地理信息系统支持下的山区遥感影像决策树分类

陈艳华[1],张万昌[2]
(1. 南京大学国际地球系统科学研究所,南京　210093;2. 中国科学院大气物理所全球变化东亚区域研究中心,北京　100029)

摘要:山区遥感影像分类是遥感研究的一大难题。本文利用一种决策树生成算法(C 4.5 算法)自动提取知识,基于知识建立决策树用于山区影像分类,并结合研究区土地利用类型与 DEM 空间统计关系的先验知识,在 GIS 空间分析的基础上进行影像分类的后处理。与传统的最大似然法分类结果相比,该方法极大地改善了山区地表覆被分类的精度,得到试验区较为可靠的遥感分类图像。

关键词:遥感影像;分类;知识;决策树;地理信息系统
中图分类号:TP 79:P 208　文献标识码:A　文章编号:1001 - 070X(2006)01 - 0069 - 06

0　引言

提高遥感数据的专题信息计算机提取精度,是遥感研究的主要方向之一。近 20 a 里,前人提出了大量的基于影像的分类算法与理论[1],但大多利用目标地物的光谱反射特性开发各种算法加以分类,而由多种因素造成的'同物异谱,异物同谱'问题,制约了基于光谱特征的统计模式分类方法精度的提高。为此,多年来,结合像元级的空间光谱特征辅以遥感信息以外的待分类区各种特征信息开发的分类算法逐渐成为一种趋势,并取得了进展[2~5]。辅助数据引入遥感图像分类这一领域出现了很多新的智能化分类方法,如专家系统[6](Expert System)、神经网络[7]、数据挖掘等。神经网络方法仍然处于发展阶段,不容易实现,而且分类结果存在着不可预测性[8,9];基于知识的专家系统分类是一种人工智能分类方法,它运用知识以及输入的数据来确定解决问题的最佳途径,而不是采用预先定义的方法[10]。当没有专业知识或专业知识不可获取时,它常常采用机器学习的方法来建立基于规则的分类系统。但这些分类方法虽然在分类精度上有了一定的提高,但是在处理同物异谱、异物同谱和山体阴影等具体问题时,特别是在山区影像分类时,还存在着一定的缺憾。

鉴于此,作者研究了一种比较通用的山区影像分类方法。该方法首先利用 DEM 对试验区的影像进行地形影响校正,减少阴阳坡植被光谱差异;然后利用 C4.5 算法自动提取知识,结合土地利用图参与训练区样本的选择与影像的分类和后处理,提高了分类精度。本文以汉江流域上游山区为实验区,系统介绍了分类方法,并讨论了该方法的分类精度。

1　试验区概况与数据源

1.1　试验区概况

试验区位于陕西省石泉县境内,汉江流域上游边界地段,北纬 32°25′~33°52′,东经 107°16′~108°20′。区内植被类型以草地、亚高山针叶林和阔叶林为主。由于受气候、水分和地形条件影响,其植被垂直地带性分布特征明显:海拔 500 m~1 600 m 的代表类型为草地和高密度草地;海拔 1 600 m~2 000 m 的代表类型为阔叶林;海拔 2 000 m~2 500 m 的代表类型为针叶林。

1.2　数据源

(1)遥感数据。根据该区的物候特征,5~8 月

收稿日期:2005 - 07 - 04;修订日期:2005 - 10 - 09
基金项目:国家重点基础研究发展规划项目(2001CB309404),中国科学院'百人计划'后动项目(8 - 047401)和教育部科学技术重点项目(2001)联合资助。

为土地利用遥感分类的最佳时期。试验选用的遥感数据是 2003 年 6 月 14 日 Landst－5 TM 影像。

(2) DEM 数据。DEM 是遥感影像在山区进行地形纠正的基础,将地面高程信息引入遥感图像分类是提高遥感分类精度的有效措施之一[11]。针对该区植被垂直分带,阴阳坡差异较大的特点,引入高程和坡度信息,可以减少遥感分类中经常出现的同谱异物现象,从而有效提高遥感影像分类精度。

(3) 1990 年土地利用现状图。数字化土地利用图的引入可以对分类结果做检验,亦可用于指导训练样本的最优选取。先将土地利用图数字化,然后与 TM 图像进行配准。

2 研究方法与技术路线

2.1 C 4.5 算法简介

C 4.5 算法是一种决策树生成算法[12]。该算法使用信息增益率来选择属性,克服了用信息增益选择属性时偏向选择取值多的属性不足,并且在树构造过程中或者构造完成之后进行剪枝,能够对连续属性进行离散化处理,采用决策树作为知识表示,最终形成产生式规则。

2.2 技术路线

首先,利用 C 4.5 算法自动提取知识,建立决策树用于影像分类;然后利用 GIS 空间叠加统计分析功能对研究区土地利用类型与 DEM 的空间关系进行知识提取,指导影像分类的后处理,得到试验区的遥感影像分类图。分类流程如图 1 所示。

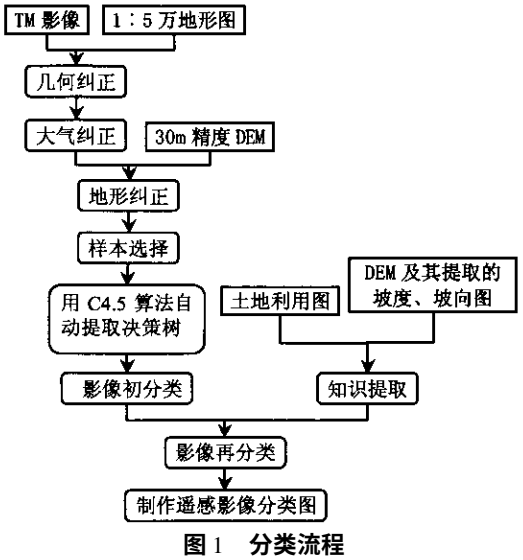

图 1 分类流程

2.3 多源数据预处理

在分类之前,首先要对各种数据进行预处理,包括遥感影像的几何精校正、大气辐射校正、地形校正以及 DEM 和土地利用图与遥感影像的配准。

本实验选取 17 个均匀分布的地面控制点(GCP)对遥感影像进行几何精校正,误差小于 0.5 个像元。为了提高精度,采用 Civco 提出的地形校正方法[13~15],基于 30m 分辨率 DEM 对影像进行地形校正,以减少地形影响。由于缺乏卫星过境时详细的大气剖面资料,无法采用 LORTRAN 或 6S 等大气校正模型,直接使用基于 TM 影像的 Gilabert 模型进行大气校正,最后得到除热红外波段之外的其它 6 个波段的地表反射率[14,16]。

2.4 决策树分类

针对实验区的特点,结合实地考察结果,确定土地覆盖/土地利用的类别为针叶林、阔叶林、灌木林、草地、高密度草地(简称高密草)、水域、裸地、水田及湿地等。但是,遥感影像局部有云和云影,所以分类时增加了云和云影两个类别。

由于试验区植被覆盖度较高,用一般的 TM743、TM741 及 TM432 等彩色合成,很难区分出各种植被的颜色和色调差异,于是对遥感影像进行主成份变换,选择主成份变换(KL 变换)后的第一、第二、第三组份进行假彩色合成。变换之后,能够明显区分出各种植被类型。

由于决策树分类法是以各像元的特征值为设定的基准值,分层逐次进行比较的分类方法。比较中所采用特征的种类与基准值对分类结果的精度有很大影响[17]。按照随机原则,结合土地利用图,对各个类别选择足够多的训练样本,并统计各类别的特征,结果如图 2 和图 3 所示。

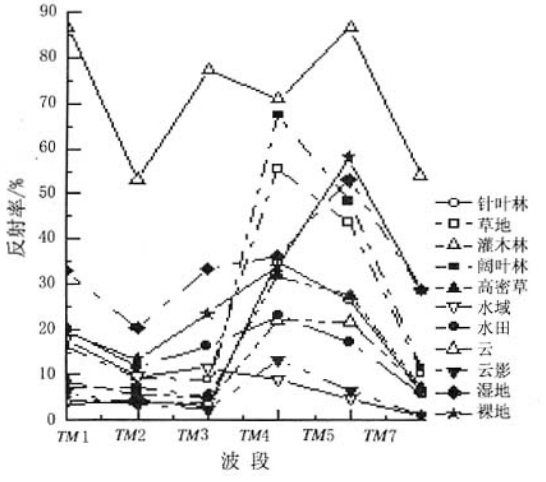

图 2 各类别的波谱特征

分析图 2、图 3 可以看出,对于针叶林、灌木林和阔叶林,单纯依据光谱特征难以获得较高的分类精

度,将 KL 变换后前三个分量参与分类,提高了各类别的光谱可分性。

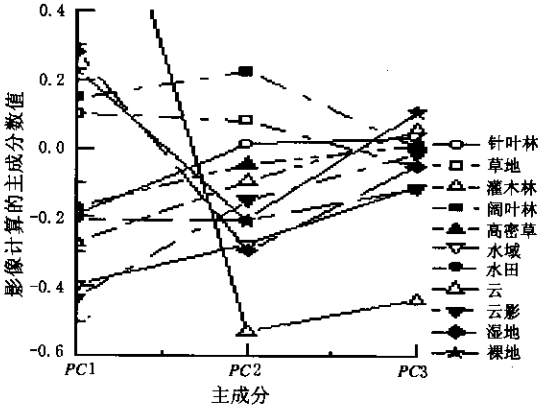

图 3　KL 变换 *PC*1、*PC*2、*PC*3 的各类别波谱特征

实验采用 C 4.5 算法对样本数据进行知识挖掘,提取分类规则,自动建立决策树,并对决策树进行精度分析。C 4.5 算法会得到多种结点个数不同、精度也不同的决策树。从结果中可以看出,结点个数达到 11 个之前,精度随着结点个数急剧上升,在达到 11 个结点时,精度达到 96.7%。之后,随着结点个数的增加,精度趋于稳定,变化不大。而且,结点个数越多,树结构越复杂,所得到的分类规则亦越复杂,分类效率也就不高。本试验只选择最终结点个数正好为分类个数的决策树(如图 4 所示),并对

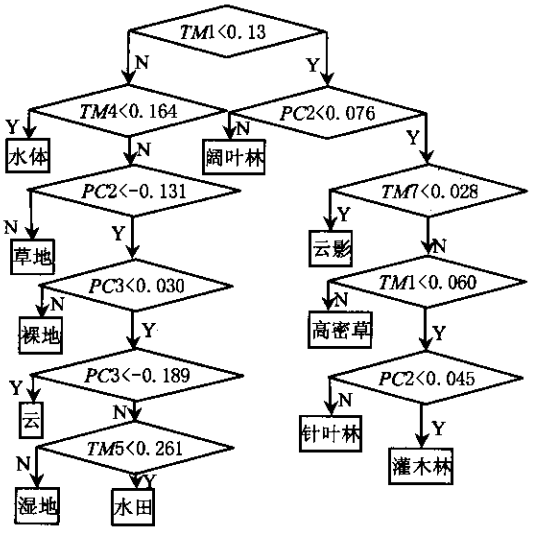

图 4　用 C 4.5 算法得到的带有 11 个结点的决策树

决策树模型进行精度评价,然后对影像进行分类,再对分类的结果利用 GIS 数据建立的知识作为辅助信息进行修正,最后得到的分类精度大大改善。

由表 1 可知,各种类别的决策树建模精度大都达到 97% 以上,阔叶林的精度相对较低,只有 84%,但也比一般的影像分类精度高,这主要是由于所选的阔叶林样本很少是纯净的,一般是与针叶林与阔叶林的混合像元造成的。

表 1　决策树模型误差矩阵精度分析

类型	针叶林	水域	裸地	水田	云	云影	草地	灌木林	阔叶林	高密草	湿地	样本总数	精度/%
针叶林	1 251	0	0	0	0	0	0	21	0	2	0	1 274	98
水域	0	1 193	0	6	0	1	0	0	0	0	1	1 201	99
裸地	0	0	1 963	0	0	0	0	0	0	0	26	1 989	98
水田	0	10	0	1 919	2	0	0	0	0	0	21	1 952	98
云	0	0	0	0	5 585	0	0	0	0	0	147	5 732	97
云影	0	0	0	77	0	3 195	0	6	0	80	1	3 359	95
草地	0	0	0	0	0	0	1 635	0	10	14	0	1 659	98
灌木林	25	0	0	0	0	22	0	717	0	83	0	847	84
阔叶林	2	0	0	0	0	0	26	0	2 014	0	0	2 042	99
高密草	20	0	0	0	0	0	0	15	0	1 478	0	1 513	98
湿地	0	0	30	84	14	0	16	0	0	2	1 675	1 821	92

总精度:96.7%

2.5　基于 GIS 知识的影像再分类

从表 1 可以看出,下面几个类别之间误分率是相对比较高的,如表 2 所示。

表 2　可能误分类

分类结果	针叶林	水体	裸地	水田	云	云影	草地	灌木林	阔叶林	高密草	湿地
可能误分类	灌木林 高密草	水田	湿地	阴影 湿地	湿地	灌木林	阔叶林 湿地	针叶林 云影	草地	阴影 草地 灌木林	裸地 水田 云

利用 1∶25 万土地利用图与 DEM 进行知识提取,综合分析土地利用类型与海拔、坡度等的统计关系,还有影像初分类中未利用的光谱信息,提取经验知识,建立影像再分类规则,对以上决策树分类结果进行改进。例如,阔叶林与草地可以通过利用海拔进行进一步分类,根据土地利用类型与海拔之间的统计关系,得知阔叶林绝大部分生长在海拔 1 200 m 以上,而草地一般都在海拔 1 200 m 以下。结合土地利用类型图,对于那些原来类型为阔叶林、海拔高于 1 200 m 而被为草地的区域再分类为阔叶林。另外,也可以利用影像初分类中未利用的光谱信息进一步对可能误分的类别进行再分类。如湿地与裸地,从图 2 可以看出,由于湿地的湿度大,水分多,其在可见光部分的反射率比裸地强,因此,湿地与裸土之间的误分可以通过判断 *TM*1 + *TM*2 + *TM*3 是否大于 0.7 来进行再分类。最终建立的类别校正规则(部分)如下:

If *Class* = 针叶林 and *LUCC* = 灌木林 and *TM*4 < 0.25 Then 灌木林;

If *Class* = 针叶林 and *LUCC* = 高密度草地 and *DEM* < 1 600 Then 高密度草地;

If *Class* = 水体 and *LUCC* = 水田 and *TM*4 + *TM*5 > 0.3 Then 水田;

If *Class* = 裸地 and *LUCC* = 湿地 and *TM*1 + *TM*2 + *TM*3 > 0.7 Then 湿地;

If *Class* = 水田 and *LUCC* ! = 水田 and *TM*4 + *TM*5 + *TM*7 < 0.25 Then 云影;

If *Class* = 水田 and *TM*5 > 0.35 Then 湿地;

If *Class* = 云 and *PC*1 < 0.6 Then 湿地;

If *Class* = 阴影 and *TM*4 + *TM*5 > 0.3 and *DEM* > 2 000 Then 灌木林;

If *Class* = 草地 and *LUCC* = 阔叶林 and *DEM* > 1 200 Then 阔叶林;

If *Class* = 草地 and *TM*3 > 0.2 Then 湿地;

If *Class* = 灌木林 and *TM*4 > 0.25 Then 针叶林;

If *Class* = 阔叶林 and *LUCC* = 草地 and DEM < 1 200 Then 草地;

If *Class* = 高密度草地 and *TM*4 < 0.15 or *DEM* > 1 600 Then 云影;

If *Class* = 高密度草地 and *TM*4 > 0.45 and *DEM* > 1 500 Then 阔叶林;

If *Class* = 湿地 and *TM*1 + *TM*2 + *TM*3 < 0.7 Then 裸地;

If *Class* = 湿地 and *PC*1 < 0 Then 水田;

If *Class* = 湿地 and (*PC*1 > 0.6 or *DEM* > 1 000) Then 云。

其中,*Class* 是决策树分类图;*LUCC* 是土地利用图;*TM*1、*TM*2、*TM*3、*TM*4、*TM*5、*TM*7 分别代表遥感影像的 6 个波段像元亮度值;*PC*1 是 KL 变换后的第一分量,*DEM* 是数字高程值。最后根据土地利用图,采用综合分析的方法,假定分类图中的云和云影覆盖的地方土地利用状况没有发生变化,直接取土地利用图中的类型。这样利用决策树建模型精度分析表来找出容易相互误分的类别,在满足一定精度要求的前提下,建立的类别校正规则数量比两两组合建立的规则要少得多,实用得多,而且适应性更强。

2.6 精度分析

为了检验这种分类方法的有效性,将常规最大似然法分类结果与之做了精度比较。本试验采用土地利用分类图作为测试依据,就每个类型选取测试样本,总共取 2 374 个样本,并对两类分类结果进行精度分析(表 3、表 4)。

表 3 检测样本的误差矩阵及精度——基于知识的分类方法

类别	针叶林	水域	裸地	水田	草地	灌木林	阔叶林	高密草	湿地	合计	生产者精度/%
针叶林	239	0	0	0	0	72	34	2	0	347	69
水域	0	128	2	16	0	0	0	0	14	160	80
裸地	0	3	191	10	21	4	0	2	13	244	78
水田	0	23	0	157	0	0	0	0	10	190	83
草地	0	0	3	0	338	0	0	84	0	425	80
灌木林	19	0	0	0	0	110	17	16	0	163	67
阔叶林	78	0	0	0	0	5	282	24	0	389	72
高密草	6	0	2	0	36	13	2	235	0	294	80
湿地	0	21	12	6	1	0	0	0	122	162	75
合计	342	175	210	189	396	204	335	363	159	1 802	
用户精度/%	70	73	91	83	85	54	84	65	77		

总体精度 = 76%, *Kappa* = 0.725

表 4　检测样本的误差矩阵及精度——最大似然分类法

类别	针叶林	水域	裸地	水田	草地	灌木林	阔叶林	高密草	湿地	合计	生产者精度/%
针叶林	178	0	0	0	0	132	34	3	0	347	51
水域	0	122	6	15	0	0	0	0	17	160	76
裸地	0	9	175	10	22	0	0	0	28	244	72
水田	0	21	0	144	0	0	0	0	25	190	76
草地	0	0	10	0	315	0	0	100	0	425	74
灌木林	28	0	0	0	0	97	20	18	0	163	60
阔叶林	76	0	0	0	0	4	250	59	0	389	64
高密草	12	0	2	0	45	17	9	209	0	294	71
湿地	0	23	21	1	1	0	0	0	116	162	72
合计	294	175	214	170	383	250	313	389	186	1 606	
用户精度/%	60	70	82	85	82	39	80	54	62		

总体精度 = 68% ,*Kappa* = 0.632

从表 3、表 4 分析结果可以看出,基于知识的分类都要比常规最大似然法分类的精度要高得多。就总体精度而言,基于知识分类法的分类精度约为 76% ,而传统的最大似然法的分类精度约为 68%。尽管两种分类法的分类精度都不太高,但是基于知识的分类法总体精度要比最大似然法的总体分类精度约高 8%。Kappa 系数说明,基于知识的分类方法优于最大似然法(插页彩片 20)。

3　结论

本文从 GIS 辅助遥感影像分类的角度出发,将 GIS 数据引入 RS 影像分类分析技术中,从 GIS 数据发现知识辅助影像分类,以改进遥感影像分类的精度、可信度和效率。与传统的最大似然分类方法比较表明,基于知识的分类方法可以获得更高的分类精度。但是校正规则的建立是基于利用本地区的 DEM 与原来土地利用图的统计关系等一些先验知识建立的经验性规则,并不能直接用于其它地区。要建立一种普适的校正规则有待进一步研究。现在各种数据异常丰富,如何充分利用 GIS 数据库提供的丰富地理数据,进而挖掘和发现 GIS 数据库中的深层知识(如地物空间关联规则、地物空间分布规律等)以更好地发挥 GIS 数据在遥感影像分类中的作用,以及利用新的遥感数据源(如高光谱、多角度和雷达等遥感数据)进行分类将是主要的研究方向之一。

致谢：南京大学国际地球所赵登忠博士、朱求安硕士、郑光硕士和中国科学院兰州寒旱所李海英博士参与了本论文的讨论并提出了宝贵的意见；南京大学城市资源系刘永学老师提供软件,给予了极大的帮助,在此一并表示诚挚的谢意！

参考文献

[1] 赵英时,等. 遥感应用分析原理与方法[M]. 北京：科学出版,2005.

[2] 徐冠华,鞠洪波,李志清. 遥感图像判读的专家系统及其应用[A]. 徐冠华. 再生资源遥感研究(平泉区)[M]. 北京：科学出版社,1988,38 – 46.

[3] 龙 晶. 用局部结构法改善 TM 图像的分类精度[A]. 孙司衡. 再生资源遥感研究(新疆区)[M]. 北京：林业出版社,1991.

[4] 术洪磊. 基于知识的遥感影像分类与制图综合方法研究[D]. 北京：北京大学,1995.

[5] 王杰生. 遥感图像应用处理中的一个分类新算法——模拟目视分辨法[J]. 环境遥感,1992,7(2)：126 – 137.

[6] Goodenough D G, Goldberg M, Pluckett G, et al. An Expert System for Remote Sensing[J]. IEEE Transactions on Geoscience and Remote sensing,1987,25(3)：349 – 359.

[7] Huang X, Jensen J R. A Machine – learning Approach to Automated Knowledge – base Building for Remote Sensing Image Analysis with GIS Data[J]. Photogrammetric Engineering and Remote Sensing,1997,63(10)：1185 – 1194.

[8] Bruzzone L, Conese C, Maselli F. Multisource Classification of Complex Rural Areas by Statistical and Neural – net – work Approaches[J]. Photogrammetric Engineering and Remote Sensing,1997,63(5)：523 – 533.

[9] Skidmore A K, Turner B J, Brinkhof W. Performance of a Neural Network：Mapping Forests Using GIS and Remotely Sensed Data[J]. Photogrammetric Engineering and Remote Sensing,1997,63(5)：501 – 504.

[10] Moninger W R. ARCHER：A Prototype Expert System for Identifying Some Meteorological Phenomena[J]. Journal Atomos. Ocean technol, 1988(5)：144 – 148.

[11] 沙占江,曾永年,马海州,等. 遥感和 GIS 支持下的龙羊峡库区土地沙漠化动态研究[J]. 中国沙漠,2000,20(1)：48 – 50.

[12] Quinlan J R. C 4.5：Programs for Machine Learning[M]. San Mateo, CA：Morgan Kaufmann,1993.

[13] 领耀文. 基于数字遥感图像的民勤绿洲 20 年变化研究[J]. 干旱区研究,2002,19(1)：69 – 74.

[14] ZHANG Wanchang, Yamaguichi Y, Ogaw K. Evaluation of the Effect of Preprocessing of the Remotely Sensed Data on the Actual

Evaporation, Surface Soil Moisture Mapping by an Approach Using Landsat, DEM and Meteorological Data [J]. Geocarto Inter, 2000, 15(4): 57 – 67.

[15] Civco D L. Topographic nomalization of Landsat Thematic Mapper digital imagery [J]. PR&RS, 1989, 55(9): 1303 – 1309.

[16] Gilabert M A. An atmospheric correction method for the automatic retrieval of surface reflectance from TM images [J]. Int. J. Remote Sens, 1994, 15(10): 2065 – 2086.

[17] 日本遥感研究会编. 遥感精解[M]. 刘勇卫, 贺雪鸿, 译. 北京: 测绘出版社, 1993.

GIS SUPPORTED DECISION TREE CLASSIFICATION OF REMOTE SENSING IMAGES IN MOUNTAINOUS AREAS

CHEN Yan – hua[1], ZHANG Wan – chang[2]

(1. *International Institute for Earth System Science, Nanjing University, Nanjing* 210093, *China*; 2. *START Regional Center for Temperate East Asia, Institute of Atmospheric Physics, CAS, Beijing* 100029, *China*)

Abstract: Remotely sensed data based land use/cover classification, especially in mountainous areas, is a difficult problem that has long drawn attentions among researchers. This paper presents a synthetic approach using C4.5 algorithm to automatically derive classification knowledge with the purpose of constructing a model of decision tree for the final classification of the image. Statistical relationships of the land – use pattern with DEM were analyzed through spatial analysis function of GIS to provide extra knowledge for the post classification processes, which improves the precision of final classification by enhancing the characteristics of the trial zones in the image. According to a classification experiment on the rugged terrain over the upstream of Hanjiang River Basin where the land use/cover ground survey data are available, the proposed approach is far superior to the traditional maximum likelihood classification method.

Key words: Remote sensing images; Classification; Knowledge; Decision tree; Geography information system

第一作者简介: 陈艳华(1983 –), 男, 硕士研究生, 主要从事遥感信息提取及遥感和 GIS 在水文学中的应用研究。

(责任编辑: 肖继春)

(上接第 55 页)

THE APPLICATION OF IMU/DGPS – SUPPORTED PHOTOGRAMMETRY

GUO Da – hai[1,2], WU Li – xing[1], WANG Jian – chao[2], ZHENG Xiong – wei[2]

(1. *China University of Mining & Technology, Beijing* 100083, *China*; 2. *China Aero Geophysical Survey and Remote Sensing Center for Land and Resources, Beijing* 100083, *China*)

Abstract: The direct measurement of exterior orientations using the integrated IMU/DGPS system proves to be an effective means. Experiments show that the accurate performance of direct orientation measurement is sufficient for forming orthophotos and small or medium scale maps. Aimed at solving the problems existent in the IMU/DGPS orientation module, this paper deals with the integrated IMU/DGPS system and the elements of IMU/DGPS – based photogrammetry and, on such a basis, emphatically analyzes the overall system calibration.

Key words: IMU; GPS; Photogrammetry

第一作者简介: 郭大海(1966 –), 男, 博士研究生, 主要从事航空遥感信息获取、数据处理及地理信息系统方法技术与应用研究。主持国家 863、国土资源大调查等国家级、省部级科技项目 9 项, 航空遥感摄影等勘查项目 20 余项。在国内外学术刊物和学术交流会上公开发表论文近 10 篇。

(责任编辑: 刁淑娟)