融合多尺度局部特征与深度特征的双目立体匹配

王旭初[1,2*]，刘辉煌[2]，牛彦敏[3]

[1]重庆大学光电技术及系统教育部重点实验室，重庆 400040；

[2]重庆大学光电工程学院，重庆 400040；

[3]重庆师范大学计算机与信息科学学院，重庆 401331

摘要 针对立体匹配中不适定区域难以找到精确匹配点的问题，提出一种融合多尺度局部特征与深度特征的立体匹配方法。特征融合阶段包括两部分，其一是融合不同尺度下 Log-Gabor 特征和局部二值模式特征组合的浅层次特征，其二是将多尺度浅层融合特征和卷积神经网络提取的深度特征进行级联，形成既包含语义信息又包含结构化信息的特征图像。通过在极线垂直方向添加不同强度的噪声来构造正负样本，减小图像中极线对齐欠准带来的误差。将该方法与两种变体方法(改变或舍弃部分模块)在 KITTI 数据集进行对比实验，结果表明各模块设置具有合理性；与一些经典方法相比，所提方法取得了有竞争力的匹配性能。

关键词 机器视觉；立体匹配；多尺度局部特征融合；浅层次特征；孪生网络；卷积神经网络

中图分类号 TP391.41 **文献标志码** A **doi**：10.3788/AOS202040.0215001

Binocular Stereo Matching by Combining Multiscale Local and Deep Features

Wang Xuchu[1,2*], Liu Huihuang[2], Niu Yanmin[3]

[1] *Key Laboratory of Optoelectronic Technology and Systems of Ministry of Education, Chongqing University, Chongqing* 400040, *China*;

[2] *College of Optoelectronic Engineering, Chongqing University, Chongqing* 400040, *China*;

[3] *College of Computer and Information Science, Chongqing Normal University, Chongqing* 401331, *China*

Abstract In this study, a method is proposed based on multiscale local and deep features to address the difficulty associated with finding exactly matching pixels from the ill-posed regions in stereo matching. The feature fusion stage comprises two parts. First, the shallow features with different scales, including the Log-Gabor features and the local binary pattern features, are fused. The second part integrates the multiscale shallow fused features and deep features via a convolutional neural network and forms the final feature image, which contains both the semantic and structural information. Further, a positive and negative sample construction method is proposed by adding some noise in the vertical direction to reduce the error that can be attributed to imprecise epipolar alignment in an image. The proposed method is compared with two variant methods (changing or discarding of some modules) with respect to the KITTI datasets. The experimental results validate the effectiveness of the module settings with respect to the proposed method. This method also achieves competitive matching results when compared with those achieved using some classical methods.

Key words machine vision; stereo matching; multi-scale local feature fusion; shallow features; siamese network; convolutional neural network

OCIS codes 150.0155; 330.1400; 200.4260

收稿日期：2019-07-17；**修回日期**：2019-08-02；**录用日期**：2019-09-02

基金项目：国家自然科学基金(61971076)、重庆市基础与前沿研究计划(cstc2016jcyjA0317)

* **E-mail**：xcwang@cqu.edu.cn

1 引　言

双目立体视觉是计算机视觉领域的重要分支，它通过模拟人的视觉系统来处理现实世界，主要应用场景包括三维重建、自动驾驶和目标检测与识别等方面。立体匹配作为双目立体视觉的核心步骤，通过计算双目摄像机左右视图对应像素点的视差，进而估计出三维场景中物体的深度。在双目立体匹配中，因为存在镜头畸变、照明变化和遮挡等诸多影响图像外观的因素，所以决定两幅视角不同的图像内容是否匹配的问题具有挑战性[1]。

根据算法约束的作用范围，传统的立体匹配方法可以分为三类[1]：全局匹配，局部匹配，以及二者相结合形成的半全局匹配。全局匹配算法[2]通过建立全局能量函数最小化全局能量函数得到最优视差值。其中，图割、置信传播[3]和动态规划[4]等是用来求解能量最小化的典型算法，计算复杂度较高，且未必能找到全局最优解。局部匹配算法[5]通过约束像素周围小区域，对每个像素点计算一个能完全覆盖弱纹理区域、内部深度连续的窗口，进而对窗口内的视差值作加权平均。代表性方法包括自适应窗口立体匹配[6]、超像素过分割[7-8]、自适应权值立体匹配[9-10]和多窗体立体匹配[11]。这类算法仅利用像素点邻域的灰度、颜色和梯度等信息计算匹配代价，计算复杂度低，多数方法能达到实时的要求，但是对弱纹理区域、重复纹理区域、视差不连续和遮挡区域匹配效果不理想。半全局块匹配[12]作为逐像素匹配的方法，采用互信息来评价匹配代价，并通过动态规划算法在一维平滑约束中实现最优路径的搜索。以上方法都是手动设计代价函数，学习数据特征之间的线性组合，难以表达图像中丰富的信息。

近些年，基于深度学习的方法在图像特征提取方面显现出较大的优势，已被逐渐应用于双目立体匹配。Žbontar 等[13]率先使用卷积神经网络(CNN)来计算立体匹配代价，并用半全局匹配来细化代价。他们提出的 MC-CNN 架构包含一层卷积层和多层全连接层，将采用 Softmax 计算出的两个输入图像块之间的相似度作为输出。在此基础上，该方法设计了快速架构和准确架构[14]，其中将快速架构作为一种孪生网络用于实现卷积层的权重共享，并输出余弦相似度。准确架构采用多个含有修正线性单元的全连接层来代替余弦相似性，进一步提高了网络性能。

目前基于 CNN 的立体匹配方法大致分为三类。第一类以孪生网络为基础，利用卷积层提取图像块特征并计算相似度，并采用一系列手工优化策略。为了计算两个图像块的相似度，Zagoruyko 等[15]提出了孪生网络模型、伪孪生网络模型和双通道模型来学习和比较图像块并输出相似值。Chen 等[16]以多尺度图像块为输入，对两个并行子网络的输出进行融合后构建决策准则。然而，由于输入图像块尺寸相对较小，为了保持尺度不变，网络中没有采用池化层和下采样层。尽管 CNN 具有很强的特征表示能力，但上述结构往往是局部的，对邻域信息的利用很有限。因此，这些方法严重依赖于后续手工优化，如基于交叉的代价聚合[17]、半全局匹配(SGM)[18]、左右一致性检查和各种滤波方法。第二类方法除了利用孪生网络学习对应关系的特征表示外，还引入条件随机场(CRF)模型，结合语义信息进行优化。例如，Pal 等[19]训练线性 CRF 模型来进行立体匹配，其方法可以自动学习比标准手工调节 MRF 模型更为丰富的模型参数，能够更好地描述图像梯度和视差跳跃之间的关系，进而减小视差误差。Knobelreiter 等[20]提出一种端到端训练的混合 CNN-CRF 模型，仅用浅层 CNN 训练，但与其他采用后处理的方法得到的效果相当。第三类方法采用 CNN 估计立体匹配置信度，进而对视差求精得到视差结果。例如，Seki 等[21]提出一种利用训练的视差图像块与中间视差之差计算立体视觉置信度的 CNN 网络，进而用置信度调制半全局匹配参数。Poggi 等[22]利用基于深度图像块网络的局部一致性假设，改进了传统的置信度测量，得到了较好的匹配结果。

尽管 CNN 方法与传统方法相比在速度和准确度方面取得了重要突破，但大多数深度学习方法依旧很难在不适定区域(例如遮挡区域，重复纹理区域，弱纹理区域和反光表面等)找到精确的匹配点。因此，当前基于 CNN 的立体匹配方法越来越趋向于寻找一种丰富原始图像信息的方式，其中一些研究试图结合特征融合去优化匹配代价卷和视差图。例如 Güney 等[23]提出的 Displet 框架是将语义分割方法应用到立体匹配算法中，根据物体通常呈现规则结构而非任意形状的先验，利用车辆的 3D 模型解决反射和无纹理区域中的匹配模糊问题。Kendall 等[24]提出一种 GC-Net(Greenland climate network)结构，采用编码解码结构融合多尺度的特征信息去调整匹配代价，只是反卷积往往难以恢复低层已经丢失的特征。Brandao 等[25]认为池化层损失的细节信息可以通过反卷积弥补，而且其感知域

越大，得到的匹配结果越准确。Park 等[26]在 CNN 网络的决策层之前加入逐像素金字塔池化模块来扩大感受域，进而对图像进行特征融合而不会丢失细节信息。然而，该模块被附加到全连接层的后端，需要针对每个可能的视差标签重新计算，计算成本较高。

虽然以上方法采用各种不同的特征融合方式来表达图像信息，但都不可避免地损失了图像的部分信息，并带来了较高的计算代价，不利于快速准确地计算视差值。而且，由于图像中存在场景复杂、物体大小不一和光照变化等问题，采用上述方法一般难以减小这些问题带来的误匹配。针对上述问题，本文提出一种结合多尺度局部特征和深度特征的立体匹配方法，将多尺度旋转不变 Log-Gabor 局部二值模式(LG-LBP)特征和 CNN 提取的深度特征进行融合，既学习图像中浅层次的结构化信息，又学习图像中深层次的语义信息，在更全面表示图像丰富信息的同时，通过尽量减少细节信息损失来降低双目立体匹配的难度。在训练过程中，为了减小图像中不严格极线对齐带来的误差，本文方法通过在极线垂直方向添加不同强度的噪声来构造正负样本。

本文主要贡献：1）将双目图像进行多尺度多方向浅层次特征和深层次特征融合，通过提供丰富特征表示来降低双目立体匹配在不适定区域找到精确匹配点的困难。由于不适定区域主要体现为双目图像中存在物体多样化、大小不一和场景复杂等特点，结合多尺度多方向 Log-Gabor 滤波器和 LBP 算子将图像分解为若干子图像，不仅能丰富图像浅层细节信息，也能缓解 CNN 由于可能不具备旋转不变性[27]而无法学习旋转不变性特征的问题。此外，利用传统方法提取图像的多尺度特征可能在保持准确匹配的同时，使得后续 CNN 网络更小巧，网络复杂度低。2）提出一种双目立体匹配正负样本构造方法，通过在竖直方向加上一定程度的噪声扰动来构造正负样本，使得左右图像中对应像素点不一定位于水平极线。目前图像块匹配算法中正负样本的构造是基于极线校正后水平极线严格对齐这一假设，但是往往由于相机成像环境光照复杂，相机内外参数计算存在误差等，极线校正结果存在一定的误差。因此，可以一定程度上减小极线校正中误差所带来的误匹配，并且这样的随机性能增加样本的多样性，进而学习到泛化性能更好的网络，提高左右图像块的匹配正确率。

2　本文方法

2.1　方法概述

本文方法总体框架如图 1 所示。首先，将左右视图送入到特征融合模块，提取图像的不同尺度多方向的 LG-LBP 特征，进行特征融合得到多尺度特征图。然后将多尺度 LG-LBP 特征和 CNN 提取的深度特征进一步级联，从而丰富原始图像的信息表示。随后，将融合后的特征图像块送入到 MC-CNN-fast 网络中进行相似度匹配计算，训练得到最终的网络模型。为了精细化网络预测的粗略视差图，对其进行代价聚合，采用 WTA(winner-take-all)策略得到左右视差图，最后采用左右一致性检测、亚像素增强和图像滤波进行后处理。

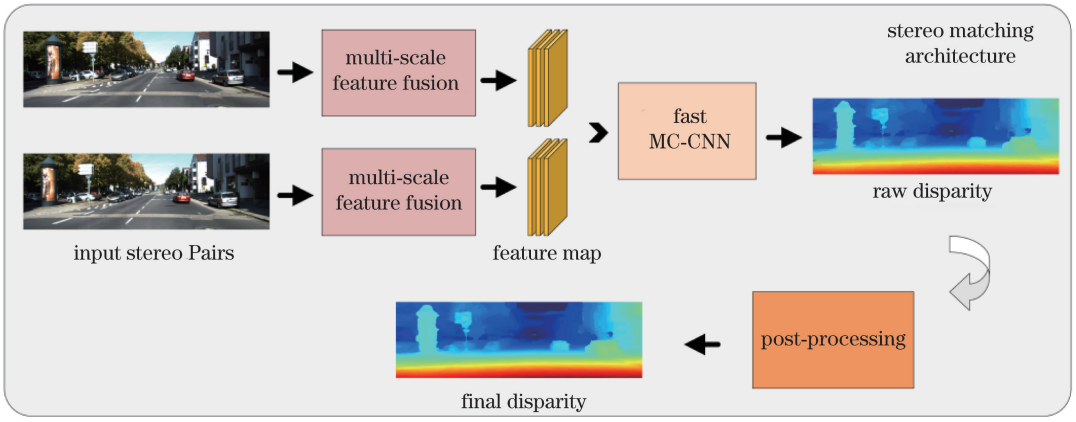

图 1　本文提出的双目立体匹配方法总体框架

Fig. 1　Overall architecture of proposed binocular stereo matching method

2.2　双目立体匹配正负样本构造

在双目立体匹配中，需要计算左图上的某点 $p(x,y)$ 和右图上的某点 $q(x-d,y)$ 的匹配代价，进而找到相对应的匹配像素点，其中 d 为所有在考虑范围的视差。数据集的构造普遍采用文献[2]的方法，正样本中心像素点的坐标在右图像上表示为

$$q(x-d+o_{\mathrm{pos}},y), \tag{1}$$

式中：o_{pos}为$[-V_{\mathrm{pos}},V_{\mathrm{pos}}]$范围内的一个随机数，其中$V_{\mathrm{pos}}$、$-V_{\mathrm{pos}}$分别是正样本中心像素点距离$q(x-d,y)$的上下限。实验表明，当$V_{\mathrm{pos}}$设置为4时立体匹配得到的效果最好。同理，将负样本设计为

$$q(x-d+o_{\mathrm{neg}},y), \tag{2}$$

式中：o_{neg}为$[-Z_{\mathrm{low}},Z_{\mathrm{high}}]$范围内的一个随机数，其中$Z_{\mathrm{high}}$、$Z_{\mathrm{low}}$分别是负样本中心像素点距离$q(x-d,y)$的上下限，$Z_{\mathrm{low}}$和$Z_{\mathrm{high}}$分别设置为4和18。由于“优良”匹配和近似“优良”匹配的匹配代价均被设置为很小时，交叉代价聚合的表现更好，因此o_{pos}没有被设置为0。

上述数据集构造方式是基于极线校正没有误差的假设，但是由于相机镜头畸变，成像环境光照复杂，以及相机内外参数计算不准确，极线校正往往存在误差。因此，对上述构造数据集的方式进行改进，通过在竖直方向上加上一定的噪声扰动来构造正负样本，使得提取的图像块不一定位于绝对对齐的水平极线上。通过将右图像块的中间像素点设置为以下形式得到正样本：

$$q(x-d+o_{\mathrm{pos}},y+h_{\mathrm{pos}}), \tag{3}$$

同理，负样本设计为

$$q(x-d+o_{\mathrm{neg}},y+h_{\mathrm{neg}}), \tag{4}$$

式中：h_{pos}和h_{neg}分别是正负样本在竖直方向上添加的噪声值，为$[-E_{\mathrm{low}},E_{\mathrm{high}}]$中任意的随机数，该范围的上下限根据实验确定，其他参数与之前的构造方法一致，其中E_{low}、E_{high}分别是正负样本中心像素点距离$q(x-d,y)$的垂直方向上的上下限。采用这样的构造方法可以通过扩大搜索空间来减小极线校正带来的误差，也能在一定程度上增加网络的泛化性能。本文所提出的数据集构造方式如图2所示。

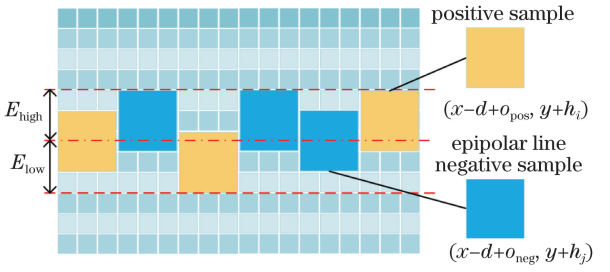

图2　正负样本的构造方式

Fig. 2　Construction of positive and negative samples

2.3　多尺度局部特征融合

2.3.1　Log-Gabor滤波器

Log-Gabor滤波器是在对数频率尺度上传递函数为高斯函数的滤波器，除具备Gabor滤波器的多通道及多分辨率等优点外，还具备两个主要特点：一是无直流分量，带宽可不受限制；二是传递函数在高频端延长性好，弥补了Gabor滤波器过度表达低频分量而对高频分量表达不足的缺点，因此Log-Gabor滤波器能更加真实地反映自然纹理图像的频率响应[28]。Log-Gabor滤波器可以在频域中通过极性可分离函数直接构造，其中极性可分离函数包括径向对数高斯函数$G(f)$和周向高斯函数$G(\theta)$，f为角频率，θ为方向角。滤波器变换函数可以表示为

$$G(f,\theta)=G(f)G(\theta)=\exp\left\{-\frac{[\ln(f/f_0)]^2}{2\ln^2(\sigma_f/f_0)}\right\}\exp\left[\frac{-(\theta-\theta_0)^2}{2\sigma_\theta^2}\right], \tag{5}$$

式中：f_0为中心频率；θ_0为滤波器的初始方向角；σ_f决定滤波器的径向带宽B，$B=2\sqrt{2/\ln 2}\,\left|\ln(\sigma_f/f_0)\right|$；$\sigma_\theta$决定带宽角度$\Delta\Omega$，$\Delta\Omega=2\sigma_\theta\sqrt{2\ln 2}$。根据(5)式，不同尺度和方向参数即可表示多个滤波器。其中，尺度和频率的关系可定义为$f_0=1/[\lambda_0 s^{(n-1)}]$，$\lambda_0$为最小尺度滤波器的波长，$s$为连续滤波器尺度因子，$n$为当前尺度。当频域中的尺度为$n$，方向为$\theta$的Log-Gabor滤波器为$G_n(f,\theta)$，待滤波图像为$I(f,\theta)$时，Log-Gabor滤波结果$P_n(f,\theta)$表示为

$$P_n(f,\theta)=\mathcal{F}^{-1}\{\mathcal{F}[I(f,\theta)]G_n(f,\theta)\}, \tag{6}$$

式中：$\mathcal{F}$和$\mathcal{F}^{-1}$分别为二维傅里叶变换和傅里叶逆变换。

由于Log-Gabor滤波器具有较好的尺度选择特性，采用不同尺度的多组Log-Gabor滤波器对图像进行卷积，可以表现图像的多尺度细节。然而，由于Log-Gabor滤波器易受图像旋转影响，因此不太适用于场景复杂和目标多元的图像特征提取任务。为此，将尺度为n的D个方向的滤波器分别与图像进行卷积，以图像实部作为输出特征图，即令

$$I_n^r=\mathrm{Re}\{P_n^r(f,\theta)\}, \tag{7}$$

式中：$P_n^r(f,\theta)$为Log-Gabor滤波器处理得到的输出；r为滤波器角度序列编号，表示$0\sim D$之间的整数序列；Re为取实部操作；I_n^r为处理后的特征图像。最后逐像素求取特征图的最大值：

$$I_n(x,y)=\max\{I_n^r(x,y)\,|\,r=0,1,\cdots,D\}。 \tag{8}$$

采用不同尺度和不同方向的Log-Gabor滤波器得到的特征图像如图3所示，其中$s\in\{1,2,4,8\}$表示不同尺度，而$D\in\{4,8\}$表示滤波器方向个数，角度序列为$\pi r/4$。左边一列图像是利用不同尺度

4 个方向的滤波器对图像进行卷积之后的最大值响应，尺度越小图像的细节越明显，体现出多分辨率特性，即变焦能力。而右边一列图像是利用不同尺度 8 个方向滤波器对图像进行卷积之后取最大值得到的，可以发现，8 个方向的滤波器的最大值响应对比度更强，特征细节更为明显，可以更好地描述场景复杂和目标多元的图像信息。在图像特征响应最大的方向提取图像特征，无需根据图像的方向性指定滤波器的某个方向。综合考虑，本文选择 3 种尺度和 8 个方向的 Log-Gabor 滤波器作为卷积模板。

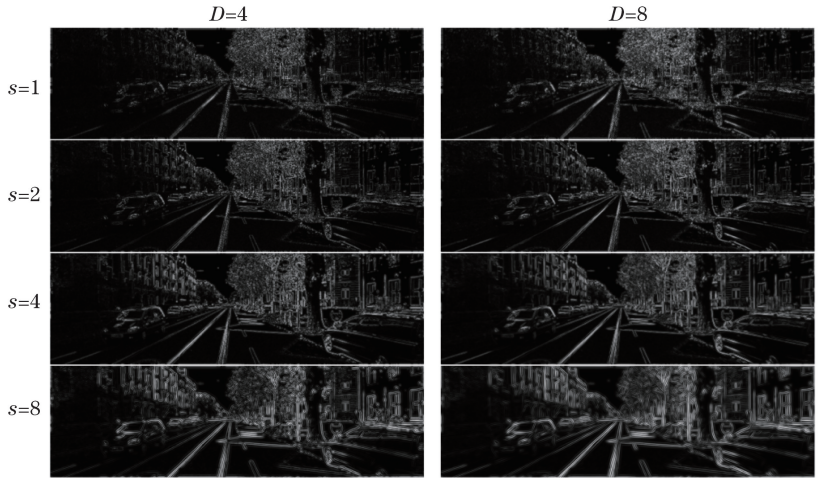

图 3　不同尺度和方向的 Log-Gabor 滤波器卷积后的图像特征

Fig. 3　Image features convoluted by Log-Gabor filter with different scales and directions

2.3.2　旋转不变均匀 LBP

LBP 方法最早由 Ojala 等[29]提出，主要用来描述图像局部纹理特征，目前该方法成为纹理分类和人脸识别领域主要的特征提取方法之一，并在图像处理和计算机视觉领域受到越来越多的关注。LBP 用于测量像素与其周围像素之间的差异，并捕获局部图像纹理的空间结构，有利于描述各种复杂场景下的图像。对给定像素点周围的相邻像素进行阈值化获得 LBP 码的计算式为

$$C_{\mathrm{LBP\_}P,R}=\sum_{p=0}^{P-1}K(g_p-g_c)2^P, \tag{9}$$

$$L(x')=\begin{cases}1, & \text{if } x'>0\\ 0, & \text{if } x'\leqslant 0\end{cases}, \tag{10}$$

式中：g_c 为中心像素的灰度值；g_p（邻域内每个像素点对应的标号 $p=0,\cdots,P-1$）为半径 R 圆域内像素的灰度值，P 为邻域像素点个数。

操作算子 $C_{\mathrm{LBP\_}P,R}$ 通过中心像素点周围的 P 个像素产生 2^P 个不同的二进制模式，但是图像一旦旋转，相邻像素将沿着围绕中心像素的圆周相应地移动，因此将导致不同的 $C_{\mathrm{LBP\_}P,R}$ 值。为了克服旋转而获得唯一的 LBP 值，研究者提出了旋转不变 LBP，对应表达式为

$$C^{\mathrm{r}\_i}_{\mathrm{LBP\_}P,R}=\min[\mathrm{ROR}(C_{\mathrm{LBP\_}P,R},i)\,|\,i=0,1,\cdots,P-1], \tag{11}$$

式中：ROR(x,i)表示将 P 位数 x 沿时钟方向移动 i 次。而算子 $C^{\mathrm{r}\_i}_{\mathrm{LBP\_}P,R}$ 作为特征检测器用于保持图像中与旋转不变图案对应的某些微小特征。由于提取的二进制特征的最小值是不变的，因此采用最小值作为提取的 LBP 特征。

为了进一步提高 LBP 特征描述子的旋转不变性能，并降低其特征维数，可以将描述子 $C^{\mathrm{r}\_i}_{\mathrm{LBP\_}P,R}$ 扩展为旋转不变均匀描述子 $C^{\mathrm{riu2}}_{\mathrm{LBP\_}P,R}$，将旋转不变 LBP 模式进一步分为均匀旋转不变模式和非均匀模式，具体定义为

$$C^{\mathrm{riu2}}_{\mathrm{LBP\_}P,R}=\begin{cases}\sum_{p=0}^{P-1}K(g_p-g_c), & U(C_{\mathrm{LBP\_}P,R})\leqslant 2\\ P+1, & \text{otherwise}\end{cases}, \tag{12}$$

式中：$U(\cdot)$表示 LBP 模式中圆周上相邻的两个二元值 0/1（或 1/0）的转移次数。旋转不变均匀模式仅为 $P+1$ 类，所有非均匀模式归为 1 类，因此用于表示整幅纹理图像的 $C^{\mathrm{riu2}}_{\mathrm{LBP\_}P,R}$ 直方图矢量特征仅 $P+2$ 维，显著低于其他 LBP 描述子。采用不同半径和邻域像素点的旋转不变均匀 LBP 算子提取的特征图如图 4 所示，其中 $R\in\{1,2,3\}$，$P\in\{8,16,24\}$，滤波器的方向选择 8 个，尺度 s 分别取 1,2,4。可以看出，当半径和邻域像素点个数较小时，图像特征保留下来的细节信息更丰富。综合考虑特征丰富程度和计算效率，本文采用 $R=3$，$P=24$ 的 LBP 算子。

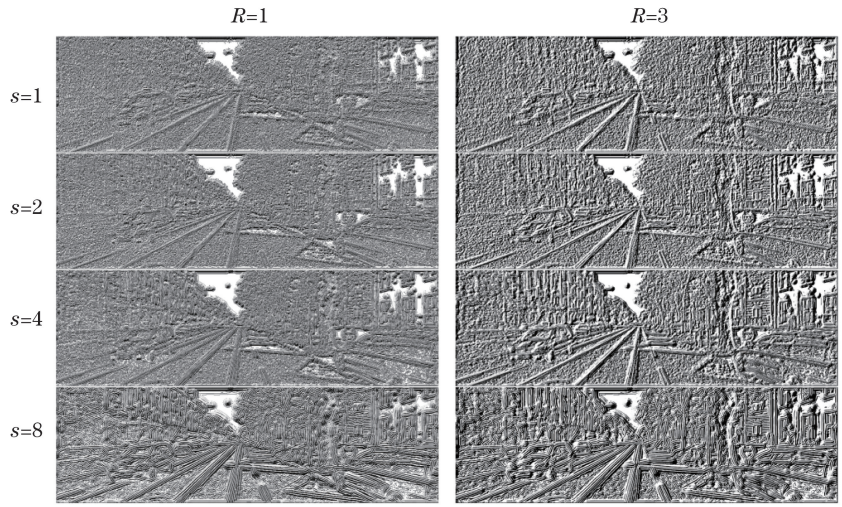

图 4　旋转不变均匀 LBP 图像特征

Fig. 4　Rotation invariant uniform LBP image features

2.3.3　多尺度旋转不变特征融合结构

考虑到 Log-Gabor 滤波器的多尺度特性以及 LBP 算子的光照不变性和旋转不变性，本文提出一种多尺度局部特征融合框架，将多尺度的 LG-LBP 特征与 CNN 抽象化的深度特征进行融合，从而在一定程度上解决图像中旋转变化、尺度变化和光照变化的问题，通过将浅层的结构化信息和深层的语义信息进行融合，能丰富图像的局部和全局信息，进而在一定程度上解决不适定区域的问题。

本文提出的多尺度局部特征融合框架如图 5 所示，该框架主要分为上下两个部分。上半部分是多尺度 LG-LBP 特征融合结构，用于提取三种不同尺度的特征图像，每种尺度的特征图像依次由 Log-Gabor 滤波器和 LBP 算子计算得到，Log-Gabor 滤波器用于提取不同尺度的特征，而 LBP 算子用于捕获局部图像纹理的空间结构，最后将三种尺度的特征图像堆叠组成三通道的特征图像。下半部分是卷积神经网络特征提取结构，其输入是从原始图像上提取的图像块，经过四个卷积层进行卷积操作，从而提取出逐层抽象的深层次特征。最终将图像的深层次特征和多尺度的 LG-LBP 特征进行融合，得到既包含浅层次结构信息又包含深层次语义信息的特征表示，从而丰富原图像的信息表达。

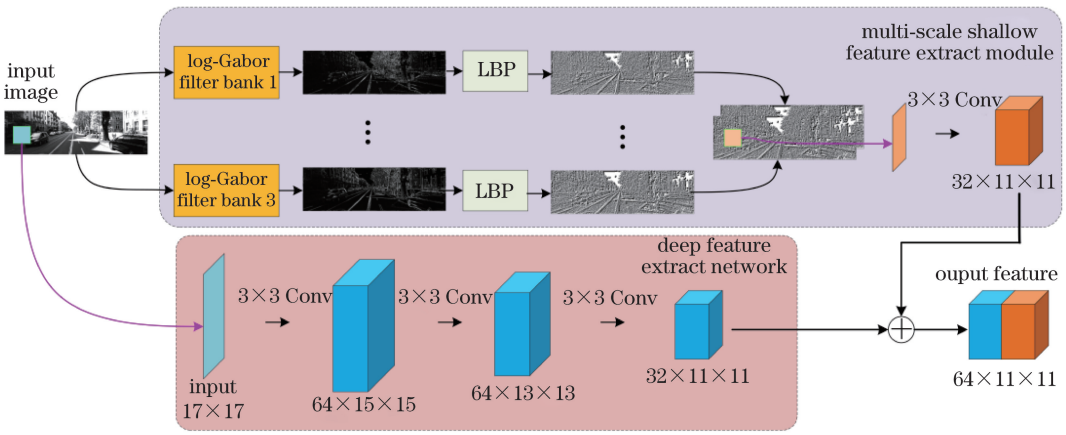

图 5　多尺度特征图像块融合模块

Fig. 5　Fusion module of multi-scale feature image blocks

2.4　MC-CNN-fast 网络

在基于 CNN 的立体匹配方法中，Žbontar 等[14]提出的快速卷积神经网络和准确卷积神经网络颇具代表性，快速卷积神经网络以孪生网络为基础计算图像块的相似度，而准确神经网络加入全连接的决策层形成一个二分类网络。由于网络结构的不同，快速卷积神经网络相对而言匹配速度快，但是准确率不如准确卷积神经网络，因此在准确度要求

高的场合不具有优势。为了弥补快速卷积神经网络在准确度方面的不足，本文在网络前端加上多尺度局部特征融合框架，在尽可能保证速度的同时提升准确度。基于 MC-CNN 的快速卷积神经网络结构如图 6 所示，多个卷积层和 ReLU 层组成特征提取器，通过 Normalize 层对特征进行归一化后计算点积，最终输出两个图像块的相似度分数。具体的步骤如下。

1）分别采用 4 层卷积层提取左右图像块的不同特征信息，并且左右支路实现参数共享。

2）经过 Normalize 层后联结左右特征信息，利用点积操作计算特征图像块的相似度。其中，损失函数使用 Hinge Loss 函数，表达式为

$$L(y',y'')=\max\{m+y''-y',0\},\quad(13)$$

式中：y'为正样本的得分，y''为负样本的得分，两者差值表示两种预测结果的相似关系；m 为间隔。本文根据实验进行设置。

3）根据网络输出的相似性分数判断图像块是否相似，进而用于后续的立体匹配。

4）对相似性分数以反比例形式构造代价函数，图像块越相似代价越小，反之，图像块相似性越小，代价越大。

5）采用交叉代价聚合、半全局匹配、亚像素增强和滤波等对代价函数进行后处理，得到最终的视差图。

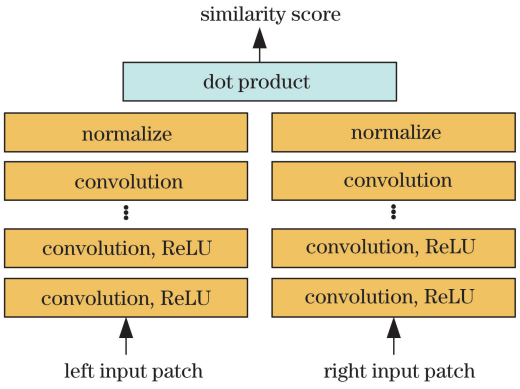

图 6　基于 MC-CNN 的快速网络架构

Fig. 6　Fast network architecture based on MC-CNN

3　实验结果及分析

3.1　数据集与实验平台

3.1.1　KITTI 数据集

KITTI 数据集[30-31]由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，是目前最大的自动驾驶场景下计算机视觉算法评测数据集。该数据集包含多个分辨率为 375 pixel×1242 pixel 的道路图像对，由固定在车顶上的两个摄像头拍摄，每张图像中最多包含 15 辆车和 30 个行人，还有各种程度的遮挡与截断。其中 KITTI2012 数据集中包括 194 个带有稀疏标签视差的训练图像对，以及没有标签视差的 195 个测试图像对，本文将其中的 160 个图像对用于训练网络，另外 34 个图像对用于验证网络。KITTI2015 数据集包括 200 个具有稀疏标签视差的训练图像对，以及没有标签视差的 200 个测试图像对，本文将其中的 160 个图像对用于训练网络，另外 40 个图像对用于验证网络。由于颜色信息对提高立体匹配的效果没有显著作用，因此实验中均将彩色图像转换为灰度图像，并且所有图像都经过预处理，即用图像减去平均值并除以其像素灰度值的标准偏差进行灰度值归一化。

3.1.2　实验平台

本文所有实验均在 2.0 GHz Intel CPU、96 GB RAM、NVIDIA Tesla K80、Windows7 64 位 PC 机上进行，编程环境采用 Anaconda 5.0.1(Python 3.5)、TensorFlow1.4。

3.2　本文方法相关参数设置

3.2.1　训练参数设置

本文采用小批量梯度下降的方式进行训练，训练和验证的过程重复一定次数，其中训练时的批尺寸设置为 128，动量设置为 0.9。在训练过程中，使用动态的学习率有助于加速收敛，以较少的迭代次数得到更优的模型精度，本文通过不断尝试不同参数值，最终将网络初始学习率设置为 0.03，为了适应不同阶段的权值修正幅度，迭代后期逐渐减小学习率，当训练到第 5000 个 epoch 时，损失函数接近收敛。训练损失曲线和验证损失曲线如图 7 所示。

3.2.2　CNN 损失函数中的间隔选择

本文 CNN 损失函数中的间隔 m 用来衡量正负样本相似度得分的差异，当正样本的相似度得分与负样本的相似度得分之差大于 m 时损失为 0，目的是为了使正样本与锚样本更加靠近，而负样本与锚样本更加远离。为了得到合适的 m 值，本文在 KITTI2012 和 KITTI2015 数据集上训练 4000 次，得到不同 m 值对应的平均错误率如图 8 所示。从图 8 中可知，当 m 为 0.2 或 0.3 时获得了较好的匹配误差率，因此将间隔 m 设置为 0.2。

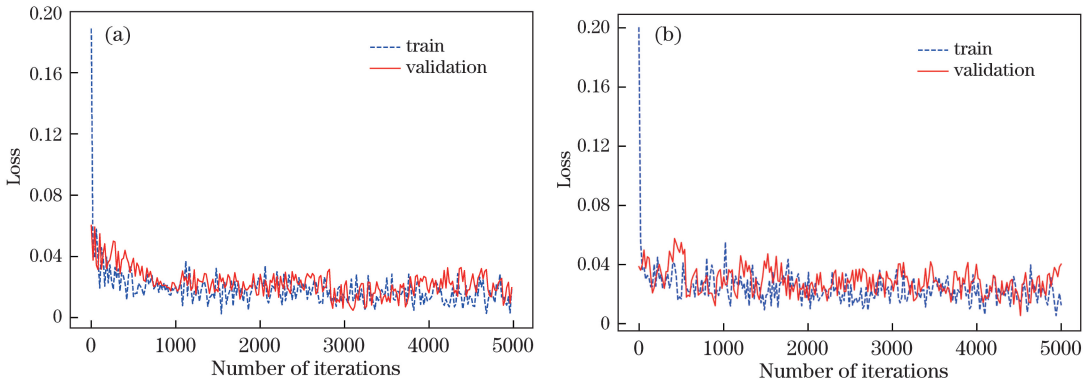

图 7　两个数据集中的训练过程损失曲线。(a) KITTI2012 dataset；(b) KITTI2015 dataset

Fig. 7　Loss curves of training process on two datasets. (a) KITTI2012 dataset; (b) KITTI2015 dataset

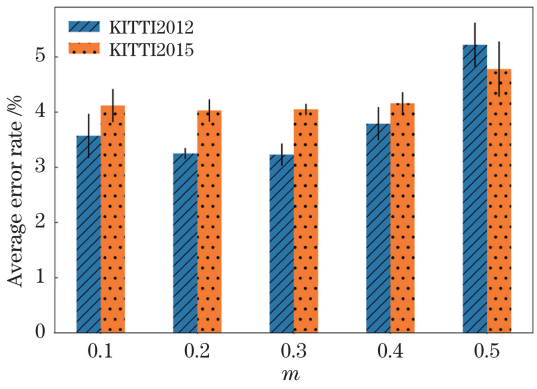

图 8　不同间隔 m 下两个数据集中的平均错误率

Fig. 8　Average error rate for different margins on KITTI2012 and KITTI2015 datasets

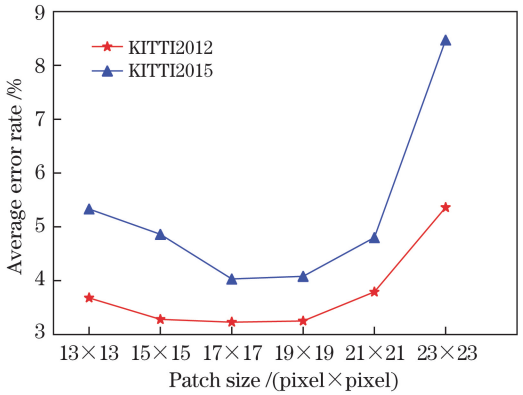

图 9　不同图像块尺寸下两个数据集中的平均错误率

Fig. 9　Average error rate for different image patch sizes on KITTI2012 and KITTI2015 datasets

3.2.3　图像块尺寸设置

为了确定输入图像块尺寸，本文尝试多种大小不同的图像块(13 pixel×13 pixel，15 pixel×15 pixel，17 pixel×17 pixel，19 pixel×19 pixel，21 pixel×21 pixel，23 pixel×23 pixel)，实验中得到的模型平均错误率如图 9 所示，可以看出，本文实验中图像块应设置为 17 pixel×17 pixel。

3.2.4　噪声容限值设置

因为本文提出在图像的竖直方向加上一定的噪声来提取图像块，所以噪声的上下限的设置尤为重要。在实验中，为了得到最佳的噪声限值，采用不同参数获得训练数据并放入网络中进行学习，搜索的范围为[−5，5]。不同的噪声限值对匹配结果的影响如图 10 所示。从图 10 可知，本文选择的最优噪声容限值为 2 pixel。

此外，为了对比加上噪声提取正负样本的方法和原始的方法，从训练过程进行分析，可以得到不同方法下模型的收敛性能，如图 11 所示。点线表示原始的正负样本提取方法，实线表示在图像竖直方向加上一定的噪声来提取正负样本的方法。可以发现，采用本文方法构造的正负样本在网络训练过程中能收敛得更好，最终在训练集和验证集上得到的损失值更小，而且在验证过程中收敛曲线更平滑，说明在一定程度上增强了网络的泛化性能。

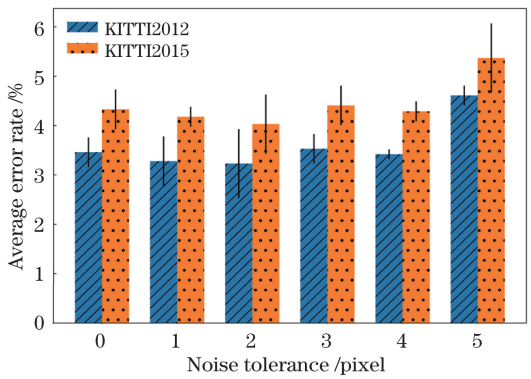

图 10　不同噪声容限下两个数据集中的平均错误率

Fig. 10　Average error rate under different noise tolerances on KITTI2012 and KITTI2015 datasets

3.3　实验结果及讨论

本文提出的方法主要包括多尺度局部特征融合

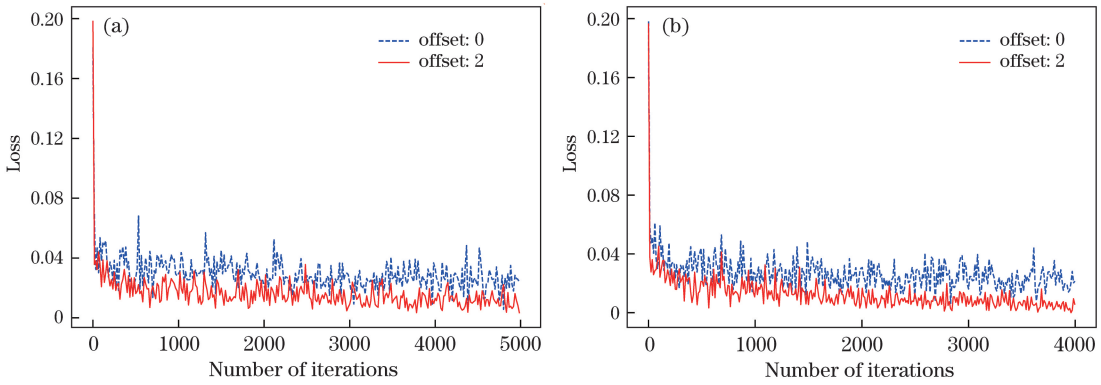

图 11　不同样本构造方法下两个数据集中的训练损失曲线。(a) KITTI2012；(b) KITTI2015

Fig. 11　Training loss curves under different sample construction methods on two datasets. (a) KITTI2012; (b) KITTI2015

模块和 MC-CNN-fast 模块，其中多尺度局部特征融合模块包括 LG-LBP 特征融合部分和卷积特征提取部分。为了验证各模块设置以及正负样本构造方式的有效性，本文对所提出方法进行拓展以产生两种变体方法。为了验证多尺度局部特征融合模块对立体匹配结果的影响，设置变体方法 1 为 LG-LBP CNN：采用多尺度特征融合模块和 MC-CNN-fast 相结合以提取特征。为了评估提出的训练数据构造方法对立体匹配结果的影响，设置变体方法 2 为 Noise CNN：采用 MC-CNN-fast 作为特征提取器，并采用本文提出的正负样本构造方法生成正负样本。其中，变体方法 1 将噪声偏移量 offset 设置为 0，变体方法 2 将噪声偏移量 offset 设置为 2。

此外，为了更全面地评估本文提出方法，实验中还引入了 MC-CNN-fast 和 SGM 进行对比，均按照本文中的设置进行实验，得到在不同数据集上的平均错误率对比结果，如表 1、2 所示。其中评价指标为视差值与基准视差值之差大于 k(k=2，3，4，5)个像素的像素点所占比例。从表 1、2 可知，本文提出的方法在立体匹配中匹配错误率明显低于变体方法 1 和变体方法 2。考虑到变体方法 1 是采用多尺度局部特征融合模块和 MC-CNN-fast 作为特征提取器，相比于 MC-CNN-fast 方法，总体的平均匹配错误率较大，这可能是由于多尺度的 LG-LBP 特征更侧重于图像的细节纹理信息，而忽略了全局信息。考虑到变体方法 2 是采用 MC-CNN-fast 作为特征提取器，正负样本的构造采用加噪声的方法，以此来纠正极线校正中的误差，相比于 MC-CNN-fast 网络，也能一定程度上提高网络的预测性能。本文提出的方法和其他代表性方法对比可以发现，本文提出的框架相对于其他算法在匹配误差率方面具有一定优势，能精确地实现双目视图的立体匹配。

表 1　各算法的视差结果平均错误率对比(KITTI2012)

Table 1　Average error rate comparison of disparity results of different algorithms (KITTI2012)

Algorithm	Average error rate /%			
	>2 pixel	>3 pixel	>4 pixel	>5 pixel
SGM[18]	6.52	5.36	4.38	3.82
MC-CNN-fast[14]	5.02	3.27	2.61	2.11
LG-LBP CNN	5.62	4.81	4.00	3.29
Noise CNN	**4.98**	3.25	2.62	2.14
Our method	5.03	**3.23**	**2.59**	**2.10**

表 2　各算法的视差结果平均错误率对比(KITTI2015)

Table 2　Average error rate comparison of disparity results with different algorithms (KITTI2015)

Algorithm	Average error rate /%			
	>2 pixel	>3 pixel	>4 pixel	>5 pixel
SGM[18]	10.37	7.13	5.54	4.71
MC-CNN-fast[14]	7.64	4.11	3.01	2.53
LG-LBP CNN	9.06	6.31	4.45	4.14
Noise CNN	7.61	4.10	3.01	**2.51**
Our method	**7.58**	**4.03**	**2.98**	2.58

通过上述对比实验，并采用交叉代价聚合等后处理方法输出图 12 所示的视差图。变体方法 1 在整体上出现了较多的匹配错误，然而图像部分细节优化明显且局部物体轮廓突出，进而体现了本文多尺度特征融合模块在特征提取方面的优势，同时也说明了多尺度提取的特征能丰富 CNN 提取的深度特征，更好地表达图像的信息。变体方法 2 采用加噪声的正负样本构造方法，在视差图中的细节部分

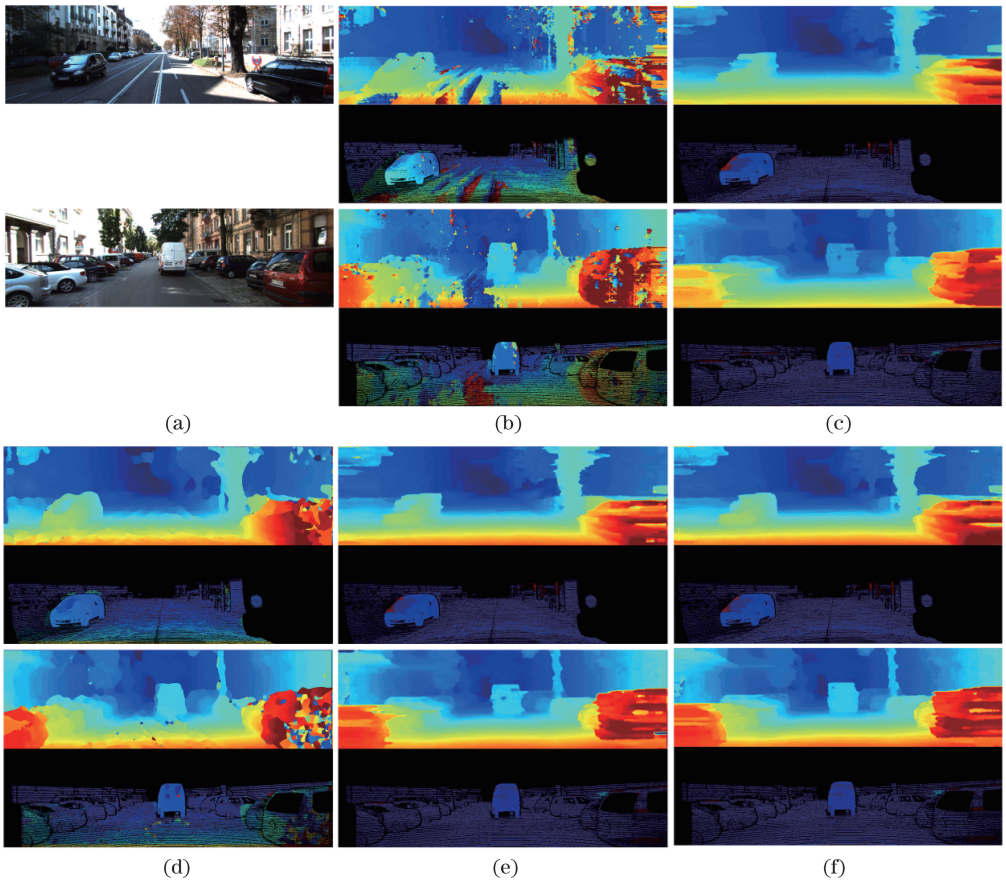

图 12　各种方法得到的最终视差结果。(a) 原输入图；(b) SGM；(c) MC-CNN-fast；(d) LG-LBP CNN；(e) noise CNN；(f) our method

Fig. 12　Disparity results obtained by various methods. (a) Original input image; (b) SGM; (c) MC-CNN-fast; (d) LG-LBP CNN; (e) noise CNN; (f) our method

能体现其作用，相对来说误匹配较少。此外，本文方法相对于变体方法在视差图中能直观显示出较好结果，在图像不适定区域能得到符合真实状况的细节。而相对于其他方法，本文方法得到的视差图同样能在不适定区域较好刻画视差细节。

在 KITTI2012 数据集和 KITTI2015 数据集上得到的视差主观结果分别如图 13 和图 14 所示。其中误差图表示实验中得到的视差值和基准视差值在各个像素点的差距，并以像素灰度值表示出来。从展示的图中可以看到，在不同数据集中本文提出的方法在视差图表现上都有较好的结果，而且对于不同的场景，面临不同的光照环境和旋转变化，以及在诸多不适定区域，本文得到的视差图都有一定的消除，这也在一定程度上说明多尺度旋转不变特征融合模块能丰富图像的信息，对不同的图像特征而言能达到信息互补的作用。

从算法时间复杂度上考虑，本文在 KITTI2015 数据集上进行训练和测试，得到各个算法的时间复杂度如表 3 所示。从中可以发现，MC-CNN-fast 方法网络训练时收敛最快，说明小网络和轻量级的模型参数有利于网络的快速收敛。本文方法在训练时间复杂度上和 MC-CNN-fast 相差不大，说明提取 LG-LBP 特征的过程耗时较短。对于测试而言，基于深度学习的方法相对于传统方法 SGM 具有明显优势。而本文方法在测试过程中不具有太大优势，主要源于 LG-LBP 特征提取需要消耗一定时间，但是相对于 MC-CNN-fast 方法，其在尽可能保证速度的同时提高了算法的匹配性能，在视差图的细节体现上具有一定的竞争性。

4　结　　论

本文结合多尺度 Log-Gabor 滤波器与旋转不变均匀 LBP 算子，提出了一种多尺度局部特征融合立体匹配方法。该方法以多尺度 Log-Gabor 滤波器组卷积输入图像，并借助旋转不变均匀 LBP 算子提取图像特征，进而捕获到图像的局部纹理信息和

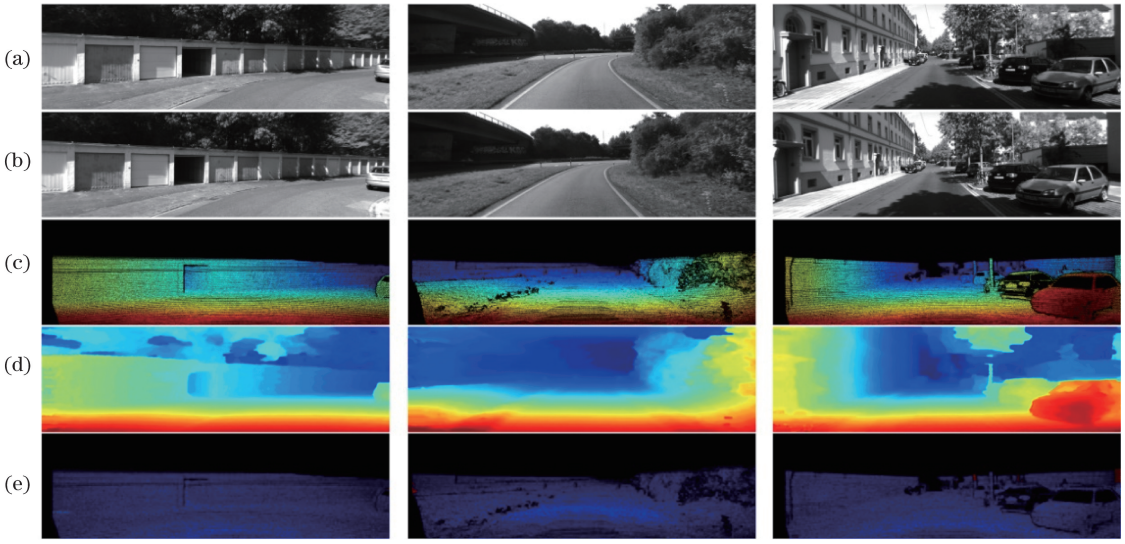

图 13　本文方法在 KITTI2012 数据集上进行立体匹配得到的视差图。(a) 原输入左图；(b) 原输入右图；(c) 基准视差；(d) 视差图；(e) 误差图

Fig. 13　Disparity maps of stereo matching obtained by proposed method on KITTI2012 dataset. (a) Original left input image; (b) original right input image; (c) ground truth; (d) disparity map; (e) error graph

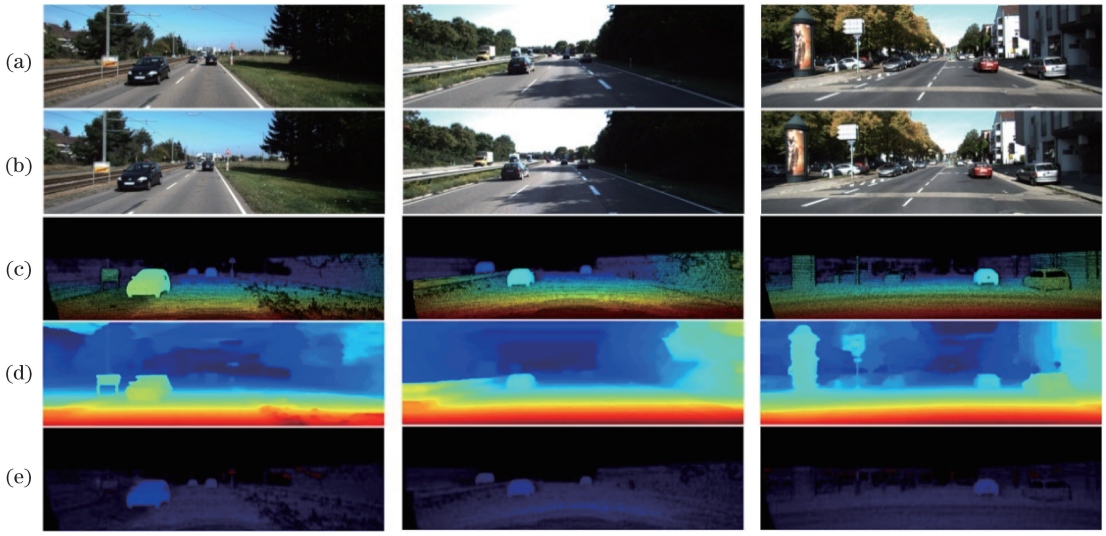

图 14　本文方法在 KITTI2015 数据集上进行立体匹配得到的视差图。(a) 原输入左图；(b)原输入右图；(c) 基准视差；(d) 视差图；(e) 误差图

Fig. 14　Disparity maps of stereo matching obtained by proposed method on KITTI2015 dataset. (a) Original left input image; (b) original right input image; (c) ground truth; (d) disparity map; (e) error graph

表 3　各算法的训练和测试时间

Table 3　Time of each algorithm in training and testing processes

Algorithm	SGM[18]	MC-CNN-fast[14]	LG-LBP CNN	Noise CNN	Our method
Train runtime /h	-	5.6	5.8	5.6	6.5
Test runtime /s	14	1.52	2.01	1.76	2.06

全局信息。此外，利用卷积神经网络提取图像的深层次特征，并将多尺度的 LG-LBP 特征和卷积特征进行级联，形成最终既包含深层次语义信息特征又包含浅层结构信息的特征图像。实验结果表明，本文方法在立体匹配不适定区域能得到较好的匹配结果，相比一些传统方法和部分深度学习方法有一定

优势，而且能在一定程度上减小由图像的旋转、尺度和光照变化，以及相机畸变带来的误差，进而得到细节较好的立体视差图。

参 考 文 献

[1] Xiao J S, Tian H, Zou W T, et al. Stereo matching based on convolutional neural network[J]. Acta Optica Sinica, 2018, 38(8): 0815017.
肖进胜, 田红, 邹文涛, 等. 基于深度卷积神经网络的双目立体视觉匹配算法[J]. 光学学报, 2018, 38(8): 0815017.

[2] Geiger A, Roser M, Urtasun R. Efficient large-scale stereo matching[M] // Kimmel R, Klette R, Sugimoto A. Computer vision-ACCV 2010. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2011, 6492: 25-38.

[3] Wang P, Wu F. A local stereo matching algorithm based on region growing[M] // Zhang W, Yang X, Xu Z, et al. Advances on digital television and wireless multimedia communications. Communications in computer and information science. Berlin, Heidelberg: Springer, 2012, 331: 459-464.

[4] Chang X F, Zhou Z, Wang L, et al. Real-time accurate stereo matching using modified two-pass aggregation and winner-take-all guided dynamic programming[C] // 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, May 16-19, 2011, Hangzhou, China. New York: IEEE, 2011: 73-79.

[5] Ma N, Men Y B, Men C G, et al. A small baseline stereo matching method based on extended phase correlation[J]. Acta Electronica Sinica, 2017, 45(8): 1827-1835.
马宁, 门宇博, 门朝光, 等. 基于扩展相位相关的小基高比立体匹配方法[J]. 电子学报, 2017, 45(8): 1827-1835.

[6] Yang Y Y, Wang H B, Liu B. A new stereo matching algorithm based on adaptive window[C] // 2012 International Conference on Systems and Informatics (ICSAI2012), May 19-20, 2012, Yantai, China. New York: IEEE, 2012: 1815-1819.

[7] Ma R H, Zhu F, Wu Q X, et al. Dense stereo matching algorithm based on image segmentation[J]. Acta Optica Sinica, 2019, 39(3): 0315001.
马瑞浩, 朱枫, 吴清潇, 等. 基于图像分割的稠密立体匹配算法[J]. 光学学报, 2019, 39(3): 0315001.

[8] Liu Y, Li Q W, Huo G Y, et al. Local binary description combined with superpixel segmentation refinement for stereo matching[J]. Acta Optica Sinica, 2018, 38(6): 0615003.
刘艳, 李庆武, 霍冠英, 等. 结合局部二进制表示和超像素分割求精的立体匹配[J]. 光学学报, 2018, 38(6): 0615003.

[9] Li P X, Liu P F, Cao F D, et al. Weight-adaptive cross-scale algorithm for stereo matching[J]. Acta Optica Sinica, 2018, 38(12): 1215006.
李培玄, 刘鹏飞, 曹飞道, 等. 自适应权值的跨尺度立体匹配算法[J]. 光学学报, 2018, 38(12): 1215006.

[10] Chen D M, Ardabilian M, Chen L M. A fast trilateral filter-based adaptive support weight method for stereo matching[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(5): 730-743.

[11] Chen J, Cai C H, Li C H. A multi-window stereo matching algorithm in rank tranform domain[C] // 2012 IEEE 11th International Conference on Signal Processing, October 21-25, 2012, Beijing, China. New York: IEEE, 2012: 997-1000.

[12] Hirschmuller H. Accurate and efficient stereo processing by semi-global matching and mutual information[C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 8624110.

[13] Žbontar J, LeCun Y. Computing the stereo matching cost with a convolutional neural network[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 1592-1599.

[14] Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of Machine Learning Research, 2016, 17: 1-32.

[15] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 4353-4361.

[16] Chen Z Y, Sun X, Wang L, et al. A deep visual correspondence embedding model for stereo matching costs[C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 972-980.

[17] Zhang K, Lu J B, Lafruit G. Cross-based local stereo matching using orthogonal integral images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(7): 1073-1079.

[18] Hirschmüller H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions

on Pattern Analysis and Machine Intelligence, 2008, 30(2): 328-341.

[19] Pal C J, Weinman J J, Tran L C, et al. On learning conditional random fields for stereo[J]. International Journal of Computer Vision, 2012, 99(3): 319-337.

[20] Knobelreiter P, Reinbacher C, Shekhovtsov A, et al. End-to-end training of hybrid CNN-CRF models for stereo[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1456-1465.

[21] Seki A, Pollefeys M. Patch based confidence prediction for dense disparity map[C]// Procedings of the British Machine Vision Conference 2016, September 19-22, 2016, York, UK. UK: BMVA Press, 2016: 23.

[22] Poggi M, Mattoccia S. Learning to predict stereo reliability enforcing local consistency of confidence maps[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 4541-4550.

[23] Güney F, Geiger A. Displets: resolving stereo ambiguities using object knowledge[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 4165-4175.

[24] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 66-75.

[25] Brandao P, Mazomenos E, Stoyanov D. Widening siamese architectures for stereo matching[J]. Pattern Recognition Letters, 2019, 120: 75-81.

[26] Park H, Lee K M. Look wider to match image patches with convolutional neural networks[J]. IEEE Signal Processing Letters, 2017, 24(12): 1788-1792.

[27] Lenc K, Vedaldi A. Understanding image representations by measuring their equivariance and equivalence[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 991-999.

[28] Walia E, Verma V. Boosting local texture descriptors with Log-Gabor filters response for improved image retrieval[J]. International Journal of Multimedia Information Retrieval, 2016, 5(3): 173-184.

[29] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[30] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 3354-3361.

[31] Menze M, Geiger A. Object scene flow for autonomous vehicles[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 3061-3070.