基于视觉词典的深度图生成算法

刘杰平*, 周华盛, 余朗衡, 丁树浩, 梁亚玲
华南理工大学电子与信息学院, 广东 广州 510641

摘要 针对从二维彩色图像中恢复深度信息的问题,提出一种基于视觉词典的深度图生成算法。采用基于数据驱动的方法,从包含深度图的深度图像库中找出图像中各种空间结构对应的深度信息,得到由空间结构相似的图像块组成的初始视觉单词;采用难例挖掘方法找到视觉单词的难例负样本,更新视觉单词分类器,获得最优的分类效果;利用视觉单词分类器和视觉单词组成的视觉词典对目标图像进行多尺度检测,得到对应的深度图并进行边缘保持平滑滤波。实验结果表明,该算法生成的深度图符合目标图像的深度变化,在主观视觉效果和各种客观评价指标上都有显著提高。

关键词 机器视觉; 深度图; 机器学习; 视觉单词; 视觉词典; 难例挖掘
中图分类号 TP391 **文献标识码** A **doi**: 10.3788/AOS201838.0915004

Depth Map Generation Algorithm Based on Visual Dictionary

Liu Jieping*, Zhou Huasheng, Yu Langheng, Ding Shuhao, Liang Yaling
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510641, *China*

Abstract In order to recover depth information from two-dimensional color image, a visual-dictionary-based depth map generation algorithm is proposed. A data-driven method is used to find depth information of various spatial structures from depth map library, so as to obtain initial visual words which consist of image patches with similar structure. Hard example mining method is used to find hard negative examples of visual word, and visual word classifier is updated to get best classification result. Visual dictionary composed of visual word classifiers and visual words is used to detect target image at multiple scales to get corresponding depth map, to which edge-preserving smoothing filter will be applied. Experimental results show that depth maps generated by the proposed algorithm match depth change of target images, and has a good improvement in both subjective visual effects and objective evaluation indexes.

Key words machine vision; depth map; machine learning; visual word; visual dictionary; hard example mining
OCIS codes 150.1135; 100.2960; 200.3050

1 引 言

三维(3D)显示可以为观察者提供比二维(2D)更逼真和沉浸式的视觉体验,而 3D 资源的匮乏制约着 3D 显示技术的发展。如果能利用现有的 2D 资源,采用 2D 转 3D 的技术对 3D 资源进行补充,即可有效解决 3D 资源短缺的问题。深度图生成算法是 2D 转 3D 的关键技术,因此,对深度图生成算法的研究尤为必要。

深度图生成是从 2D 图像中估计深度信息,主要有提取深度线索和利用机器学习算法训练模型两类方法。Wang 等[1]结合图像语义和场景先验理论,在通过暗通道生成深度图后进一步优化得到更好的深度图;Jung 等[2]使用线跟踪算法建立图像的等高线图,根据相对高度线索原理,赋予等高线之间的区域相应的深度值;丁伟利等[3]则将图像中不同封闭区域的边缘和颜色信息作为深度线索,以此估计城市道路图像中的深度;何建梅等[4]通过融合特征点密度与边缘信息建立了新的聚焦测度,并为深度线索实现了特定场景的深度估计。

为了克服深度线索只是图像与深度关系的局部总结的片面性,一些学者提出了基于机器学习的深

收稿日期: 2017-12-13; **修回日期**: 2018-04-10; **录用日期**: 2018-04-28
基金项目: 国家自然科学基金(61471173,61701181)、广东省自然科学基金(2016A030313455,2017A030325430)
* **E-mail**: eeliujp@scut.edu.cn

度图生成算法，利用深度图像库进行建模与学习，Hoiem 等[5]将训练图像库的图像分为多个图像块，提取每个图像块的颜色、位置、纹理等特征后进行训练分类器，并根据相机成像原理推导出不同深度下的景物尺寸关系[6]，统一不同位置和不同结构的景物的尺度，改善了深度图质量；Saxena 等[7-9]将 Make3D 图像库作为训练集，每幅图像分成小的图像块，分别提取块的局部特征和相对特征，局部特征对应图像块本身的深度线索，相对特征表示图像块间的相对深度，然后利用马尔可夫随机场进行模型学习，从而对目标图像的深度图进行预测；Konrad 等[10]通过统计深度值分别在颜色、位置、运动矢量上各个量化区间的平均值，利用查表法计算目标图像的加权深度值；许路等[11]结合人工提取特征与卷积神经网络自动提取特征提出了一种基于深层卷积神经网络的深度估计方法；吴寿川等[12]采用双向递归的视频序列信息传递机制估计红外视频的深度；姚广顺等[13]根据深度值的范围量化雷达的距离数据，将深度估计问题转化为像素级分类问题进行模型训练；Konrad 等[14]计算目标图像的梯度直方图(HOG)[15]描述子与图像库中各个图像的 HOG 描述子的欧氏距离，从中选出 k 幅最近邻的图像并将深度图进行融合；Xu 等[16]采用 k 近邻方法生成初始深度图后，利用聚类算法分割图像的结果改进初始深度图。

基于机器学习的深度图生成算法具有普适性，该类算法生成深度图的质量取决于图像特征的有效性及数据库的丰富程度。然而，提取有效图像特征的算法通常具有较高的复杂度，且需要存储和维护庞大的数据库，这无疑是对设备的巨大挑战。

为克服目前深度图生成算法特征提取复杂，设备储存开销大等问题，本文提出一种基于视觉词典的深度图生成方法，该方法通过训练得到的深度视觉单词(视觉单词)构造深度视觉词典(视觉词典)，用于检测目标图像中视觉单词，匹配得到对应的深度信息，最终生成深度图。

2　本文方法

2.1　算法框架

为了从图像中提取完整的视觉语义，挖掘图像世界中的基本视觉元素，学者们进行了大量的研究。Sivic 等[17]在大规模图像集中进行尺度不变特征变换(SIFT)关键点检测，并对检测到的 SIFT 关键点进行聚类，使之产生具有一定视觉意义的聚类单元，对应文本中的单词。SIFT 关键点包含的视觉意义可能相对低级，为了得到更高层次的视觉语义，Russell 等[18]通过图像分割挖掘具有共性的中等规模的图像区域，图像中存在多种基本的结构元素，如简单的点、线结构，也有门窗、桌椅等复杂的空间结构，相似空间结构的图像块对应相似的深度信息[19]。

本文设计的深度图生成算法中视觉词典由各个视觉单词组成，每个视觉单词代表一个特定的空间结构，并有对应的深度信息和分类器，可用于目标图像中检测视觉单词，匹配对应的深度信息，完成局部的深度估计；当目标图像各个区域的深度估计完成后，即可生成目标图像最终的深度图。该算法无需提取复杂特征，且不需要保存大量的彩色图像和深度图，能够有效节省数据存储和维护的开销。

本文算法分为深度视觉词典训练和深度图生成两部分。

在训练阶段，首先，从图像库中找出图像中各种空间结构对应的深度信息，得到初始视觉单词；然后，利用难例挖掘方法找到视觉单词的难例负样本，同时更新视觉单词、深度信息及分类器，建立视觉词典。

在深度图生成阶段，对目标图像进行高斯金字塔分解，利用视觉词典进行多尺度视觉单词检测，更新深度图。当金字塔所有层级的深度图更新完成后，进行滤波，达到保持边缘和平滑物体内部的目的，得到最终的深度图，算法框图如图 1 所示。

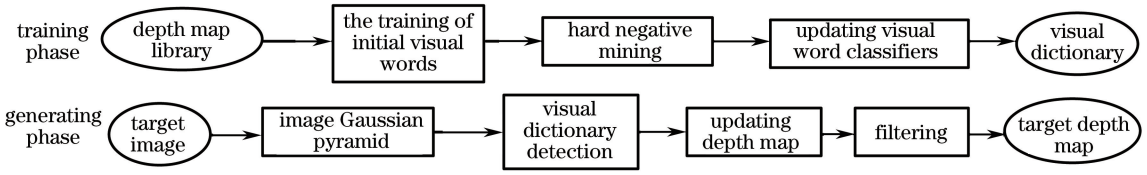

图 1　本文算法框图

Fig. 1　Block diagram of the proposed algorithm

2.2　视觉词典生成算法

2.2.1　视觉单词

视觉单词是视觉词典的基本单元，表示图像中某种空间结构及对应的深度信息。视觉单词具有以下性质：1）视觉单词在图像中必须有较高的出现频率，即常见性；2）各个视觉单词应该有明显的区分度；3）视觉单词应包含一定的深度信息，即信息量。

从深度图像库中训练初始视觉单词的过程如图2所示。

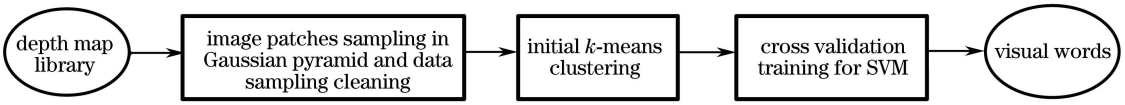

图2　初始视觉单词训练

Fig. 2　Training of initial visual words

2.2.2　图像高斯金字塔块采样与数据清洗

当景物与照相机的距离改变时，图像中的对应物体大小和清晰度也随之改变。因此，为了保证采样在尺度和空间分布上的完备性，使采样结果具有常见性，采样之前先对图像进行高斯金字塔分解[20]。在各个金字塔尺度下使用一个固定大小的采样窗口，对图像进行大量随机采样，由于采样得到的图像块数目很大，必然包含空间位置高度重叠以及信息量低的图像块。

为保证采样结果的区分度和信息量达到要求，需要清洗图像块样本的数据，去除高度重叠和信息量低的图像块。对于空间位置高度重叠的图像块，计算图像块样本之间的余弦距离，若余弦距离小于预设阈值，即认为是重叠块，只保留其中一个图像块，可采用公式表示为

$$\cos(\theta)=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\cdot\|\boldsymbol{B}\|}=\sum_{u=1}^{z}\boldsymbol{A}_u\boldsymbol{B}_u\Big/\left(\sqrt{\sum_{u=1}^{z}\boldsymbol{A}_u^2}\cdot\sqrt{\sum_{u=1}^{z}\boldsymbol{B}_u^2}\right),\tag{1}$$

式中 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 为两个图像块的像素组成的向量，z 为图像块中像素点的数目。

信息量低是指图像块中包含很少的结构变化，如墙壁部分。随机采样的图像块均有对应的深度块，图像块上的结构变化信息均会反映在深度块上，计算深度块的方差，若深度块的方差小于预设阈值，表示深度块上的深度信息量较少，其对应的图像块包含很少的结构变化，应该删除对应的图像块，可采用公式表示为

$$\mathrm{Var}(\boldsymbol{A})=\frac{1}{z}\sum_{u=1}^{z}(\boldsymbol{A}_u-\mu)^2,\tag{2}$$

式中 $\mu=\frac{1}{z}\sum_{u=1}^{z}\boldsymbol{A}_u$ 为深度块 $\boldsymbol{A}$ 中像素的灰度平均值。

通过在不同的图像-深度对上多次实验发现，当设定余弦距离阈值为0.5，方差阈值为10时，能够保证图像块之间的重叠面积较小，同时有效避开平坦区域，经过数据清洗后的图像块样本如图3中方框所示。从图中可见，在不同层级的高斯金字塔中采样得到的图像块样本包含了较多的空间结构变化，而且空间重叠部分较少。

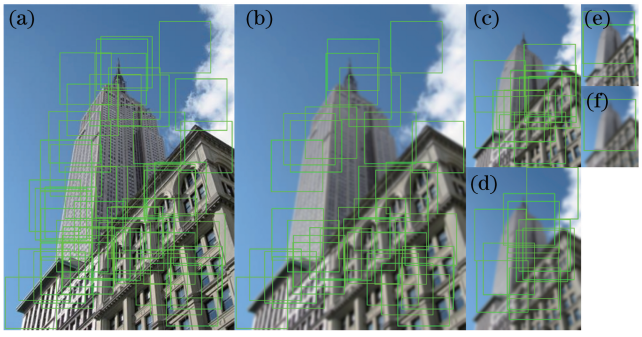

图3　数据清洗后的高斯金字塔图像块样本。(a)第一级第一层，方差为 σ；(b)第一级第二层，方差 $\sqrt{2}\sigma$；(c)第二级第一层，方差 2σ；(d)第二级第二层，方差 $2\sqrt{2}\sigma$；(e)第三级第一层，方差 4σ；(f)第三级第二层，方差 $4\sqrt{2}\sigma$

Fig. 3　Sample image Gaussian pyramid patches after data cleaning. (a) Layer1, octave1, variance is σ; (b) layer2, octave1, variance is $\sqrt{2}\sigma$; (c) layer1, octave2, variance is 2σ; (d) layer2, octave2, variance is $2\sqrt{2}\sigma$; (e) layer1, octave3, variance is 4σ; (f) layer2, octave3, variance is $4\sqrt{2}\sigma$

2.2.3　初始视觉单词训练

由于没有已标注的训练深度图像集，因此采用无监督学习的方法训练视觉单词，本文采用 k-means 算法对图像块进行初始化聚类，得到具有初步结构意义的聚类单元。设置视觉单词的常见性阈值为 m，即每个聚类单元必须包含 m 个以上的图像块，否则不能用于训练。

但是，k-means 算法的聚类数需要人工设定，导致聚类结果具有很大的随机性。另外，该算法只能使用简单的相似度计量，如欧氏距离、马氏距离等，产生的类别边界较简单。所以，k-means 聚类的结果不能直接作为视觉单词。为了进一步获得更加稳健的视觉单词，本文采用支持向量机(SVM)作为视觉单词的分类器，对于可能产生的过拟合问题，采用交叉验证进行训练，同时为了提高视觉单词类别内的纯净度，只保留类别内 SVM 分数最高的前 m 个图像块。交叉验证训练中的 SVM 分类器利用公式可表示为

$$f_\beta(x)=\boldsymbol{\beta}\cdot\boldsymbol{\Phi}(x),\tag{3}$$

式中 $\boldsymbol{\beta}$ 为分类器模型参数向量，x 为任意图像块样本，$\boldsymbol{\Phi}(x)$为由图像块 x 计算得到的 HOG 描述子向量。(3)式得到的 f_β 即为 SVM 分数，f_β 值越高，表示图像块 x 属于正样本的概率越大。在交叉验证时，分类器模型参数向量 $\boldsymbol{\beta}$ 需要根据目标函数更新优化，目标函数为

$$L(\boldsymbol{\beta})=\|\boldsymbol{\beta}\|^2/2。\tag{4}$$

初始视觉单词的训练步骤如下。

输入：正样本为采样深度图像库中室内图像或建筑物图像得到的图像块集合 R，分为数量相等的两个集合 R_1 和 R_2；负样本为随机采样室外自然图像得到的图像块集合 N，分为数量相等的两个集合 N_1 和 N_2。

初始化：对 R_1 进行 k-means 聚类，得到的结果为 K_1，以 K_1 中的各类 K_{1j}(j 为类序号)为起点，进行交叉训练，每次迭代的具体步骤如下。

1) 将 K_{1j} 包含的图像块作为正样本，在 N_1 随机抽取等数量的图像块作为负样本，训练得到 SVM 分类器 $\boldsymbol{G}_{1j}$。

2) 利用 SVM 分类器 $\boldsymbol{G}_{1j}$ 检测 R_2 中的图像块，得到正响应样本集合 K_{2j}，每个样本集合只保留前 m 个图像块。

3) 将 K_{2j} 包含的图像块作为正样本，在 N_2 中随机抽取等数量的图像块作为负样本，训练得到 SVM 分类器 $\boldsymbol{G}_{2j}$。

4) 利用 SVM 分类器 $\boldsymbol{G}_{2j}$ 检测 R_1 中的图像块，得到正响应样本集合 K_{1j}，同样只保留前 m 个图像块。

5) 若步骤 2)和 4)中得到的 K_{2j} 和 K_{1j} 与上次迭代过程中检测出的 K_{2j} 和 K_{1j} 相同，结束迭代；否则，返回步骤 1)。

输出：合并 K_{2j} 和 K_{1j}，得到具有相似空间结构的图像块类 K_j，组成第 j 个初始视觉单词，即每个初始视觉单词中有 $2m$ 个图像块。

2.2.4　挖掘难例负样本

在交叉验证训练时，采用的正、负样本(室内外图像块)差异较大，导致训练得到的 SVM 超平面较为宽松，可能会将空间结构不同和深度信息差别较大的室内图像块划分为一个视觉单词。针对此缺点，按照传统改进思路，可将目标视觉单词中的图像块作为正样本、其他视觉单词的图像块作为负样本训练 SVM 分类器。但在实际操作中，负样本数量巨大，优化过程很慢，结果亦不够理想。

为了进一步提高 SVM 分类器的可靠性，同时保证整体算法的执行效率，采用难例挖掘算法[21]对 SVM 分类器进行优化。其思路为：使用初始负样本集训练一个模型，然后收集被模型错误分类的负样本，形成新的难例负样本集，重复该过程多次，最终得到一个稳定的模型。由于难例挖掘算法每次训练只使用少量样本，训练速度很快，当迭代次数足够多时，分类效果能够有效提高。具体做法如下：

用带有目标视觉单词类标记的图像块样本集 D 训练(3)式的模型参数向量 $\boldsymbol{\beta}$，其中 $D=\{\langle x_1,y_1\rangle,\langle x_2,y_2\rangle,\cdots,\langle x_i,y_i\rangle,\cdots,\langle x_n,y_n\rangle\}$，$n$ 为图像块样本集 D 的容量，i 为图像块样本 x_i 的序号，$y_i\in\{-1,1\}$为图像块样本 x_i 的类标记，$y_i=1$ 表示 x_i 属于目标视觉单词类别；$y_i=-1$ 表示 x_i 不属于目标视觉单词类别。在训练过程中，若 $y\cdot f_\beta(x)<0$，表示图像块 x 被分类器错分；若 $y\cdot f_\beta(x)>1$，表示图像块 x 被分类器正确划分；若 $0<yf_\beta(x)<1$，表示图像块 x 位于 SVM 分类间隔内，即图像块样本集 D 中包含的线性不可分的“特异点”，为了强化分类器的检测效果，将训练的目标函数修改为

$$L_D(\boldsymbol{\beta})=\frac{1}{2}\|\boldsymbol{\beta}\|^2+S\cdot\sum_{i=1}^{n}\xi_i,\tag{5}$$

式中 ξ_i 为 SVM 的松弛变量，目的是使图像块位于分类间隔外，常数 S 为控制松弛变量的相对权值。

将图像块样本集 D 分成两个样本集合：难例样本集合 $H(\boldsymbol{\beta},D)$和一般图像块样本集合 $E(\boldsymbol{\beta},D)$。

其中，$H(\boldsymbol{\beta},D)$为被分类器错误分类或位于 SVM 分类间隔内的图像块样本集合，$E(\boldsymbol{\beta},D)$为被分类器正确分类并位于分类间隔外的图像块样本集合。两者可分别定义为

$$H(\boldsymbol{\beta},D)=\{\langle x,y\rangle\in D\mid y\cdot f_{\beta}(x)<1\},\tag{6}$$

$$E(\boldsymbol{\beta},D)=\{\langle x,y\rangle\in D\mid y\cdot f_{\beta}(x)<1\},\tag{7}$$

令 $\boldsymbol{\beta}^{*}(D)=\arg\min_{\boldsymbol{\beta}}L_{D}(\boldsymbol{\beta})$表示从样本集 D 中训练得到的最优分类器模型参数 $\boldsymbol{\beta}$。由于 $L_{D}(\boldsymbol{\beta})$是严格凸的，所以 $\boldsymbol{\beta}^{*}(D)$唯一[18]。根据 SVM 的原理，对于大量图像块样本数据组成的训练集 D，能够找到一个较小的图像块样本集 $C\subset D$，使得 $\boldsymbol{\beta}^{*}(D)=\boldsymbol{\beta}^{*}(C)$。

设 $C_{1}\subset D$ 是初始图像块样本集合，重复以下步骤训练分类器模型和更新集合：1）令 $\boldsymbol{\beta}_{t}=\boldsymbol{\beta}^{*}(C_{t})$，即用 C_{t} 训练一个模型参数 $\boldsymbol{\beta}_{t}$；2）若 $H(\boldsymbol{\beta}_{t},D)\subseteq C_{t}$，即 $\boldsymbol{\beta}_{t}$ 在样本集 D 上获得的难例都已包括在 C_{t} 中，停止迭代，训练结束，返回模型参数 $\boldsymbol{\beta}_{t}$；否则继续；3）对任意图像块 $x\in C_{t}$ 且 $x\in E(\boldsymbol{\beta}_{t},C_{t})$和 $y=-1$，将 x 从 C_{t} 中移除；4）对任意图像块 $x\notin C_{t}$ 且 $x\in H(\boldsymbol{\beta}_{t},D)$，将 x 加入到 C_{t} 中。

$\boldsymbol{\beta}_{t}$ 为第 t 次迭代的分类器模型参数向量，C_{t} 为第 t 次迭代的图像块样本集合。步骤 3）保证 C_{t} 中只含有视觉单词的正样本和难例负样本；步骤 4）通过向集合中加入新的难例负样本扩大集合。训练结束后，将最终的分类器模型参数向量 $\boldsymbol{\beta}$ 代入(3)式，便可获得最终的图像块 SVM 分类器。

2.2.5　生成视觉词典

通过训练得到多个代表结构元素的图像块类后，每种结构元素均有相应的 SVM 分类器。这些图像块类满足视觉单词的三个原则：常见性、区分度和信息量，此时图像块类即为本文的视觉单词。利用图像块相应的深度块，采用图像融合技术即可生成视觉单词对应的深度块。

在迭代训练过程的检测步骤中，图像块均有相应的 SVM 分数，分数越高代表图像块属于该视觉单词的置信度越高，因此，以 SVM 分数为权值，将视觉单词内深度块的加权平均作为视觉单词的深度块，即

$$\delta_{\mathrm{word}}=\sum_{v=1}^{2m}f_{\beta}^{v}\delta_{v}\Big/\sum_{v=1}^{2m}f_{\beta}^{v},\tag{8}$$

式中 v 为视觉单词中图像块的序号，f_{β}^{v} 为第 v 个图像块的 SVM 检测分数，δ_{v} 为第 v 个图像块对应的深度块。

为检测图像中的空间结构所对应的视觉单词，需要使用视觉单词的分类器，视觉单词分类器相当于视觉单词的“检索目录”，利用所有的视觉单词及其分类器构成深度视觉词典。

2.3　深度图生成

生成深度图的第一步是初始化深度图。本文生成下近上远的初始深度图，即处于深度图上方的像素点具有更大的深度值。以深度图左上角为坐标原点，点(p,q)的深度值 $V(p,q)$可表示为

$$V(p,q)=(w-p)/(w\cdot 2^{o}),\tag{9}$$

式中 w 为深度图高度，o 为深度图的量化比特数。

视觉单词以不同的分辨率和清晰度出现在图像中，其检测和深度图的更新应该在目标图像的高斯金字塔中进行，具体步骤如下：

1）对目标图像和初始深度图进行高斯金字塔分解，从金字塔的最高级开始检测，并对深度图进行更新。

2）在高斯金字塔的同一级中，从该级最高层图像开始，使用视觉词典进行视觉单词检测，对检测出视觉单词的区域，利用该视觉单词的深度块更新深度图的对应区域，同层操作完成后，继续对下一层图像进行检测和更新，直至完成该级所有层的检测和更新。

3）对下一级图像进行检测和更新，由于不同级的图像分辨率不同，需要预先对深度图进行放大处理。

4）重复步骤 2）～3）直至高斯金字塔所有层级的视觉单词检测和深度图更新完成。

5）采用导向滤波器[22]对更新完成后的深度图进行边缘保持平滑滤波。

在同级不同层的视觉单词检测过程中，若下层检测出视觉单词的区域与上层的检测结果出现重叠，则利用下层的深度图更新直接覆盖重叠区域，因为下层的清晰度更高，检测结果的精度更高。

3　仿真实验与分析

为验证算法的可行性和有效性，在公共数据集上进行本文算法和对比算法的实验仿真。实验中使用的训练集正样本为 NYU Depth v2 深度图像数据库中的 1200 幅图像-深度对，负样本为 Flickr 上随机下载的 6000 幅室外自然图像，测试图像为 NYU Depth v2 深度图像数据库中除训练时所用正样本外的 249 幅图像-深度对。所有实验均在相同的硬件平台（Intel（R） Core（TM） i5-3470 CPU

@3.20 GHz，8GB 内存）和软件平台（OpenCV＋Qt）上进行。

3.1　视觉单词训练仿真实验

进行视觉单词训练时，高斯金字塔分解包括 3 级，每级 2 层，随机采样的窗口大小为 80×80，方差为 $\sigma=0.5$，单幅图像随机采样数目为 350，视觉单词的常见性阈值 $m=8$。图 4 所示为本文方法训练得到的视觉单词中的 5 个图像块及深度块。

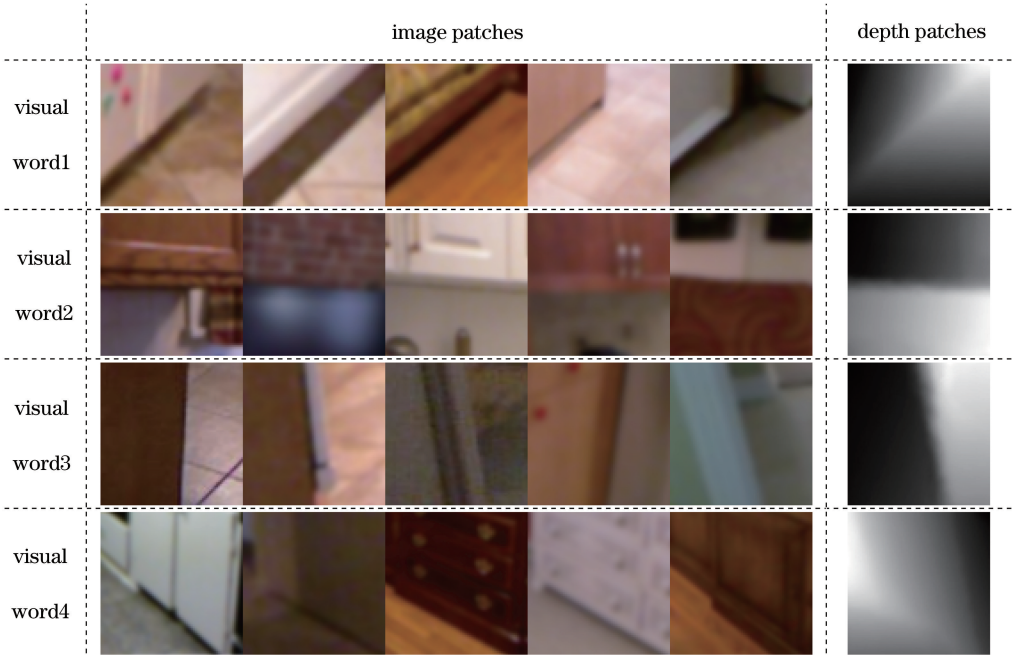

图 4　视觉单词训练结果示例

Fig. 4　Examples of visual word training results

由图 4 可知，视觉单词中的图像块具有较好的结构相似性，且包含丰富的深度信息，该训练结果证明本文算法能够从深度图库挖掘出代表不同空间结构元素的视觉单词，可组成应用于生成深度图的视觉词典。

3.2　深度图生成仿真实验

为了验证算法的有效性，将本文算法的实验结果与文献[2]算法、文献[14]算法和文献[16]算法生成的深度图进行对比，并从主观效果和客观效果两方面进行评价。

在主观效果方面，图 5 所示为真实深度图以及 4 种算法生成的深度图。

由图 5 可知，文献[2]算法生成的深度图中有比较明显的分层或分块结构，和真实深度图相比深度估计显得突兀而不自然，且不符合图像的景深变化趋势，这是因为该算法以图像的等高线作为深度线索；文献[14]算法得到的深度图像在局部结构上有所不同，未能体现 2D 图像中各个物体之间的深度变化，由于其以整幅图像为单位进行全局搜索，因此与真实深度图之间存在明显的偏差；从整体看，文献[16]该算法应用基于超像素的密度聚类算法对图像进行分割，生成的深度图比较符合真实深度图的景深变化，但在细节部分有待完善，特别是一些边缘的斑点导致深度变化不连续；本文算法在多个尺度检测目标图像的空间结构，逐步对深度图进行更新，对目标图像的深度有较细致的估计，因此生成的深度图较符合目标图像的深度变化。纵观 4 种算法的实验结果，本文算法生成的深度图在不同物体间的边缘有明显的深度差，而在同一物体内部具有平滑一致的深度，与真实深度图最接近，在视觉效果上明显优于其他三种算法。

在客观效果方面，采用峰值信噪比（PSNR）作为评价深度图质量的客观指标，表 1 所示为对应图 5 深度图的 PSNR 值。

从表 1 可以看出，本文算法生成的深度图 PSNR 值最高，以 image1 图像为例，本文算法的 PSNR 比文献[2]、文献[14]和文献[16]算法分别高 2.23 dB、3.23 dB 和 2.83 dB。实验结果表明，与其他三种算法比较，本文算法生成的深度图更接近真实深度图。

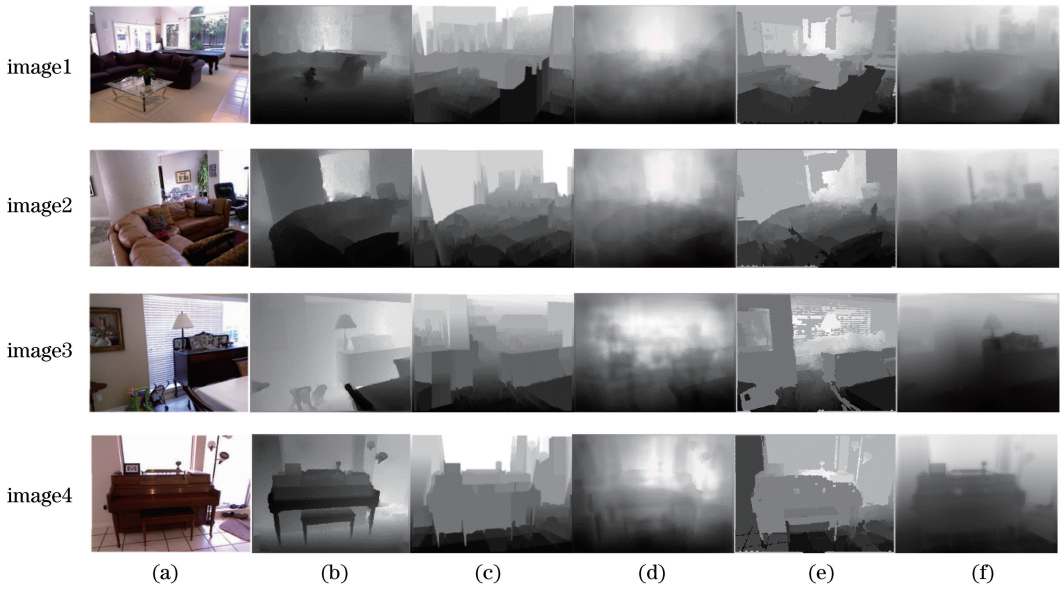

图 5　4 种算法生成的深度图。(a) 2D 彩色图像;(b)真实深度图;(c)文献[2]算法;(d)文献[14]算法;(e)文献[16]算法;(f)本文算法

Fig. 5　Depth maps generated by four algorithms. (a) 2D color images; (b) ground truth depth maps; (c) Ref. [2] algorithm; (d) Ref. [14] algorithm; (e) Ref. [16] algorithm; (f) proposed algorithm

表 1　4 种算法生成的深度图的 PSNR 值

Table 1　PSNR values of depth maps generated by four algorithms　　dB

Image	Ref. [2] algorithm	Ref. [14] algorithm	Ref. [16] algorithm	Proposed algorithm
Image 1	26.05	25.05	25.45	28.28
Image 2	25.00	24.97	25.05	25.54
Image 3	29.94	29.96	31.35	31.43
Image 4	26.41	26.64	26.85	27.98

为了进一步验证本文算法的有效性,分别计算 249 幅测试图像在 4 种算法下生成深度图的 PSNR、相对误差(Rel Err)和均方根误差(RMSE)的平均值,结果如表 2 所示。

表 2　4 种算法生成的深度图的结果对比

Table 2　Results comparison of depth maps generated by four algorithms

Parameter	Ref. [2] algorithm	Ref. [14] algorithm	Ref. [16] algorithm	Proposed algorithm
PSNR /dB	26.48	26.36	26.20	28.36
Rel Err	0.74	0.49	0.72	0.32
RMSE	12.10	12.26	12.49	9.47

从表 2 可以看出,本文算法生成的深度图的 PSNR 均值、相对误差均值和均方根误差均值都最优,PSNR 均值分别比文献[2]、文献[14]和文献[16]算法高 1.88 dB、2.00 dB 和 2.16 dB,相对误差均值分别比文献[2]、文献[14]和文献[16]算法低 0.42、0.17 和 0.40,均方根误差的均值分别比文献[2]、文献[14]和文献[16]算法低 2.63、2.79 和3.02,充分说明本文算法优于其他三种算法。

综上所述,因为本文算法充分考虑了图像的空间结构与深度信息之间的对应关系,且在利用视觉词典对图像进行空间结构检测时,匹配了对应区域的深度信息。所以生成的深度图的场景结构和物体深度准确,与真实深度图更接近。另外,多尺度检测方法以及导向滤波器的应用,使得本文算法不仅可以降低单次检测的误差,有效反映目标图像的深度信息,而且能够较好地保持物体的边界和相对位置关系。

4　结　　论

提出了一种基于深度视觉词典的深度图生成算法,该算法通过构造深度视觉词典对目标图像进行多尺度的视觉单词检测,匹配对应的深度信息,最终生成深度图。该算法只需要用到一种图像特征,训练得到的视觉单词占储小,克服了目前深度图生成算法特征提取复杂,设备储存开销大等问题。实验结果表明,本文算法能够获得场景结构明显,物体边界显著,物体位置更准确且深度变化较连续的深度图。

然而,本文算法仍然存在不适用的情况,例如当目标图像中含有镜像时,视觉单词会对镜像进行深度估计,导致生成的深度图结果不够理想,这也是目前深度图生成算法普遍存在的缺陷,其原因在于视觉词

典不足以反映图像与深度图之间深层次的映射关系。若要解决该问题，需要深入研究视觉词典的原理，并改进视觉词典模型，使之能够分析单目图像和深度图之间的内在联系，进一步提高算法的普适性。

参 考 文 献

[1] Wang K, Dunn E, Tighe J, *et al*. Combining semantic scene priors and haze removal for single image depth estimation[C]// IEEE Winter Conference on Applications of Computer Vision (WACV), 2014: 800-807.

[2] Jung Y J, Baik A, Kim J, *et al*. A novel 2D-to-3D conversion technique based on relative height-depth cue[J]. Proceedings of SPIE, 2009, 7273: 72731U.

[3] Ding W L, Li Y, Wang W F, *et al*. Depth estimation of urban road image based on contour understanding[J]. Acta Optica Sinica, 2014, 34(7): 0715001.
丁伟利, 李勇, 王文锋, 等. 基于轮廓特征理解的城市道路图像深度估计[J]. 光学学报, 2014, 34(7): 0715001.

[4] He J M, Qiu J, Liu C. Fusing feature point density and edge information for scene depth estimation[J]. Laser & Optoelectronics Progress, 2017, 54(7): 071101.
何建梅, 邱钧, 刘畅. 融合特征点密度与边缘信息的场景深度估计[J]. 激光与光电子学进展, 2017, 54(7): 071101.

[5] Hoiem D, Efros A A, Hebert M. Automatic photo pop-up[J]. ACM Transactions on Graphics, 2005, 24(3): 577-584.

[6] Hoiem D, Efros A A, Hebert M. Recovering surface layout from an image[J]. International Journal of Computer Vision, 2007, 75(1): 151-172.

[7] Saxena A, Sun M, Ng A Y. Learning 3D scene structure from a single still image[C]// IEEE International Conference on Computer Vision (ICCV), 2007: 1-8.

[8] Saxena A, Sun M, Ng A Y. 3D reconstruction from sparse views using monocular vision[C]// IEEE International Conference on Computer Vision (ICCV), 2007: 1-8.

[9] Saxena A, Sun M, Ng A Y. Make3D: learning 3D scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840.

[10] Konrad J, Wang M, Ishwar P. 2D-to-3D image conversion by learning depth from examples[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR), 2012: 16-22.

[11] Xu L, Zhao H T, Sun S Y. Monocular infrared image depth estimation based on deep convolutional neural networks[J]. Acta Optica Sinica, 2016, 36(7): 0715002.
许路, 赵海涛, 孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计[J]. 光学学报, 2016, 36(7): 0715002.

[12] Wu S C, Zhao H T, Sun S Y. Depth estimation from monocular infrared video based on bi-recursive convolutional neural network[J]. Acta Optica Sinica, 2017, 37(12): 1215003.
吴寿川, 赵海涛, 孙韶媛. 基于双向递归卷积神经网络的单目红外视频深度估计[J]. 光学学报, 2017, 37(12): 1215003.

[13] Yao G S, Sun S Y, Fang J A, *et al*. Depth estimation of night driverless vehicle scene based on infrared and radar[J]. Laser & Optoelectronics Progress, 2017, 54(12): 121003.
姚广顺, 孙韶媛, 方建安, 等. 基于红外与雷达的夜间无人车场景深度估计[J]. 激光与光电子学进展, 2017, 54(12): 121003.

[14] Konrad J, Wang M, Ishwar P, *et al*. Learning-based, automatic 2D-to-3D image and video conversion[J]. IEEE Transactions on Image Processing, 2013, 22(9): 3485-3496.

[15] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005: 886-893.

[16] Xu H H, Jiang M Y, Li F. Depth estimation algorithm based on data-driven approach and depth cues for stereo conversion in three-dimensional displays[J]. Optical Engineering, 2016, 55(12): 123106.

[17] Sivic J, Zisserman A. Video google: a text retrieval approach to object matching in videos[C]// IEEE International Conference on Computer Vision, 2003, 2: 1470-1472.

[18] Russell B C, Freeman W T, Efros A A, *et al*. Using multiple segmentations to discover objects and their extent in image collections[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2006: 1605-1614.

[19] Herrera J L, del-Blanco C R, García N. A novel 2D to 3D video conversion system based on a machine learning approach[J]. IEEE Transactions on Consumer Electronics, 2016, 62(4): 429-436.

[20] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[21] Felzenszwalb P F, Girshick R B, McAllester D, *et al*. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[22] He K M, Sun J, Tang X O. Guided image filtering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(6): 1397-1409.