第 41 卷 第 6A 期
2014 年 6 月

计 算 机 科 学
Computer Science

Vol. 41 No. 6A
June 2014

有向网络重叠社区的快速划分算法

李莉杰 陈端兵 王冠楠

(电子科技大学互联网科学中心 计算机科学与工程学院 成都 611731)

摘 要 随着社会的发展,数据量越来越大,网络规模也在迅速增长。作为一种研究网络结构的有效方法,社区划分对于深刻认识超大规模网络有重要的意义。在分析研究有向网络的非重叠社区划分算法和无向网络的重叠社区划分算法的基础上,提出了一种有向网络重叠社区划分的快速算法。算法根据节点的有向权值和归属度进行社区划分,并分析了有向权值和归属度对划分结果的影响,在此基础上得到了一组最优的有向权值和归属度参数。使用 2 个实际网络和 1 个人工构建网络对算法的性能进行了测试并与已有算法进行了对比。实验结果表明,所提出的算法能够有效地划分出有向网络中的重叠社区。

关键词 有向网络,社区划分,模块度,重叠社区

中图法分类号 TP301.6 **文献标识码** A

Fast Algorithm for Overlapping Community Detection in Directed Networks

LI Li-jie CHEN Duan-bing WANG Guan-nan

(Web Sciences Center, School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract With the development of the society, the scale of data and networks are increasing rapidly. As one of the effective ways of studying network structure, community detecting is a significant issue in deep understanding of large scale networks. Based on the methods of non-overlapping community detecting in directed networks and overlapping community detecting in undirected networks, an overlapping community detecting algorithm in directed networks was proposed in this paper. The basic idea of this method is to divide a network according to two parameters: directed weight and belonging degree of nodes. The influences of these parameters on partition are discussed and the optimal parameters are obtained. The performance of the proposed algorithm was tested and compared with other algorithms on two real networks and one computer-generated network. Experimental results show that the algorithm presented in this paper is rather efficient to detect overlapping communities of directed network.

Keywords Directed networks, Community detection, Modularity, Overlapping community

1 引言

现实社会中很多复杂系统都可以用复杂网络描述。复杂网络除了具有小世界、无标度等特征外,社区结构也是其最重要的结构特征之一。充分挖掘复杂网络中的社区结构对于深刻认识复杂网络有重要的意义。

经典的社区划分算法有随机游走法[1]、K-L 算法[2]、谱分析法[3]、GN 算法[4]、基于模块度优化的方法[5,6]、遗传算法[7]以及函数公式法[8]等。文献[1]用一个随机游走的概率流来代替实际网络中的信息流,并使用压缩的概率流代表网络分解成的模块,结果表示为图的简单形式,突出了结构和结构之间关系的规律性。GN 算法[4]根据边介数不断对网络进行划分,最后得到具有层级结构的社区,此方法不需要预先知道社区个数,克服了 K-L 算法和谱分析算法的不足,但计算复杂度较高。Leicht 和 Newman[5]使用最大化模块度的思想,实现对社区的划分。文献[6]通过优化模块度函数来划分社区。文献[7]对 Newman 的模块度进行了改进以适用于有向网络,并提出了有向网络重叠社区划分的遗传算法。算法中每一个染色体用一个矩阵来表示,矩阵的行号对应社区编号,列号对应网络中点的编号,用模块度作为适应度函数。此算法需要不断地进行交叉、遗传和变异,多次迭代而得到一个较优的社区划分,收敛速度慢,算法的时间复杂度比较高,此外它还需要预先设定网络中社区的个数。Sanjeev 等人[8]通过求网络的邻接矩阵的最大特征值来划分社区。Newman[9]将已有的社区划分方法例如谱分析法映射到图的最小割方法中,并使用最大似然估计将全局搜索变为局部搜索,实现对无向网络的社区划分。

最近 Blondel[10]等人提出的快速算法,首先将网络中每一个节点看成一个社区,然后考虑每一个节点 i 的邻居 j;将 i 暂时放入邻居 j 所在的社区中,计算模块度,如果模块度增

本文受国家自然科学基金(60973069,90924011,60903073,60973120),华为高校合作基金(YBCB2011057)资助。

李莉杰(1988—),女,硕士生,主要研究方向为数据挖掘,E-mail:lilijie.1988@163.com;陈端兵(1971—),男,副教授,硕士生导师,主要研究方向为复杂网络、信息传播与控制;王冠楠(1990—),男,硕士生,主要研究方向为社会计算。

加，则将社区 i 合并到社区 j 中，否则保持不变，重复这个过程，直到所有社区不再变化；之后，再构建一个上层网络，网络中一个节点相当于上一步中的一个社区，重复社区合并和网络重建的过程，直到模块度不再增加为止。该算法简单，复杂度较低，但是它仅考虑了无向网络中非重叠社区的划分。

现实中的很多复杂网络，如社交关系网络、神经网络等，社区之间是有重叠的。Chen 等人[11]引入点的强度和点的归属度，实现了对无向带权网络的重叠社区的快速划分算法。Liu 等人[12]在文献[11]的基础上提出了点的有向权值，并使用点的有向权值和归属度实现了有向带权网络非重叠社区的快速划分。

本文在陈端兵等人[13]重叠社区的两段策略的基础上，结合点的有向权值和归属度，实现了对有向网络重叠社区的划分。算法用 2 个实际的网络和 1 个人工构造的网络对算法性能进行了测试，并与文献[6]中的 GAs 算法进行了对比，实验结果表明，本文的算法在时间和效果上都要优于 GAs。

2 基本概念与定义

2.1 有向权值

在有向网络中，节点的出度和入度会影响节点在社区中的地位，即是否处于社区的中心位置。为了衡量节点的地位，引入有向权值 D_p：

$$D_p=\alpha D_{ip}+(1-\alpha)D_{op} \tag{1}$$

其中，D_{op} 为节点 p 的出度，D_{ip} 为节点 p 的入度，α 为可调参数，调整出度与入度在有向权值中的权重。

2.2 归属度

节点归属度用来衡量这个点属于某一个社区的程度。点 p 相对于社区 c 的归属度 $B(p,c)$ 定义为：

$$B(p,c)=\frac{\alpha D_{ipc}+(1-\alpha)D_{opc}}{D_p} \tag{2}$$

其中，D_{ipc} 为社区 c 中其它点指向 p 的边数，D_{opc} 为从 p 指向社区 c 中其它点的边数。例如，图 1 中小圆圈为社区 c 中的节点，正方形为社区外的节点，与点 p 相连的边共 8 条，点 p 的出度为 5，入度为 3，其中社区 c 中有 2 条边指向节点 p，即 $D_{ipc}=2$；有 3 条边是从节点 p 指向社区 c 中其他节点，即 $D_{opc}=3$；假设参数 $\alpha=0.4$，那么 $D_p=0.4\times5+0.6\times3=3.8$，$B(p,c)=(0.4\times2+0.6\times3)/3.8\approx0.68$。

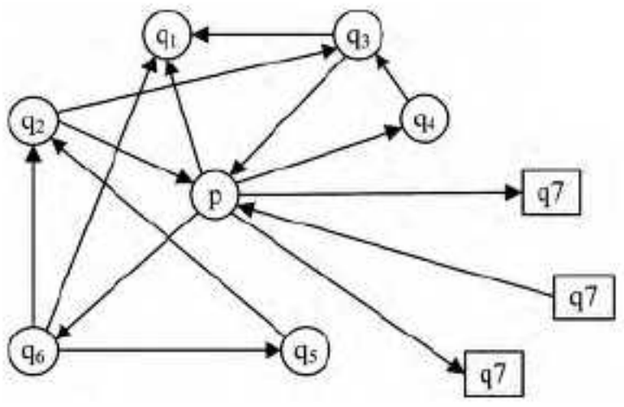

图 1 有向网络的社区划分示意图

2.3 模块度

为了衡量社区划分的质量，Newman 与 Girvan[14]引入了社区划分的模块度。Newman 等人提出的模块度是基于无向网络的，Nicosia 等人[7]在文献[14]的基础上，对模块度进行改进，提出了针对有向网络重叠社区的模块度计算公式，如式(3)—式(9)所示。本文采用 Nicosia 提出的模块度计算公式对社区划分效果进行评价。

$$Q_{ov}=\frac{1}{m}\sum_{c\in C}\sum_{i,j\in V}\beta_{l(i,j),c}A_{ij}-\frac{\beta^{out}_{l(i,j),c}k_i^{out}\beta^{in}_{l(i,j),c}k_j^{in}}{m} \tag{3}$$

$$\beta_{l(i,j),c}=F(B(i,c),B(j,c)) \tag{4}$$

$$F(B(i,c),B(j,c))=\frac{1}{(1+e^{-f(B(i,c))})(1+e^{-f(B(j,c))})} \tag{5}$$

$$f(x)=2px-p,p\in R \tag{6}$$

$$\beta^{out}_{l(i,j),c}=\frac{\sum_{j\in V}F(B(i,c),B(j,c))}{|V|} \tag{7}$$

$$\beta^{in}_{l(i,j),c}=\frac{\sum_{i\in V}F(B(i,c),B(j,c))}{|V|} \tag{8}$$

$$A_{ij}=\begin{cases}1, & \text{存在边}\langle i,j\rangle\\ 0, & \text{否则}\end{cases} \tag{9}$$

其中，m 代表网络中连边数，V 代表网络中的节点集合，$|V|$ 为网络中节点数。

3 算法思想与分析

3.1 算法描述

本文所提算法的基本思想是：每次迭代选择不属于任何社区且有向权值最大的节点作为新社区的中的一个点，将其邻居节点加入社区并按照归属度进行调整，直到所有的节点至少属于一个社区为止，最后合并社区。具体的算法实现步骤如下：

Step 1 将网络中的叶子节点标记为 true，其它节点标记为 false。

Step 2 选择具有最大的有向权值并且标记为 false 的点作为社区 c 的初始节点，并将其邻居加入社区 c。

Step 3 调整社区，对社区 c 中每一个点计算其归属度，如果归属度大于给定的阈值，将其保留在社区中，否则将其移出社区，调整后的社区记为 c'。

Step 4 向社区中添加节点。将社区 c' 的邻居节点暂时放入社区并计算与社区 c' 的归属度，如果大于归属度阈值，则添加该邻居到社区 c' 中。重复这个过程，直到所有的归属度大于阈值的点全部添加到社区 c' 中，此时得到一个更大的社区，记为 c''。

Step 5 将社区 c'' 中每一个点的标记设置为 true。

Step 6 重复 Step 2—Step 5，直到所有非叶子节点至少属于一个社区。

Step 7 合并社区。如果上述步骤划分出的两个社区之间有重叠节点，则计算这两个社区合并后的模块度。如果模块度增加，则合并，否则不合并。

Step 8 将叶子节点加入到与之相连的点所在的社区。

图 2 是一个人工构造的 virtual 网络的社区划分结果，共划分了 5 个社区，分别用不同形状来表示，社区划分的模块度为 0.6469。图中灰色节点表示重叠节点，其中节点 5 是社区 1、社区 2、社区 4 的重叠节点，节点 11 是社区 1 与社区 2 的重叠节点，节点 31 是社区 1 与社区 4 的重叠节点，节点 35、37、38 是社区 4 与社区 5 的重叠节点，节点 21 是社区 2 与社区 3 的重叠节点。

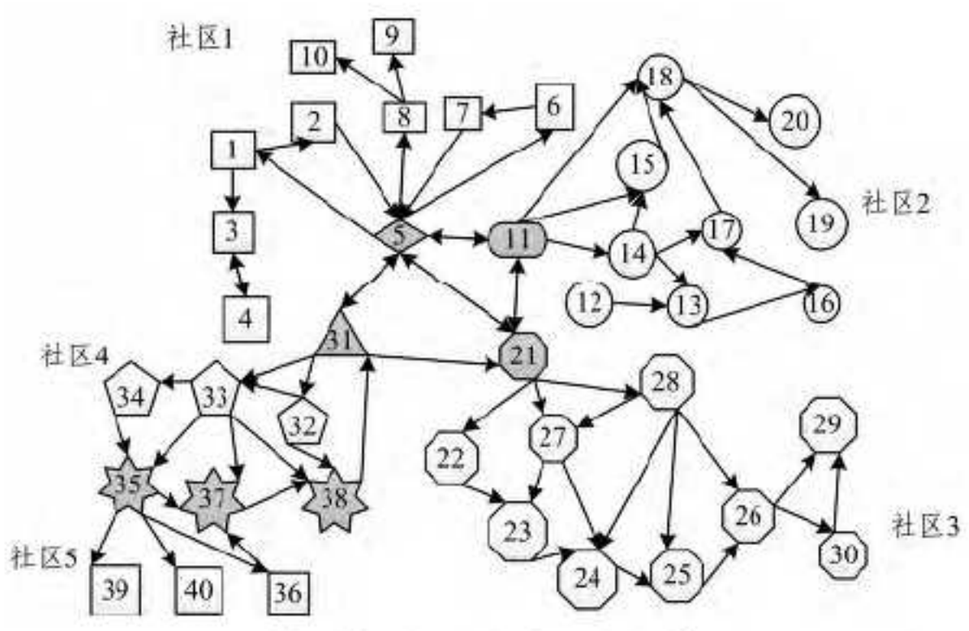

图 2 virtual 网络的社区划分，灰色节点为社区间的重叠节点

3.2 算法分析

设网络中点的个数为 n。在 Step 1 中，网络 G 去除一些末端的点，例如图 2 中的 4 号、9 号、10 号等节点，得到 G'，时间复杂度为 $O(n)$。剩余的节点数为 n'，算法接下来的步骤都是对这 n' 个点进行计算。假定在 Step 7 之前，网络 G' 被划分为 k 个社区，包含的节点数分别为 $m_1, m_2, \cdots, m_k$，那么划分的时间复杂度为 $O(\sum_{i=1}^{k} m_i) = O(kn')$。由于计算社区划分模块度的时间复杂度为 $O(n^2)$，因此在 Step7 中，社区合并的时间复杂度是 $O(k^2 n^2)$。因此，算法的时间复杂度是 $O(k^2 n^2)$。

4 实验结果

4.1 实验数据

本文使用 2 个实际网络 wiki-vote[15] 和 p2p-Gnutella06[16,17] 和 1 个用 Gephi 生成的人工有向网络 virtual(包含 40 个点，连线概率为 0.095) 对本文算法进行了测试。wiki-vote 是维基百科截止到 2008 年的一个的职位选举投票网络，p2p-Gnutella06 是 2002 年 8 月 6 号 Gnutella 点对点的文件分享网络。这 3 个网络的基本信息如表 1 所列。

表 1 3 个数据集的基本信息

数据集	点数	边数	最大度	最大出度	最大入度
virtual	40	72	13	6	7
wiki-vote	7115	103689	1167	893	457
p2p-Gnutella06	8717	31525	115	113	64

4.2 实验结果与分析

运行环境为 C#.net，所用计算机的内存为 3GB，主频为 3GHz。网络的划分效果用模块度来衡量，并与文献[7]中遗传算法(GAs)进行了比较，如表 2 所列。通过多次实验，将 GAs 的迭代次数设为 10，染色体个数设为 10，并根据文献[10]算法的结果，将 wiki-vote 网络的社区个数设为 30，p2p-Gnutella06 网络的社区个数设为 40。结合实验分析结果，当 $\alpha=0,9$ 和 $B(p,c)=0.3$ 时，划分效果较好。

表 2 wiki-vote 与 p2p-Gnutella06 的划分结果

数据集	本文算法		GAs[6]	
	模块度	时间(s)	模块度	时间(s)
virtual	0.6469	<1	0.6430	5
wiki-vote	0.5418	150	0.5403	26242
p2p-Gnutella06	0.6257	18	0.5685	18414

从表 2 可以看出，本文算法比遗传算法 GAs 快，划分效果比 GAs 好；并且本文算法在划分过程中不需要提前设置网络中社区的个数，是一种自适应的社区划分，而 GAs 算法需要提前设置好社区的个数。

算法中会影响划分结果的参数是出度与入度的比例权值参数 α 和归属度 $B(p,c)$ 的阈值 t。图 3 和图 4 比较了在 wiki-vote 网络下，不同的 α 和 t 对划分结果的影响。

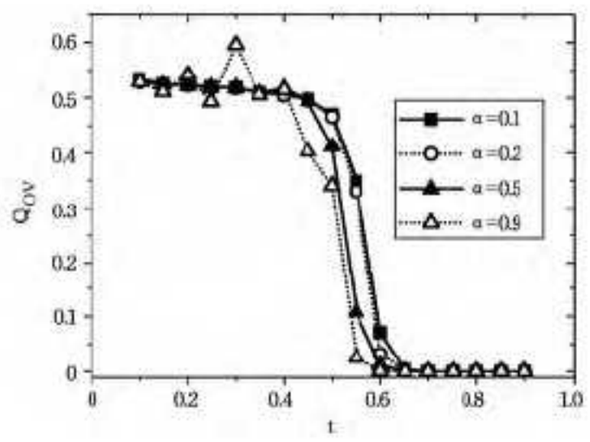

图 3 归属度阈值 t 对模块度的影响

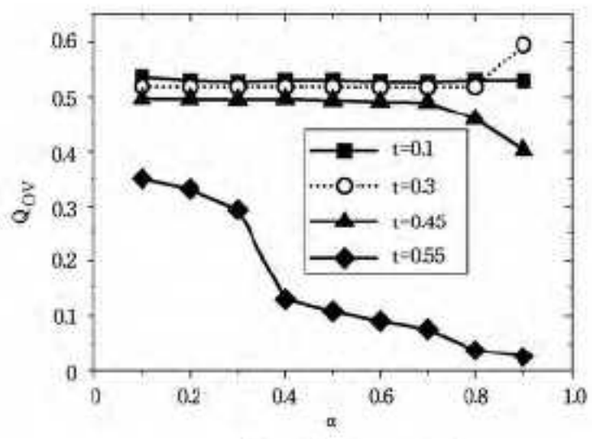

图 4 α 对模块度的影响

从图 3 和图 4 可以看出参数 α 与阈值 t 对划分结果有一定的影响。从图 3 可以看出，对于同一个 α，不同的阈值 t 对于模块度的影响较大，虽然在 $t \leqslant 0.5$ 和 $t \geqslant 0.6$ 时，模块度相差不是很大；但是 t 在 0.5 左右有较大的跳跃。从图 4 中可以看出，当 $t=0.1, 0.3, 0.45$ 时，不同的 α 对于模块度的影响不是很大；但是当 $t=0.55$ 时，α 的影响比较明显，入度对模块度的影响比较大。这是因为，对于社区中某个点，入度代表其它点主动与该点的联系强度。wiki-vote 是一个投票网络，收到的回复越多，说明这个点越重要，在其周围越容易形成社区(例如，在社交网络中，收到别人的回复表示得到他人的关注，这样的节点会成为社区团体的中心点，相对于其他的节点，它的周围更容易形成一个社区)，所以节点出度与入度对于节点的地位都有影响，而且入度的影响更大一些。因此，算法在选择参数时，参数 α 的取值应该偏小一些，同时，归属度的阈值 t 取值应在 $[0.1, 0.5]$。

结束语 社区划分是复杂网络结构分析的一个重要方面。本文在已有的社区划分算法基础上，基于节点的有向权值和归属度提出了有向网络重叠社区的划分方法，与 GAs 算法相比，计算速度有极大的提升，划分效果也有一定程度的改善。实验结果表明节点的出度、入度和归属度阈值都会影响社区划分的效果，其中入度对社区划分的影响较大，当归属度阈值小于 0.5 时可以得到较好的社区划分效果。现实中很多网络除了具有方向性之外，还具有典型的时序性，如社交网络、通讯网络等，如何对时序网络[18]进行划分，将是下一步需要研究的问题。

参 考 文 献

[1] Rosvall M, Bergstrom T C. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1123

[2] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49(2):

291-307

[3] Barnes E R. An Algorithm for partitioning the nodes of a graph [J]. SIAM J. Alg. Disc. Meth, 1982, 4(3): 541-550

[4] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 99: 7821-7826

[5] Leicht E A, Newman M E J. Community structure in directed networks[J]. Physical Review Letters, 2008, 100(11): 118703

[6] Newman M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582

[7] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009: 1742-5468

[8] Chauhan S, Girvan M, Ott E. A network function-based definition of communities in complex networks[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2012, 22(3): 033129

[9] Newman M E J. Community detection and graph partitioning [J]. Europhys. Lett., 2013, 103: 28003

[10] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in larger networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10: P10008

[11] Chen D B, Shang M S, Lv Z H, et al. Detecting overlapping communities of weighted networks via a local algorithm[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(19): 4177-4187

[12] Liu H T, Qin X, Yun H F, et al. A Community Detecting Algorithm in Directed Weighted Networks[J]. Lecture Notes in Electrical Engineering, 2011, 98: 11-17

[13] 陈端兵, 尚明生, 李霞. 重叠社区划分的两阶段策略[J]. 计算机科学, 2013, 40(1): 225-228

[14] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113

[15] Leskovec J, Huttenlocher D, Kleinberg J. Predicting Positive and Negative Links in Online Social Networks [C]// Proceedings of the 19th international conference on World Wide Web. ACM, 2010: 641-650

[16] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 2

[17] Ripeanu M, Foster I, Iamnitchi A. Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design[J]. IEEE Internet Computing Journal, 2002, 6: 50-57

[18] Holme P, Saramäki J. Temporal networks[J]. Physics Reports, 2012, 519: 97-125

(上接第 226 页)

均分数(Mean Opinion Scores, MOS)和融合图像客观评价结果都归一化到[0,1]。图 5 为 FVIF 客观评价结果与主观评价结果的一致性测试结果。

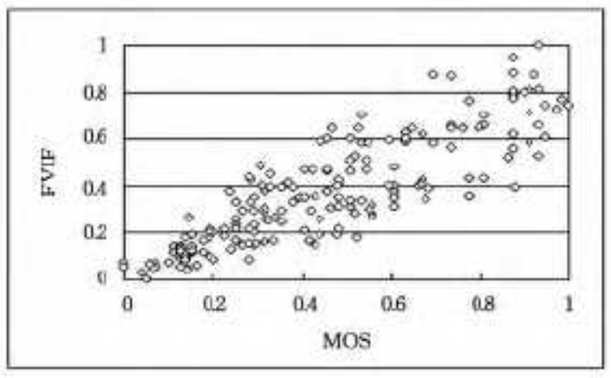

图 5 FVIF 客观评价结果与主观评价结果的一致性测试结果

分别测试各种客观评价结果与主观评价结果的线性相关系数 PLCC、均方根误差(Root Mean Square Error, RMSE)及斯皮尔曼等级相关系数 SRCC[9],如表 1 所列。

表 1 客观评价方法的性能比较

评价方法	性能参数		
	PLCC	RMSE	SRCC
$Q^{AB/F}$	0.24	0.37	0.12
FSSIM	0.31	0.32	0.48
$IE^{AB/F}$	0.42	0.29	0.51
$MI^{AB/F}$	0.08	0.58	0.01
FVIF	0.87	0.15	0.91

图 5 及表 1 中的实验数据表明,采用 FVIF 方法的评价结果真实地反映了融合图像的视觉质量,与主观评价的结果一致性最好。FVIF 评价方法是基于图像自然场景模型和图像信号模型,并考虑人眼视觉失真的影响,更能真实反映 HVS 从融合图像中提取的图像源中继承的图像信息量的大小,而其它客观评价方法注重于评价图像源与融合图像像素灰度值或统计直方图之间的平均差异,其评价结果有时与主观视觉质量相悖。

结束语 本文提出了基于视觉信息保真度的融合图像质量客观评价方法,并构造融合图像的评价指标 FVIF,该方法能够对融合图像的质量进行客观评价。实验表明,该方法的评价结果与主观评价结果的基本一致,性能优于其它评价方法。FVIF 方法对自动图像处理、实时图像处理及融合算法的评价均具有重要的实用价值。

参 考 文 献

[1] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error measurement to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[2] Xydeas C S, Petrovic V. Objective image fusion performance measure[J]. Electronics Letters, 2000, 36(4): 308-309

[3] Li Shu-tao, Kang Xu-dong, Hu Jian-wen, et al. Image matching for fusion of multi-focus images in dynamic scenes[J]. Information Fusion, 2013, 14: 147-162

[4] Yang C, Zhang J, Wang X, et al. A novel similarity based quality metric for image fusion[J]. Information Fusion, 2008, 9(2): 156-160

[5] Sheikh H R, Bovik A C. Image information and visual quality [J]. IEEE Trans. Image Process., 2006, 15(2): 430-444

[6] Wang Z, Bovik A C. A universal image quality index [J]. IEEE Signal Processing Letters, 2002, 9(3): 81-83

[7] Portilla J, Strela V, Wainwright M J, et al. Image denoising using scale mixtures of Gaussians in the wavelet domain[J]. IEEE Trans. Image Process., 2003, 12: 1338-1351

[8] Sheikh H R, Bovik A C, de Veciana G. An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics[J]. IEEE Trans. Image Process., 2005, 14(12): 2117-2128

[9] Wang Zhou, Li Qiang. Information Content Weighting for Perceptual Image Quality Assessment[J]. IEEE Transactions on Image Processing, 2011, 20(5): 1185-1198