武汉大学学报 信息科学版
GEOMATICS AND INFORMATION SCIENCE OF WUHAN UNIVERSITY

局部物体块匹配的图像匹配算法

李钦, 游雄, 李科, 汤奋, 王玮琦

引用本文:
李钦, 游雄, 李科, 汤奋, 王玮琦. 局部物体块匹配的图像匹配算法[J]. 武汉大学学报・信息科学版, 2022, 47(3): 419-427.
LI Qin,YOU Xiong,LI Ke,TANG Fen,WANG Weiqi. Image Matching Based on Local Object Matching[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(3): 419-427.

DOI:10.13203/j.whugis20190364 文章编号:1671-8860(2022)03-0419-09

局部物体块匹配的图像匹配算法

李 钦[1] 游 雄[1] 李 科[1] 汤 奋[1] 王玮琦[1]

1 信息工程大学地理空间信息学院,河南 郑州,450002

摘 要:构建具有较强表达能力的图像特征是图像匹配应用的核心环节。训练孪生神经(Siamese)特征提取网络构建图像局部特征,通过图像局部特征的匹配解决整体图像匹配的问题。在图像匹配过程中,首先检测图像中包含的物体块,采用特征提取网络构建各物体块的特征表达,然后计算各物体块间的相似度,组成图像对相似矩阵,最后基于相似矩阵构建图像匹配模型。实验结果表明,所提网络构建的物体块特征可以有效匹配图像中相同物体块,区分不同物体块,且相比于已有方法,所提图像匹配算法具有更强的匹配性能,能够高效准确地解决图像匹配问题。

关键词:图像匹配;Siamese网络;特征提取网络;相似矩阵;匹配物体块

中图分类号:P237 **文献标志码**:A

图像是对真实复杂世界的映射成像,其本质上是由若干像素排列组合而成,图像特征则是对复杂的图像信息进行抽象、简化的表达[1],在此基础上才能进一步完成各种视觉任务,如图像检索[2]、图像识别[3]等。图像匹配算法一般通过构建整幅图像的特征表达计算图像相似度[4],根据图像相似度数值判断匹配结果。整体来讲,图像匹配算法可以分为两大类[5]:(1)在提取图像局部特征的基础上构建整幅图像的向量表达,如基于词汇带模型(bag of words, BoW)的图像表达方法[6]。(2)基于深度学习的图像匹配方法,该方法主要是利用孪生神经(Siamese)网络[7]进行图像匹配,网络输入为两张图像,输出即为图像匹配结果。文献[8]提出了一种端到端的基于Siamese网络的物体块匹配模型,即MatchNet。该网络将特征提取与相似度计算过程融合在一起,一定程度上限制了其在视觉任务中应用的灵活性。文献[9]基于Siamese网络提取物体块特征,通过挖掘一些硬样本来提高网络训练的效率。由于该Siamese网络仅包含3个卷积层,难以充分描述物体块中的非线性结构信息,因此,物体块特征的表达能力存在一定的局限。文献[10]基于深度卷积神经网络将初始物体块投影至特征空间,使得在特征空间中匹配物体块的距离较小,非匹配物体块的距离较大。文献[3, 11]直接利用Siamese网络进行整幅图像的匹配,基于深度卷积网络构建整幅图像的特征表达,该图像特征具有很强的泛化能力。

一般的图像匹配方法通过直接对整幅图像进行特征表达来完成图像匹配,由于一些匹配图像中存在大量的不匹配内容,这些内容不可避免地参与了图像的匹配计算,很大程度上影响了图像匹配结果。图1所示为一组匹配图像,红色方框内的区域为两张图像中的匹配物体块,红色方框外的大部分区域都是非匹配的无关内容,且这些无关内容占据了图像的主体,如果直接从整幅图像出发来判断图像是否匹配,这些占据图像主体的不匹配内容会大大干扰图像匹配的结果,一定程度上限制了匹配性能。而相较于整幅图像的匹配,单纯地进行物体块匹配会更加容易。

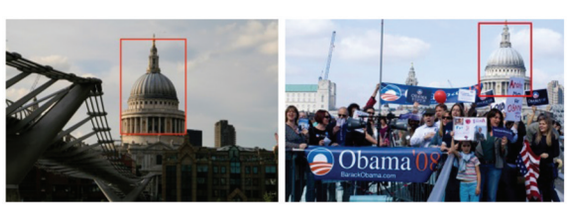


图1 匹配图像及其匹配物体块

Fig.1 Matched Images and Consistent Object Patches

因此,本文从局部物体块入手,基于Siamese

收稿日期:2019-09-29

项目资助:国家自然科学基金(41871322);河南省科技创新项目(142101510005)。

第一作者:李钦,博士生,主要从事深度学习与机器视觉研究。leequer20419@163.com

通讯作者:游雄,博士,教授。 youarexiong@163.com

特征提取网络构建物体块的本质表达，通过在图像中匹配相同的物体块判定整幅图像是否匹配。

1 局部物体块匹配

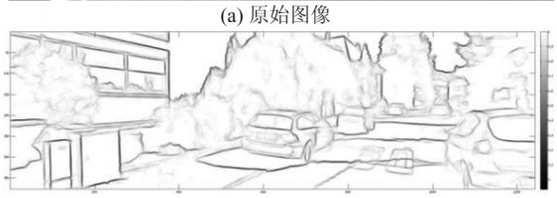
本文将整幅图像的匹配问题转化为图像中局部物体块的匹配，整体的图像匹配流程包括采用边缘盒算法[12]检测图像中包含的物体块、采用Siamese网络提取物体块深度特征、采用相互匹配机制检测图像对中的匹配物体块。其中，局部物体块深度特征提取是图像匹配的核心环节，本文采用对比机制对深度特征提取网络进行训练，训练结果使得匹配物体块间的特征距离较小，非匹配物体块间的特征距离较大。

1.1 物体块检测

本文采用边缘盒算法[12]检测图像中的物体块，具体过程为：

1)边缘图像生成。对图像进行边缘检测[13]，获取稠密的边缘图像，如图2所示，图2(a)为原始图像，图2(b)为其边缘图像。由于边缘图像比较紧密，利用非极大值抑制算法(non maximum suppression，NMS)对其进行稀疏化处理。

(a) 原始图像



(b) 边缘图像

图2 边缘图像示意图

Fig.2 Demonstration of the Edge Image

2)边缘组构建。在稀疏的边缘图像中，将近乎在一条直线上的边缘点组合起来构建边缘组，具体过程为：任意选取一个边缘点为起始点，不断寻找与其8连通的边缘点，累加两两边缘点之间的方向角度差值，直至累加值大于$\pi/2$，将寻找过的边缘点设置为一个边缘组，最终将整幅图像上所有的边缘点聚合成若干边缘组。

3)边缘组聚合。对于图像中生成的若干边缘组，计算相邻边缘组间的相似度，若相邻边缘组之间的均值夹角接近于边缘组的方向，说明这两个边缘组具有较高的相似度，属于同一个物体，对具有较高相似度的边缘组进行聚合。聚合后的边缘组包含了物体的外边缘信息，根据聚合后边缘组所在的区域范围生成对应物体的外包围盒。

在物体块检测的基础上，为了构建各物体块的特征表达，需要调整物体块尺寸(64×64像素)以满足特征提取网络的输入要求。为了减少物体块缩放对特征提取的影响，本文对物体块的原始尺寸进行约束，公式为：

$$64\times 64<W\times H<256\times 256 \tag{1}$$

$$0.5<W/H<2.0 \tag{2}$$

式中，W、H分别为物体块的宽度与高度。

通过控制物体块的长度与宽度，剔除尺寸过小或者过大，以及长宽差异较大的物体块。图3所示为检测到的物体块外包围盒示意图，物体块尺寸适中、形状方正，在缩放过程中产生的形变也相对较少，为构建具有较强表达能力的物体块特征提供了良好的条件。

图3 检测到的物体块外包围盒

Fig. 3 Bounding Boxes of the Detected Objects

1.2 Siamese特征提取网络

物体块特征旨在表达区域物体的本质性的不变信息，其不因物体在图像中呈现形态的变化而变化，同时不同物体的特征又存在着本质的区别。物体特征这种既能维持自身稳定，又能与其他物体进行有效区分的能力，也叫做特征的表达能力[14]。

本文利用Siamese网络提取物体块深度特征，基于公共数据集训练特征提取网络，网络输入为局部物体块，输出为构建的物体块特征。训练过程中，同时输入两张物体块，通过完全相同、权值共享的特征提取网络生成各自特征描述符；结合物体块标签构建网络误差函数，使得匹配物体块间的特征距离尽可能小，非匹配物体块间的特征距离尽可能大。

1.2.1 Siamese特征提取网络结构

本文Siamese网络结构如图4所示，其中网络输入为两组固定尺寸(64×64像素)的灰度物体块，每组物体块数量为n，D_1、D_2分别为两组物体块通过特征提取网络构建的特征向量。另外，根

据两张物体块的匹配真值确定物体块相似度标签(若物体块匹配,则标签为1,反之为0),训练过程中基于两组物体块特征向量与真值标签构建误差函数。

图4中Siamese网络包含两个完全相同的通道,每个通道是一个包含7个卷积层、2个全连接层的深度卷积网络,其网络各层参数如表1所示。

卷积操作(Conv)是特征提取网络的核心,通过对输入图像的层层卷积可以获得不同类型、不同层次的特征表达,这些特征可以充分反映图像中隐含的抽象语义信息。

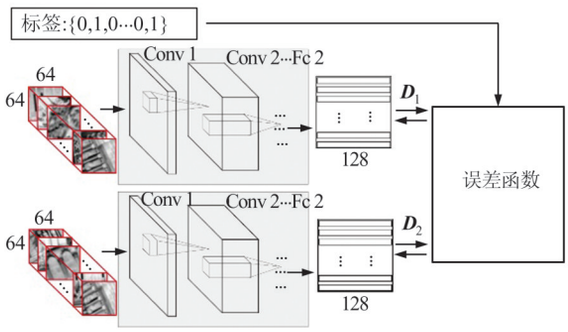


图4　Siamese网络结构

Fig.4　Architecture of the Siamese Network

表1　特征提取网络中的各层参数

Tab.1　Layer Parameters of the Feature Extraction Network

各层名称		输入特征图 尺寸/像素,数量	卷积核尺寸(长×宽×通道数)/数量/步长	反余弦函数/纠正线性单元+池化(采样步长)	输出特征图 尺寸/像素,数量
卷积层	Conv 1	64×64,1	3×3×1/32/1	T/R+MaxPool(2)	32×32,32
	Conv 2	32×32,32	3×3×32/64/1	T/R+MaxPool(2)	16×16,64
	Conv 3	16×16,64	3×3×64/64/1	T/R	16×16,64
	Conv 4	16×16,64	3×3×64/128/1	T/R	16×16,128
	Conv 5	16×16,128	3×3×128/128/1	T/R+MaxPool(2)	8×8,128
	Conv 6	8×8,128	3×3×128/256/1	T/R	8×8,256
	Conv 7	8×8,256	3×3×256/256/1	T/R+MaxPool(2)	4×4,256
全连接层	Fc 1	4 094(4×4×256)		T/R	1 024
	Fc 2	1 024		T/R + l2_norm	128

非线性操作(T/R)将卷积输出的特征值限缩到特定的区间,本文分别利用反余弦函数与纠正线性单元(rectified linear unit, ReLU)构建特征提取网络,探索各自模型构建特征的表达能力。

部分卷积层(Conv 1、Conv 2、Conv 5、Conv 7)包含降采样操作(MaxPool),该操作使得特征图尺寸不断缩小,降采样的结果使得输出特征图上的每个像素覆盖了原始图像上更大区域,这实质上是对图像空间区域结构的高效整合。

通过对输入图像进行层层卷积操作最终生成256张4×4像素的特征图,对输出特征图进行拉直组合生成4 096维的物体块初始描述向量。为了进一步对特征向量进行精炼简化,构建两层的全连接网络对描述符进行降维,生成128维的物体块特征向量,对第二全连接层的输出进行归一化操作(l2_norm),使所得的特征描述符模长为1。

1.2.2　误差函数

本文采用特征向量间的夹角余弦作为物体块相似度,由于特征提取网络输出的特征向量模长为1,特征向量点乘结果即为其夹角余弦,计算公式为:

$$S=\cos<\boldsymbol{D}_1,\boldsymbol{D}_2>=\boldsymbol{D}_1\cdot\boldsymbol{D}_2{}^{\mathrm{T}} \tag{3}$$

式中,$\boldsymbol{D}_1$、$\boldsymbol{D}_2$即为特征提取网络输出的特征向量;S为特征向量间的夹角余弦,余弦值越大,表明描述向量间的夹角越小,物体块越相似。

根据特征提取网络中非线性函数的不同,本文采用不同的误差函数训练网络构建各自模型,包括:

1)模型T。采用反余弦函数对各层卷积输出进行非线性处理,将网络各层输出限缩至[−1,1],此时,物体块相似度的取值范围为$S\in[-1,1]$,构建误差函数为:

$$L=-\frac{1}{2N}\sum_{i=1}^{N}\left(L_i\cdot\left\|1-S_i\right\|_2^2+(1-L_i)\cdot\left\|1+S_i\right\|_2^2\right) \tag{4}$$

式中,S_i为第i组物体块相似度(即特征向量夹角

余弦)；L_i为第i组物体块的训练标签(物体块匹配，标签为1，反之为0)；N为训练过程中每个批次的样本数量。

通过不断迭代训练使得匹配物体块间的特征距离尽可能地趋于1，而不匹配物体块间的相似度尽可能地趋于−1。

2)模型R。基于ReLU函数将各卷积层输出的负值归零，此时输出的特征向量各维度均为正数，物体块相似度$S\in[0,1]$，其阈值范围与网络标签相对应，因此基于交叉熵构建误差函数L为：

$$L=-\frac{1}{2N}\sum_{i=1}^{N}(L_i\cdot\log(S_i)+(1-L_i)\log(1-S_i)) \tag{5}$$

采用模型R进行网络训练，使得匹配物体块间的特征距离趋于1，而不匹配物体块间的相似度趋于0。

1.3 基于图像对相似矩阵的物体块匹配

对于图像对中的每一张图像，在物体块检测的基础上，调整各物体块尺寸至64×64像素，利用特征提取网络构建各物体块特征，每张图像上物体块特征向量堆叠而成的特征矩阵标记为：$F_1\in\mathbf{R}^{M\times 128}$，$F_2\in\mathbf{R}^{N\times 128}$，其中$M$、$N$分别为两张图像上检测到的物体块数目，计算两张图像中物体块间的相似度，构建相似矩阵为：

$$\boldsymbol{S}=F_1\cdot F_2{}^{\mathrm{T}} \tag{6}$$

式中，$F_2{}^{\mathrm{T}}$为F_2的转置；相似矩阵$\boldsymbol{S}\in\mathbf{R}^{M\times N}$，$\boldsymbol{S}$中的任意一个元素$s_{ij}$代表第一张图像中第$i$个物体块与第二张图像中第$j$个物体块间的相似度。由于余弦函数在0附近值域变化较小，为了让物体块间的相似度差异更加明显，将相似矩阵变换为$\boldsymbol{S}_A=\arccos\boldsymbol{S}$，这样物体块间的相似度变为其向量间的夹角，向量夹角越小，物体块越相似。

$\boldsymbol{S}_A$中的第i行元素集合为：$R_i=\{s_{ij},j=1,2\cdots N\}$，第$j$列元素集合为：$C_j=\{s_{ij},i=1,2\cdots M\}$，设$T_M$为特征向量间的夹角匹配阈值，对于$\boldsymbol{S}_A$中的任意一个元素$s_{pq}$，若满足以下条件：

$$s_{pq}<T_M \tag{7}$$

$$s_{pq}=\min C_q \tag{8}$$

$$s_{pq}=\min R_p \tag{9}$$

即s_{pq}小于该匹配阈值，且在其所在行列上均是最小值，则其对应物体块(第一张图像上的第p个物体块与第二张图像上的第q个物体块)相互匹配。若两张图像中存在匹配物体块，表明其包含一致性的内容，然而由于物体块存在误匹配的情况，非匹配图像中也可能检测到较少的匹配物体块，简单地根据图像中是否存在匹配物体块判定图像匹配结果会存在大量错误匹配的情况。一般来讲，两幅图像中包含的匹配物体块数目越多，判定图像匹配的置信度越高。在图像匹配实践中，设置匹配物体块数目阈值，当两张图像中的匹配物体块数目大于该阈值时，则判定图像匹配。

2 特征提取网络与图像匹配实验

为了检验本文物体块特征的表达能力，利用训练好的特征提取网络构建测试物体块特征，计算物体块相似度，分析测试结果。另外，为了验证本文图像匹配算法在图像匹配实践中的可行性，构建测试数据集，利用本文方法进行图像匹配实验，同时对比已有方法，分析本文算法的图像匹配性能。

本文实验平台为64位的Ubuntu 16.04 LTS，16 GB内存，Xeon 3.2 GHz处理器，实验代码基于Python 2.7编写，相关深度学习模型的训练是基于TensorFlow1.1.0框架，图形处理器为Nvidia Titan XP。

2.1 Siamese特征提取网络训练实验

为了训练特征提取网络，本文采用多视角立体数据集(multi-view stereo dataset，MVS)，该数据集包含1.5×10^6张尺寸为64×64像素的灰度物体块与500 kB个空间点，每张物体块都是从不同视角观测某个空间点获取的。任意两张物体块即可组成一组训练样本，若两张物体块观测的是相同的空间点，即为匹配物体块(正样本)，反之为不匹配物体块(负样本)。

数据集包含3组场景数据：自由神像(Statue of Liberty，LY)、巴黎圣母院(Notre Dame，ND)、约塞米蒂半圆体(half dome in Yosemite，YO)。分别从3组数据中选取20万组训练样本(正负样本各10万组)与1万组测试样本(正负样本各5 000组)，且测试样本不包含在训练样本中，对3组训练样本数据进行随机组合作为训练数据，包括：(1)LY+ND(测试集为YO)；(2)LY+YO(测试集为ND)；(3)YO+ND(测试集为LY)；(4)LY+ND+YO(测试集为LY+ND+YO)。

训练过程中，将所有的训练数据遍历51次，每次遍历分为1 000个批次，按照每批次400组样本(正负样本各200组，对于训练数据集LY+YO+ND，每批次包括600组样本)输入网络。采用随机梯度下降(stochastic gradient descent，

SGD)对误差函数进行优化。

分别对模型 R 与模型 T 进行训练,选取误差最小时的网络模型进行保存,所得模型的输入为原始灰度物体块,输出为模长为 1 的 128 维特征向量。

图 5 所示为基于训练后模型计算的 2 000 组测试样本相似度与样本初始相似度对比图,其中 2 000 组样本是从 ND 测试数据集中随机选取的(正负样本各 1 000 组),模型 T 与模型 R 为基于 LY+YO 训练数据集训练所得。图 5(a)中训练前正负样本的初始相似度数值并没有明显的区分,而图 5(b)、5(c)中正样本相似度数值明显高于负样本,表明基于训练后的两种模型均能使匹配物体块间的距离变小(相似度数值较大),而非匹配物体块间的距离变大(相似度数值较小)。

为了分析本文模型相较于已有模型的优劣,将其与已有的物体块匹配网络 CNN3[9]、CNN7[10]、MatchNet[8]进行对比,在相同的训练数据上构建物体块匹配模型。为了量化评估各自模型特征的表达能力,构建测试样本召回率-准确率(P-R)曲线,计算模型 P-R 曲线的曲线下面积(area under the curve, AUC),表 2 汇总了相应曲线的 AUC 值。

相比于网络 CNN3 与 CNN7,本文模型(模型 R 与模型 T)在所有测试集上均取得更好的匹配性能,表明本文特征具有更强的表达能力。MatchNet 通过训练端到端的物体块匹配模型,有效地将物体块特征提取与相似度计算过程融合在一起,使得其模型匹配性能优于模型 R。

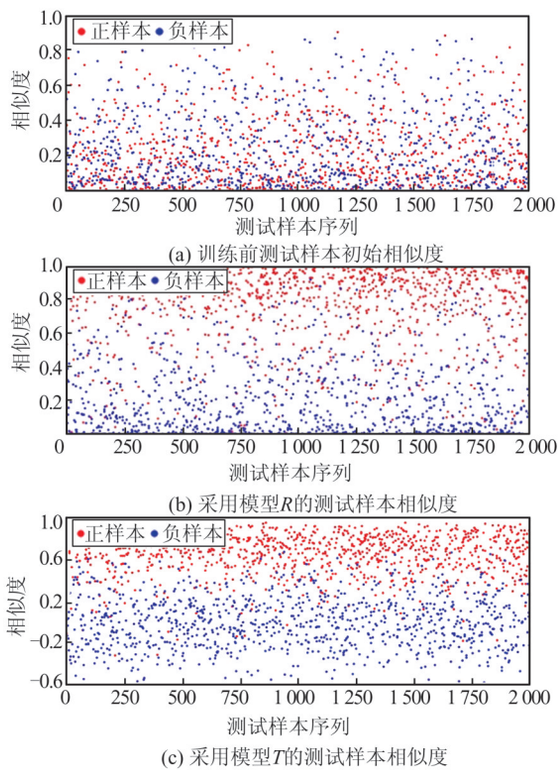


图 5　基于训练后模型的测试样本相似度与样本初始相似度对比图

Fig.5　Comparison Between Initial Sample Similarities and Test Sample Similarities Based on Well Trained Model

表 2　模型 AUC 值对比表

Tab.2　Comparison of AUC

训练集	测试集	AUC 值				
		CNN3[9]模型	CNN7[10]模型	MatchNet[8]模型	模型 T	模型 R
LY+ND	YO	0.787 1	0.756 5	0.947 4	0.953 8	0.913 0
LY+YO	ND	0.838 5	0.852 7	0.961 4	0.969 5	0.944 2
YO+ND	LY	0.862 6	0.815 8	0.954 7	0.959 4	0.933 2
LY+YO+ND	LY+YO+ND	0.830 9	0.841 4	0.971 5	0.972 0	0.944 9

模型 T 的特征匹配性能明显优于模型 R,表明基于反余弦函数构建特征提取网络优于 ReLU 函数,这是因为本文基于余弦距离度量物体块相似度,余弦距离的自然区间为[−1,1],采用 ReLU 函数进行非线性操作将特征值中的负值归零,使得物体块相似度的取值区间限缩为[0,1],而反余弦函数顺应了物体块相似度应有的取值规律,更大的相似度变化区间为区分不同特征提供了充分的差距范围,因此相应网络取得了最佳的匹配性能。

实验结果表明,Siamese 网络(模型 T)构建的物体块特征具有更强的表达能力,在不同的测试数据上均取得了更好的匹配效果,其在有效地匹配正样本的同时,也可以较好地区分负样本。

2.2　图像匹配实验

为了验证本文图像匹配算法的可行性,以及相比于已有算法的优势,设计对比实验,分别采用基于 BoW 模型的方法、基于 MatchNet 模型的方法与本文方法进行图像匹配实验。

本文利用 KITTI 数据集[15]的测程(Odometry)数据中的第一组图像序列(编号 00)构建测试图像数据集,该序列包含 4 541 张尺寸为 1 240×

376像素的图像。序列中的任意两张图像即可组成一组测试样本，结合图像序列的位姿信息，若两张图像拍摄点位置小于5 m且拍摄方向夹角小于π/6，则其为匹配样本（正样本），若拍摄点位置大于100 m，则其为非匹配样本（负样本）。随机选取正、负样本各500组，分别利用不同方法进行图像匹配实验。

1）基于BoW模型的图像匹配方法。以图像序列中的4 541张图像作为训练数据，构建训练数据中的尺度不变特征变换（scale invariant feature transform，SIFT）特征词典，词汇数目设为1 000，进而构建基于BoW模型的图像描述符，图像描述符点乘结果即为图像间的相似度数值（描述符模长为1）。

2）基于MatchNet模型检测匹配物体块的方法。对于测试样本，在物体块检测的基础上，基于MatchNet模型计算两两物体块间的相似度，采用相互匹配机制检测图像对中的匹配物体块，统计匹配物体块数目。

3）本文方法。对于每一组测试样本，结合本文算法检测每幅图像中的物体块；基于模型T（训练集为LY+ND+YO）提取各物体块特征；计算各物体块间相似度，组成相似矩阵，基于相互匹配机制检测样本中的匹配物体块，记录样本包含的匹配物体块数目。

图6所示为3种方法在测试样本中进行图像匹配的实验结果图。图6(a)表明基于BoW模型的图像匹配方法使得正样本整体上具有更高的相似度。基于MatchNet模型的方法与本文方法根据样本中的匹配局部物体块度量相似度，匹配物体块数目越多，相似度越高。图6(b)中虽然正样本整体上包含更多的匹配物体块，但是大部分负样本中依然包含一些匹配物体块，虽然MatchNet网络在物体块匹配实验中取得了很好的效果，但是在本文图像匹配实验数据中却存在很多误匹配的情况，表明MatchNet网络泛化能力较弱，针对实验样本中的街道场景，匹配性能有所下降。图6(c)中大部分负样本的匹配物体块数目为0，表明相比于MatchNet模型，本文特征提取网络具有更强的泛化能力，构建的物体块特征具有较强的辨识能力，可以有效区分不匹配的物体块。

另外，图6(c)中仍有部分负样本包含较少的匹配物体块，这是由于现实世界的复杂性，实验样本中的场景图像包含各种各样的物体，即使是非匹配图像对依然存在一些相似的物体块，如统一规划建设的居民楼、风格相似的窗户等，这些最终导致了物体块的误匹配。

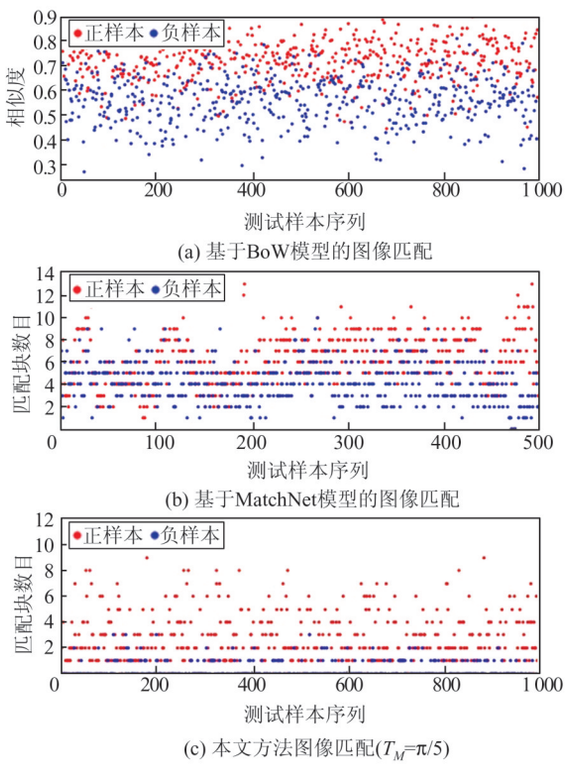


图6 3种方法实验结果图

Fig.6 Matching Results of 3 Methods

图7所示为利用本文方法在一组正样本中获取的匹配物体块结果，两张图像中黄色与绿色对应区域为检测到的匹配物体块，表明本文算法准确匹配到了两张图像中包含一致性内容的物体块。

图7 一组正样本与其匹配物体块

Fig.7 A Positive Sample and the Matched Objects

对比分析各种方法的匹配结果，统计各方法在不同相似度阈值条件下样本的召回率与准确率，绘制P-R曲线对比图，如图8所示。

由于匹配物体块数目阈值为非负整数，因此基于MatchNet模型的方法与本文方法的P-R曲

线上只有部分点有意义(星号标记的离散点)。图8表明,$T_M=\pi/4$时,本文算法的图像匹配效果较差,这是由于夹角匹配阈值较低,本文方法将部分负样本中视觉上相似的物体块判定为匹配物体块;当$T_M=\pi/5$、$\pi/6$时,本文方法的图像匹配结果明显优于BoW模型与MatchNet模型,表明本文通过匹配图像中的局部物体块,有效消除了图像中无关内容的干扰,取得了更好的匹配效果。

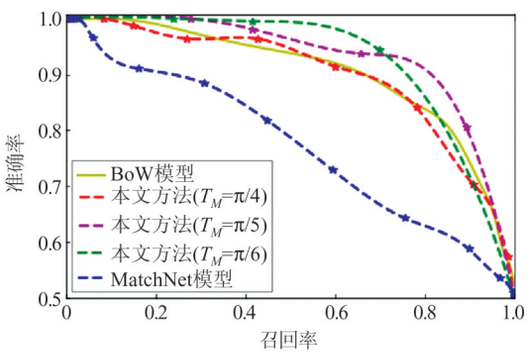


图8　P-R曲线对比图

Fig.8　Comparison of the P-R Curves

在一些动态匹配应用中,匹配过程需要实时快速地完成,如视觉导航[16]中需要实时根据图像匹配结果进行定位,在即时定位与地图构建(simultaneous localization and mapping,SLAM)中的闭环检测[17-18]应用中,需要实时地对当前图像与过往图像进行匹配,判断当前位置是否形成环路,进而进行全局优化。因此,匹配效率是评价图像匹配性能的又一重要指标,统计各方法在测试样本中的平均匹配耗时,如表3所示。表3表明本文方法的平均匹配耗时最低,每分钟可以处理约30对图像,达到了很高的匹配效率。

表3　各种方法平均匹配耗时对比/ms

Tab.3　Consumed Time for the Comparison Methods/ms

方法	平均匹配耗时
Bow模型	256.9
MatchNet模型	81.8
本文方法	31.8

MatchNet网络将特征提取与相似度计算融合到一起,极大限制了其应用的灵活性,在图像匹配应用中,物体块需要两两组合作为网络输入,各物体块的特征提取过程被重复多次运行,大大提高了其匹配耗时。不同于MatchNet模型,本文方法将物体块特征提取与相似度计算过程分开,每个物体块的特征提取过程仅运算一次,且多个物体块间的相似度计算可以转换为矩阵点乘运算一次完成,提高了匹配效率,为实时准确的图像匹配提供了良好的基础。

实验结果表明,本文通过构建具有较强表达能力的局部物体块特征,匹配图像中包含一致性内容的物体块,进而完成整幅图像的匹配,在测试样本中可以取得较好的匹配结果,且匹配效率极高,基本满足实时准确图像匹配应用的要求。

3　结　语

本文针对图像匹配中存在大量无关内容干扰的问题,将图像匹配问题转化为局部物体块的匹配,首先利用边缘盒算法在图像中检测物体块;其次基于Siamese特征提取网络构建物体块的特征表达;进而计算各物体块相似度,生成图像对相似矩阵,通过分析相似矩阵确定图像中的匹配物体块;最终根据图像中包含的匹配物体块数目判断图像是否匹配。Siamese特征提取网络构建的物体块特征具有较强的表达能力,该特征可以有效匹配图像中相同物体块、区分不同物体块,同时本文图像匹配算法相比于已有方法具有更好的匹配性能,可以高效准确地匹配包含一致性内容的图像对,区分不匹配的图像对。

参　考　文　献

[1] Li Qin, You Xiong, Li Ke, et al. Deep Hierarchical Feature Extraction Algorithm [J]. *Pattern Recognition and Artificial Intelligence*, 2017, 30(2): 127-136 (李钦, 游雄, 李科, 等. 图像深度层次特征提取算法[J]. 模式识别与人工智能, 2017, 30(2): 127-136)

[2] Xu Xuemei, Zhou Lichao, Yang Bingchu, et al. CIFO: A Retrieval Method for Color Images with Salient Object [J]. *Geomatics and Information Science of Wuhan University*, 2015, 40(1): 53-58 (许雪梅, 周立超, 杨兵初, 等. CIFO:针对显著对象的彩色图像检索方法[J]. 武汉大学学报·信息科学版, 2015, 40(1): 53-58)

[3] Yang Yubin, Lin Hui. Outliers Detection of Multi-beam Data Based on Bayes Estimation [J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(2): 87-92 (杨育彬, 林珲. 利用天文观测图像对空间碎片目标进行自动识别与追踪[J]. 武汉大学学报·信息科学版, 2010, 35(2): 87-92)

[4] Roy K, Mukherjee J. Image Similarity Measure Using Color Histogram, Color Coherence Vector, and So-

bel Method[J]. *International Journal of Science and Research*, 2013, 2(1): 538-543

[5] Melekhov I, Kannala J, Rahtu E. Siamese Network Features for Image Matching [C]//International Conference on Pattern Recognition, Cancun, Mexico, 2017

[6] Sivic J, Zisserman A. Video Google: A Text Retrieval Approach to Object Matching in Videos[C]// International Conference on Computer Vision, Nice, France, 2003

[7] Chopra S, Hadsell R, LeCun Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification[C]//IEEE International Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005

[8] Han N X, Leung T, Jia Y, et al. MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching [C]// IEEE International Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015

[9] Simo-Serra E, Trulls E, Ferraz L, et al. Discriminative Learning of Deep Convolutional Feature Point Descriptors [C]//International Conference on Computer Vision, Santiago, Chile, 2015

[10] Melekhov I, Kannala J, Rahtu E. Image Patch Matching Using Convolutional Descriptors with Euclidean Distance[C]//Asian Conference on Computer Vision, Springer, Cham, 2016

[11] Lin T Y, Cui N Y, Belongie S, et al. Learning Deep Representations for Ground-to-Aerial Geolocalization [C]//IEEE International Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015

[12] Zitnick C L, Dollar P. Edge Boxes: Locating Object Proposals from Edges[C]//European Conference on Computer Vision, Springer, Cham, 2014.

[13] Piotr Dollár, Zitnick C L . Structured Forests for Fast Edge Detection[C]//International Conference on Computer Vision, Sydney, Australia, 2013

[14] Li Qin, Li Ke, You Xiong, et al. Place Recognition Based on Deep Feature and Adaptive Weighting of Similarity Matrix [J]. *Neurocomputing*, 2016, 199(C): 114-127

[15] Geiger A , Lenz P , Urtasun R . Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite [C]// IEEE International Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012

[16] Dudek G, Jugessur D. Robust Place Recognition Using Local Appearance Based Methods[C]//IEEE International Conference on Robotics and Automation, San Francisco, USA, 2000

[17] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System[J]. *IEEE Transactions on Robotics*, 2017, 31(5): 1147-1163

[18] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-Time Single Camera SLAM [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6): 1052-1067

Image Matching Based on Local Object Matching

LI Qin[1] *YOU Xiong*[1] *LI Ke*[1] *TANG Fen*[1] *WANG Weiqi*[1]

1 Institute of Geospatial Information, Information Engineering University, Zhengzhou 450002, China

Abstract: Objectives: Image matching is a basic step in many vision applications. Aiming at the challenging problem in image matching that the consistent contents between a matched image pair generally occupy much less regions, we convert the whole image matching to local object matching. And the Siamese structure based feature extraction network is employed to produce the discriminative features for local objects. The proposed feature could effectively match the consistent objects between image pairs, and further complete the whole image matching task. **Methods**: We solve the image matching task by matching the consistent objects within image pairs. The local object patches are firstly detected based on edge boxes algorithm, and the proper objects, which satisfy the inputs of the feature extraction network, are selected according to the size of the detected object patches. Then the feature extraction network is constructed based on the Siamese structure, and the network is trained based on comparison mechanism. The training process makes the feature distances of the consistent objects close with each other, and those of the inconsistent objects far

from each other. Thus, the object features could effectively match the consistent objects and distinguish the inconsistent. Finally the object distances are calculated to constitute the similarity matrix, and the consistent objects are detected based on the mutual matching mechanism, the image matching task is ultimately completed according to the number of the consistent objects. **Results**: We predict whether or not two images are matched by detecting the consistent objects between the image pair, and the core of the whole method is to construct the discriminative features for local objects. The experimental results demonstrate that the proposed Siamese structure based feature extraction network is capable of producing the high representative features, which could effectively match the consistent objects and distinguish the inconsistent objects. Comparing with the existing networks, the proposed feature extraction network could achieve better matching performance on the test datasets. In the image matching experiments, the proposed method could outperform the other approaches. In addition, the proposed method only describes the critical objects within images, which greatly reduces the data volume. Thus, it could achieve high efficiency. **Conclusions**: The core of image matching is to decide whether two images contain the consistent objects, and the background contents in two images are useless in image matching practice. However, the existing methods generally describe the overall image as a whole, and all the image contents need to be process, which increases the data volume and limits the matching performance. Instead of representing the whole image directly, we convert the image matching to local object matching. If there exist several consistent objects between an image pair, the images can be predicted as matched. Only the critical objects within images are described, the irrelevant contents, out of the object regions, is not involved in the object matching, which actually elements the impact of the inconsistent contents, and make the object matching more accurate. As the useless contents are not described, the computation cost is greatly reduced, and the high matching efficiency could be achieved. The experimental results indicate that the local object based method could effectively solve the image matching task with high efficiency.

Key words: image matching; Siamese network; feature extraction network; similarity matrix; corresponding object pairs

First author: LI Qin, PhD candidate, specializes in deep learning and computer vision. E-mail: leequer20419@163.com

Corresponding author: YOU Xiong, PhD, professor. E-mail: youarexiong@163.com

Foundation support: The National Natural Science Foundation of China (41871322); the Project of Science and Technology Innovation of Henan Province (142101510005).

引文格式: LI Qin, YOU Xiong, LI Ke, et al. Image Matching Based on Local Object Matching[J]. Geomatics and Information Science of Wuhan University, 2022, 47(3): 419-427. DOI: 10.13203/j.whugis20190364(李钦, 游雄, 李科, 等. 局部物体块匹配的图像匹配算法[J]. 武汉大学学报·信息科学版, 2022, 47(3): 419-427. DOI: 10.13203/j.whugis20190364)