基于多级 Sigmoid 神经网络的城市交通场景理解*

谭论正[1],夏利民[1],夏胜平[2]
(1. 中南大学 信息科学与工程学院,湖南 长沙 410075;
2. 国防科技大学 ATR 重点实验室,湖南 长沙 410073)

摘 要:交通场景的理解是交通监控、汽车安全辅助驾驶的基础。提出一种基于多级 Sigmoid 神经网络的城市交通环境理解方法。将5个3D结构特征与物体外观特征相结合表征城市交通环境,为了提高交通环境识别率,采用多级 Sigmoid 神经网络(MSNN)进行图像分割与识别。在公共测试视频数据库 CamVid dataset 进行实验,实验结果表明了该方法的有效性。

关键词:空间结构特征;城市交通场景;多级 Sigmoid 神经网络
中图分类号:TP391 **文献标志码**:A **文章编号**:1011-2486(2012)04-132-06

Urban traffic scene understanding based on multi-level sigmoidal neural network

TAN Lunzheng[1], *XIA Limin*[1], *XIA Shengping*[2]
(1College of Information Science and Engineering, Central South University, Chang sha 41007, China;
2. ATR Key Lab, National University of defense Technology, Changsha 410073, China)

Abstract: Urban traffic scene understanding is the basis of traffic monitoring and safety driving assistant system. A novel approach to understanding urban traffic scene captured from a car-mounted camera is proposed based on multi-level Sigmoidal neural network. Five 3D structure features were combined with the appearance features to represent the urban traffic environment and the recognition accuracy of traffic environment was improved by utilizing multi-level Sigmoidal neural network (MSNN) to segment and recognize the input images. Tested by the public CamVid dataset, the experimental results demonstrate the efficiency of the proposed approach.

Key words: 3D structure; urban traffic scene; multi-level sigmoidal neural network

城市交通场景理解是交通监控、汽车安全辅助驾驶的重要组成部分,具有重要的研究意义[1-3],目前,针对城市交通环境的视频图像的研究已成为图像处理的一个研究热点。然而,由于城市交通环境复杂(场景中通常包括大量的物体:车辆、道路、行人、树木、建筑物、交通标志、天空等),使得城市交通场景理解非常困难。

交通场景理解主要是识别场景中的物体,因此目前交通场景理解的研究主要集中在场景分割和场景中物体的识别,如 TextonBoost[4] 是一种利用纹理、布局和上下文的联合建模进行交通场景分割与多类别的物体识别的方法;条件随机场(CRF)[5-6]通常用来结合图像的纹理、颜色、位置等信息进行交通场景图像分割。然而这些方法都没有利用物体的运动特征,Brostow[7] 提出了一种利用运动结构粒子云图进行交通场景图像分割和识别的方法,通过分析3D粒子云图得到物体的空间结构特征,利用这些空间结构特征进行图像的语义分割。此方法具有较高的鲁棒性,可以适应各种环境条件,缺点是分割的图像缺乏清晰的边界;Sturgess[8] 在此基础上提出结合物体的外观特征和运动结构特征进行交通场景理解的方法,此方法大大提高了图像的分割效果,但该方法对路旁柱子分割效果较差,并且未对路面标记作出分割。文献[9]在对城市交通场景图像进行分块的基础上,采用分层的动态 CRF 模型分离出物体和背景[10],并利用车载双目摄像机输入的图像计算物体的空间深度信息,该算法主要功能集中在分辨各种路面方向标记,以及识别车辆、行人等车前方物体,该方法对理解城市道路有重要的应用意义,其不足在于分割的图像效果比较粗糙。

本文算法结合物体的运动特征和外观特征,

* **收稿日期**:2011-09-19
基金项目:国家863计划项目(2009AA11Z205);国家自然科学基金项目(50808025);国家教育部博士点基金项目(20090162110057)
作者简介:谭论正(1981—),女,湖南株洲人,博士研究生, E-mail:tanlunzheng@126.com;
夏利民(通信作者),男,教授,博士,博士生导师, E-mail: xlm@mail.csu.edu.cn

进行图像分割和物体识别。首先提取 5 个关键的城市交通环境的 3D 结构特征,并将这些特征映射到 2D 图像中,结合物体外观特征,采用多级 Sigmoid 神经网络(MSNN)进行图像分割、识别。实验结果表明该方法可以达到理想的图像分割效果,特别是对路面标记的分割与识别,为汽车辅助驾驶提供了重要的信息。

1　城市交通场景特征提取

1.1　交通环境的 3D 结构特征

使用单目车载摄像头获取的视频序列 I,根据结构来自于运动(SFM)[11]产生 3D 粒子云图。首先,对单帧图像 $i \in I$ 进行色彩转换为 CIE $L^* a^* b^*$ 色彩模式,并将 a^* 和 b^* 颜色通道的值置为 0,得到灰度图像。由于角点信息在刚体运动中具有稳定的特征,本文选择具有高可靠性和稳定性同时实效性较高的 SURF(Speeded Up Robost Features)[12]来计算角点的特征描述符。利用连续的视频帧之间的相似性,我们在连续 5 帧图像中采用实时的归一化相关匹配算法进行角点的 SURF 特征描述符匹配,产生稳定特征点的 2D 运动轨迹,在此基础上进行摄像机自标定[11]。本文中的摄像机固定在汽车上,仅发生刚体运动(非限定运动):旋转是绕转动轴的旋转,平移是沿轴方向的平移。刚体运动的旋转矩阵和平移矩阵可以由 2D 特征点的运动轨迹来求取。

摄像机自标定以后,可以计算 2D 图像中所有像素点的三维坐标值 $W(x,y,z)$。为了提取交通场景的关键特征,我们按文献[7]的方法计算空间点 W 的 5 个 3D 结构特征:

(1) 点 W 相对于摄像机的垂直距离 f_H

以汽车的垂直方向为摄像机的 y 轴方向,空间中每个点 W 相对于摄像机中心 C_y 的高度:$f_H(W) = W_y - C_y$。

(2) 距摄像机路径水平方向最近的距离 f_C

道路具有一定的宽度,通过对整个视频序列计算其两旁的物体(如建筑、树木等)距摄像机路径最近的距离。$C(t)$ 代表摄像机的中心,$f_C(W) = \min_t \| W - C(t) \|$。

(3) 表面的方向

对每帧中的 3D 点 W 进行 2D Delaunay 三角化,得到近似的表面方向 fo_x, fo_y。

(4) 密度

$f_D(t)$ 代表图像 i 中的特征点在连续几帧图像中运动轨迹的密度。快速运动的物体相对于静止物体具有稀释的点云密度,例如建筑和树木的关键点的密度高于道路和天空。

(5) 射影误差 f_R

射影误差用来度量已估计出的 3D 空间中的点的齐次坐标射影到 2D 图像平面的上影像点的齐次坐标与特征点 (u_t, v_t) 的差值,射影误差 $f_R(W) = \ln[1 + q(W)]$,$q(W)$ 为由 3D 到 2D 的射影运算。射影误差用于显示由于场景中物体的自运动产生的射影误差。此特征可以用来区分运动物体与静止背景。

我们用 f_C 判断场景中物体距离汽车的水平距离(f_C 的值较大,则可能是道路以外物体,较小则是路中障碍物);f_H 判断物体的高度;f_R 的值可用来区分场景中静止的物体与运动的物体(f_R 的值大是运动的物体,小是静止的物体);密度 f_D 可判断场景中的物体是否快速运动(f_D 的值越小,物体运动越快)。例如 f_C、f_H、f_D 的值大,而 f_R 的值小,则此物体可能为道路两旁的建筑物。如果 f_C、f_H 的值较小,则可能是道路中的障碍物,同时如果 f_D 的值小且 f_R 的值大,则判断是道路上运动的汽车;否则,如果 f_D 的值比较大且 f_R 的值比较小,则可能是道路上速度较慢的自行车或行人。

1.2　3D 到 2D 的摄影机投影

将得到的 3D 特征点的空间结构特征投影到 2D 的摄影机图像中,如图 1 所示。定义空间梯形塔 $p(x,y)$ 为由 3D 空间映射到 2D 平面 $r(x,y)$ 的空间部分,2D 平面中的 $r(x,y)$ 是以特征点 (x,y) 为中心的矩形。为了减少稀疏的空间点阵可能带来的误差,我们在一个空间范围内求 2D 点的空间结构特征。

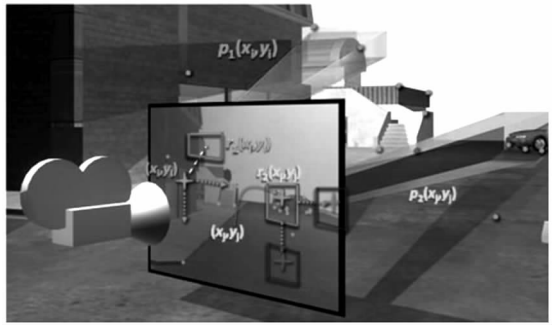


图 1　3D 射影到 2D

Fig. 1　Projecting from 3D to 2D

(1) 对 $p(x,y)$ 中粒子的高度 f_H、距摄像机最近的距离 f_C、射影误差 f_R 分别求和,得到 2D 图像中特征点 (x,y) 的相应空间特征值:

$$F_T(x,y) = \sum_{W \in p(x,y)} f_T(W) \quad T \in \{H,C,R\} \quad (1)$$

(2) 表面方向是通过直接对 2D 矩形 $r(x,y)$ 的所有图像点进行三角网格化,求和得到表面方

向 $\boldsymbol{F}_{Ox}$,$\boldsymbol{F}_{Oy}$:

$$F_{O_x}(x,y) = \sum_{(x',y')\in r(x,y)} f_{O_x}(x',y') \quad (2)$$

同理求出 $\boldsymbol{F}_{Oy}$。

(3) 2D 图像中特征点(x,y)的运动密度是其对应的 3D 空间金字塔 $\boldsymbol{p}(x,y)$中所有粒子的数量和:

$$F_D(x,y) = |\{W\in p(x,y)\}| \quad (3)$$

1.3 物体外观特征

本文选取物体纹理特征以及 9 维梯度方向直方图(HOG)作为物体的外观特征参与图像分割。

1.3.1 物体纹理特征

采用文献[13]的方法提取纹理特征:将图像都转换为 CIE $L^*a^*b^*$ 色彩模式下的灰度图像。使用 17 维滤波器组(包括多尺度高斯滤波、x 和 y 方向高斯微分以及拉普拉斯高斯滤波)将纹理转换到不同尺度和方向上的纹理向量,然后对纹理向量进行 K 均值聚类(K 值较大),需要使用训练图像进行特征训练建立 T 维的纹理直方图($T<<K$),聚类过程中采用 Mahalanobis 距离度量特征间的距离。

为每个类定义一个高斯分布的直方图均值$\overline{H}_c$以及对角协方差 β_c(采用 Mahalanobis 距离度量),那么该类中每个物体的纹理直方图 H 都靠近该类的直方图均值$\overline{H}_c$。定义 $\theta=(\overline{H},\beta)$为判断物体是哪类的概率。

$$P(\theta) = \prod_{i=1}^{T} N(\overline{H}_i^{\frac{1}{2}} \mid \mu,(\lambda\beta_i)^{-1}) g(\beta_i \mid a,b) \quad (4)$$

$$P(H \mid \theta) = \prod_{i=1}^{T} N(H_i^{\frac{1}{2}} \mid \overline{H}_i^{\frac{1}{2}},\beta_i^{-1}) \quad (5)$$

其中 $\mu=0$,$\lambda=0.1$,$a=0.01$,$b=0.01$,g 为 gamma 分布,H_i 为直方图中第 i 位。使用 c 个类别的人工标记好的训练图像建立新的 T 维纹理直方图 $H_1,\cdots,H_N$,N 个训练区域,每个区域有一个纹理直方图,定义 R_c 是标记为类别 c 的区域,θ_c 为类别 c 的概率:

$$P(\{H_n\} \mid c) = \prod_{c=1}^{C} \int \prod_{n\in R_c} P(H_n \mid \theta_c) P(\theta_c) \mathrm{d}\theta_c \quad (6)$$

使用 8 类交通场景相关的纹理(建筑、树木、天空、汽车、自行车、路面、柱子、行人),采用来自 CamVid 的人工标记的图像进行训练,建立 50 维的纹理直方图。测试图像的每个像素经过滤波器组滤波,选择其最近的聚类中心的纹理特征作为该像素的纹理。

1.3.2 梯度方向直方图(HOG)

对灰度图中的每个像素构造一个 9 维梯度方向直方图 Hog(n)。首先对每个像素的 8 邻域利用 Sobel 梯度算子进行梯度计算:

$$\begin{aligned}\frac{\partial f}{\partial x} &= f(x-1,y+1)+2f(x,y+1)\\ &\quad +f(x+1,y+1)-f(x-1,y-1)\\ &\quad -2f(x,y-1)-f(x+1,y-1)\\ \frac{\partial f}{\partial y} &= f(x-1,y-1)+2f(x-1,y)\\ &\quad +f(x-1,y+1)-f(x+1,y-1)\\ &\quad -2f(x+1,y)-f(x+1,y+1)\end{aligned}$$

$$G[f(x,y)] = \sqrt{\left[\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2\right]} \quad (7)$$

然后进行直方图统计的方向单元划分,将梯度方向在 0°~360°量化为 9 级(故 $1\leqslant n\leqslant 9$),每 20°为一级。对像素点的 8 邻域内的每个像素的梯度模作加权,得到其 9 维梯度直方图。

2 基于多级 Sigmoid 神经网络图像分割与识别

我们利用多级 Sigmoid 神经网络(MSNN)进行图像分割与物体识别。与其他分类器学习机比较,神经网络分类器具有很多优势,例如自适应性、非线性逼近、易于训练、高选择性、联想记忆以及对特征空间进行任意形状的分割的能力等。然而,一般的神经网络模型由于其神经元采用标准的 Sigmoid 函数,导致其分类准确度低,一般只能进行 2 类别分类,通用能力差。为了弥补此缺陷,让神经元产生多元反应,进行多类分类,我们采用多级 Sigmoid 神经网络,如图 2 所示。

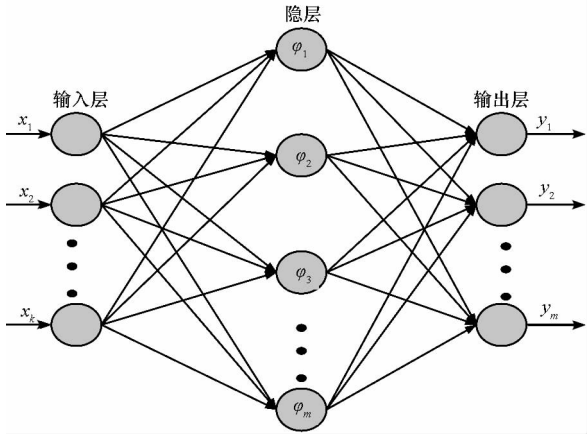


图 2 多级 Sigmoid 神经网络

Fig. 2 Multi-level Sigmoidal neural network

采用的神经网络为三层神经网络,包括输入层、隐含层、输出层,对于输入信号,要先向前传播

到隐含层节点,经激活函数后,再把隐含层节点的输出信号传播到输出节点,最后给出输出结果。节点的激励函数一般为标准的 Sigmoid 函数:

$$f(x) = \frac{1}{1 + e^{-\beta x}} \tag{8}$$

其中 β 为常量,由于该类型的神经网络只能进行 2 分类,为了进行多分类,因此采用多级 Sigmoid 神经网络,激活函数为[14-15]:

$$\varphi_r(x) \leftarrow f(x) + (\gamma - 1) f(c), \quad 1 \leqslant \gamma \leqslant K \tag{9}$$

其中 K 为层级数,c 为常量(代表类的带宽)。

输出层节点是隐层基函数的输出进行线性加权组合,即输出层的输出 y_j 为:

$$y_j = \sum_{i=1}^{m} \omega_{ij} \varphi_i(x), \quad j = 1, 2, \cdots, p \tag{10}$$

式中 ω_{ij} 为隐层第 i 个节点到输出层第 j 个节点之间的连接权系数,m 为隐层节点个数。对输出值进行归一化,取归一化后最大的为该节点的类别:

$$g_l(\boldsymbol{x}, \boldsymbol{\theta}_l) = \frac{\exp y_l(\boldsymbol{x}, \boldsymbol{\theta}_l)}{\sum_{j=1}^{J} \exp y_j(\boldsymbol{x}, \boldsymbol{\theta}_j)}$$

若

$$Label(X) = \operatorname{argmax}(g_j(x)) \tag{11}$$

则决策 $X \in L_j$,即输入属于第 j 类。

我们用 CamVid dataset[16] 中 350 幅人工标记好的图像输入神经网络对网络进行训练。先提取图像的纹理、梯度直方图以及像素点的空间结构信息($F_C, F_H, F_R, F_{Ox}, F_{Oy}, F_D$),利用 BP 算法对多级 Sigmoid 神经网络进行学习,最终该网络能将城市交通场景图像的像素分成 13 个类(空集、建筑、汽车、树木、柱子、旁路、栅栏、符号标记、路面标记、自行车、道路、行人、天空),即在完成图像分割的同时,进行物体识别。

3　实验与结果

使用公共测试视频数据库 CamVid 进行本文的实验。此数据库视频是使用车载单目摄像机在拥挤的城市交通环境中,驾驶员按照平常的驾驶习惯自然驾驶的情况下拍摄的。视频图像为彩色图像,图像像素大小为 960 × 720,帧率为 30 帧/秒。包括 3 段白天和 1 段夜晚环境,共计 10 多分钟。该数据库每 30 帧进行一次人工图像分类标记(用不同的颜色代表不同的物体),标记类别达 32 种。我们将 700 余幅人工标记好的图像的 50% 作为训练图像,另外 50% 作为测试图像。本文实验使用 13 种城市交通场景常见的物体(建筑、汽车、树木、柱子、旁路、栅栏、符号标记、路面标记、自行车、道路、行人、天空、空集),进行图像标注,其对应的颜色如图 3 所示。

图 3　13 种物体类的标记颜色

Fig. 3　List of the 13 object class names and their corresponding colours used for labelling

实验使用 CPU 为 Intel 酷睿 2 四核处理器(3GHz)、4G 内存、独立显卡的计算机,算法使用 C + + 和 OpenCV 实现。每个像素包含 65 维特征图像纹理、9 维梯度直方图以及像素点的空间结构信息($F_C, F_H, F_R, F_{Ox}, F_{Oy}, F_D$),输出为 13 个类(建筑、汽车、树木、柱子、旁路、栅栏、符号标记、路面标记、自行车、道路、行人、天空、空集),对每个像素进行标记,为了便于理解,在像素标记的基础上还标出了类别名字。使用 CamVid 数据库中人工标记的图像对多级 Sigmoid 神经网络进行训练。

为了对比算法的性能,挑选了 3 个白天一个夜晚环境的测试片段(每个片段包括 30 帧连续的图像以及数据库中对应的人工标记图像),将本文算法与文献[7](简称 Brostow 算法)和文献[8](简称 Sturgess 算法)的方法进行图像分割对比实验。图 4 是原始测试图像(30 帧连续视频图像中最后 1 帧),图 5 为人工标记结果,图 6 (a)、(b)、(c)分别为文中方法、文献[7]、文献[8]的方法的分割与识别结果。

可以看到,本文方法结合物体的运动结构特征和外观特征,使用标准数据库中人工标记好的图像训练多层 Sigmoid 神经网络,达到较好的图像分割效果,并利用图像的 HOG 特征对路面标记作出分割和拟合。Brostow 算法仅仅依靠图像的运动结构信息进行图像分割,由于选取图像特征较少,图像分割粗糙,效果较差。Sturgess 算法利用条件随机场将运动结构和图像外观特征结合起来进行图像分割,其图像分割效果有了很大改善;但是由于使用 CRF 图像分割机制,该方法对柱子、路面标记等较窄的物体的分割效果较差,该方法也未对路面标记作出分割。本文在路面标记的分割与识别方面做出了改进,可用于汽车辅助驾驶研究中识别路面标记的导向,进而预测驾驶行

为,因此本文的工作对于汽车辅助驾驶具有重要 的意义。

图 4 原始测试图像

Fig. 4 Test image

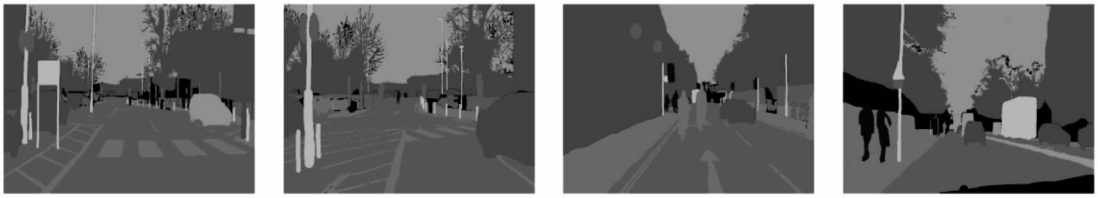
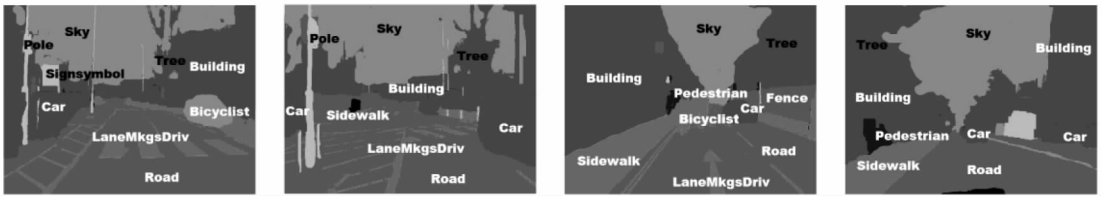


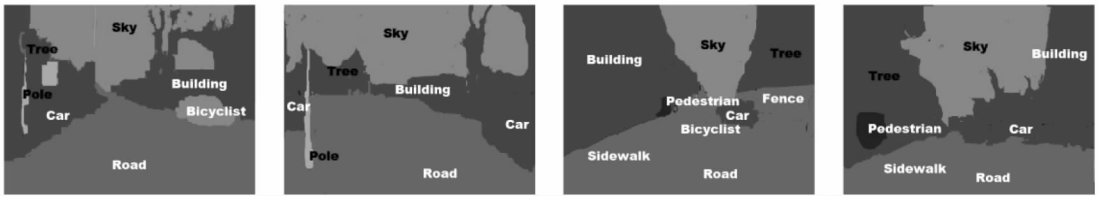
图 5 人工标记结果

Fig. 5 Ground truth labelling image



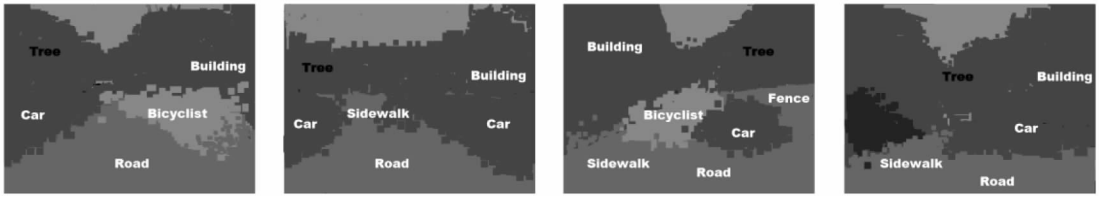
(a) 本文算法分割的结果

(a) Results of our approach



(b) Sturgess 算法分割的图像

(b) Results of Sturgess' method



(c) Brostow 算法分割的图像

(c) Results of Brostow' method

图 6 各种分割方法的分割结果比较

Fig. 6 Segmentation results of different method

进一步,我们采用 350 幅人工标记的图像作为评价图像分割准确性的判断标准,采用式(12)评价指标 PCR 来评估图像分割效果,综合 350 幅测试图像的 PCR 平均值为该算法图像分割的准确性。比较结果如表 1 所示。从表 1 看出,Sturgess 算法正确率为 83.8%,而 Brostow 算法仅为69.1%,本文方法达到 88.7%,其分割效果得到大大提高。

$$\mathrm{PCR} = \frac{\text{正确分类的像素数目}}{\text{总的像素数目}} \tag{12}$$

表 1　三种分割方法的正确率比较

Tab. 1　Comparison of three sementation methods

Method	EER	PCR
Proposed	0.113	0.887
Sturgess 算法	0.162	0.838
Brostow 算法	0.309	0.691

4　结　论

本文提出了一种结合交通场景运动结构特征和外观特征进行交通场景理解的方法。该算法首先提取 5 个关键的城市交通场景三维结构特征,并结合物体的纹理特征以及 HOG 特征,利用多级 Sigmoid 神经网络进行交通场景分割与物体识别,实验结果表明,本文算法在道路路面标记的分割上有了显著提高,这对于辅助驾驶具有重要的现实意义。

在进一步的研究中,我们将尝试在本文基础上引入一种算法用于分析驾驶员在城市交通场景中的驾驶行为,并建立学习机制,在此基础上预测具体的交通场景中的驾驶行为。

参考文献(References)

[1] Medici P, Caraffi C, Cardaralli E, et al. Real-time road signs classification[C]//IEEE International Conference on Vehicular Electronics and Safety, 2008:253 – 258.

[2] Fleyeh H. Road and traffic sign color detection and segmentation—a fuzzy approach[C]//Conference on Machine Vision Applications, May 16 – 18, 2005, Tsukuba Science City, Japan 2005:124 – 127.

[3] Heracles M, Martinelli F, Fritsch J. Vision – based behavior prediction in urban traffic environments by scene categorization [C]//BMVC, Wales, UK, 2010:1 – 11.

[4] Shotton J, Winn J. TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation[C]//ECCV, Graz, Austria May 7 – 13, 2006 (1):1 – 15.

[5] Ladicky L, Russell C, Kohli P, et al. Associative hierarchical CRFs for object class image segmentaiton[C]//CVPR, June 20 – 25, 2009, Florida, USA.

[6] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labelling sequence data [C]//ICML, 2001:282 – 289.

[7] Brostow G J, Shotton J, Fauqueur J, et al. Segmentation and recognition using structure from motion point clouds[C]//ECCV, Marseille, France, October 12 – 18, 2008(1):44 – 57.

[8] Sturgess P, Alahari K, Ladicky L, et al. Combining appearance and structure from motion features for road scene understanding[C]//BMVC, London, Sep 2009.

[9] Ess A, Mueller T, Grabner H, et al. Segmentation-Based urban traffic scene understanding[C]// BMVC, London, England, 2009: 1 – 11.

[10] Wojek C, Schiele B. A dynamic CRF model for joint labelling of object and scene classes[C]//ECCV, Marseille, France, October 12 – 18, 2008.

[11] Hartley R, Zisserman A. 计算机视觉中的多视图几何[M]. 韦穗,等译. 合肥:安徽大学出版社, 2002:345 – 350.
Hartley R, Zisserman A. Multiple view geometry in computer vision[M]. Hefei: Anhui University press, 2002:345 – 350. (in Chinese)

[12] Bay H, Tuytelaars T, VanGool L J. SURF: speeded up robust features[C]. ECCV, Graz, Austria, 2006:404 – 417.

[13] Winn J, Criminisi A. Object categorization by learned universal visual dictionary[C]//ICCV, Beijing, China, 2005.

[14] Dutta P, Bhattacharyya S, Dasgupta K. Multi-scale object extraction using a self organizing neural network with a multi-level beta activation function. [C]//ICISIP, 2004:139 – 142

[15] Bhattacharyya S, Dutta P, Minka U. Object extraction using selforganizing neural network with a multi-level sigmoidal transfer function[C]//ICAPR, Kolkata, India, 2003:435 – 438.

[16] Brostow G J, Fauqueur J, Cipolla R. Semantic object classes in video: high-denition ground truth database[J]. Pattern Recognition Letters, 2009, 30(2): 88 – 97.