基于BI-LSTM-CRF模型的中文分词法

张子睿，刘云清
（长春理工大学 电子信息工程学院，长春 130022）

摘 要：递归神经网络能够很好地处理序列标记问题，已被广泛应用到自然语言处理（NLP）任务中。提出了一种基于长短期记忆（LSTM）神经网络改进的双向长短期记忆条件随机场（BI-LSTM-CRF）模型，不仅保留了LSTM能够利用上下文信息的特性，同时能够通过CRF层考虑输出标签之间前后的依赖关系。利用该分词模型，通过加入预训练的字嵌入向量，以及使用不同词位标注集在Bakeoff2005数据集上进行的分词实验，结果表明：BI-LSTM-CRF模型比LSTM和双向LSTM模型具有更好的分词性能，同时具有很好地泛化能力；相比四词位，采用六词位标注集的神经网络模型能够取得更好的分词性能。

关键词：中文分词；BI-LSTM-CRF；词位标注
中图分类号： TP391 **文献标识码**：A **文章编号**：1672-9870(2017)04-0087-06

Chinese Word Segmentation Based on Bi-directional LSTM-CRF Model

ZHANG Zirui, LIU Yunqing
(School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022)

Abstract: Recurrent neural network had been broadly applied to natural language processing (NLP) problems, because they deal well with the problem of sequence labeling. In this paper, we propose to use bidirectional LSTM CRF (BI-LSTM-CRF) model for Chinese word segmentation, which is based on long short-term memory (LSTM) units. This model not only can keep the contextual information in both directions, but also through the CRF layer to consider the dependency between the output tag. By using different tag set and adding pre-trained character embeddings, and using the model in the Bakeoff2005 data set on the word segmentation experiment results show that: BI-LSTM-CRF model has better segmentation performance than LSTM and bidirectional LSTM model, and has good generalization ability; Compared with the four-tag-set, the neural network model with the six-tag-set can achieve better segmentation performance.

Key words: Chinese word segment; BI-LSTM-CRF; tag set

由于中文写法的特性，决定了词与词之间没有类似空格的显式标志来进行分割，因此中文分词问题就成了中文自然语言处理中面临的首要基础性工作。近些年，特别是从国际中文分词评测活动Bakeoff开展以来，中文自动分词技术发生了重大的变化和进步[1]。中文自动分词的研究方法主要分为三种：基于词表的方法；基于传统统计模型的方法；基于深度学习神经网络的方法。

基于词表的分词方法基本上是20世纪80年代或者更早一些时候提出来的，其中刘源[2]做出了一些总结性的工作，介绍了包括正向最大匹配法，逆向最大匹配法，双向扫描法等16种不同的基于词表的分词方法。而基于词表的分词方法由于对词表的依赖性很大，在针对中文分词难点中的命名实体识别，未登陆词识别方面表现很差。同时词表的构建受到相关针对领域的限制，适应性较差。

基于传统统计模型的分词方法，自从Bakeoff比赛开展后，就出现了大量相关工作的论文。最常用

收稿日期：2017-05-26
基金项目：吉林省科技攻关项目（No.20160204003GX）
作者简介：张子睿（1993-），男，硕士研究生，E-mail：zzirui@163.com
通讯作者：刘云清（1970-），男，博士，教授，Email：mzliuyunqing@163.com

的方法是将中文分词问题看作是序列标注问题，如Xue等人[3]使用最大熵算法实现由字构词模型，将汉字标注成为4位标注集(B,M,E,S)其中之一，然后利用标注规则进行分词。Peng等人[4]通过构建线性链条件随机场(CRF)来进行分词。Tseng等人[5]在文献[4]的基础上利用条件随机场和字标注的方法构建分词系统取得了比较不错的分词效果。Stanford最初发布的汉语分词工具亦是基于文献[5]的成果实现。之后，同样是利用条件随机场模型但是却是构建6词位标注(B,B_1,B_2,M,E,S)和6特征模板(TMPT－6)的分词系统被黄昌宁等人[6]提出。而基于传统统计模型的分词方法无论是最大熵，条件随机场，还是它们的联合模型[7]等，其性能都受到了特征设定的局限。在实际使用中往往出现特征过多的情况使得模型过拟合。

基于深度学习神经网络的分词方法，能够从原始数据中自主学习，从而找到更深层次，更加抽象的特征。由于这些特征不是人为设定，这样使它规避了传统统计模型分词方法的局限所在，因此能够在自然语言处理任务中获得更好的表现。Collobert等人[8]就是利用神经网络这一优势，抛弃了传统的人工特征设定及提取的方式，从而在英文语句标注方面获得了不错的成绩。Zheng 等人[9]按照文献[8]中提出的思路，首次将神经网络运用到中文分词任务上，并且用感知器算法替换了原本神经网络的训练算法，加速了整个训练过程。但文献[9]只是简单的通过设置窗口来获得上下文向量，针对这一点文献[10]增加了标签的嵌入和张量转换层，想以此加大上下文特征之间的关联，但是所取得的效果略逊于传统方法。之后，Chen 等人[11-12]先后提出了GRNN(gated recursive neural network)模型和基于GRNN模型改进的具有上下文记忆单元的LSTM(long short-term memory)模型来进行中文分词，取得了与传统统计模型分词方法相当的成绩。而LSTM模型只能对记住过去的上下文信息，无法对未来的上下文信息进行处理，因此Yao等人[13]提出了双向LSTM模型来进行中文分词，进一步的提升分词的准确率。

这些深度神经网络模型的分词方法均是基于4词位标注方法，无法全面表达词语中每一个词的词位信息，并且对于神经网络模型训练的结果仅仅只是做了维特比来进行分词，没有进一步来深化处理。针对这些问题，本文做出了如下工作：

(1)应用基于LSTM神经网络的双向LSTM-CRF模型到中文分词任务上。这样既利用了双向LSTM模型能够保存上下文信息的优势，同时也利用了CRF层从句子层面考虑前后标注之间的影响，而不是只对神经网络层的输出进行简单的动态规划处理。最后的实验结果表明，BI-LSTM-CRF模型对比双向LSTM模型具有更好的分词性能。

(2)在神经网络分词模型上分别采用了4词位标注与6词位标注方法，实验结果表明采用6词位标注的分词模型具有更加优越的性能表现

1 模型建立

通过介绍LSTM单元和CRF模型来分别构建LSTM神经网络、BI-LSTM神经网络以及BI-LSTM-CRF模型。其中BI-LSTM-CRF模型在文献[14]中第一次被提出用做处理序列标注问题。

1.1 LSTM单元

LSTM单元能够在不保留冗余上下文信息的同时解决长期依赖问题，因此被广泛的运用在自然语言处理任务中。LSTM记忆单元的基本结构由三个门结构和一个细胞状态组成，如图1所示。

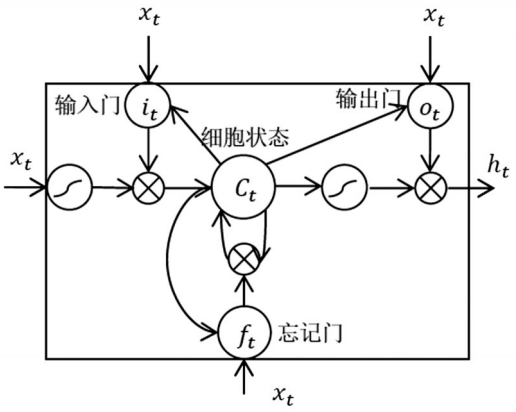

图1 LSTM记忆单元结构图

从图1中可以看出，LSTM记忆单元中细胞状态的保存与更新由输入门，忘记门和输出门决定。其中输入门控制将新的信息中哪些部分保存到细胞状态中，忘记门决定历史细胞状态的保留信息，输出门控制全部更新后的细胞状态哪些部分被输出。LSTM单元具体工作流程用以下公式表示：

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i)$$
$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f)$$
$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o)$$
$$C_t=f_t\cdot C_{t-1}+i_t\cdot\tanh(W_C\cdot[h_{t-1},x_t]+b_C)$$
$$h_t=o_t\cdot\tanh(C_t)$$

其中，i_t，f_t，o_t，C_t 分别表示 t 时刻输入门，忘记门，输出门和细胞状态的输出，x_t 和 h_t 表示 t 时刻时的输入向量和隐藏层向量。σ 表示sigmoid激活函数，W 和 b 分别表示权重矩阵和偏置向量，它们的下标代表它们所属归类，如 W_i 和 b_i 表示它们属于输入门结构中的权值矩阵和偏置向量。

LSTM神经网络通过LSTM记忆单元来保存前面的上下文信息，解决了RNN算法出现的长距离依赖问题，其结构图如图2所示。其中输入层为 X，即语料库中每个字构造成的特征向量。隐藏层为 h，A 表示每一个LSTM记忆单元。输出层为 y，即每个字对应的标注。

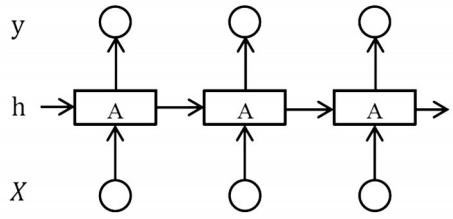

图2　LSTM神经网络结构

但是LSTM不能考虑到未来的上下文信息，比如“他们”这个词，若无法考虑到“他”后面的上下文信息，那么就会把“他”单独分词。这也是LSTM神经网络作为分词模型的局限所在。

1.2　BI-STM神经网络

BI-LSTM神经网络的思路来自于双向RNN模型[15]，其结构图如图3所示。它拥有两个不同方向的并行层，前向层与后向层的运行方式和常规神经网络的运行方式相同。这两个层分别从句子的前端和末端开始运行，因此能存储来自两个方向的句子的信息。这样BI-LSTM既能保存前面的上下文信息，也能同时考虑到未来的上下文信息，从而使其在中文分词中拥有更好的表现。

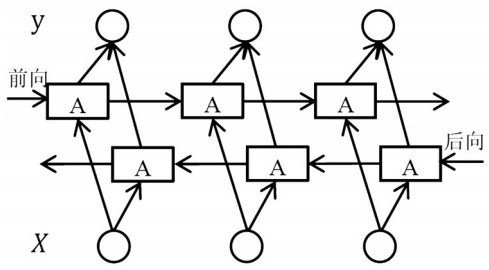

图3　BI-LSTM神经网络结构

1.3　CRF模型

在中文分词中，一般有两种序列标注的思想。第一种是从时间顺序上考虑，从前向后预测出序列标注分布，比如最大熵马尔可夫模型(MEMM)[16]。第二种是从句子层面考虑序列标注问题，每个字标注的状态不仅仅是受到时间顺序上的从前向后的影响，同时也会被未来的状态所影响，比如线性条件随机场模型(CRF)[17]。

这种考虑前后状态影响的序列标注思想和BI-LSTM模型利用前后上下文特征的思想很相似，同时在文献[17]和文献[13]中分别被证明这类模型性能优于只考虑单方面影响的模型。CRF的链式结构图如图4所示。

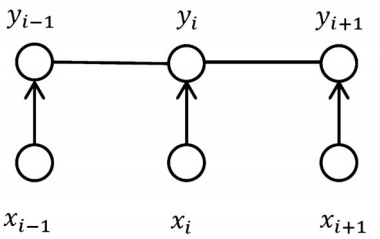

图4　CRF链式结构图

那么在给定观察序列 $X=(x_1, x_2, \ldots, x_n)$ 时，标注序列 y 的概念可以定义为：

$$\exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_k u_k s_k(y_i, X, i))$$

其中，$t_j(y_{i-1}, y_i, X, i)$ 为概率转移函数，表示对于序列 X 在其标注为 y_{i-1} 和 y_i 之间的转移概率。$s_k(y_i, X, i)$ 为状态函数，表示对于序列 X 其 i 位置的标记为 y_i 的概率。λ_j，u_k 分别对应着相应函数的权重。

1.4　BI-LSTM-CRF模型

BI-LSTM-CRF模型是将BI-LSTM网络和CRF模型结合起来，即在BI-LSTM网络的隐藏层后加一层CRF线性层，模型结构如图5所示。该模型通过双向LSTM层很好地结合了上下文的特征，并且经由CRF层有效地考虑了句子前后的标签信息。

由1.3小节可知，对比单独的BI-LSTM神经网络，该模型还需要一个标注之间的状态转移矩阵作为CRF层的参数。通过引入状态转移矩阵 A，然后设定矩阵 P 为双层LSTM网络的输出。其中 $A_{i,j}$ 表示时序上从第 i 个状态转移到第 j 个状态的概率，$P_{i,j}$ 表示输入观察序列中第 i 个词为第 j 个标注的概率。则观察序列 X 对应的标注序列 $y=(y_1, y_2, \ldots, y_n)$ 的预测输出为：

$$s(X, y)=\sum_{i=1}^{n}(A_{y_i, y_{i+1}}+P_{i, y_i})$$

利用动态规划算法可以很好地计算标签序列的推理，具体算法请参阅文献[17]。

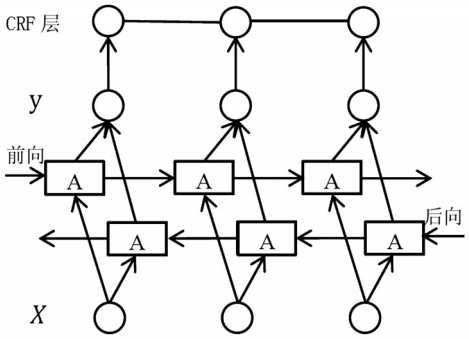

图5 BI-LSTM-CRF模型结构

2 训练方法

为了将中文分词问题转化成序列标注问题，就需要将分词中的每一个字进行标注。在深度学习分词研究工作中，常用的标注集是四词位(B,M,E,S)，分别表示一个分词的开始词位，中间词位，结束词位以及以单独一个字构成的分词。而对比四词位标注，文献[6]中提出的六词位(B,B_1,B_2,M,E,S)标注方法更能有效地表达出字在词语中的词位信息。因此本文对分词语料库文本分别进行四词位标注和六词位标注的预处理，从而对比最后的分词效果。以四词位标注为例，基于本文构建的分词模型训练流程图如图6所示。

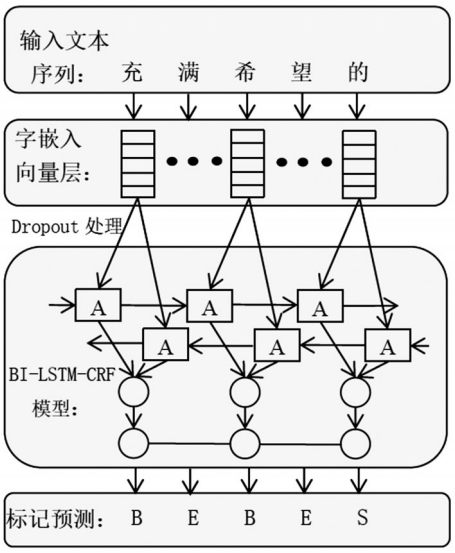

图6 BI-LSTM-CRF分词模型训练流程图

2.1 文本向量化

如图6的向量层所示，输入分词模型的文本序列首先需要分布式向量化。这里采用的是文献[18]提出的word2vec方法，同时该方法在文献[13]中被证明对描述字级别特征有效，能够提升分词的准确率。通过word2vec方法，训练语料库中每一个字可以转化成一个长度为d的空间向量。而对于每一个输入分词模型的字，其对应的分布式向量由该字和该字上下文的字向量拼接而成。

2.2 Dropout方法

Dropout[19]是当前流行的避免神经网络过拟合的方法之一。Dropout方法直接作用在神经网络结构上，随机选择部分单元连同它们的输入输出连接，都暂时从网络中删除它们。其中单元被选中暂时删除的概率可以在训练的时候人为设定。如图6所示，在BI-LSTM-CRF模型输入之前，对最终的字嵌入向量层应用dropout方法。

3 实验

3.1 实验环境，数据集及评价指标

本实验所有的模型全部使用NVIDIA GeForce GTX 1070显卡来训练，相比用CPU训练时间长达好几天的情况下，使用该型号GPU训练模型的时间缩短为3个小时左右。

实验采用Bakeoff2005提供的数据集来评估构建的分词模型，它包括PKU Corpus，MSRA Corpus，AS Corpus和CityU Corpus。为了公正的评估模型的分词性能，实验采用了SIGHAN规定的“封闭测试”原则，其评价指标有：准确率(P)，召回率(R)，综合指标值(F)。

3.2 实验设计及结果分析

由于文本向量化的维度大小及使用dropout方法设定的百分比都会在一定程度上影响最终的分词结果，因此最终使用MSRA Corpus作为语料库，基于4词位标注方法，以基础的LSTM神经网络为模型，分别设定嵌入字向量维度大小为50，100，200，300，dropout百分比为20%，50%，其实验结果见表1。通过表1可知，在字嵌入向量维度为200，dropout大小为20%时，LSTM分词模型取得最佳性能。

通过表1的测试结果，可以设定接下来分词模型实验均是在字嵌入维度为200，dropout百分比为20%的条件下进行。记4词位标注为4-tag，6词位标注为6-tag，以LSTM，BI-LSTM，BI-LSTM-CRF为分词模型，以4-tag与6-tag为标注，分别在PKU，MSRA，AS和CityU语料库上进行分词性能测试。其实验结果见表2，其中模型1-6分别表示LSTM（4-tag），LSTM（6-tag），BI-LSTM

(4-tag)，BI-LSTM (6-tag)，BI-LSTM-CRF (4-tag)和BI-LSTM-CRF(6-tag)。

表1　在不同模型参数下LSTM模型分词性能

模型参数（字嵌入维度,dropout百分比）	准确率（P）	召回率（R）	综合指标（F）
(50,20%)	0.932	0.928	0.930
(50,50%)	0.924	0.918	0.921
(100,20%)	0.940	0.935	0.937
(100,50%)	0.933	0.930	0.932
(200,20%)	0.948	0.944	0.946
(200,50%)	0.943	0.939	0.940
(300,20%)	0.936	0.931	0.933
(300,50%)	0.930	0.926	0.928

通过表2可知,采用6词位标注的模型在分词性能上优于采用4词位标注的模型。对比LSTM以及BI-LSTM模型,BI-LSTM-CRF模型在分词性能评价的各个指标上也有着更加优秀的表现。表3为BI-LSTM-CRF模型与前人在分词领域研究的对比。其中Bakeoff-best表示2005年度Bakeoff测评的最好成绩,(Chen,2015)为文献[11]中的最好结果,其使用了LSTM模型并且在文本向量化过程中加入了双字符嵌入向量,(Yao,2016)为文献[13]中的最好结果,其使用了BI-LSTM模型并且堆叠了BI-LSTM模型的层数为3层。而在本文中并未使用复杂的嵌入向量和增加神经网络模型层数的方法就达到了不错的分词性能,证明了BI-LSTM-CRF分词模型的优越性。

表3　BI-LSTM-CRF模型与前人研究模型性能对比

模型	PKU	MSRA
Bakeoff-best	0.950	0.964
(Chen,2015)	0.965	0.974
(Yao,2016)	0.965	0.976
BI-LSTM-CRF(6-tag)	0.965	0.971

4　结论

本文的工作是在LSTM神经网络的基础上建立BI-LSTM-CRF分词模型,采用不同词位标注并加入预训练字嵌入向量的方法来进行中文分词。通过在Bakeoff 2005数据集上进行的分词性能实验测试以及与前人的研究工作结果对比得出:在同样的分词模型下,六词位标注比四词位标注具有更好的分词性能;在中文分词任务中,对比LSTM模型和BI-LSTM模型,BI-LSTM-CRF模型在各项指标上具有更优秀的性能表现。研究结果还表明了BI-LSTM-CRF分词模型在简体中文和繁体中文中均具有良好的性能,更容易推广到自然语言处理中其他序列标注问题上,具有一定的泛化能力。

参考文献

[1] 黄昌宁,赵海.中文分词十年回顾[J].中文信息学报,2007,21(3):8-19.

[2] 刘源.信息处理用现代汉语分词规范及自动分词方法[M].北京:清华大学出版社,1994.

[3] Nianwen Xue. Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing,2003,8(1):29-48.

[4] Peng F,Feng F,Mccallum A. Chinese segmentation and new word detection using conditional random fields[C]. Proceedings of Coling,2004:562-568.

[5] Tseng H,Chang P,Andrew G,et al. A conditional random field word segmenter for sighan bakeoff 2005[C]. Proc of the Fourth SIGHAN Workshop on Chinese Language Processing,2005:168-171.

[6] Zhao H,Huang C N,Li M,et al. An improved Chinese word segmentation system with conditional random field[C]. Proceedings of the Fifth Sighan Workshop on Chinese Language Processing,2006:162-165.

[7] 刘一佳,车万翔,刘挺,等.基于序列标注的中文分词、词性标注模型比较分析[J].中文信息学报,2013,27

表2　不同标注下LSTM,BI-LSTM以及BI-LSTM-CRF分词模型性能

模型	PKU			MSRA			AS			CityU		
	P	R	F	P	R	F	P	R	F	P	R	F
1	0.940	0.935	0.937	0.948	0.944	0.946	0.947	0.939	0.943	0.944	0.935	0.940
2	0.945	0.937	0.941	0.957	0.943	0.950	0.950	0.943	0.946	0.947	0.941	0.944
3	0.954	0.945	0.950	0.961	0.954	0.957	0.956	0.950	0.953	0.953	0.946	0.950
4	0.959	0.952	0.956	0.966	0.959	0.962	0.961	0.955	0.958	0.958	0.951	0.954
5	0.963	0.958	0.960	0.970	0.963	0.966	0.968	0.962	0.965	0.960	0.952	0.956
6	0.968	0.963	0.965	0.974	0.968	0.971	0.972	0.968	0.970	0.965	0.958	0.961

(4):30–36.
[8] Collobert R, Weston J, Bottou L. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493–2537.
[9] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]. Conference on Empirical Methods in Natural Language Processing, 2013:647–657.
[10] Pei W, Ge T, Chang B. Max–margin tensor neural network for Chinese word segmentation[C]. Meeting of the Association for Computational Linguistics, 2014:293–303.
[11] Chen X, Qiu X, Zhu C, et al. Gated recursive neural network for Chinese word segmentation[C]. Proc of Annual Meeting of the Association for Computational Linguistics, 2015:1744–1753.
[12] Chen X, Qiu X, Zhu C, et al. Long short–term memory neural networks for Chinese word segmentation[C]. Conference on Empirical Methods in Natural Language Processing, 2015:1197–1206.
[13] Yushi Yao, Zheng Huang. Bi–directional LSTM recurrent neural network for Chinese word segmentation[C]. International Conference on Neural Information Processing, 2016:345–353.
[14] Huang Z, Xu W, Yu K. Bidirectional LSTM–CRF models for sequence tag–ging[OL]. http://arxiv.org/pdf/1508.01991v1.pdf, 2015.
[15] Hochreiter S, Schmidhuber J. Long short–term memory[J]. Neural Computation, 2012, 9(8):1735–1780.
[16] Mccallum A, Freitag D, Pereira F C N. Maximum entropy markov models for information extraction and segmentation[C]. Proc of ICML, 2000:591–598.
[17] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]. Proc of ICML, 2002, 3(2):282–289.
[18] Tomas Mikolov, Kai Chen, Greg Corrado, et al. Efficient estimation of word representations in vector space[C]. International Conference on Learning Representations, 2013:1388–1429.
[19] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929–1958.

(上接第86页)

应的阈值，接着分别在前景直方图和背景直方图相应地阈值区间进行直方图均衡化。最后把均衡化后的前景直方图和背景直方图合并作为最后输出图像。经过对几种算法结果的对比分析可知，本文算法不仅对图像的增强效果好，而且对原始图像亮度的维持效果也非常好。

参考文献

[1] Menotti D, Najman L, Facon J. Multi histogram equalization methods for contrast enhancement and brightness preserving[J].IEEE Transactions on Consumer Electronics, 2007, 53(3):1186–1194.
[2] Sengee N, Choi H.Brightness preserving weight clustering histogram equalization[J].IEEE Transactions on Consumer Electronics, 2008, 54(3):1329–1337.
[3] 陈永亮，王华彬，陶亮. 自适应动态峰值剪切直方图均衡化[J]. 计算机工程与应用，2015，51(1)：167–171.
[4] Ibrahim H, Kong N S P. Brightness preserving dynamic histogram equalization for image contrast enhancement[J].IEEE Transactions on Consumer Electronics, 2007, 53(4):1752–1758.
[5] 张龙涛，孙玉秋.基于模糊熵改进的直方图匹配算法研究[J].西南大学学报：自然科学版，2016，38(4)：124–129.
[6] 马超玉.光照不均匀条件下图像增强算法研究[D].长春：长春理工大学，2014.
[7] 陈驰.光学相关图像增强技术研究[D].长春：长春理工大学，2012.
[8] 刘春香.基于DSP图像增强系统的设计与实现[D].长春：长春理工大学，2010.