基于双向递归卷积神经网络的单目红外视频深度估计

吴寿川[1]，赵海涛[1]，孙韶媛[2]

[1]华东理工大学信息科学与工程学院，上海 200237；

[2]东华大学信息科学与技术学院，上海 201620

摘要 考虑到红外视频的深度特征具有单帧图像的独特性和视频全局的连续性，在单目红外视频深度估计问题上提出一种基于双向递归卷积神经网络(BrCNN)的深度估计方法。BrCNN 在卷积神经网络(CNN)能够提取单帧图像特征的基础之上引入循环神经网络(RNN)传递序列信息机制，使其既具有 CNN 良好的图像特征提取能力，能够自动提取视频中每一帧图像的局部特征，又具有 RNN 良好的序列特征提取能力，能够自动提取视频中每一帧图像所包含的序列信息，并向后递归传递这种信息。采用双向递归的视频序列信息传递机制来估计红外视频的深度，提取到的每一帧图像的特征都包含了视频前后文的序列信息。实验结果表明，相对于传统 CNN 提取单帧图像特征进行的估计，使用 BrCNN 能够提取更具有表达能力的特征，估计出更精确的深度。

关键词 机器视觉；双向递归卷积；深度估计；单目红外视频；深度神经网络

中图分类号 TP391 **文献标识码** A

doi：10.3788/AOS201737.1215003

Depth Estimation from Monocular Infrared Video Based on Bi-Recursive Convolutional Neural Network

Wu Shouchuan[1], Zhao Haitao[1], Sun Shaoyuan[2]

[1] *School of Information Science and Engineering, East China University of Science and Technology, Shanghai* 200237, *China*;

[2] *School of Information Science and Technology, Donghua University, Shanghai* 201620, *China*

Abstract For depth estimation from monocular infrared video, a method based on bi-recursive convolutional neural network (BrCNN) is proposed considering the uniqueness of a single frame and the continuity of the entire infrared video. BrCNN introduces the sequence information transfer mechanism of recurrent neural network (RNN) on the basis of the single frame feature extracted by the convolutional neural network (CNN). Thus, BrCNN possesses the feature extraction ability of CNN for a single image, which can automatically extract the local features of each frame in the infrared video, and the sequence information extraction ability of RNN, which can automatically extract the sequence information contained in each frame of the infrared video and recursively transfer this information. By introducing the bi-recursive sequence information transfer mechanism to estimate the depth of monocular infrared video, features extracted from each image containing the context information. The experimental results show that BrCNN can extract more expressive features and estimate the depth from the infrared video more precisely than the traditional CNN, which estimate the depth by extracting the feature of a single frame.

Key words machine vision; bi-recursive convolution; depth estimation; monocular infrared video; deep neural network

OCIS codes 150.1135; 100.2960; 350.4600

1 引 言

生成图像和视频深度图是计算机视觉领域中的一个基本课题，其中红外视频深度估计问题作为当前热

收稿日期：2017-06-21；**收到修改稿日期**：2017-08-08

基金项目：国家自然科学基金(61375007)、上海市科委基础研究项目(15JC1400600)

作者简介：吴寿川(1993—)，男，硕士研究生，主要从事深度神经网络与计算机视觉方面的研究。

E-mail：scwu@mail.ecust.edu.cn

导师简介：赵海涛(1974—)，男，博士，教授，主要从事模式识别与计算机视觉方面的研究。

E-mail：haitaozhao@ecust.edu.cn(通信联系人)

点受到了广泛的关注。红外图像与视频的成像特性决定了其难以从局部细节入手而是通过提取纹理、梯度等特征来进行深度估计。不过，相对于单帧静止的红外图像，红外视频拥有更多可利用的帧间先验信息等动态序列特征。基于此，本文在提取红外视频中每一帧图像特征的同时加入了视频的序列信息，以提取更加丰富且具有表达能力的特征。实验结果表明，基于加入序列信息之后提取到的视频深度特征可以估计出更精确的红外视频深度。

针对图像和视频的深度估计问题，前人已经做了大量的工作，包括使用传统机器学习方法[1-5]、深度学习方法[6-8]以及几何结构方法[9-11]。在彩色图像的深度估计中，通过基于 k 近邻(KNN)算法得到图像的先验深度的策略被广泛使用。这类方法[1-4]首先需要建立一个原图像和深度标签图像的先验池，当要估计某一图像的深度时，首先从该先验池中选择与待估计图像相似的图像，然后用这些相似图像的深度融合成待估计图像的先验深度，最后在先验深度的基础上提取图像特征(如颜色、位置、纹理等)来估计图像的深度。先验深度的加入，降低了图像深度估计的难度，同时也提高了估计的精度。在单目红外图像深度估计中，席林等[5]借助逐步线性回归和独立成分分析(ICA)获得图像的深度特征，通过支持向量机(SVM)模型学习到图像的深度。

卷积神经网络(CNN)现已在图像分类[12-14]和目标检测[15-17]等领域获得极大的成功。近来，越来越多的研究工作将其引入到图像深度估计中。许路等[6]将人工提取特征与 CNN 自动提取的特征相结合，基于得到的图像特征估计图像深度，取得了较好的估计结果。Eigen 等[7]提出了一种多尺度的卷积网络用于估计图像深度，该网络有两种特征提取机制，一是从图像全局角度出发提取图像整体特征，二是在全局特征的基础之上精细提取图像各区域的局部特征，通过将两种特征相结合，最终在测试数据上得出满意的结果。Liu 等[8]提出了一种基于超像素分割的彩色图像深度估计方法，通过将图像进行超像素分割，然后使用 CNN 分别提取各个超像素上的特征，估计出每一个超像素的深度。该方法视每一个超像素具有同样的深度，减少了实际估计过程中的计算量，提高了可运用的能力。

循环神经网络(RNN)[18-20]在自然语言处理等序列处理问题上取得了很好的效果。考虑到红外视频深度估计这一问题中既包含对单帧图像的深度特征提取，又要综合视频整体的序列特征，基于此，本文在深度 CNN 的基础之上引入 RNN 的序列特征提取递归机制，得到双向递归卷积网络(BrCNN)用于估计红外视频深度。BrCNN 包含 3 个过程：1)通过卷积操作提取红外视频中每一帧图像的局部信息；2)分别通过正向和反向递归传递，将视频的整体信息传至每一帧图像；3)综合每一帧图像已提取的局部信息和视频整体的信息，得到红外视频的深度特征并用于深度估计。在考虑每一帧图像特征的同时，加入了视频序列特征，提高了特征全局表达的能力，降低了图像中因噪声带来的特征不稳定性，并在实验中提高了红外视频深度估计的准确率。

2　递归卷积机制

卷积操作对于单幅图像来说是一个常用且有效的特征提取方式，已广泛地运用在计算机视觉的各个研究领域。深度 CNN 就是将多层的卷积操作堆叠在一起，通过采样层降低数据的维度，最后将提取到的特征输入进分类器得到结果。在红外视频中如果第 k 帧图像的第 i 个像素用 s_i^k 表示，那么卷积操作提取的特征可表示为

$$h_i^k = \sum_{o\in[-l/2,l/2]^2} \langle (s_i^k + o), (w + o) \rangle , \tag{1}$$

式中 h_i^k 表示卷积操作提取到的第 k 帧图像中第 i 个像素的特征，w 表示卷积操作的卷积核，o 表示该像素周围大小为 $l\times l$ 的方块，$\langle a, b\rangle$ 表示 a，b 元素的乘积。使用不同的卷积核能够提取到一幅图像不同的特征。深度 CNN 在训练数据上能够学习到提取不同特征的卷积核。

红外视频不仅拥有单帧图像的局部信息，而且也包含大量的序列信息。LSTM(Long short-term memory)[18]是 RNN 一个非常成功的变体，已广泛地运用在序列处理问题上。基于此，在递归卷积操作中加入了 LSTM 的设计哲学，使得每一帧图像的特征中都蕴含视频整体信息。若 h_i^k 为红外视频中第 k 帧图像中

第 i 个像素所提取到的局部特征，那么递归卷积提取到的特征为

$$x_i^k=\mu\cdot x_i^{k-1}+\nu\cdot h_i^k,\tag{2}$$

式中 x_i^k 表示通过递归卷积操作提取到的红外视频中第 k 帧图像中第 i 个像素的特征。在实验中令 $x_i^0=0$，即在视频第一帧图像之前的视频序列信息设为 0。μ、ν 为控制视频的序列特征与当前图像的局部特征之间的比例系数。递归卷积操作并没有将全部量的特征输出，而是将各个维度上的特征激活，再输出一定比例的激活后的特征，若其比例系数为 ω，则可表示为

$$y_i^k=\omega\cdot f_{\mathrm{relu}}(x_i^k),\tag{3}$$

式中 $f_{\mathrm{relu}}(x)$ 为深度学习中常用的激活函数，其表达式为

$$f_{\mathrm{relu}}(x)=\begin{cases}0, x\leqslant 0\\ x, x>0\end{cases}。\tag{4}$$

该激活函数可以有效减少深度神经网络层数较多时的梯度消失现象，且计算简单，反向传播中梯度恒定。通过类似于(1)式中的卷积操作来学习得到递归卷积中的系数

$$[\mu,\nu,\omega]=\sigma\left\{\sum_{o\in[-l/2,l/2]}\langle([x_i^k,s_i^k]+o),(w+o)\rangle\right\},\tag{5}$$

式中 σ 是值域为(0,1)的 sigmoid 函数，其表达式为

$$\sigma(x)=\frac{1}{1+\exp(-x)}。\tag{6}$$

通过学习红外视频在每一个像素点的序列特征和图像中每一个像素的局部特征，得到这些系数，保证了各个特征之间的平衡，以获得更加优良的特征。

以上是单方向的递归卷积提取红外视频特征的全过程。采用双向的递归卷积提取红外视频特征，即先提取视频沿着时间轴从前到后方向的序列特征(正向序列特征)，再提取视频沿着时间轴从后往前方向的序列特征(反向序列特征)，将二者等权相加，得到最后的视频整体的序列特征，其公式为

$$F=0.5\times F_{\mathrm{f}}+0.5\times F_{\mathrm{b}},\tag{7}$$

式中 F 为总的序列特征，F_{f}、F_{b} 分别为正向序列特征和反向序列特征。图 1 展示了双向递归卷积提取红外视频特征的原理，每一个递归卷积层需要 3 个输入和 3 个输出。输入为累积到前一帧图像的双向视频序列信息和当前图像；输出为加入当前图像信息之后的双向视频序列信息和递归卷积层提取到的特征。其中，黄色部分表示沿着时间轴从前到后提取特征，绿色部分表示沿着时间轴从后往前提取特征。每一层 BrCNN 都会提取双向的红外视频序列特征，这样每提取一帧视频的深度特征时都引入了视频前后的信息，从而得以增强提取到的特征的表达能力。

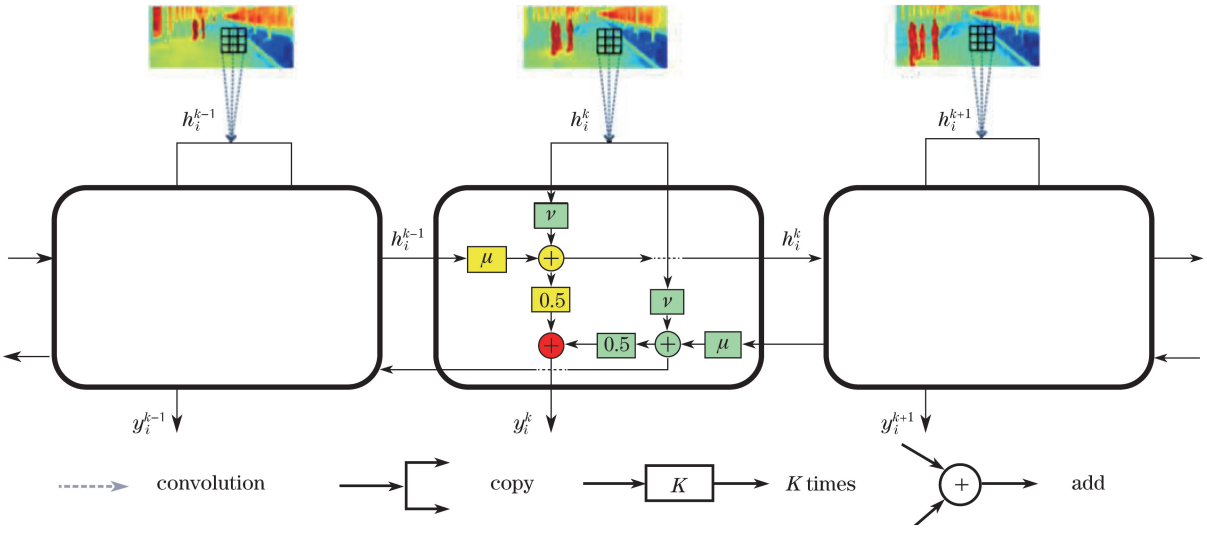


图 1　双向递归卷积原理

Fig. 1　Mechanism of bi-recursive convolution

3　双向递归卷积神经网络

3.1　网络结构

目前流行的深度卷积网络的结构，如 AlexNet[12]、VggNet[13]、ResNet[14]等，都是采用多个卷积层提取特征，然后将提取到的特征输入进一个分类器得到最终的网络输出。本文采用类似的思想，堆叠 3 个双向卷积层来提取特征。图 2 展示了实验中所采用的网络结构，该网络将传统 CNN 卷积层中的信息双向流动到红外视频的前后帧图像。

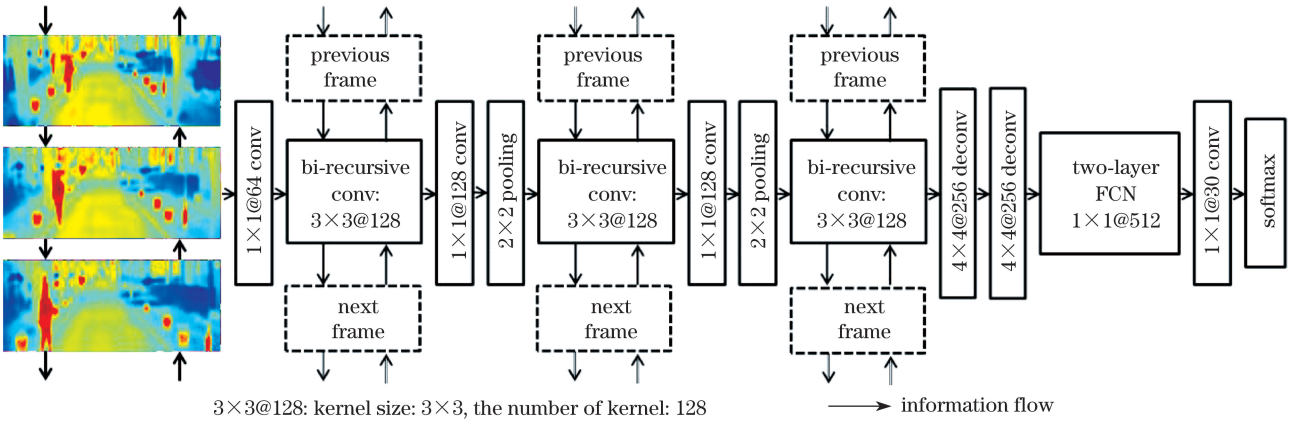


图 2　BrCNN 示意图

Fig. 2　Diagram for BrCNN

在网络的底部采用了一层卷积核数目为 64、大小为 1×1 的卷积层，将原红外视频中每一个像素映射到特征空间以输入双向递归卷积层；在网络的中部是 3 层双向递归卷积，每层卷积都采用卷积核数目为 128、大小为 3×3 的卷积来提取局部特征，以及网络中所需要的各个系数。这些卷积层都进行了填充操作，以使图像在提取特征前后的大小不变。在两个递归卷积层中间加入隔离层，该层是由一个卷积核数目为 128、大小为 1×1 的卷积层和一个大小为 2×2 的最大采样层组成。隔离层减少了相邻 2 个递归卷积层所提取到的特征的关联性，使得递归卷积层尽量提取到多样的特征。在 3 层递归卷积提取特征之后，引入 2 个反卷积层(反卷积核的数目为 256，大小为 4×4，步长为 2×2，填充为 1)，将提取到的特征图像分辨率调整到与红外视频相一致。

当特征提取完毕后，在网络顶部放置一个两层的全卷积网络(FCN)[21]用于估计红外视频中每一个像素点的深度。传统 CNN 在最后需要全连接层，全连接层将卷积层提取到的特征映射到一维空间里用于后续的分类或回归，该分类或回归属于图像级别。本文采用 FCN 进行逐个像素点的深度估计，将传统卷积网络解决的任务从图像级别引入到像素级别，可以对图像中每一个像素进行分类或回归。估计红外视频每一个像素点的深度属于像素级别的任务，鉴于此，本文采用这种全卷积神经网络对像素深度进行分类训练，其中，两层网络的卷积核数目均为 512，大小均为 1×1。采用 Softmax 分类器分类得到红外视频中每一个像素的深度，其表达式为

$$P(\xi=j\mid \boldsymbol{z})=\frac{z_j}{\sum_{i=1}^{C} z_i}。\tag{8}$$

式中 $\boldsymbol{z}$ 为分类训练得到的特征向量；z_j 表示特征向量 $\boldsymbol{z}$ 在第 j 维的数值；C 表示类别数，同时也是向量 $\boldsymbol{z}$ 的维数；$P(\xi=j\mid \boldsymbol{z})$ 表示在训练得到特征向量 $\boldsymbol{z}$ 这一条件下输入 ξ 属于第 j 类的概率。取概率最大的类别为当前输入的分类类别。

这样就得到了所提出的 BrCNN 架构。

3.2　估计策略

红外视频深度是一个连续的值，一般采用回归方法进行估计，而本文首先将红外视频的深度离散化，再使用分类的方法估计深度。这么做有 2 个优点：1)实际中红外视频的场景是多变的，那么在误差范围内估计视频中每一个像素所在的深度层次比估计一个具体的深度值更加容易、有效；2)分层次估计红外视频的深度

可以控制不同深度层次深度的精度，精度要求较高的地方可以分割得稠密，而精度要求不高的地方则可以分割得稀疏。本文在红外视频深度的对数空间中通过平均分割深度来获得分类标签，表示为

$$t_i^k = \mathrm{floor}\left[\frac{\ln d_i^k - \ln d_{\min}}{\ln d_{\max} - \ln d_{\min}} \times N\right], \tag{9}$$

式中 t_i^k 为红外视频第 k 帧图像中第 i 个像素的分类标签；floor[•]为向下取整函数；d_i^k 为第 k 帧图像中第 i 个像素的真实深度；$d_{\max}$ 和 $d_{\min}$ 分别为红外视频中最大和最小的深度；N 为分类类别数，即将红外视频的深度分割成 N 个深度层级。

将待估计深度的红外视频输入已训练好的 BrCNN，得到视频中每一个像素的分类标签，再将其换算成估计的深度，公式为

$$d = \exp\left[\frac{t}{N} \times (\ln d_{\max} - \ln d_{\min}) + \ln d_{\min}\right]。 \tag{10}$$

视频中标签为 t 的像素估计深度为 d。图 3 是将一帧图像的深度分割成不同层次的效果图，灰度图像中颜色越深表示距离越近，深度值越小，反之则深度值越大。将深度层数分割得越多、越精细，则越与真实深度相近，但估计难度就越大。实验中，将红外视频的深度分割成 30 个层级，取得了较好的估计结果。

图 3　将红外视频真实深度分割成不同深度层数的效果。(a)红外视频；(b)真实图像；(c) 10 层；(d) 20 层；(e) 30 层

Fig. 3　Effect of dividing ground truth depth of infrared video into different depth levels. (a) Infrared video; (b) ground truth; (c) 10 layers; (d) 20 layers; (e) 30 layers

4　实验与结果分析

4.1　数据采集

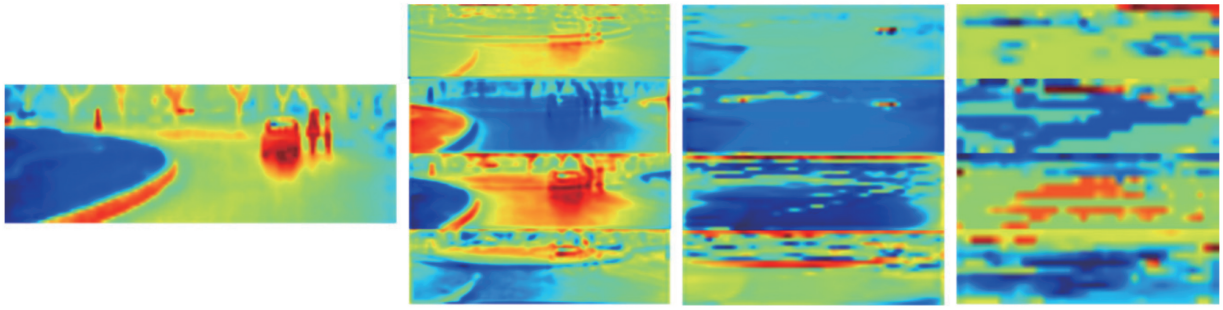
采用的数据集由包含 4144 帧户外场景的红外视频及其深度图像组成，分辨率为 40×144，数据由搭载着红外摄像机和测距雷达的汽车拍摄而成。因为红外摄像机与测距雷达的位置不完全相同，所以红外视频及其深度图像会有微小的视角上的差异，但这不影响本文实验。数据集分为 2 部分，由 3152 帧红外视频及其深度图像组成训练集，而余下的 992 帧数据则组成测试集。在训练集上训练本文算法，然后在测试集上测试结果，并与其他算法相比较。

4.2　训练过程

本文实验均是在装有单个 GTX 1070 显卡的计算机上使用深度学习库 PyTorch 进行训练和测试的。采用学习率为 0.0001 的 Adam[22]算法优化网络模型。Adam 算法具有 3 个优点：1)善于处理非平稳优化目标；2)对计算机的内存要求较小；3)可以对不同的参数计算自适应的学习率。实验中将连续的 8 帧红外视频作为一个序列数据输入给 BrCNN，采取批处理方式优化网络模型，每次输入 2 个数据优化网络。BrCNN 的输入数据是大小为(2,8,1,40,144)的张量，其中“2”表示每次输入的 2 个数据，“8”表示连续的 8 帧图像作为一个序列输入进网络，“1”表示红外图像是单通道图像，图像数据仅有一层，“40,144”表示红外视频的分辨率为 40×144。BrCNN 的输出数据是大小为(2,8,30,40,144)的张量，其中“30”表示将深度分成 30 个层级，其张量值表示深度属于对应层级的概率，取概率最大值的层级作为当前像素所在的深度层级，其他维度与输入数据表示的含义相同。模型在训练数据上共训练 500 个周期，即在训练集上训练 21900 次，取得了比较好的收敛效果。

4.3　结　果

图 4 分别展示一帧视频经过三个递归卷积层之后的输出结果。可以看出，不同层递归卷积层提取到了不同角度的特征。第一层递归卷积层主要提取了一些场景的背景信息，突出了不同层次、不同深度变化等信息；第二层递归卷积主要提取了更加综合的边缘、纹理等信息；第三层提取了更加抽象的综合信息。经过三层网络的特征提取，特征输入全卷积估计网络，估计出红外视频的深度。



(a)　　(b)　　(c)　　(d)

图 4　各个递归卷积层的输出。(a)红外视频；(b)第一个递归卷积层输出；(c)第二个递归卷积层输出；(d)第三个递归卷积层输出

Fig. 4　Output of different bi-recursive convolutional layers. (a) Infrared video; (b) output of the first bi-recursive convolutional layer; (c) output of the second bi-recursive convolutional layer; (d) output of the third bi-recursive convolutional layer

采用不同精度和误差度量方式评价本文算法和其他算法的估计结果。目前比较常用的精度和误差度量方法有平均相对误差(MRE)[1,3,7-8]、均方根误差(RMSE)[1,3,7-8]、对数误差(LE)[1,3,8]、不同阈值的精度(ζ)[6-8]，分别表示为

$$E_{\mathrm{mre}}=\frac{1}{M}\sum_{i=1}^{M}\frac{|d_i^*-d_i|}{d_i},\tag{11}$$

$$E_{\mathrm{rmse}}=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(d_i^*-d_i)^2},\tag{12}$$

$$E_{\mathrm{LE}}=\sum_{i=1}^{M}|\lg d_i^*-\lg d_i|,\tag{13}$$

$$\zeta=\max\left(\frac{d_i^*}{d_i},\frac{d_i}{d_i^*}\right)<1.25^{1,2,3},\tag{14}$$

式中 d_i^* 为估计的第 i 个像素的深度，d_i 为第 i 个像素的真实深度，M 为测试集中红外视频的总像素点个数。

表 1 给出了本文提出的 BrCNN 模型与其他一些经典模型的比较结果。可以看出，针对红外视频深度估计问题，BrCNN 相对于传统的 CNN(如 AlexNet、VGG16、Res34)能够获得比较好的估计结果。引入红外视频的序列信息可以提高红外视频深度估计的精度，降低误差。表 1 中表示精度的 3 栏数据越大代表估计结果越好，而表示误差的 3 栏数据越小则代表估计结果越好。

表 1　BrCNN 模型与其他经典网络模型的深度估计结果的对比

Table 1　Comparison of depth estimation results between BrCNN model and other classical network models

Method	Accuracy			Error		
	$\zeta<1.25$	$\zeta<1.25^2$	$\zeta<1.25^3$	MRE	RMSE	LE
BrCNN	0.762	0.901	0.956	0.214	10.201	0.083
AlexNet[12]	0.737	0.873	0.908	0.258	10.797	0.090
VGG16[13]	0.719	0.868	0.912	0.274	10.842	0.093
Res34[14]	0.721	0.870	0.921	0.275	10.782	0.089

图 5 选择 2 个场景下连续 4 帧视频估计的深度图像来比较不同算法的估计效果。可以看出，AlexNet、VGG16、Res34 等网络模型只能估计出总体的场景深度，对于一些细节特征，如道路上的行人、突然出现在场景中的杂物等，并不能精确地估计出来；而本文提出的 BrCNN 因加入了红外视频的序列特征，因此对于红外视频深度估计问题，能够估计出更多的细节信息，得到更好的估计结果。

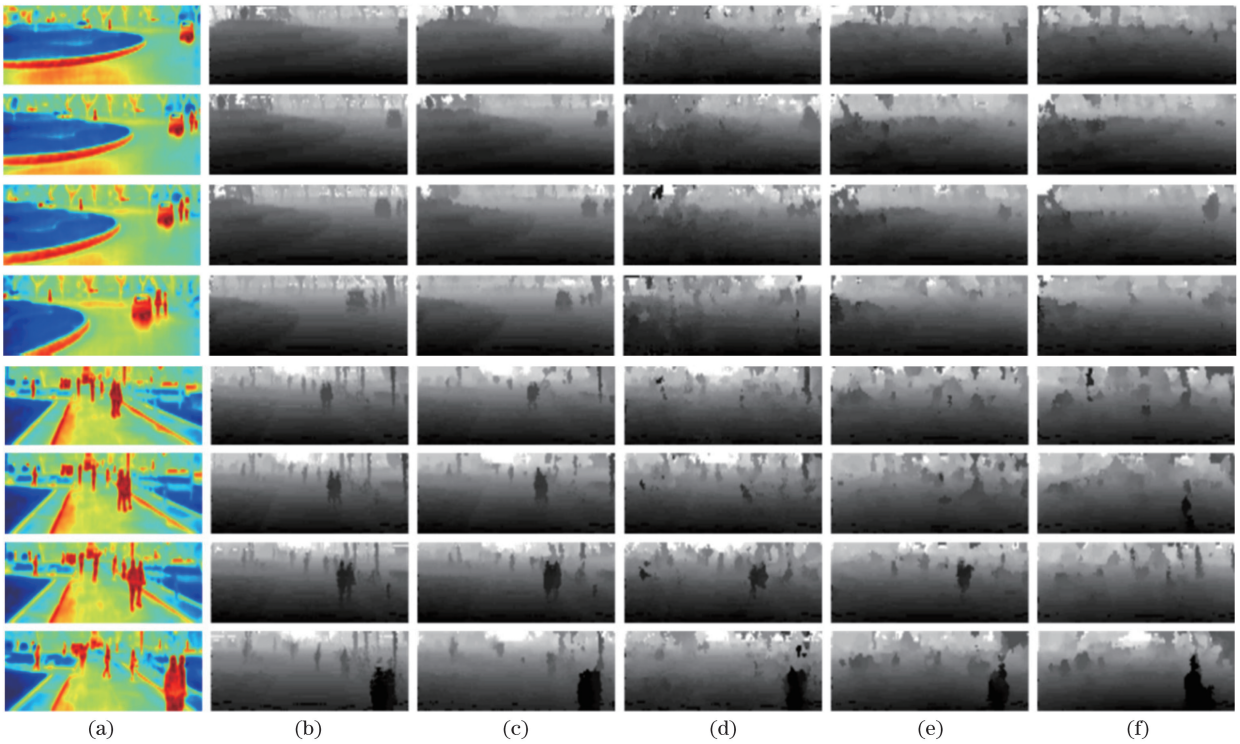


图 5　不同模型下的深度估计结果。(a)红外视频;(b)真实深度;(c) BrCNN 估计的深度;(d) AlexNet 估计的深度;(e) VGG16 估计的深度;(f) Res34 估计的深度

Fig. 5　Depth estimation results of different models. (a) Infrared video; (b) ground truth depth; (c) depth estimated by BrCNN; (d) depth estimated by AlexNet; (e) depth estimated by VGG16; (f) depth estimated by Res34

4.4　模型分析

BrCNN 在提取每一帧图像局部特征的同时加入了红外视频的序列信息,提高了特征的表达能力。但随着红外视频中场景的变化,序列信息对特征的表达能力起到了不同的作用。本文选取了一段连续的包含 470 帧图像的红外视频,计算得到任意相邻 2 帧图像差分的 2 范数,并度量了该段视频在 BrCNN 与 VGG16[13]模型下的 MRE,结果如图 6 所示。

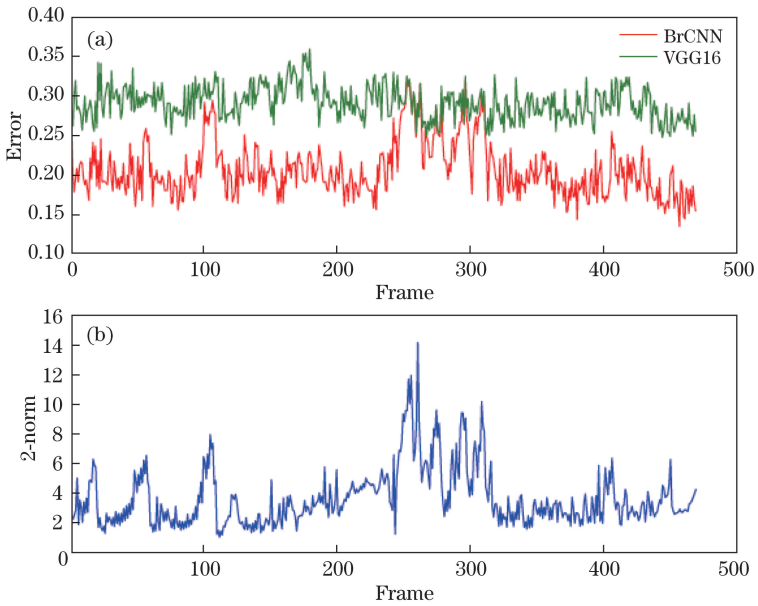
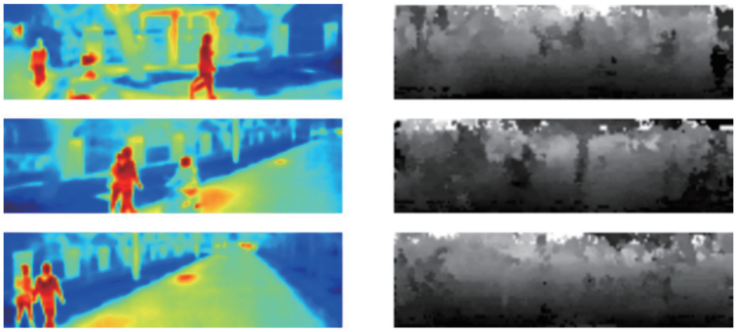


图 6　(a) BrCNN 与 VGG16 模型的平均相对误差;(b) 相邻 2 帧图像差分 2 范数

Fig. 6　(a) Mean relative error of BrCNN and VGG16 model; (b) 2-norm of difference between adjacent frames

从图 6 可以看出，当红外视频帧间差分的 2 范数较小时，BrCNN 的估计误差要小于 VGG16 模型，即 BrCNN 提高了红外视频深度的估计精度；但当帧间差分 2 范数较大的时候（如 250～300 帧之间），BrCNN 的估计误差与 VGG16 模型相似，此时红外视频的序列信息并不能提高特征的表达能力。

图 7 从帧间差分 2 范数较大的区间内取出几帧连续的图像，并使用 BrCNN 估计出其深度。可以看出，这是一段路口急转弯处，帧与帧之间的差异较大，此时视频的序列信息对于特征的提取并没有帮助，BrCNN 只能与传统的网络模型（如 VGG16 模型）取得精度类似的估计结果。



(a)　　(b)

图 7　帧间差分 2 范数较大时的深度估计结果。(a)连续 3 帧红外视频；(b) BrCNN 的估计结果

Fig. 7　Depth estimation results of larger 2-norm between frame differences.
(a) Continuous three frames of infrared video; (b) estimation results of BrCNN

5　结　　论

针对红外视频深度估计问题，在传统卷积网络能够很好地提取单帧图像特征的基础上，引入 RNN 有效提取序列特征的机制，得到的 BrCNN 能够有效地提取红外视频特征，并在深度估计中获得了较好的效果。该网络吸收了 CNN 优良的单幅图像特征提取能力，在递归传递视频序列信息机制下，对每一帧图像都能提取到包含红外视频上下文信息的特征。本文采用了双向递归机制，保证了每一个特征都蕴含视频整体序列信息，同时保持了视频序列信息的稳定性。实验中对比该网络与其他一些经典的网络结构的红外视频深度估计结果，表明该网络在红外视频深度估计上具有优秀的建模能力，能够有效解决红外视频深度估计问题。

参　考　文　献

[1] Karsch K, Liu C, Kang S B. Depth transfer: Depth extraction from video using non-parametric sampling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2144-2158.

[2] 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.

[3] Kong N, Black M J. Intrinsic depth: Improving depth transfer with intrinsic images[C]. IEEE International Conference on Computer Vision, 2015: 3514-3522.

[4] Saxena A, Chung S H, Ng A Y. 3D depth reconstruction from a single still image[J]. International Journal of Computer Vision, 2008, 76(1): 53-69.

[5] Xi Lin, Sun Shaoyuan, Li Linna, *et al*. Depth estimation from monocular infrared images based on SVM model[J]. Laser & Infrared, 2012, 42(11): 1311-1315.
席林, 孙韶媛, 李琳娜, 等. 基于 SVM 模型的单目红外图像深度估计[J]. 激光与红外, 2012, 42(11): 1311-1315.

[6] Xu Lu, Zhao Haitao, Sun Shaoyuan. 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.

[7] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]. Advances in Neural Information Processing Systems, 2014: 2366-2374.

[8] Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5162-5170.

[9] Zhang G F, Jia J Y, Hua W, *et al*. Robust bilayer segmentation and motion/depth estimation with a handheld

camera[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 603-617.

[10] Akhter I, Sheikh Y, Khan S, *et al*. Trajectory space: A dual representation for nonrigid structure from motion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(7): 1442-1456.

[11] Ha H, Im S, Park J, *et al*. High-quality depth from uncalibrated small motion clip[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 5413-5421.

[12] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems, 2012: 1097-1105.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. International Conference on Learning Representations, 2015:1-14.

[14] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[15] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[16] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[17] Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[18] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[19] Cho K, Van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. Computer Science, 2014.

[20] Chung J, Gülçehre C, Cho K, *et al*. Gated feedback recurrent neural networks[J]. Computer Science, 2015: 2067-2075.

[21] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[22] Kingma D, Ba J. Adam: A method for stochastic optimization[C]. 3rd International Conference for Learning Representations, 2015.