基于语义目标匹配的三维跟踪注册方法

安喆[1**]，徐熙平[1*]，杨进华[1]，刘洋[1]，闫宇轩[2]
[1]长春理工大学光电工程学院，吉林 长春 130022；
[2]长春理工大学高功率半导体激光国家重点实验室，吉林 长春 130022

摘要 提出了一种基于语义目标匹配的三维跟踪注册方法。通过改进的单发多框检测(SSD)深度卷积神经网络对图像进行语义分割，获取场景中不同目标的像素级语义分割结果。在求取相机姿态的目标函数时，融合了图像的灰度约束与几何约束对相机的姿态进行估计。所提方法减小了特征点的缺乏或误匹配问题对三维跟踪注册算法性能的影响，且能够适应不同结构的场景。研究结果表明，该方法的误差不超过 2.2 pixel，基本满足了实时性的要求。

关键词 测量；增强现实；语义分割；相机姿态估计；三维跟踪注册
中图分类号 TP391 **文献标识码** A **doi**：10.3788/AOS201838.1212002

Three-Dimensional Tracking Registration Method Based on Semantic Object Matching

An Zhe[1**], Xu Xiping[1*], Yang Jinhua[1], Liu Yang[1], Yan Yuxuan[2]
[1] *School of Optoelectronic Engineering, Changchun University of Science and Technology, Changchun, Jilin* 130022, *China*;
[2] *State Key Laboratory of High Power Semiconductor Lasers, Changchun University of Science and Technology, Changchun, Jilin* 130022, *China*

Abstract A three-dimensional (3D) tracking registration method is proposed based on the semantic object matching. The improved single-shot multi-box detector (SSD) deep convolution neural network is used to segment images semantically and thus the pixel level semantic segmentation results for different objects in the scene are obtained. To solve the object function of the camera pose, the camera pose is estimated by the combination of the gray and the geometric constraints of images. The proposed method not only reduces the influence of the lack or mismatch of feature points on the performance of 3D tracking registration algorithm, but also it can adapt to the scenes with different structures. The research results show that the error of this proposed method is less than 2.2 pixel, which basically satisfies the requirement of real-time.

Key words measurement; augmented reality; semantic segmentation; camera pose estimation; three-dimensional tracking registration
OCIS codes 120.4820; 100.3020; 110.2960; 170.3010

1 引 言

增强现实(AR)技术以真实场景信息为基础，以虚拟对象作为补充信息，通过对真实世界场景的增强，提升了人类对外界的感知能力[1-4]。AR 系统一般包括光学系统、场景目标识别、三维跟踪注册、实时交互等几个方面的内容[5-6]。三维跟踪注册技术[7]作为 AR 系统的关键技术之一，实现了真实世界场景与计算机产生的虚拟信息的融合与配准。

三维跟踪注册是指将虚拟物体正确放置在现实场景的三维坐标中，并且获取指定时间下虚拟物体的二维投影，使虚拟物体与现实世界融合一致。摄像头拍摄的图像是三维世界在二维平面的映射，因此，三维注册估计摄像机在世界坐标系下的位姿，并用该位姿来获取虚拟物体在当前视角下的投影图像，再通过图像合成完成三维注册。现有的三维跟

收稿日期：2018-06-22；**修回日期**：2018-07-20；**录用日期**：2018-07-25
基金项目：国家自然科学基金(61605016)
* **E-mail**：xxp@cust.edu.cn；** **E-mail**：2016200046@mails.cust.edu.cn

踪注册方法一般分为有标识[8-10]和无标识[11-13]两种。有标识的三维跟踪注册方法在场景中放置人工标识物，或通过识别自然标识进行注册，通常选取具有最佳对比度且较容易被检测到的标记，然后将此类已知的模型放置在场景中进行跟踪。通过这种方式使检测和跟踪的方式更简单，且精度更为可靠。比较成熟的标记设计为采用圆形或方形的标记，这两种形状都可以很方便地在图像中检测到。圆形标记有一个形心点，而正方形有四个角点，因此根据六自由度恢复相机姿态的理论，需要三个对应点即可实现对相机姿态的跟踪。但是，此类方法完全依赖于标识信息，若标识不在视角区域内，则无法完成注册。无标识的三维跟踪注册方法解决了这一问题，此方法一般基于场景中的特征点实现虚实融合。其基本过程为：首先提取图像特征点，并对图像间的特征点进行匹配；然后经过最小化特征点的重投影误差估计出图像帧之间的相对位姿变换矩阵，即相机的姿态矩阵；再通过对场景目标的识别，实现虚拟信息与真实环境的融合[14-15]。但是，提取特征点需要一定时间，且容易出现特征点误匹配的情况，需要经过去除冗余点的计算过程。这对系统的实时性是一个挑战，而且也无法完全消除误匹配。此外，特征点的提取损失了大量图像信息，且在纹理较少的环境中，无法提取足够数量的特征点，三维跟踪注册的性能受到影响。

针对上述问题，提出了一种基于语义目标匹配的相机姿态估计方法，以实现虚拟图像与现实场景之间的三维跟踪注册。所提算法利用了场景目标识别过程中得到的目标像素分类结果。采用基于改进的单发多框检测(SSD)语义分割网络[16]对场景图像进行分割，获得了较好的分类效果。通过计算灰度与深度约束结合的目标函数，获取了图像帧之间的相对位姿变换矩阵，最终实现了虚拟图像与现实场景的三维跟踪注册。所提算法避免了基于特征点的三维跟踪注册方法会损失大量图像信息的问题，并且省去了特征提取、特征匹配及误匹配消除等步骤，不受场景纹理复杂度的影响，在保证实时性的同时提升了三维跟踪注册的精度。

2　算法基本原理

所提的基于语义目标匹配的三维跟踪注册算法选取了 KITTI 数据集[17]对改进的 SSD 网络进行训练。首先利用深度卷积网络对输入的图像对进行语义分割，输出场景中各类目标的像素分类结果，这样就获取了周围环境的内容信息。得到目标的分类信息后，结合双目相机获取的深度图像，采用灰度与深度约束联合估计的方式，得到相机的姿态估计结果。最后将虚拟信息与现实场景进行融合。

2.1　基于改进 SSD 的语义分割网络

与传统的语义分割方法不同，采用学习的方法对网络进行训练更加方便，不需要人为设计算法，并且不需要事先提取特征。采用课题组之前设计的改进 SSD 网络对图像进行语义分割，在训练时，输入的图像大小为 300 pixel×300 pixel，图片来自于 KITTI 数据集，卷积过程采用空洞卷积算法[18]。激活函数选取了修正线性单元(ReLU)函数，卷积操作后得到用于检测识别的特征图，然后利用双线性插值的方法对特征图进行上采样操作，逐层恢复特征图大小。最后通过非最大抑制(NMS)策略筛选结果，即根据得分矩阵和目标区域的坐标信息，找到置信度比较高的边界框。通过对网络的训练，可以同时得到目标识别与像素分割的结果，为相机姿态的估计提供环境信息。

2.2　灰度-几何约束的三维跟踪注册

影响三维跟踪注册精度的一个重要因素是相机的姿态估计，常用的方法是根据图像之间的灰度约束建立相机姿态估计的目标函数，并对函数进行优化，得到变换矩阵。这种方式对纹理环境较为敏感，但易受光照环境的影响。另一种常用的方法是根据目标之间的几何约束关系对相机姿态进行估计，根据获取的深度信息将二维图像转化为三维点云，并建立约束方程，对相机姿态进行直接估计，该方法受光照的影响较小。因此，将灰度-几何约束结合，并在目标区域间进行目标函数的计算，也可以得到相机的姿态矩阵。

灰度约束是基于图像间的灰度不变假设原理。灰度不变假设即理论上，在同一空间中的两幅图像，对应像素点的灰度值应相同。设空间中任意一点 $\boldsymbol{p}$ 在三维坐标轴上的坐标分别为 x，y，z，则 $\boldsymbol{p}=\begin{pmatrix}x\\y\\z\end{pmatrix}$。点 $\boldsymbol{p}$ 在关键帧图像中的像素坐标 $\boldsymbol{p}_{\mathrm{f}}$ 的求解公式为

$$\boldsymbol{p}_{\mathrm{f}}=(u_{\mathrm{f}},v_{\mathrm{f}})^{\mathrm{T}}=\frac{1}{Z_1}\boldsymbol{K}\boldsymbol{p}, \tag{1}$$

式中 $(u_{\mathrm{f}},v_{\mathrm{f}})$ 为空间点 $\boldsymbol{p}$ 在相机成像平面上的坐标；Z_1 为在关键帧图像下的深度；$\boldsymbol{K}$ 为相机的内参矩阵；T 代表求转置。关键帧与当前帧之间的旋转矩

阵为 $\boldsymbol{R}$，平移向量为 $\boldsymbol{t}$，则此时点 $\boldsymbol{p}$ 在当前帧中的像素坐标 $\boldsymbol{p}_{\mathrm{b}}$ 为

$$\boldsymbol{p}_{\mathrm{b}}=(u_{\mathrm{b}},v_{\mathrm{b}})^{\mathrm{T}}=\frac{1}{Z_2}\boldsymbol{K}(\boldsymbol{R}\boldsymbol{p}+\boldsymbol{t}),\tag{2}$$

式中 Z_2 为此时点 $\boldsymbol{p}$ 在当前图像帧的深度，$(u_{\mathrm{b}},v_{\mathrm{b}})$ 为当前帧在相机成像平面上的坐标值。用 I_{f} 表示关键帧图像中某一像素的灰度值，I_{b} 为当前帧图像中对应像素点的灰度值，则像素间的灰度残差 ε_{g} 可表示为

$$\varepsilon_{\mathrm{g}}=I_{\mathrm{b}}(\boldsymbol{p}_{\mathrm{b}})-I_{\mathrm{f}}(\boldsymbol{p}_{\mathrm{f}})。\tag{3}$$

若在语义分割后得到了 k 类目标的分割结果，对于 k 类目标，每一类目标间的某一像素点的灰度残差之和应最小，有

$$\varepsilon_{\mathrm{g}}=\varepsilon_{\mathrm{g}_1}+\varepsilon_{\mathrm{g}_2}+\cdots+\varepsilon_{\mathrm{g}_k},\tag{4}$$

式中 $\varepsilon_{\mathrm{g}i}(i=1,2,\cdots,k)$ 为第 i 类目标的灰度残差。

影响图像语义分割准确性的因素主要集中在目标的边缘像素部分，而边缘正是具有明显梯度变化的像素部分。因此，为了保证算法的精度，在(4)式中加入梯度变化约束，用来减小像素级语义分割对相机姿态估计的影响，若某一像素点的灰度残差用 $\varepsilon_{\mathrm{gt}}$ 表示，则(4)式改写为

$$\varepsilon_{\mathrm{g}}=\varepsilon_{\mathrm{g}_1}+\varepsilon_{\mathrm{g}_2}+\cdots+\varepsilon_{\mathrm{g}_k}+\varepsilon_{\mathrm{gt}}。\tag{5}$$

目标边缘与明显梯度变化重合的像素用明显梯度变化的像素代替。根据双目相机的深度恢复原理，可以方便地恢复出空间中三维点的坐标。对于关键帧图像中的点 $\boldsymbol{p}_{\mathrm{f}}$，在当前帧图像的 $(u_{\mathrm{b}},v_{\mathrm{b}})$ 处寻找三维坐标值与法向量 $\boldsymbol{n}_{\mathrm{f}}$ 相近的点作为匹配点，几何残差的表达式为

$$\varepsilon_{\mathrm{d}}=(\boldsymbol{p}_{\mathrm{b}}-\boldsymbol{p}_{\mathrm{f}})\cdot\boldsymbol{n}_{\mathrm{f}}。\tag{6}$$

对于 k 类目标，其几何残差之和为

$$\varepsilon_{\mathrm{d}}=\varepsilon_{\mathrm{d}_1}+\varepsilon_{\mathrm{d}_2}+\cdots+\varepsilon_{\mathrm{d}_k},\tag{7}$$

式中 $\varepsilon_{\mathrm{d}i}(i=1,2,\cdots,k)$ 为第 i 类目标的几何残差。

故基于灰度-几何约束的总残差为

$$\varepsilon=\varepsilon_{\mathrm{g}}+\varepsilon_{\mathrm{d}}。\tag{8}$$

因此，需要进行优化的目标函数为

$$E=\mathrm{argmin}\sum\|\varepsilon\|^2。\tag{9}$$

采用高斯牛顿迭代优化算法对(9)式进行求解，即可得到相机的姿态变换矩阵。在相机姿态估计的过程中，不可避免地会产生累积误差，也会出现相机抖动或遮挡时目标对象移出场景区域的情况，导致三维跟踪注册出现误差。因此采用图优化的方法[19]减小误差累积，以保证算法的稳定性。经过上述方法，即可将虚拟图像注册到真实环境中。

3　实验过程与分析

3.1　语义分割与姿态估计

为了验证所提三维跟踪注册算法的性能，对双目相机采集的图像进行测试，计算机内存为 8 GB。为了保证算法的实时性，采用了图形处理器(GPU)加速的方法，实验中使用的 GPU 型号为 GTX1060。算法首先对 KITTI 数据集的图像进行训练，第一次使用训练集中的全部样本进行训练，时间消耗为 126.7 s，此后所用时间均在 94 s 左右。训练时初始学习率为 0.001，权重惩罚项为 0.0005，动量项为 0.9，一次处理图像批量为 32。经过网络训练可以得到场景语义分割的结果，如图 1(a)所示。结合图像的深度信息，将带有语义目标的二维图像，转化为三维点云，如图 1(b)所示。

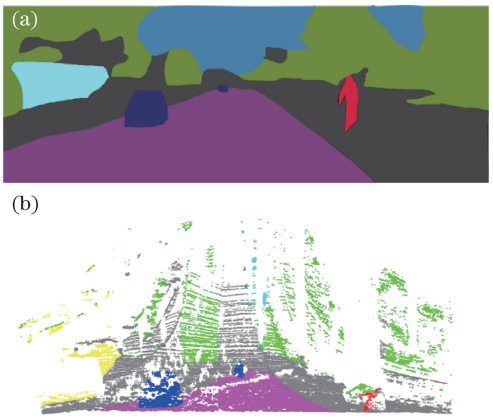

图 1　语义分割结果示例。(a)语义分割图像；(b)带有语义目标的点云恢复结果

Fig. 1　Result of semantic segmentation. (a) Semantic segmentation image; (b) point cloud retrieval result with semantic objects

图 1(a)中不同颜色代表了不同种类的目标。由于在场景中存在天空等无穷远处的点，其深度无法进行估计，因此在恢复三维点云时，将这些数据点剔除。采用 2.2 节给出的方法得到带有语义信息的三维点云，结合下一帧的图像进行相机的姿态估计，图 2 所示为真实轨迹与所提算法的估计结果。

由图 2 可知，所提算法估计的结果与真实轨迹比较接近，基本满足了对算法的精度要求。

3.2　算法实时性分析

为了评定算法的实时性，将所提算法与基于扩展卡尔曼滤波(EKF)-即时定位与地图构建(SLAM)的虚实注册方法[20]进行对比，所提算法的每帧图像各步骤消耗的平均处理时间以及算法总时间见表 1。

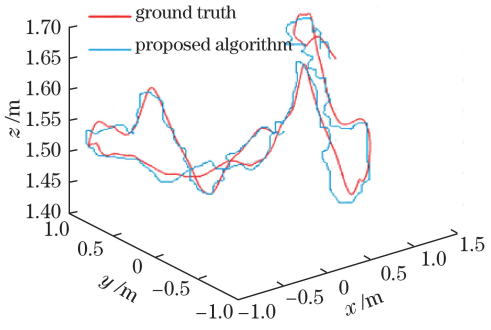

图 2　所提算法的估计结果与真实轨迹

Fig. 2　Estimated result by proposed algorithm and real trajectory

由表 1 可知，所提算法与基于 EKF-SLAM 的方法相比，所需时间较少，算法本身不仅能够适应纹理较少的环境，且满足了实时性的要求。

3.3　三维跟踪注册的精度测试与应用

通过计算像素点的图像坐标与用投影矩阵重投影后坐标的均方根(RMS)误差，衡量所提三维跟踪注册算法的精度。选取了 600 帧图像进行实验，分别计算了绕 x，y，z 轴旋转的角度误差以及沿 x，y，z 轴平移的误差，图 3 所示为基于语义目标约束的三维跟踪注册方法的误差与其他方法的对比结果。

表 1　每帧图像的三维跟踪注册的平均处理时间

Table 1　Average processing time of 3D tracking registration for each frame　　s

Method	Image preprocessing	Camera pose estimation	Feature point extraction and matching	Semantic information extraction	Camera pose calculation	Rendering registration	Total
Method based on EKF-SLAM	1.2	19.05	6.1		9.6	1.2	37.15
Proposed method	1.2	19.33		5.4	9.7	1.2	36.83

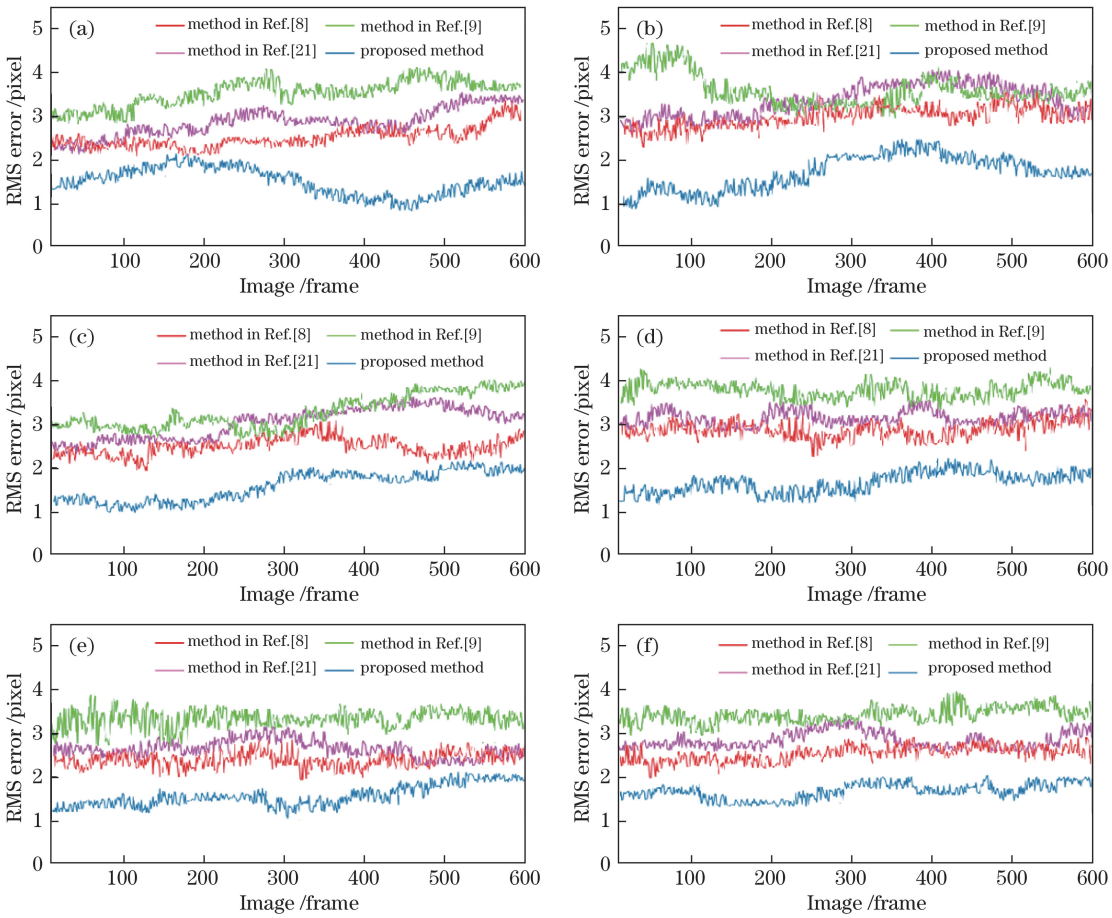

图 3　各类算法给出的误差对比。(a)沿 x 轴旋转；(b)沿 y 轴旋转；(c)沿 z 轴旋转；(d)沿 x 轴平移；(e)沿 y 轴平移；(f)沿 z 轴平移

Fig. 3　Error comparison among different methods. (a) Rotating along x-axis; (b) rotating along y-axis; (c) rotating along z-axis; (d) moving along x-axis; (e) moving along y-axis; (f) moving along z-axis

实验测试结果表明，所提三维跟踪注册算法的误差在旋转或平移变化的状态下，平均在 2.2 pixel 以下，基本满足了精度需求，且与其他方法相比精度更高。将所提算法应用到 AR 型车载平视显示器(AR-HUD)上，可以使驾驶员在行车过程中看到导航等信息，而不用频繁转换视野。现有的车载平视显示器(HUD)[22]只是将虚拟信息投射到视线前方，在一定程度上会使驾驶员分散注意力，而采用所提算法可将虚拟信息与真实环境更好地融合，避免了由注意力分散导致的交通事故。图 4 所示为将所提三维跟踪注册算法应用于 AR-HUD 的图像。

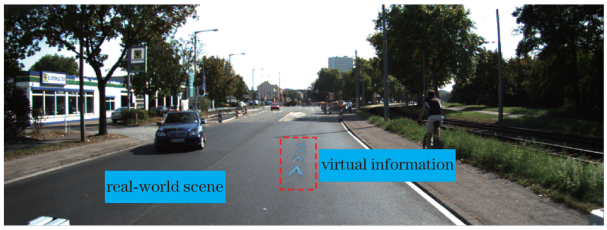

图 4　所提算法应用于 AR-HUD

Fig. 4　Application of proposed algorithm to AR-HUD

4　结　　论

针对 AR 系统中的三维跟踪注册问题，提出了一种语义目标匹配的三维跟踪注册方法。采用深度卷积神经网络对场景进行像素级语义分割。在对应目标之间建立灰度-几何约束，并融合了梯度变化明显像素的约束，建立了目标约束方程，对相机姿态进行了估计，获取了三维跟踪注册算法的注册矩阵。研究结果表明，算法在满足实时性的同时，精度高于其他方法的，三维跟踪误差平均在 2.2 pixel以下，且算法可以应用于 AR-HUD 系统中，有利于行车安全。

参　考　文　献

[1] Rehman U, Cao S. Augmented-reality-based indoor navigation: A comparative analysis of handheld devices versus google glass[J]. IEEE Transactions on Human-Machine Systems, 2017, 47(1): 140-151.

[2] Joo-Nagata J, Abad F M, Giner J G-B, *et al*. Augmented reality and pedestrian navigation through its implementation in m-learning and e-learning: Evaluation of an educational program in Chile[J]. Computers & Education, 2017, 111: 1-17.

[3] Kress B, Starner T. A review of head-mounted displays (HMD) technologies and applications for consumer electronics[J]. Proceedings of SPIE, 2013, 8720: 87200A.

[4] Maisto M, Pacchierotti C, Chinello F, *et al*. Evaluation of wearable haptic systems for the fingers in augmented reality applications[J]. IEEE Transactions on Haptics, 2017, 10(4): 511-522.

[5] Montero A, Zarraonandia T, Diaz P, *et al*. Designing and implementing interactive and realistic augmented reality experiences[J]. Universal Access in the Information Society, 2017, 36(4): 1-13.

[6] Tsai C W. The applications of augmented reality for universal access in online education[J]. Universal Access in the Information Society, 2017, 35(3): 1-3.

[7] Yu H B, Ho H. System designs for augmented reality based ablation probe tracking[C]. Pacific-Rim Symposium on Image and Video Technology, 2017: 87-99.

[8] Khan D, Ullah S, Yan D M, *et al*. Robust tracking through the design of high quality fiducial markers: An optimization tool for AR ToolKit[J]. IEEE Access, 2018, 6: 22421-22433.

[9] Lin H C K, Su S H, Wang S T, *et al*. Influence of cognitive style and cooperative learning on application of augmented reality to natural science learning[J]. International Journal of Technology and Human Interaction, 2015, 11(4): 41-66.

[10] Zhang G, Chen H S, Ye Y D. A LoG operator based markerless augmented reality algorithm: LoG-PTAMM[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(9): 1577-1586.
张格, 陈昊升, 叶阳东. 一种基于 LoG 算子的无标识增强现实算法: LoG-PTAMM[J]. 计算机辅助设计与图形学学报, 2016, 28(9): 1577-1586.

[11] Ng-Thow-Hing V, Bark K, Beckwith L, *et al*. User-centered perspectives for automotive augmented reality[C]. IEEE International Symposium on Mixed and Augmented Reality-Arts, Media, and Humanities, 2013: 13-22.

[12] Hayashi T, Uchiyama H, Pilet J, Saito H. An augmented reality setup with an omnidirectional camera based on multiple object detection[J]. Proceedings of the 20th International Conference on Pattern Recognition, 2010: 3171-3174.

[13] Kong S H, Haouchine N, Soares R, *et al*. Robust augmented reality registration method for localization of solid organs′ tumors using CT-derived virtual biomechanical model and fluorescent fiducials[J]. Surgical Endoscopy & Other Interventional Techniques, 2017, 31(7): 2853-2871.

[14] Streckel B, Koch R. Lens model selection for visual tracking[C]. Joint Pattern Recognition Symposium, 2005: 41-48.

[15] Skrypnyk I, Lowe D G. Scene modelling, recognition

and tracking with invariant image features[C]. Third IEEE and ACM International Symposium on Mixed and Augmented Reality, 2004: 110-119.

[16] An Z, Xu X P, Yang J H, *et al*. Design of augmented reality head up display system based on image semantic segmentation[J]. Acta Optica Sinica, 2018, 38(7): 0710004.
安喆, 徐熙平, 杨进华, 等. 结合图像语义分割的增强现实型平视显示系统设计与研究[J]. 光学学报, 2018, 38(7): 0710004.

[17] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3354-3361.

[18] Zhang K. Research on a rapid fusion method for remote sensing images based on an improved atrous wavelet decompsition[D]. Zhengzhou: Henan University, 2016: 66-72.
张凯. 基于改进 atrous 小波分解的遥感影像快速融合方法的研究[D]. 郑州: 河南大学, 2016: 66-72.

[19] Kümmerle R, Grisetti G, Strasdat H, *et al*. G2o: A general framework for graph optimization[C]. IEEE International Conference on Robotics and Automation, 2011: 3607-3613.

[20] Liang C, Wang L, Liu H Y. Real-time vision SLAM algorithm based on extend Kalman filtering[J]. Computer Engineering, 2013, 39(8): 231-234, 238.
梁超, 王亮, 刘红云. 基于扩展卡尔曼滤波的实时视觉 SLAM 算法[J]. 计算机工程, 2013, 39(8): 231-234, 238.

[21] Park H S, Min W P, Won K H, *et al*. In-vehicle AR-HUD system to provide driving-safety information[J]. ETRI Journal, 2013, 35(6): 1038-1047.

[22] Liu S G. Research on interaction design of automobile HUD-based on user′s research and analysis [J]. China Packaging, 2018, 38(6): 56-58.
刘双广. 对汽车 HUD 的交互设计研究——基于用户的研究与分析[J]. 中国包装, 2018, 38(6): 56-58.