一种基于图像区域系综分类的室外场景理解方法

张 敏 刘利雄 贾云得
(北京理工大学计算机科学与工程系,北京 100081)

摘 要 多层感知机分类器是一种有效的数据分类方法,但其分类性能受训练样本空间的限制。通过多层感知机分类器系综提高室外场景理解中图像区域的分类性能,提出了一种自动识别室外场景图像中多种景物所属概念类别的方法。该方法首先提取图像分割区域的低层视觉特征,然后基于系综分类方法建立区域视觉特征和语义类别的对应关系,通过合并相同标注区域,确定图像中景物的高层语义。对包含 5 种景物的 150 幅图像进行测试,识别率达到了 87%。与基于多层感知机方法的实验结果相比,本文提出的方法取得了更好的性能,这表明该方法适合于图像区域分类。此外,系综方法还可以推广到其他的分类问题。

关键词 图像理解 室外场景 特征提取 多层感知机 系综分类
中图法分类号:TP391 **文献标识码**:A **文章编号**:1006-8961(2004)12-1443-06

An Outdoor Scene Understanding Method Based on Ensemble Classification of Image Regions

ZHANG Min, LIU Li-Xiong, JIA Yun-De
(*Department of Computer Science and Engineering, Beijing Institute of Technology, Beijing* 100081)

Abstract Even multi-layer perception (MLP) classifier has been an efficient method of data classification, and the performance is often limited by the training samples space. In this paper, the MLP classifiers ensemble is used to improve the performance of image region classification in understanding of outdoor scene and a scheme for automated recognition concept classes of objects in outdoor scene images by image region classification is presented. First, the low-level visual features are extracted from the segmented image region, and then the ensemble classifiers are used to establish corresponding relationship between the visual features of image region and semantic class. Finally, the high-level semantic class of each object in an image is formed by combining the region with same label. The method has been evaluated on 150 images including five objects and recognition rate is around 87%. The experimental results show that the proposed method that has better performance compared to MLP-based method is suitable for image regions classification. Moreover, this ensemble method appears to generalize to other classification problems.

Keywords image understanding, outdoor scene, feature extraction, multi-layer perception, ensemble classification

1 引 言

赋予计算机能够鲁棒地解释场景语义信息的能力是计算机视觉研究的主要目标[1]。为了解释场景的内容,得到对场景的语义描述,需要建立高层场景概念描述与低层图像视觉特征之间的联系,在对场景图像处理和分析的基础上,识别和定位场景中包含的景物以及它们之间的空间关系[2]。对于室外场景,这无疑是非常困难的,因为室外场景中的景物可以有任意的大小、形状和位置,光照条件的变化,以及存在大量的遮挡,导致了同一景物类的不同实体存在较大的可变性,这些引起了人们极大的研究热情。

多种方法被用来对室外场景进行语义理解。单调树的多级分层方法[3]用场景的低层视觉特征直接进行表示景物的语义特征,通过对风景图像的 9 种

收稿日期:2004-03-04;改回日期:2004-05-19

景物进行语义提取实现室外场景理解;基于自顶向下和学习的控制策略分割图像的方法[4]针对室外场景光照条件变化大的特点,对室外场景天空、树木、道路和地面 4 种常见的景物进行识别和标注;自动图像分类系统[5]能够标注道路和城市场景中的景物,这种方法采用的是自下而上的识别策略和基于神经网络的分类方法,并取得了很好的效果;结合景物相互关系的知识,利用局部爬山算法标注自然场景图像的方法[6],在速度上取得很大的提高。

这些研究表明将场景语义理解问题分解为一些子问题能够更容易地解决,如图像分割、区域特征描述和区域分类(如图 1 所示)。

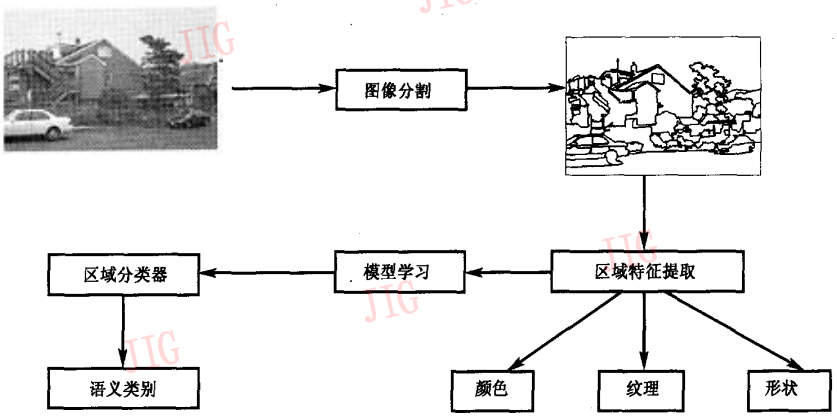

图 1 室外场景理解系统的体系结构图

首先将输入场景图像分割为一些封闭区域,理想情况下这些区域对应于场景中的某个景物或部分景物;然后经过特征提取,每个区域描述为一个特征集合,该集合表示区域的低层视觉属性,这些特征集合描述将作为学习机的输入;最后利用学习算法对图像区域的分类,标注出各区域的语义类别。其中对分割后的图像区域分类是整个场景语义理解的关键步骤,本文提出对图像区域进行分类的方法实现特定室外场景图像的语义理解,区域的分类利用了系综(ensemble)分类方法,通过对训练样本的学习得到所有景物类别的归纳结果,根据训练的模型确定待分类区域的语义类别。

假设这类场景的图像包含有限数目的景物类型,通常有天空、建筑、树木、道路和车辆等景物。实际中,这几种景物比较常见,一般的室外场景中都包含这些景物。由于场景中景物的复杂性,同一类中样本变化很大,单一分类器不能很好地区分不同类型的景物,其推广性能受到样本空间的限制,为了解决这个困难,利用 Bagging 和 Boosting 技术[7],设计多层感知机的系综,在构造系综分类器的基础上,建立室外场景图像理解系统,识别和定位场景图像中的景物,确定它们之间的相互关系。对多层感知机进行系综是因为系综分类器的性能要超过其中任何一个单独的分类器[8],并且对于多层感知机,训练时是采用梯度下降法寻找局部最优的网络权值,产生的分类器是次优的,在进行分类器系综时,改变算法的参数会生成另一个次优的分类器,系综方法在某种程度上可以提高单个多层感知机分类器的推广性能。需要说明的是该方法并不仅限于这几种景物,通过增加学习的景物类型可以识别更多的景物。

2 系综分类算法

分类器系综是一个分类器的集合,对样本的分类结果是通过将单独每个分类器的分类结果以某种方式组合进行决策的[9]。本文提出利用 Bagging 和 Boosting 技术,构造多层感知机分类器的系综,前者每个单独的多层感知机的训练是通过 bootstrap 随机地选择训练样本实现的,后者对每个多层感知机训练样本的选择是根据样本错误分类的概率分布。

2.1 多层感知机

使用的多层感知机分类器是一个标准的 3 层神经网络,包括一个线性的输入层,一个非线性的隐藏层和一个线性的输出层。多层感知机的训练采用了误差反向传播(BP)算法,BP 算法主要包括前馈计算和反向学习。假定训练样本集合 $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1,\cdots,N}$

($\boldsymbol{x}_i \in \mathbf{R}^d$是输入向量,$\boldsymbol{y}_i \in \mathbf{R}^c$ 是期望输出的类别),输入层有 d 个结点,每个结点对应输入向量的一个特征,隐藏层的结点数目为 h,输出层的结点数目为 c,对于分类问题,除了期望类别对应的输出位置设置 1,其他位置设置为 0。w_{ji}表示输入层结点 i 到隐藏层结点 j 的连接权,w_{kj}表示隐藏层结点 j 到输出层结点 k 的连接权,这里 i,j,k 分别对应输入层、隐藏层和输出层结点的索引。

对于前馈计算,隐藏层每个结点的净激活(net activation)为

$$n_j = w_{j0} + \sum_{i=1}^{d} w_{ji} x^i \tag{1}$$

x^i 是输入向量 $\boldsymbol{x}$ 的第 i 个坐标,非线性的隐藏层采用双曲正切函数 tanh,每个隐藏层结点的输出 h_j 为

$$m_j = \tanh(n_j) \tag{2}$$

输出层每个结点的净激活为

$$n_k = w_{k0} + \sum_{j=1}^{h} w_{kj} m_j \tag{3}$$

对输出层用"胜者全取"(logsoftmax)的方法确定用于分类的输出 O_k

$$o_k = \log \frac{\exp(n_k)}{\sum_{j=1}^{c} \exp(n_j)} \tag{4}$$

然后计算误差,$e_k = y^k - o_k$,y^k 是期望输出向量 $\boldsymbol{y}_i$ 的相应位。

多层感知机的反向学习基于梯度下降法,连接权初始化为随机值,然后向误差减小的方向调整,通过反向传播计算相应结点的梯度。输出层的梯度为

$$\delta_k = e_k \cdot o_k \cdot (1 - o_k) \tag{5}$$

隐藏层到输出层的连接权更新规则是

$$\Delta w_{kj} = \lambda \cdot \delta_k \cdot m_j \tag{6}$$

λ 是学习速率,表示连接权的相对变化尺度。隐藏层的梯度为

$$\delta_j = m_j \cdot (1 - m_j) \cdot \sum_k \delta_k w_{kj} \tag{7}$$

输入层到隐藏层的连接权更新规则是

$$\Delta w_{ji} = \lambda \cdot \delta_j \cdot x_i \tag{8}$$

因此,进行迭代计算时用 $w = w + \Delta w$ 更新网络连接权。迭代前馈计算和反向学习这两步,直到满足一定的停止标准,需要指出的是学习速率 λ 在迭代的过程中是逐渐减小的。

2.2　多层感知机系综

为了解决单一多层感知机分类器的推广性能受到训练样本空间的限制,对多层感知机进行了系综。图 2 是多层感知机系综的结构。训练阶段,每个单独的多层感知机的训练是通过独立的训练样本进行的,所有的分量 MLP(multi-layer perception)采用某种聚集策略进行组合,测试阶段,用所有的 MLP 对测试样本同时进行分类,然后用相应的聚集策略对分类结果进行决策。

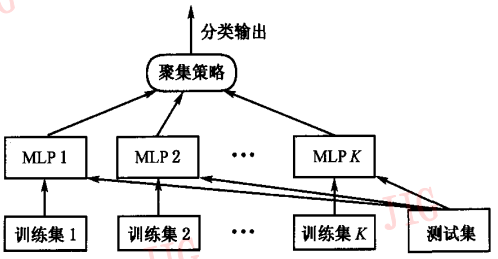

图 2　多层感知机系综分类器的结构

首先用 Bagging 算法构造多层感知机系综,多个独立的多层感知机通过 bootstrap 随机地选择训练样本进行训练,这些 MLP 的聚集是通过投票(voting)的方式实现的。假定训练样本集合 $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1,\cdots,N}$,为了构造 K 个独立多层感知机的系综,就需要 K 个训练样本集合。从统计的观点出发,为了得到更高的分类结果,应该使这些训练样本集合尽量的互不相同,因此通过 bootstrap 从原始的训练集合中随机地选择样本,并将这一过程重复多次,直到产生 K 个独立的训练样本集合,原始训练集合中的样本在新的训练集合中可以出现多次,也可以不出现。这样,每个训练集合被用来训练相应的多层感知机。

Bagging 算法的聚集策略通过投票方式对 MLP 系综进行决策。假定 $h_k(k=1,2,\cdots,K)$是第 k 个 MLP 的决策函数,$l_c(c=1,2,\cdots,C)$是第 c 类的标注,$N_c = num\{k \mid h_k(x) = l_c\}$,那么,对测试样本 $\boldsymbol{x}$,MLP 系综 $g(\boldsymbol{x})$的决策规则是

$$g(\boldsymbol{x}) = \arg\max_c N_c \tag{9}$$

同 Bagging 算法一样,Boosting 算法中每个多层感知机的训练也是要用不同的训练集合,但训练样本的选择方式不同于 Bagging 算法。训练样本集合$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1,\cdots,N}$中有 N 个样本,初始化对每个样本指定一个相同的权值 $W_0^i = 1/N$,训练第 k 个多层感知机时,训练集 k 中的样本是根据第 $k-1$ 次迭代时的权值 W_{k-1}^i选择的。训练样本权值的调整是依照样本分类结果的对错进行的,错误分类样本的权值得到提高,而正确分类样本的权值就被降低,这意味着

分类比较困难的样本可以多次得到训练。整个训练过程如此进行，直到 K 生成个多层感知机，实现时使用如下 AdaBoost 算法。

初始化：训练样本 $D=\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1,\cdots,N}$，$W_0^i=1/N$

For $k=1$ to K

根据 W_{k-1}^i 选择的训练集训练多层感知机 M_k

训练误差 $E_k=\sum_{i=1}^{N}W_{k-1}^i$，$i$ 为分类错误的样本

$\alpha_k=-\frac{1}{2}\ln[E_k/(1-E_k)]$

$W_k^i=\frac{W_{k-1}^i}{Z_k}\times\begin{cases}\exp(-a_k),\text{样本 } i \text{ 分类正确}\\ \exp(a_k),\text{样本 } i \text{ 分类错误}\end{cases}$

(Z_k 是规一化系数)

End

分量 MLP 分类器为 $h_k(x)=M_k$，系综采用加权平均的聚集策略，根据 AdaBoost 算法生成的权值 α_k，对于测试样本 $\boldsymbol{x}$ 所属的类别，最后的 MLP 系综的决策规则 $g(\boldsymbol{x})$ 是根据各分量 MLP 的加权平均确定的

$$g(\boldsymbol{x})=\sum_{k=1}^{K}\alpha_k h_k(\boldsymbol{x}) \tag{10}$$

3 区域的特征表示和提取

通过对区域分类进行场景语义理解，是因为室外场景中景物的变化非常大，区域是有效表示场景内容的基本单位[10]，用精确的区域表示场景有以下优点：首先分割后区域数目相对于图像的像素数大为减少，便于对高层进行分析；其次，有区分力的特征不是在像素上，而是在区域上定义，因此可以对区域进行分类，进行评估的区域特征包括：(1)几何属性的度量，如形状；(2)物理属性的度量，如颜色和纹理；此外，对景物空间关系的描述和推理也有意义。

3.1 图像分割

获得精确的区域需要采用鲁棒的图像分割，通过研究大量的彩色图像分割技术[11]，在图像分割阶段，使用了均值移动(mean shift)算法[12]，该算法是一种鲁棒的特征空间分析技术，通过密度梯度估计的非参数化过程，即密度梯度上升来恢复图像概率分布中的高密度区域(对应着图像中的重要特征)。对于复杂的室外场景图像，景物的边缘会因为光照或遮挡的原因变得比较模糊，采用均值移动算法可以得到高质量的区域边缘图，并可以控制分割区域的大小[13]。

进行均值移动分割时，首先将图像像素从 RGB 空间映射到 Luv 特征空间，在特征空间的随机位置定义足够数量的搜索窗口，然后对每个窗口用均值移动算法找到高密度区域，用图像空间中的约束验证提取的中心，得到特征调色板，最后利用图像空间信息将所有的特征向量分配到特征调色板得到分割的区域。

分割的区域将要通过分类进行语义标注，理想情况下这些区域对应于某个景物或部分景物，但是场景中的景物比较复杂，并且遮挡现象严重，为了获得较好的结果，分割时对图像应该进行过分割(over-segmentation)，通常一种景物包括若干个区域。同时，为了避免出现大量的小区域，设置区域像素数目的阈值至少为 100 个像素。

3.2 区域特征提取

图像区域的特征提取是室外场景理解研究的另一个重要问题，提取特征的好坏将直接影响整个场景理解系统的性能，特征要简洁并具有较强的区分力。因此，针对室外场景中景物变化非常大的特点，采用的特征都是一些基本视觉属性，并没有用精确描述景物模型的特征，如关于形状的角点、边缘描述，而是采用区域的统计特征。

区域的颜色特征用分割后该区域所有像素 Luv 的平均值表示。区域的纹理采用 Gabor 滤波器进行特征提取[14]。选择 Gabor 滤波器提取纹理特征主要基于以下两点：它类似于人类视觉系统的某些方面，在频域中可以通过任何的椭圆形区域；Gabor 滤波可达到频域和空域相结合的优化分辨率，取得频谱信息和在图像中的位置精确性的折衷。

2 维 Gabor 函数和相应的傅氏变换为

$$g(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{-\frac{1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2+\left(\frac{y}{\sigma_y}\right)^2\right]+2\pi\mathrm{i}Wx\right\} \tag{11}$$

$$G(u,v)=\exp\left\{-\frac{1}{2}\left[\left(\frac{u-W}{\sigma_u}\right)^2+\left(\frac{v}{\sigma_v}\right)^2\right]\right\} \tag{12}$$

其中，σ_x、σ_y 和 σ_u、σ_v 是相应的滤波参数，W 是中心频率。

根据 Gabor 小波函数 $g(x,y)$ 的自相似性，通过生成函数 $g_{mn}(x,y)$ 实现 $g(x,y)$ 的扩展和旋转，可以得到区域的自相似的滤波：

$g_{mn}(x,y)=a^{-m}G(x',y')$，$a>1$，$m,n$ 为整数 (13)

其中，$x'=a^{-m}(x\cos\theta+y\sin\theta)$，$y'=a^{-m}(-x\sin\theta+y\cos\theta)$，$\theta=n\pi/S$，$S$ 是方向数目。

对于一幅图像 $I(x,y)$，Gabor 小波变换的输出为

$$W_{mn}(x,y)=\iint I(x1,y1)g^{*}_{mn}(x-x1,y-y1)\mathrm{d}x1\mathrm{d}y1 \tag{14}$$

其中,* 表示复共轭。假定区域中局部纹理具有空间一致性,则使用区域 Gabor 变换系数的均值 μ_{mn} 和标准差 σ_{mn} 表示区域的纹理特征:

$$\mu_{mn}=\iint|W_{mn}(x,y)|\mathrm{d}x\mathrm{d}y$$
$$\sigma_{mn}=\sqrt{\iint(|W_{mn}(x,y)|-\mu_{mn})^2\mathrm{d}x\mathrm{d}y} \tag{15}$$

将 μ_{mn} 和 σ_{mn} 作为区域的纹理特征,采用 3 级尺度,4 个方向就得到了 24 维的纹理特征:

$$f_t=[\mu_{00}\sigma_{00}\mu_{01}\sigma_{01}\cdots\mu_{23}\sigma_{23}] \tag{16}$$

区域的几何特征提取主要是对形状特征的提取。区域的形状特征采用了 7 个不变矩[15]进行描述,这些矩对于一些常见几何变换是不变量,如平移、缩放、旋转,因此这种矩特征可以粗略地表示物体的形状,并希望这些矩特征对于多层感知机的训练具有鲁棒性。

这样,对每个区域构造了一个 34 维的特征向量集合(如表 1 所示)。为了将这些不同类型的特征集成到一个特征向量中,对数据进行了规一化处理。

表 1　区域的特征向量表示

视觉属性	区域特征	数目
亮度	平均灰度　L	1
颜色	平均色度　u,v	2
纹理	Gabor 滤波　$G_0\cdots G_{23}$	24
形状	不变矩　$H_0\cdots H_6$	7

4　实验结果

为了验证本文所采用方法的可行性,对一个 150 幅彩色图像的图库进行实验,图像的大小为 384×256,图像区域的语义类型是通过手工进行标注的。首先采用单个 MLP 进行分类,对其中 100 幅图进行训练,使用 34 维的特征集合,构造的 MLP 分类器有 34 个输入和 5 个输出,经过训练和优化,隐藏层采用 15 个结点可以达到最优的结果,然后用 MLP 来对 50 幅图像测试集进行分类,可以达到 82%的整体精确度。分别用 Bagging 和 Boosting 构造的系综分类器进行分类,其中的基本分类器采用单个 MLP 的结构,相同的训练集合和测试集合,出于计算量的考虑,系综的数目为 10,两种方法分别达到了 85%和 87%的整体识别率。

对每一类景物区域的分类结果如图 3 所示,几乎所有的类别系综方法的分类精度都比单个 MLP 有不同程度的提高,例如树木区域类别的识别率提高了 6%,比较容易识别的天空区域也有一定程度的提高。同时,利用 Boosting 构造的系综分类器和 Bagging 相比,识别率较高,可能主要是因为采用 Boosting 技术构造的各分量分类器,在训练时逐步选用最富信息的样本进行训练,提高了分类准确率。图 4 是用 Boosting 构造的系综分类器对 3 幅测试场景的分类结果,可以看出左边场景中天空、道路和部分车辆区域基本上被识别出来了,而建筑和树木由于相互遮挡,及光照的影响,没有被完全地分类出来。中间场景的车辆由于分割采用固定的融合区域,较小部分的区域被错分类为道路,而最右边的场景由于各景物相互之间位置关系被准确地识别出来。

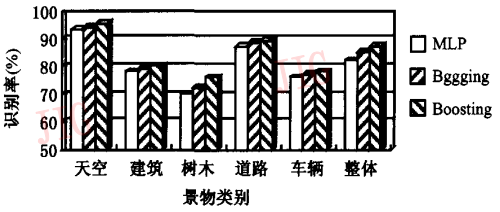

图 3　不同景物类别区域的识别率

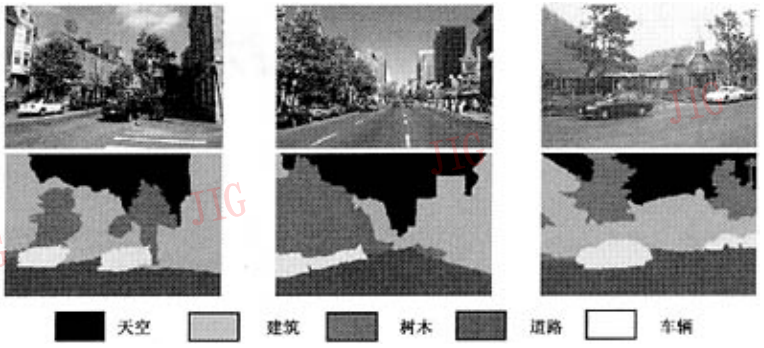

图 4　原始图像(上)的识别结果(下)

5 结 论

提出了一种实现自动解释室外场景的方法，通过对景物区域低层视觉特征的提取和样本集合的学习来训练系综分类器，完成对天空、建筑、树木、道路和车辆 5 类景物的识别。并对一个有 150 幅图，包含 5 种景物的图像数据库进行测试，其中采用 Boosting 技术的 MLP 系综达到了 87%的识别率。与单个 MLP 相比，系综方法提高了分类的准确率，实验结果表明这种方法具有可行性和应用潜力。同时该方法还可以用来分类不同场景中的景物，因此今后的工作是希望进一步细化场景中的景物类别，扩大可识别景物的范围，提高识别的精度。

参 考 文 献

1 Shapiro Linda, Stockman George. Computer Vision [M]. Englewood Cliffs, NJ, Prentice Hall, 2000:13～20.

2 Kodratoff Y, Moscatelli S. Machine learning for object recognition and scene analysis [J]. International Journal of Pattern Recognition and Artificial Intelligence, 1994, **8**(1): 259～304.

3 Song Y, Zhang A. Analyzing scenery images by monotonic tree [J]. ACM Multimedia Systems, 2003, **8**(6): 495～511.

4 Marti Joan. A new approach to outdoor scene description based on learning and top-down segmentation[J]. Image and Vision Computing, 2001, **19**(14): 1041～1055.

5 Campbell N W, Mackeown W P. Interpreting image databases by region classification[J]. Pattern Recognition, 1997, **30**(4): 555～563.

6 Hiroki Hayashi, Mineichi Kudo. Fast labelling of natural scenes using enhanced knowledge[J]. Pattern Analysis & Applications, 2001, **4**(1): 20～27.

7 Duda R, Hart P. Pattern Classification(2nd Edition)[M]. New York: John Wiley & Sons, 2001: 475～478.

8 Thomas G, Dietterich. Ensemble methods in machine learning [A]. In: First International Workshop on Multiple Classifier Systems[C], Cagliari, Italy, 2000:1～15.

9 Thomas G, Dietterich. Machine learning research: four current directions [J]. Artificial Intelligence Magazine, 1997, **18**(4): 97～136.

10 Luo Jiebo, Guo Chengen. Perceptual grouping of segmented regions in color images[J]. Pattern Recognition, 2003, **36**(12): 2781～2792.

11 Lucchese L, Mitra S K. Advances in color image segmentation [A]. In: Global Telecommunications Conference [C]. Rio de Janeiro, Brazil, 1999: 2038～2044.

12 Comaniciu Dorin. Mean shift: a robust approach toward feature space analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, **24**(5): 603～619.

13 Comaniciu Dorin, Meer Peter. Robust analysis of feature spaces: color image segmentation[A]. In: IEEE Conference on Computer Vision and Pattern Recognition[C]. San Juan, Puerto Rico, 1997: 750～757.

14 Manjunath B S, Ma W Y. Texture features for browsing and retrieval of image data [J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 1996, **18**(8): 837～842.

15 Hu M K. Visual pattern recognition by moment invariant[J]. IEEE Transactions on Information Theory, 1962, **8**(2):179～187.

张 敏 1975 年生。2001 年于郑州大学获计算机应用技术专业硕士学位，现为北京理工大学计算机科学与工程系博士研究生。主要研究方向为计算机视觉、机器学习、图像处理。

E-mail:zhangmin2001@bit.edu.cn

刘利雄 1974 年生。讲师。2001 年于武汉大学获计算机应用技术专业硕士学位，现为北京理工大学计算机科学与工程系在职博士研究生。主要研究方向为图像编码、计算机视觉和机器学习。

贾云得 1962 年生。工学博士、教授、博士生导师。主要研究方向为计算机视觉、媒体计算和智能系统。曾获得部委级科技进步一等奖 1 项，二等奖 1 项；获得和申请国家发明专利 12 项；发表学术论文 80 余篇，编著图书 1 册、参与编著图书 1 册。