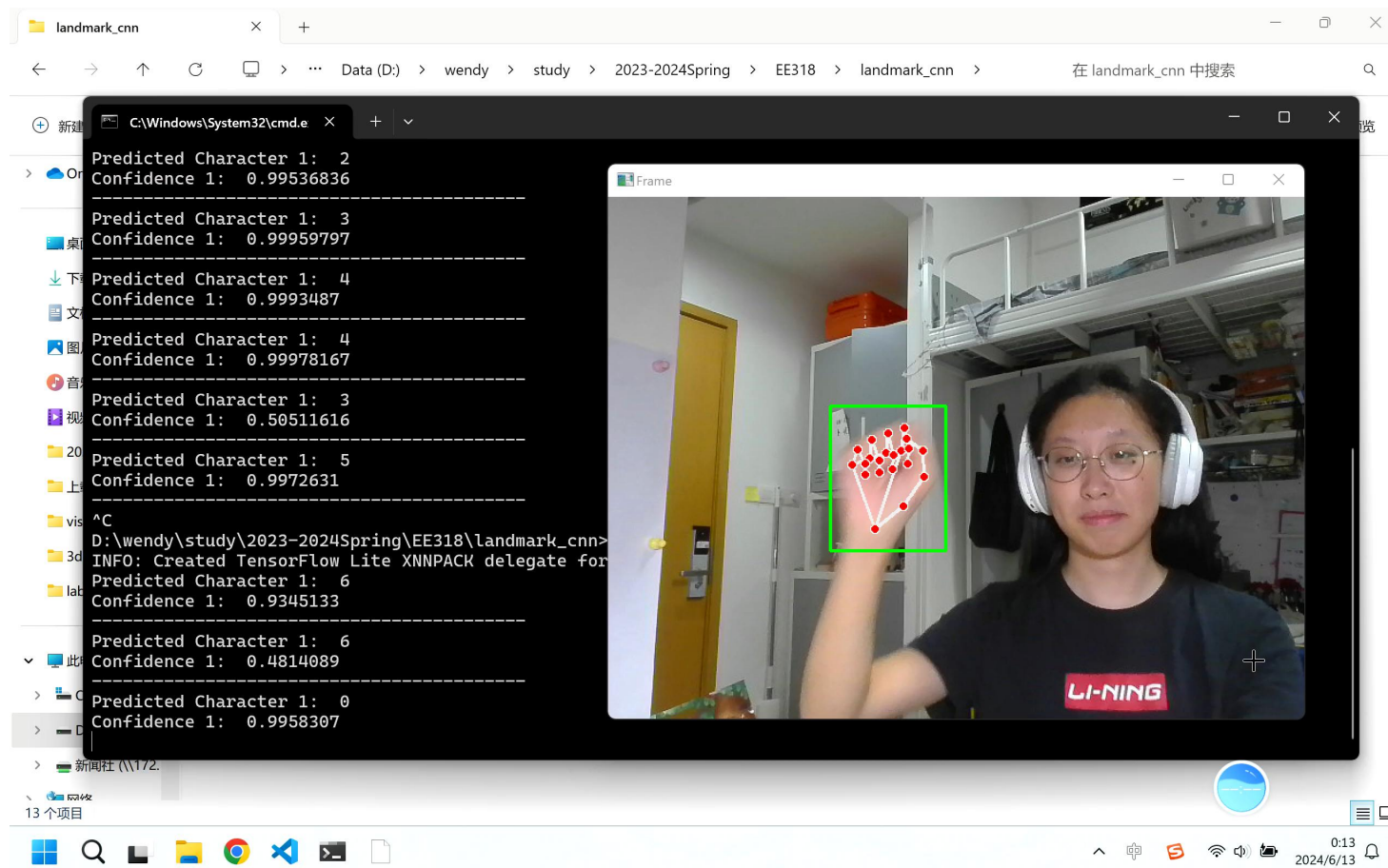
Top 1: class: J prob: 0.6614500880241394
Top 2: class: N prob: 0.3343445658683777
Top 3: class: T prob: 0.002650776645168662
^C
D:\wendy\study\2023-2024Spring\EE318\vision-transformer-hand>python predict_end.py
```



# 效果展示 (数字)

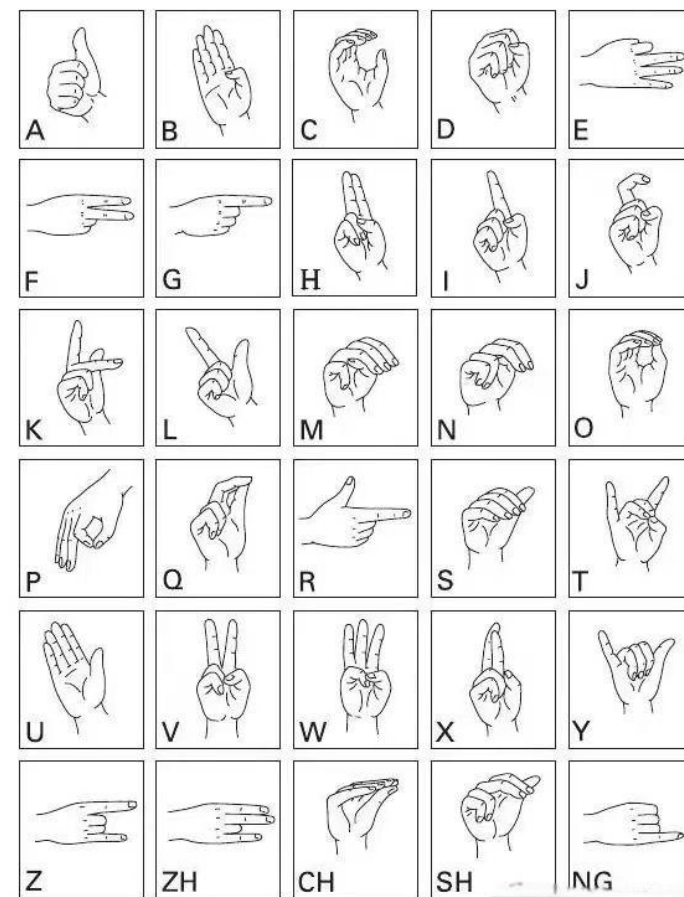
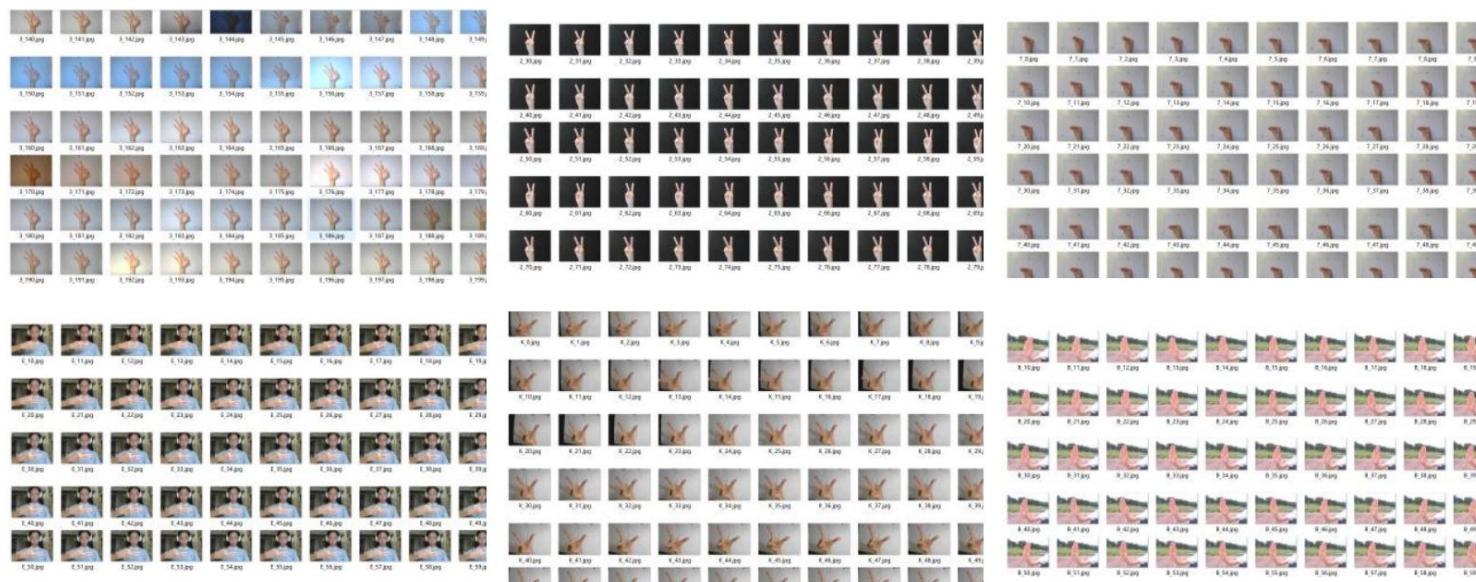
数字  
——10分类

基于CNN+  
骨骼点坐标  
信息  
准确率:  
100%



# 实施过程

- 数据集准备
- 不同背景，不同光照，不同角度



# 实施过程

- 图像数据集增强
- 水平翻转、对比度增强、亮度增强、仿射变换、错切变换、HSV增强、模糊化、增加白噪声、增加随机遮挡、增加随机图像融合



亮度增强



饱和度增强



翻转图像



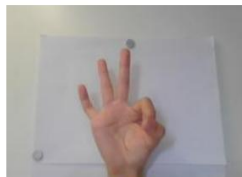
HSV 增强



仿射变换



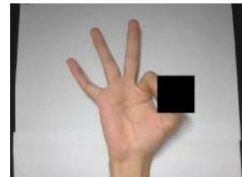
错切变换



模糊处理



加白噪声



随机遮挡

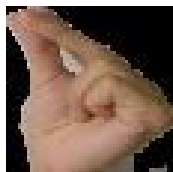


随机融合



# 实施过程

- 图像预处理
- YCrCb+OTSU→前景提取
- Mediapipe→手部位置
- 各点灰度值；各点HSV值；骨骼坐标信息。



# 实施过程

- 深度学习
- CNN：通过多层卷积和池化操作来提取图像的特征，通过全连接层进行分类或回归。

```
C:\Windows\System32\cmd.e  x  +  v
164/164 [=====] - 90s 548ms/step - loss: 0.1633 - accuracy: 0.9566 - val_loss: 1.6391 - val_acc
uracy: 0.3898 - lr: 0.0010
Epoch 2/6
164/164 [=====] - 98s 595ms/step - loss: 0.0032 - accuracy: 0.9997 - val_loss: 0.6777 - val_acc
uracy: 0.8676 - lr: 0.0010
Epoch 3/6
164/164 [=====] - 111s 676ms/step - loss: 0.0013 - accuracy: 1.0000 - val_loss: 0.0757 - val_ac
curacy: 0.9962 - lr: 0.0010
Epoch 4/6
164/164 [=====] - 115s 702ms/step - loss: 5.9634e-04 - accuracy: 1.0000 - val_loss: 0.0167 - va
l_accuracy: 0.9972 - lr: 0.0010
Epoch 5/6
164/164 [=====] - 114s 694ms/step - loss: 4.4512e-04 - accuracy: 1.0000 - val_loss: 0.0122 - va
l_accuracy: 0.9972 - lr: 0.0010
Epoch 6/6
164/164 [=====] - ETA: 0s - loss: 5.3236e-04 - accuracy: 0.9999
Epoch 00006: ReduceLROnPlateau reducing learning rate to 0.0005000000237487257.
164/164 [=====] - 114s 696ms/step - loss: 5.3236e-04 - accuracy: 0.9999 - val_loss: 0.0126 - va
l_accuracy: 0.9966 - lr: 0.0010

D:\wendy\study\2023-2024Spring\EE318\sign-language-detection-cnn-deeplearning>
```

# 实施过程

- 深度学习
- Vision Transformer: 将图像划分为一系列固定大小的图块, 并将这些图块视为序列输入到Transformer模型中。通过多头自注意力机制和前馈神经网络进行处理。
- 时间长
- 适合大数据集

```
returning results
train epoch 0] loss: 0.853, acc: 0.814: 100%| 6774/6774 [09:37<00:00, 11.74it
valid epoch 0] loss: 0.089, acc: 0.971: 100%| 847/847 [01:10<00:00, 11.98it/s
train epoch 1] loss: 0.453, acc: 0.901: 100%| 6774/6774 [09:31<00:00, 11.85it
valid epoch 1] loss: 0.166, acc: 0.959: 100%| 847/847 [01:10<00:00, 11.96it/s
train epoch 2] loss: 0.315, acc: 0.927: 100%| 6774/6774 [09:26<00:00, 11.97it
valid epoch 2] loss: 0.031, acc: 0.989: 100%| 847/847 [01:10<00:00, 11.97it/s
train epoch 3] loss: 0.217, acc: 0.947: 100%| 6774/6774 [09:24<00:00, 11.99it
valid epoch 3] loss: 0.037, acc: 0.986: 100%| 847/847 [01:10<00:00, 11.98it/s
train epoch 4] loss: 0.170, acc: 0.956: 100%| 6774/6774 [09:25<00:00, 11.98it
valid epoch 4] loss: 0.016, acc: 0.995: 100%| 847/847 [01:10<00:00, 12.03it/s
valid epoch 4] loss: 0.014, acc: 0.996: 100%| 847/847 [01:10<00:00, 11.97it/s
est accuracy: 0.9959, Test loss: 0.0135
```

# 实施过程

- 转换成语音
- 单个字符，验证成功，进入缓冲队
- 分割拼音，转换为汉字
- 文本转语音输出
- 缓冲队列，多线程语音输出，不干扰图像识别

```
请输入拼音： dianziyudianqigongchengxi  
转换后的汉字： 电子雨点气功承袭  
请输入拼音： woshiyingyiwen  
转换后的汉字： 我是应以文
```

# 结果分析

- 手工设计算法——二维向量夹角
- (0,1) & (3,4) etc.
- 不支持复杂手势，7和9无法识别
- 准确率：88.57%



```
rex@rexPC: ~/EE318/tradition
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
KeyboardInterrupt

rex@rexPC:~/EE318/tradition$ python3 test.py
2024-06-05 20:45:06.655737: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow
ow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild TensorFlow with t
he appropriate compiler flags.
2024-06-05 20:45:07.556393: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT War
ning: Could not find TensorRT
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
I0000 00:00:1717591508.885405    6880 gl_context_egl.cc:85] Successfully initialized EGL. Major
: 1 Minor: 5
I0000 00:00:1717591508.916535    6944 gl_context.cc:357] GL version: 3.2 (OpenGL ES 3.2 NVIDIA
510.47.03), renderer: NVIDIA GeForce RTX 2060/PCIe/SSE2
INFO: Created TensorFlow Lite XNNPACK delegate for CPU.
Label: 1, Total Images: 3600, Correct Predictions: 3412, Accuracy: 0.9477777777777778
Label: 3, Total Images: 3600, Correct Predictions: 1612, Accuracy: 0.4477777777777778
Label: 5, Total Images: 3600, Correct Predictions: 3583, Accuracy: 0.9952777777777778
Label: 8, Total Images: 3600, Correct Predictions: 3032, Accuracy: 0.8422222222222222
Label: 4, Total Images: 3600, Correct Predictions: 3575, Accuracy: 0.9930555555555556
Label: 6, Total Images: 3600, Correct Predictions: 3490, Accuracy: 0.9694444444444444
Label: 2, Total Images: 3600, Correct Predictions: 3481, Accuracy: 0.9669444444444445
Label: 0, Total Images: 3600, Correct Predictions: 3325, Accuracy: 0.9236111111111112
rex@rexPC:~/EE318/tradition$
```



# 结果分析

- CNN+像素
- 数字
- 表现尚可

灰度值：72.03%

```
Overall Accuracy: 0.720360180090045
Total Samples: 3998
Total Correct: 2880
Label 0 Accuracy: 0.7325
Label 0 Samples: 400
Label 0 Correct: 293
Label 1 Accuracy: 0.9725
Label 1 Samples: 400
Label 1 Correct: 389
Label 2 Accuracy: 0.8675
Label 2 Samples: 400
Label 2 Correct: 347
Label 3 Accuracy: 0.75
Label 3 Samples: 400
Label 3 Correct: 300
Label 4 Accuracy: 0.91
Label 4 Samples: 400
Label 4 Correct: 364
Label 5 Accuracy: 0.815
Label 5 Samples: 400
Label 5 Correct: 326
Label 6 Accuracy: 0.95
Label 6 Samples: 400
Label 6 Correct: 380
Label 7 Accuracy: 0.02756892230576441
Label 7 Samples: 399
Label 7 Correct: 11
Label 8 Accuracy: 0.8325
Label 8 Samples: 400
Label 8 Correct: 333
Label 9 Accuracy: 0.3433583959899749
Label 9 Samples: 399
Label 9 Correct: 137
rex@rexPC:~/EE318/grey$
```

HSV值：83.64%

```
Overall Accuracy: 0.8364182091045522
Total Samples: 3998
Total Correct: 3344
Label 0 Accuracy: 0.8975
Label 0 Samples: 400
Label 0 Correct: 359
Label 1 Accuracy: 0.83
Label 1 Samples: 400
Label 1 Correct: 332
Label 2 Accuracy: 0.905
Label 2 Samples: 400
Label 2 Correct: 362
Label 3 Accuracy: 0.765
Label 3 Samples: 400
Label 3 Correct: 306
Label 4 Accuracy: 0.915
Label 4 Samples: 400
Label 4 Correct: 366
Label 5 Accuracy: 0.8125
Label 5 Samples: 400
Label 5 Correct: 325
Label 6 Accuracy: 0.9025
Label 6 Samples: 400
Label 6 Correct: 361
Label 7 Accuracy: 0.6666666666666666
Label 7 Samples: 399
Label 7 Correct: 266
Label 8 Accuracy: 0.795
Label 8 Samples: 400
Label 8 Correct: 318
Label 9 Accuracy: 0.87468671679198
Label 9 Samples: 399
Label 9 Correct: 349
rex@rexPC:~/EE318/hsv-occlusion$
```

# 结果分析

- CNN+像素 (HSV)
- 拼音
- 2125/4091
- 无法接受

```
Predicted Label: CH, True Label: Q
1/1 [=====] - 0s 19ms/step
Predicted Label: CH, True Label: Q
1/1 [=====] - 0s 19ms/step
Predicted Label: CH, True Label: Q
1/1 [=====] - 0s 20ms/step
Predicted Label: CH, True Label: Q
Overall Accuracy: 0.519432901491078
Total Samples: 4091
Total Correct: 2125
Label A Accuracy: 0.0
Label A Samples: 135 Correct: 0
Label B Accuracy: 0.051094890510948905
Label B Samples: 137 Correct: 7
Label C Accuracy: 0.9057971014492754
Label C Samples: 138 Correct: 125
Label CH Accuracy: 0.9197080291970803
Label CH Samples: 137 Correct: 126
Label D Accuracy: 0.5507246376811594
Label D Samples: 138 Correct: 76
Label E Accuracy: 0.5079365079365079
Label E Samples: 126 Correct: 64
Label F Accuracy: 0.8359375
Label F Samples: 128 Correct: 107
Label G Accuracy: 0.7322834645669292
Label G Samples: 127 Correct: 93
Label H Accuracy: 0.14705882352941177
Label H Samples: 136 Correct: 20
Label I Accuracy: 0.9285714285714286
Label I Samples: 140 Correct: 130
Label J Accuracy: 0.2785714285714286
Label J Samples: 140 Correct: 39
Label K Accuracy: 0.9285714285714286
Label K Samples: 140 Correct: 130
Label L Accuracy: 0.8857142857142857
Label L Samples: 140 Correct: 124
Label M Accuracy: 0.4142857142857143
Label M Samples: 140 Correct: 58
Label N Accuracy: 0.05714285714285714
Label N Samples: 140 Correct: 8
Label NG Accuracy: 0.4117647058823529
Label NG Samples: 136 Correct: 56
```

I

```
Label O Accuracy: 0.03571428571428571
Label O Samples: 140 Correct: 5
Label P Accuracy: 0.8142857142857143
Label P Samples: 140 Correct: 114
Label Q Accuracy: 0.007194244604316547
Label Q Samples: 139 Correct: 1
Label R Accuracy: 0.9202898550724637
Label R Samples: 138 Correct: 127
Label S Accuracy: 0.03731343283582089
Label S Samples: 134 Correct: 5
Label SH Accuracy: 0.09420289855072464
Label SH Samples: 138 Correct: 13
Label T Accuracy: 0.8148148148148148
Label T Samples: 135 Correct: 110
Label U Accuracy: 0.5428571428571428
Label U Samples: 140 Correct: 76
Label V Accuracy: 0.4857142857142857
Label V Samples: 140 Correct: 68
Label W Accuracy: 0.9714285714285714
Label W Samples: 140 Correct: 136
Label X Accuracy: 0.9214285714285714
Label X Samples: 140 Correct: 129
Label Y Accuracy: 0.9785714285714285
Label Y Samples: 140 Correct: 137
Label Z Accuracy: 0.30158730158730157
Label Z Samples: 126 Correct: 38
Label ZH Accuracy: 0.024390243902439025
Label ZH Samples: 123 Correct: 3
rex@rexPC:~/EE318/hsv-occlusion$
```

# 结果分析

- CNN+手部骨架位置信息
- 数字：100%（该情况的最佳解）
- 拼音：78%
- 速度非常快

```
17/1 [-----]
Accuracy for class 0: 1.00
Accuracy for class 1: 1.00
Accuracy for class 2: 1.00
Accuracy for class 3: 1.00
Accuracy for class 4: 1.00
Accuracy for class 5: 1.00
Accuracy for class 6: 1.00
Accuracy for class 7: 1.00
Accuracy for class 8: 1.00
Accuracy for class 9: 1.00
Overall accuracy: 1.00
```

```
image: 1.jpg, True Label: 21
Accuracy for class A: 0.98
Accuracy for class B: 1.00
Accuracy for class C: 0.99
Accuracy for class CH: 1.00
Accuracy for class D: 0.46
Accuracy for class E: 0.84
Accuracy for class F: 0.99
Accuracy for class G: 1.00
Accuracy for class H: 0.18
Accuracy for class I: 0.80
Accuracy for class J: 0.54
Accuracy for class K: 0.61
Accuracy for class L: 0.84
Accuracy for class M: 0.97
Accuracy for class N: 0.99
Accuracy for class NG: 1.00
Accuracy for class O: 0.21
Accuracy for class P: 0.91
Accuracy for class Q: 0.31
Accuracy for class R: 0.98
Accuracy for class S: 0.00
Accuracy for class SH: 0.97
Accuracy for class T: 0.52
Accuracy for class U: 0.90
Accuracy for class V: 0.97
Accuracy for class W: 1.00
Accuracy for class X: 0.96
Accuracy for class Y: 0.99
Accuracy for class Z: 0.92
Accuracy for class ZH: 0.85
Overall accuracy: 0.78
```



# 结果分析

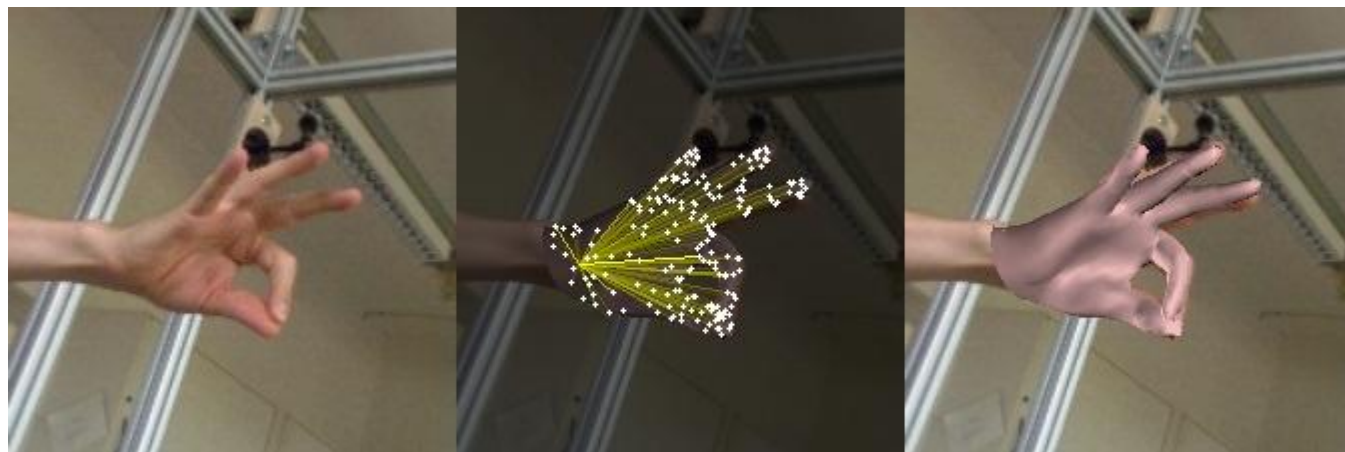
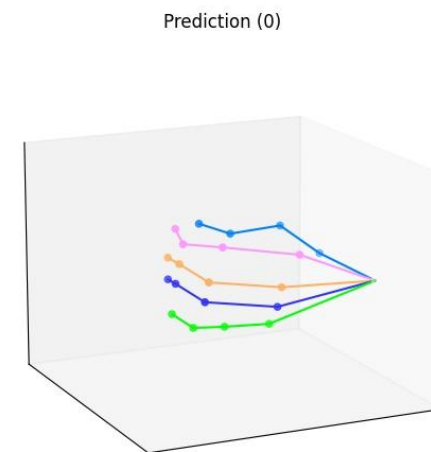
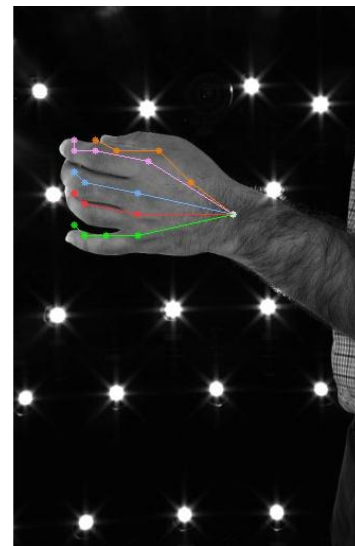
- Vision Transformer
- 数字：89.24%
- 拼音：75.00% (该情况的最佳解)
- 速度非常慢

```
0 0
7: 100.00% (100 correct out of 100 images)
1: 100.00% (100 correct out of 100 images)
3: 100.00% (100 correct out of 100 images)
5: 100.00% (100 correct out of 100 images)
9: 83.00% (83 correct out of 100 images)
8: 100.00% (100 correct out of 100 images)
4: 99.00% (99 correct out of 100 images)
6: 100.00% (100 correct out of 100 images)
2: 72.00% (72 correct out of 100 images)
0: 44.68% (42 correct out of 94 images)
Global Accuracy: 89.24%887 correct out of 994 images)
rex@rexPC:~/EE318/vit-number$
```

```
rex@rexPC:~/EE318/vit$ python3 test.py
J: 77.14% (108 correct out of 140 images)
G: 94.93% (131 correct out of 138 images)
V: 55.00% (77 correct out of 140 images)
ZH: 54.35% (75 correct out of 138 images)
U: 100.00% (140 correct out of 140 images)
M: 24.29% (34 correct out of 140 images)
E: 94.20% (130 correct out of 138 images)
B: 0.00% (0 correct out of 139 images)
NG: 95.00% (133 correct out of 140 images)
X: 100.00% (140 correct out of 140 images)
K: 61.43% (86 correct out of 140 images)
N: 8.57% (12 correct out of 140 images)
O: 98.57% (138 correct out of 140 images)
S: 90.58% (125 correct out of 138 images)
CH: 99.28% (137 correct out of 138 images)
Y: 94.29% (132 correct out of 140 images)
H: 86.33% (120 correct out of 139 images)
I: 23.57% (33 correct out of 140 images)
A: 43.57% (61 correct out of 140 images)
Z: 98.55% (136 correct out of 138 images)
SH: 24.29% (34 correct out of 140 images)
P: 100.00% (140 correct out of 140 images)
F: 95.65% (132 correct out of 138 images)
T: 95.71% (134 correct out of 140 images)
D: 100.00% (140 correct out of 140 images)
C: 100.00% (140 correct out of 140 images)
R: 96.43% (135 correct out of 140 images)
W: 100.00% (140 correct out of 140 images)
L: 55.71% (78 correct out of 140 images)
Q: 100.00% (140 correct out of 140 images)
Global Accuracy: 75.00%3138 correct out of 4184 images)
rex@rexPC:~/EE318/vit$
```

# 结果分析

- 3D姿态估计
- 准确率：0.6270
- 发现部分手势没有被正确提取坐标
- 考虑全手深度信息叠加原来的RGB信息



# 总结&展望

- 数字（简单手势）：使用骨骼信息比较预测效果优异。
- 拼音（复杂手势）：Vision Transformer模型表现良好。
- 图像数据集增强提高系统的鲁棒性和准确性。
- 丰富数据集提供不同角度、不同光照的预测。
- 前景提取算法滤除背景干扰。

# 总结&展望

- 项目搭建了汉语手语翻译系统，仅采用常见电脑手机都能支持的普通RGB相机，使用ViT网络识别汉语手语，同时处理输出语音，实现了从手势到语音的完整系统。
- 准确率高（无遮挡的情况数字100%，拼音75%），具有鲁棒性，所需设备廉价，实用性强。
- 亮点：美国手语→汉语手语，Vision Transformer在手语识别的应用，从手势到语音的完整转换系统，3D姿态估计应用于图像分类。

# 总结&展望

- 项目未来还可以拓展至使用更全面的三维姿态估计，从RGB图像还原深度信息，使用更丰富的信息增强系统的准确性。
- 项目未来还可以将成果封装成APP，实现该系统的实用功能。
- 项目可以扩展更大的训练集和测试集。

# 参考文献

- [1]张淑军,张群,李辉.基于深度学习的手语识别综述[J].电子与信息学报,2020,42(04):1021-1032.
- [2]Zhang Y, Wang J, Wang X, et al. Static hand gesture recognition method based on the vision transformer[J]. Multimedia Tools and Applications, 2023, 82(20): 31309-31328.
- [3]TANG Ao, LU Ke, WANG Yufei, et al. A real-time hand posture recognition system using deep neural networks[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(2): 1-23.
- [4]李琦.基于GCN的RGB三维手势跟踪算法研究[D].内蒙古科技大学,2023.
- [5]Devineau G, Moutarde F, Xi W, et al. Deep learning for hand gesture recognition on skeletal data[C]. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 106-113.
- [6]Zhang Y, Wang J, Wang X, et al. Static hand gesture recognition method based on the vision transformer[J]. Multimedia Tools and Applications, 2023, 82(20): 31309-31328.
- [7]Zimmermann C, Brox T. Learning to estimate 3d hand pose from single rgb images[C]. Proceedings of the IEEE international conference on computer vision. 2017: 4903-4911.
- [8]Kanazawa A, Black M J, Jacobs D W, et al. End-to-end recovery of human shape and pose[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7122-7131.

# Q&A

## 基于视觉方法的手语翻译系统

应逸雯