Software Engineering and Applications 软件工程与应用, 2022, 11(2), 320-329
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.112034

Hans汉斯

基于深度学习的盲人行路辅助软件设计

史娜维，汪华章*

西南民族大学电气工程学院，四川 成都

收稿日期：2022年3月7日；录用日期：2022年4月13日；发布日期：2022年4月20日

摘　要

随着时代的进步，科技化的社会给部分人的生活带来了极大便利，但也使视障人群的生活变得举步维艰，针对国内关于盲人导盲设备的设计存在的不足，如智能化低、精度不高、无法对障碍物进行实时警报等，本文提出了一种结合计算机视觉和自然语言处理的盲人辅助行路的软件设计方案，利用视频检测技术实现实时检测过往车辆、行人以及障碍物，分析其所处位置，并对检测出的目标进行话术丰富，再基于语音合成技术将检测到的周围目标进行实时播报，为盲人提供相应的语音辅助指导，测试结果表明，上述系统能够以低成本的方法实现提高实用性、保障用户安全并且增加测障精度，使得盲人出行与生活更加方便。

关键词

盲人出行，视频检测，语音合成

Design of Walking Assistance Software for Blind People Based on Deep Learning

Nawei Shi, Huazhang Wang*

College of Electrical Engineering, Southwest Minzu University, Chengdu Sichuan

Received: Mar. 7th, 2022; accepted: Apr. 13th, 2022; published: Apr. 20th, 2022

Abstract

With the progress of the times, the technological society has brought great convenience to the lives of some people, but it has also made the lives of visually impaired people difficult. Due to low intelligence, low precision, and inability to provide real-time alerts to obstacles, this paper proposes a software design scheme for blind people's assisted walking that combines computer vision and

*通讯作者。

文章引用: 史娜维, 汪华章. 基于深度学习的盲人行路辅助软件设计[J]. 软件工程与应用, 2022, 11(2): 320-329.
DOI: 10.12677/sea.2022.112034

natural language processing. Video detection technology is used to detect passing vehicles, pedestrians and obstacles, analyze their locations, enrich the detected targets, and then broadcast the detected surrounding targets in real time based on speech synthesis technology, providing corresponding voice assistance guidance for the blind. The test results show that the above system can be implemented in a low-cost way to improve the practicability, ensure the safety of users, and increase the accuracy of obstacle detection, making travel and life of the blind more convenient.

Keywords

Travel for the Blind, Video Detection, Speech Synthesis

1. 引言

在当代社会，盲人很难在没有辅助的情况下生活出行，而中国视障人数约占世界视障人数的 20% [1]。随着科技的发展，各国都致力于增强盲人与世界的互动，各种智能辅具层出不穷[2] [3] [4]，让盲人的出行有了多种选择。但普通的导盲杖智能化低，获取信息的范围和广度存在限制，不能达到全面的导盲需求；传统的 RFID 导盲技术[5]识别精度不高，普及性差；一些智能导盲辅具造价高，往往难以大规模投入使用，且大多设备只能对静态障碍示警，无法分析动态障碍的运动趋势，难以实现实时警报，盲人出行仍存有安全隐患。因此，开发低成本、实时性、高便携和高精度的盲人辅助行路系统有很高的现实意义。

为了关爱盲人群体，改善目前导盲工具落后的欠缺，本文提出了一种结合计算机视觉(Computer Vision, CV)和自然语言处理(Natural Language Processing, NLP)的盲人辅助行路的软件设计方案，利用视频检测技术实现实时检测过往车辆、行人以及障碍物，分析其所处位置，并对检测出的目标进行话术丰富，再基于语音合成技术将检测到的周围目标进行实时播报，为盲人提供相应的语音辅助指导，测试结果表明，上述系统能够以低成本的方法实现提高实用性、保障用户安全并且增加测障精度，使得盲人出行与生活更加方便。

2. 软件整体设计思路

2.1. 设计思路

针对目前导盲设备存在的问题，本系统主要从目标检测和语音合成两个方面进行设计。

环境信息获取的方式包括视觉感知和声音感知。视觉感知用于识别物体和获取物体位置，它依赖于物体检测。而视觉目标跟踪算法中的基于孪生网络(Siamese) [6]系列的跟踪方法主要应用于单目标跟踪，并且多目标跟踪的深度学习网络在成熟度和实时性上都难以达到“导盲”所需的标准，因此目标检测技术相较于目标跟踪技术更适用于“导盲”领域。随着算法的成熟，出现了 RCNN [7]、Faster R-CNN [8]、SSD [9]、Yolo [10]等主流算法，其中 Yolov3 [11]在识别精度和速度方面都有着优秀的表现，对小目标的检测尤为精确，处理器和 GPU 并行运算更是加快了检测速度，实时性满足本设计的需求。并且通过在 Yolov3 中加入空洞卷积(Dilated/Atrous Convolution) [12]来改进网络结构，代替下采样/上采样，在扩大感受野的同时保留输出特征图分辨率。

声音感知主要基于语音合成播报，选用端到端的 Tacotron 神经网络模型[13]，采用 seq2seq + attention 结构，以字符作为输入，将输入的文本输出为音频波形。通过两者结合实现通过实时视频输入来检测周围环境，并经过检测到的周围目标以语音播报的形式对盲人做出引导，提供语音辅助。

2.2. 软件设计流程

软件设计流程主要分为三个部分：1) 目标检测，2) 文本丰富，3) 语音合成。前期数据预处理主要包含视频流的帧抓取，经过 Yolov3 检测后的信息通过筛选和小目标抑制操作后，将清洗过的数据凝练成关键词送入文本丰富模块，经过时间间隔重复操作达到实时检测播报的效果。经过模拟盲人测试，得到每 5 秒钟截取一帧视频数据作为输入为宜，首先，因为盲人的行进速度较慢，场景内的目标变化不会产生十分剧烈的情况；其次，每句语音播报需要一定时间，五秒的间隔足以完成“目标识别－文本丰富－语音播报”的多模态流程。详细流程如图 1 软件结构所示。

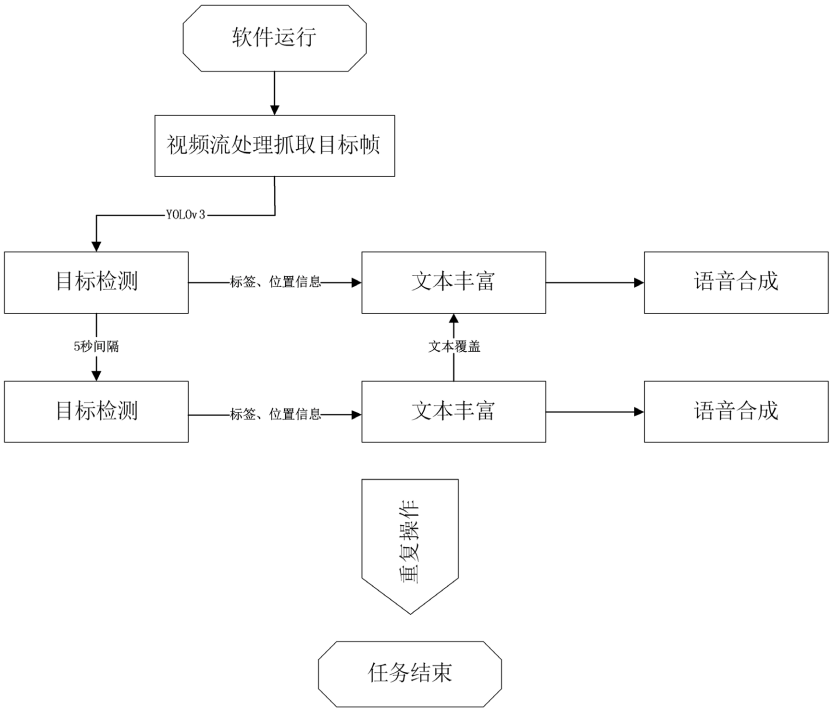

Figure 1. Software structure
图 1. 软件结构

3. 关键技术研究

3.1. 目标检测模块

为了契合本设计对于实时性和多目标检测的要求，本设计采用 Yolov3 算法，将检测问题转化为识别与回归问题。下面通过三个方面介绍 Yolov3 算法的核心思想：

1) 特征提取方式

相较于 Faster R-CNN 模型使用候选区域来提取特征，Yolov3 则选用整图进行训练，这样能够在速度加快的同时，更好的区分背景区域和目标。本设计还使用空洞卷积代替下采样/上采样，引入超参数扩张率(dilation rate)来定义卷积核处理数据时各值的间距，不仅保留了图像的空间信息，还避免了下采样那样

造成信息损失，改善了网络结构。

2) 网络预测方式

Yolov3 以端到端的检测来预测图像，将输入图像分为 S * S 个网格，相应网络会对中心落在该格内的目标进行检测，每个单元格为每个 bounding box (预测区域)预测四个参数：中心偏移量(x_i, y_i)，宽高缩放比(t_w, t_h)，根据 Yolo 模型训练完成后给出的一组标签来判断需要播报的目标及其位置信息通过逻辑回归得到置信度，用来反映当前边界框存在目标的可能性和准确度，如果当前区域有对象存在时，预测目标类别并打上标签。

3) 网络模型

Yolov3 作为一个大型的深度卷积神经网络模型，遵循 GoogleNet [14]思想，但区别在于采用了更快速精准的 Darknet-53 网络结构(包含 53 个全连接层)。当输入为一张 256 * 256 大小的图片，则在 32、16 和 8 维降采样时进行检测，使用 1 * 1 和 3 * 3 的卷积核分别用于降维和特征提取，最终输出大、中、小三个尺度，且三个尺度间存在联系。最后通过置信度大小做逻辑回归，得到预测结果。整体网络结构如图 2。

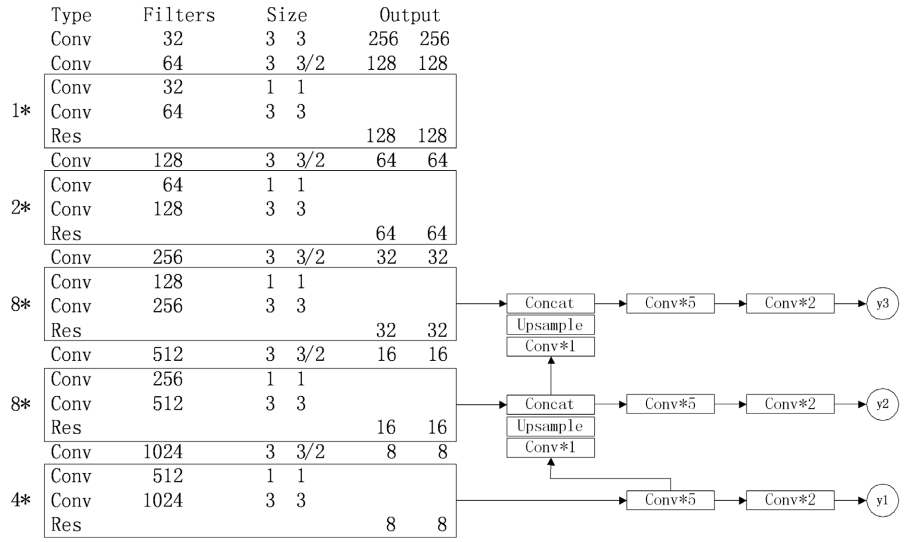

Figure 2. Yolov3 structure
图 2. Yolov3 结构

最后根据模型训练完成后给出的一组标签来判断需要播报的目标及其位置信息，首先对于视频流进行帧抓取，再通过位置信息与抓取帧宽度，判断目标分区，由目标大小辨别目标远近，并在判断过程加入对应识别物的标签作为权重。

3.2. 语音合成模块

选用端到端的 Tacotron 神经网络模型，它包含一个编码器，一个含注意力机制的解码器[15]和一个后处理器，核心是 seq2seq + attention 结构，它以字符作为输入，生成频谱图，并将其转换为波形，而后使用 Griffin-Lim 算法生成对应音频，将文本合成语音，可以仅通过<文本，声谱>数据对随机开始训练。模型结构如图 3。

其中 CBHG 模块用于提取文本特征，包含一维卷积滤波器、一个 Highway Networks 和一个双向 GRU。

输入序列先经过卷积层，它有 K 个一维卷积核，卷积核个数分别为 $C_1 - C_K$，宽度分别为 1 − K，对本身以及上下文信息进行有效建模，再将结果堆叠，沿时轴做最大池化以增加局部不变性，使用 stride = 1 来保留时间上的分辨率。后经由两个一维卷积层，输出通过残差连接与原始输入序列求和，卷积后的结果经过高速网络中提取高维特征，最后在顶部加入双向 GRU，得到最终序列特征。

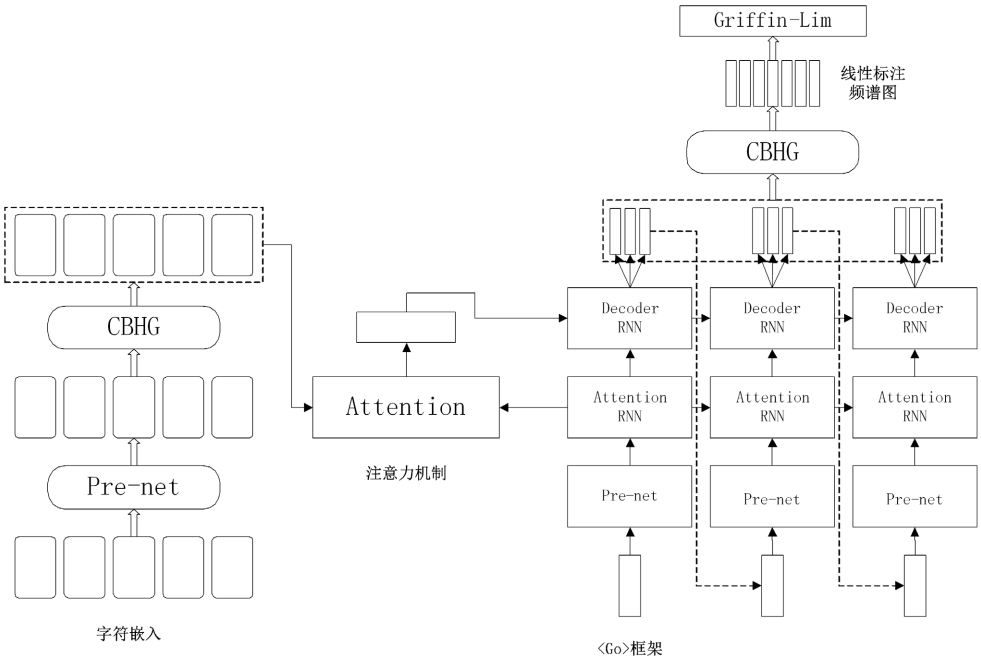

Figure 3. Tacotron model architecture
图 3. Tacotron 模型结构

编码器对输入进行非线性变换。输入是用独热向量编码后嵌入至一个连续向量中的字符序列，通过由全连接层和 dropout 组成的 pre-net 进行预处理，可加速收敛并提高泛化能力。然后将输出的连续向量送入 CBHG 子网络，得到注意模块的最终编码器表示。

解码器主要有两个模块，Attention-RNN 包含 256 个 GRU，使用 Bahdanau attention 机制，生成查询向量，做为输入经过 attention 模块生成文本向量，Decoder-RNN 为两层 residual GRU，它将前一阶段产生的查询向量和文本向量拼接作为输入，输出为输入与经过 GRU 单元输出的和。为了加快收敛速度，还通过生成 80 波段的梅尔频谱图代替直接生成语谱图，在减少计算量的同时保证了最终波形的质量。

后处理网络将 seq2seq 目标转换为可以合成为波形的目标。先由 CBHG 模块将上一层产生的梅尔频谱图转换为线性频谱图，然后利用 Griffin-Lim 算法在不改变左右和相邻的幅度谱的情况下，在一个线性频率范围内预测频谱幅度，将后处理网络的输出合成为时域语音信号。

4. 技术实现

4.1. 视频数据提取关键目标信息

首先是数据预处理过程，为了方便测试，本设计了两种输入方式：摄像头输入和 mp4 视频输入。首先将输入的视频数据赋给变量“video”；通过 video.read()将视频数据读取为一帧图片并赋值给变量“frame”；然后转变其格式(BGRtoRGB)，从而完成视频数据处理。

Yolov3 经过特征提取网络以及预选框操作等，得到一个回归框，每个回归框由 5 个预测值组成：中心偏移量(x_i, y_i)，宽高缩放比(t_w, t_h)和置信度，可以通过感受野映射回原图从而找到目标。

当前帧“frame”包含的信息会以图 4 的形式被提取。包含其标签和 4 个位置信息，而后通过位置信息 x_i、y_i 与抓取帧宽度的一半做差，得到方位值 Z，若 Z 值小于零则代表目标物位于视角的左半区；反之则代表目标位于右半区。接着通过目标大小 t_w * t_h 来判断目标物距离视角的远近，在判断过程中会加入对应识别物的标签作为权重，如 car 的权重为 0.5，person 的权重为 1，dog 的权重为 2。

```
fps= 14.64
b'person 0.87' 188 460 231 484
b'person 0.73' 190 486 229 511
b'car 0.85' 171 97 264 209
b'truck 0.65' 172 96 260 218
fps= 20.52
b'person 0.93' 187 461 231 486
b'person 0.76' 189 489 229 513
b'car 0.88' 170 98 265 211
```

Figure 4. Yolov3 information extraction display
图 4. Yolov3 信息提取展示

4.2. 文本丰富及语音合成

在文本丰富模块中，采用由代码生成固定句式，自动填充经过视频检测模块识别出的标签以及位置信息，然后保存到 txt 文档中。

在语音合成模块中，使用 THCHS-30 数据集，对其进行预处理，生成各个音频文件的梅尔频谱和线性频谱，其中 train.txt 文件中存放有 csv 格式的声谱与拼音标注对。npy 文件使用 numpy 库加载后能够得到数个多维矩阵，用于提取语音的声学特征。

最后的训练效果可以通过观察 alignment 图中的编解码器序列对齐情况，来判断学习是否收敛。

在测试环节中，由于无法直接输入汉字文本，且经由文本生成模块中生成的话术格式为 txt 格式，所以需要使用 python-pinyin 将汉字文本转换为拼音标注，然后拷贝到 eval.py 生成后缀为 wav 的音频，最后使用 python-pydub 进行音频播放。

5. 模型训练

本实验的操作系统为 Ubuntu 18.04.1，使用 NVIDIA GeForce v100 32G GPU 进行训练，程序运行框架为 tensorflow1.14.0 平台。首先使用 VOC 数据集对 Yolov3 网络进行预训练，在 V100 上经过 500 轮的迭代后保留其权重。

通过 Yolo-mark 工具标注部分针对性数据，用矩形框标记数据集图片上的物体，完成后在 img 文件夹下会生成与 jpg 文件同名的 txt 文件，里面每一行代表一个物体的类的编号，以及标记物体的坐标。如图 5 (其中每一行的数据分别代表了：<物体类的编号> <x_center> <y_center> <width> <height>)。

对于 Tacotron 模型的训练，语料库采用由清华大学开放的 THCHS-30 汉语普通话语料，数据采样率为 16 KHz，样本宽度为 16-bit，单声道。在该语料库中每条音频都有对应的两个文件，分别是后缀为 wav 的音频文件以及后缀为 trn 的标注文件。其中，不同的语言标注方法不同，如英文标注可以直接使用自己本身加标点符号，而由于中文的多样性，通常使用拼音做标注，如图 6 所示。

在训练之前，对数据进行预处理，生成各个音频文件的梅尔频谱和线性频谱，其中的 npy 文件使用

numpy 库加载，得到数个多维矩阵，用于提取语音的声学特征。得到<文本，声谱>数据对形式后开始训练。训练中解码步骤使用真实的梅尔谱图，并使用 Teacher Forcing 的方式来减少误差。

1	1 0.716797 0.395833 0.216406 0.147222
2	0 0.687109 0.379167 0.255469 0.158333
3	1 0.420312 0.395833 0.140625 0.166667

Figure 5. Labeling part of the targeted data display
图 5. 标注部分针对性数据展示

```
国务委员 兼 国务院 秘书长 罗干 民政部 部长 多吉 才 让 也 一同 前往 延安 看望 人民群众
guo2 wu4 wei3 yuan2 jian1 guo2 wu4 yuan4 mi4 shu1 zhang3 luo2 gan4 min2 zheng4 bu4 bu4 zhang3 duo1 ji2 cai2 rang4 ye3 yi4 tong2 qian2 wang3 yan2 an1 kan4 wang4 ren2 min2 qu
n2 zhong4
g uo2 uu u4 uu ui3 vv van2 j ian1 g uo2 uu u4 vv van4 m i4 sh u1 zh ang3 l uo2 g an4 m in2 zh eng4 b u4 b u4 zh ang3 d uo1 j i2 c ai2 r ang4 ii ie3 ii i4 t ong2 q ian2 uu u
ang3 ii ian2 aa an1 k an4 uu uang4 r en2 m in2 q vn2 zh ong4
```

Figure 6. Character annotation display
图 6. 字符标注展示

经过 92,000 次迭代，生成的 alignment 图如图 7 所示，图中对齐情况较好，训练基本符合预期。

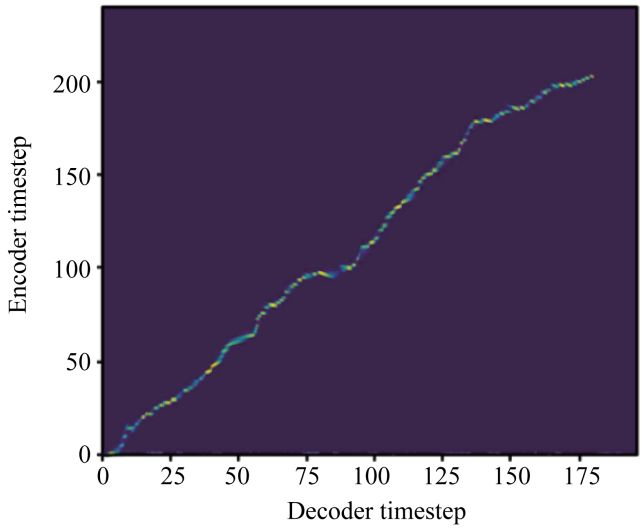

Figure 7. Alignment diagram
图 7. Alignment 图

6. 系统演示

训练结束后，在动态视频中显示对于多目标判断的标签以及准确率。每隔五秒抓取一帧进行目标识别，提取标签以及目标位置信息，当(目标尺寸 * 对应标签的权值)小于 600 时系统判定该目标过小，转换文本过程会忽略过小的目标，每次抓取帧经过判断转换操作会以覆盖的形式生成 test.txt 文本并由语音转换程序不断播报。经过测试，该系统能够实时检测静态和动态障碍物，提高佩戴者的环境感知能力，另一方面，能够对出现的障碍物等目标进行话术丰富，通过语音实时播报障碍物及其位置进行语音预警，实现人机交互。使佩戴者能够在无人引导的情况下了解环境，安全出行。

系统软件演示如图 8 所示。

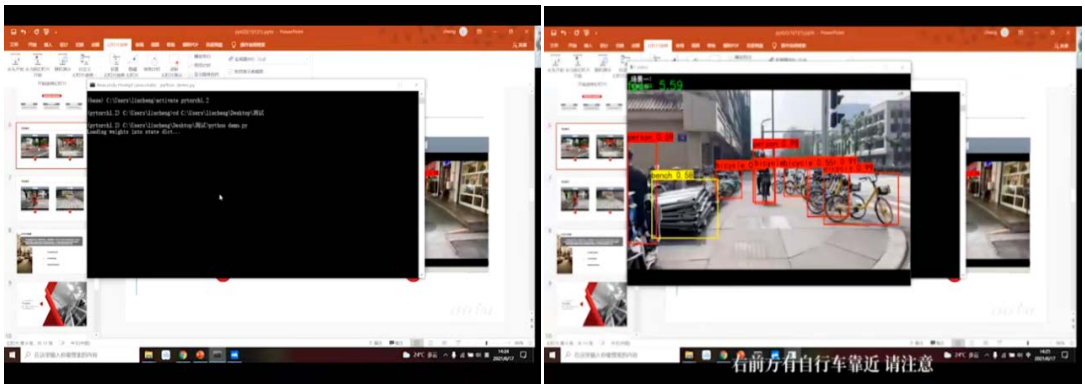

Figure 8. System software display
图 8. 系统软件显示

其中，图 9 展示了视角中出现车辆和行人的情况和其对应的转换文本。

Figure 9. Video detection block diagram and corresponding converted text
图 9. 视频检测框图与对应转换文本

图 10 展示了视角中出现车辆靠近的情况(行人目标过小被忽略)和其对应的转换文本。

Figure 10. Video detection block diagram and corresponding converted text
图 10. 视频检测框图与对应转换文本

图 11 展示了视角中出现狗的情况和其对应的转换文本。

Figure 11. Video detection block diagram and corresponding converted text
图 11. 视频检测框图与对应转换文本

7. 总结

本文提出了一种结合计算机视觉和自然语言处理的盲人辅助行路的软件设计，从视频检测算法、文本丰富、语音合成三个方面进行方案设计，利用 Yolov3 算法实现实时检测过往车辆、行人以及障碍物的功能，使用文本丰富模块对检测出的车辆等目标进行话术丰富，再使用基于 Tacotron 模型的语音合成技术对周围目标进行语音播报。三者结合，构成了盲人辅助行路系统的框架，经过测试表明，本设计方案符合低成本、实时性、高精度要求，已在第 16 届研究生电子设计竞赛中获奖，有一定的应用价值与参考意义。

参考文献

[1] 李冉, 刘正一. 人文关怀理念下的盲人助行产品设计研究[J]. 工业设计, 2021(11): 70-71.

[2] Hwang, A.D. and Peli, E. (2014) An Augmented-Reality Edge Enhancement Application for Google Glass. *Optometry and Vision Science*, **91**, 1021-1030.

[3] 何冰冰, 盛涛, 李凯鹏, 吴明明. 基于 RFID 地铁站内智能导盲系统设计[J]. 南方农机, 2020, 51(9): 106-107+116.

[4] 周浩, 吕俊燕, 杨瑞青. 三模块控制的助盲拐杖设计[J]. 电子制作, 2021(12): 87-88+65.

[5] 武墨晗, 荣学文, 范永. 导盲机器人研究现状综述[J]. 计算机工程与应用, 2020, 56(14): 1-13.

[6] 柳赟, 孙淑艳. 基于自适应模板更新的改进孪生卷积网络目标跟踪算法[J]. 计算机应用与软件, 2021, 38(4): 145-151+230.

[7] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[8] Ren, S., He, K. and Girshick, R. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Washington DC, 6 June 2017, 1137-1149.

[9] Liu, W., Anguelov, D., Erhan, D., *et al.* (2016) SSD: Single Shot MultiBox Detector. *European Conference on Computer Vision*, Amsterdam, 11-14 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[10] Redmon, J., Divvala, S.K., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[11] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv:1804.02767v1.

[12] 候少麒, 梁杰, 殷康宁, 刘学婷, 殷光强. 基于空洞卷积金字塔的目标检测算法[J]. 电子科技大学学报, 2021, 50(6): 843-851.

[13] Wang, Y., Skerry-Ryan, R.J., Stanton, D., *et al*. (2017) Tacotron: Towards End-to-End Speech Synthesis. *Proceedings of Interspeech* 2017, Stockholm, 20-24 August 2017, 4006-4010. https://doi.org/10.21437/Interspeech.2017-1452

[14] Szegedy, C., Liu, W., Jia, Y., *et al*. (2015) Going Deeper with Convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 1-9.

[15] 王庆尧. 基于强制单调注意力机制的改进 Tacotron2 语音合成方法[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2021.