文章编号:1674-7070(2010)02-0118-06

# 一种基于不平衡样本集的摩托车识别算法

文学志[1] 郑钰辉[1] 赵英男[1] 吴毅[1]

**摘要**

提出了一种基于HSV(Hue-Saturation-Value)空间的Haar小波特征和多SVM(Support Vector Machine)分类器的摩托车识别算法,以解决因样本比例不平衡所导致的对摩托车识别性能差的问题.首先在HSV颜色空间基于无符号小波系数构造特征提取算法,然后对训练数据应用所提出的样本重构方法得到若干训练子集,基于各个训练子集训练相应的SVM分类器,识别时将各SVM的输出结果进行融合即可得到最终识别结果.实验结果表明:该方法识别性能高,鲁棒性好,对于受数据的不平衡性严重影响的对象识别具有较好的应用和推广价值.

**关键词**

摩托车识别;特征提取;不平衡数据;支持向量机(SVM)

**中图分类号** TP391.41

**文献标志码** A

## 0 引言

Introduction

在基于视觉的车辆识别中,统计模式识别方法由于识别性能高、鲁棒性好及操作便捷而受到越来越多的关注,文献[1-2]介绍了利用PCA(Principal Component Aualysis,主成分分析)进行特征提取然后采用SVM(Support Vector Machine,支持向量机)或神经元网络分类器来进行车辆检测的方法. Goerick 等[3]介绍了利用局部方向编码(Local Orientation Code,LOC)提取感兴趣区域(Region of Interest,ROI)的边缘特征信息,然后将LOC的直方图输入神经元网络(Neural Network,NN)来对车辆进行检测. Papageorgiou 等[4]使用过完备的小波系数结合SVM来进行车辆及行人的检测. Schneiderman[5]使用截断的小波系数特征结合SVM进行车辆的检测. Sun 等[6-7]分别介绍了采用Gabor滤波器提取矩特征或采用Haar小波特征与Gabor特征相结合,然后利用SVM分类器来进行车辆检测的方法.本文主要关注白天情形下基于视觉的摩托车识别问题.针对当前在灰度空间基于有符号小波系数的特征提取算法得到的特征存在类间变化小、类内聚类效果较差导致分类复杂度高以及用于分类时的抗噪能力差的问题,本文提出一种基于HSV(Hue-Saturation-Value)颜色空间的Haar小波特征提取方法.同时,道路上的摩托车一方面与车辆一样易于受到外界环境如光照、道路、护栏、绿色植物、建筑物等的影响,另一方面还受路面上车辆的影响,再加上道路上行驶的摩托车数量较少,使得训练数据中摩托车(正类样本)数量远少于非摩托车(负类样本)数量,从而导致传统统计模式识别方法、识别性能尤其是对摩托车的识别受到严重影响,基于此,提出了一种样本重构方法,即首先将负类样本分成若干份,将每一份与正类样本组成子训练集,基于每一个子训练集训练相应的SVM分类器,识别阶段对每一个SVM分类器的输出结果进行融合后得到最终的分类识别结果.实验结果表明:本文方法与文献方法相比能显著提高对摩托车的识别性能.

## 1 方法描述

Method description

整个算法分为离线训练和在线识别两个阶段.离线训练阶段首先对经特征提取后的训练样本数据进行重构,然后基于重构后的训

**收稿日期** 2009-12-16

**基金项目** 国家高技术研究发展计划(863计划)项目(2006AA11Z221);国家自然科学基金(60702076)

**作者简介**

文学志,男,博士,副教授,主要研究领域为模式识别、图像处理等. wwpub@163.com

1 南京信息工程大学 计算机与软件学院,南京,210044

练样本子集训练相应的 SVM 分类器,在线识别阶段将 ROI 输入到离线训练阶段得到的各 SVM 分类器中,将各 SVM 的输出结果进行融合即可得到 ROI 的最终分类识别结果,下面分别进行详细介绍.

### 1.1 离线训练阶段

离线训练阶段包括对收集的摩托车和非摩托车样本进行预处理和特征提取,然后将特征提取得到的非摩托车样本数据进行抽样处理,将抽样数据与摩托车样本数据组成子训练集分别训练相应的 SVM 分类器,如图 1 所示. 以下先介绍图像预处理和特征提取,然后介绍对训练数据的抽样方法以及 SVM 分类器的训练方法.

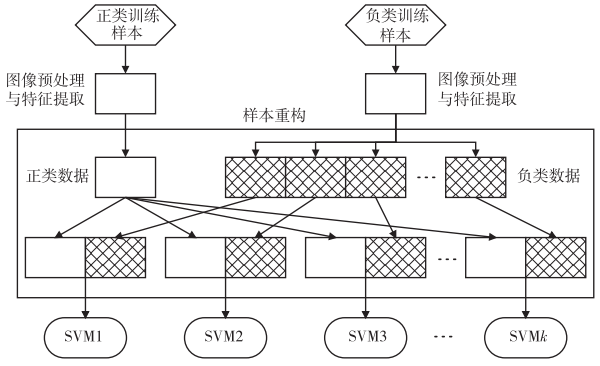


图 1 离线训练

Fig. 1 Off-line training

#### 1.1.1 图像预处理及特征提取

受摩托车外形、姿态及与自车距离的影响等使得 ROI 图像尺寸变化范围大,加上摩托车检测系统对实时性要求高,因而对所有 ROI 图像进行归一化处理是非常必要的. 首先将 ROI 的 RGB 彩色图像缩放为 32 像素 ×32 像素的大小,这个尺寸的图像基本上保证了摩托车识别所需要的信息,同时又具有不太高的维数;然后将尺寸归一化后的图像由 RGB 颜色模型转化为 HSV 颜色模型,这是因为 HSV 颜色模型比 RGB 颜色模型更接近于人们的经验和对彩色的感知,对光照的适应性更好.

由于 Haar 小波特征对图像的边缘及纹理特征信息具有多尺度、多方向描述能力,且计算速度快,本文采用 Haar 小波来实现特征提取. 对以上预处理后的图像,构造以下基于 Haar 小波的特征提取算法:

1) 基于 HSV 颜色模型的 $V$ 通道分量对 ROI 进行 5 层 Haar 小波塔式分解;

2) 对小波塔式分解得到的所有小波系数取绝对值得到系数幅值,设小波系数幅值构成的向量为 $\boldsymbol{W}$,记 $\boldsymbol{W}$ 中的元素个数为 $m$,文中 $m$ 为 1 024;

3) 求向量 $\boldsymbol{W}$ 中元素的最大值 $\boldsymbol{W}_{\max}$ 和最小值 $W_{\min}$;

4) 将向量 $\boldsymbol{W}$ 中的所有元素按以下方法归一化到[0,1],得到归一化向量 $\boldsymbol{W}_{\text{norm}}$:

```
For i = 1 To m
    {
    W_norm[i] = (W[i] - W_min)/(W_max - W_min);
    }
Endfor
```

5) 对 $\boldsymbol{W}_{\text{norm}}$ 按以下方法进行阈值化处理,即 $\boldsymbol{W}_{\text{norm}}$ 中大于或等于某个阈值 $T$ 的元素值保持不变,否则将其置为零;

```
For i = 1 To m
    {
    If(W_norm[i] < T)
    {
    W_feature[i] = 0;
    }
    Else
    {
    W_feature[i] = W_norm[i];
    }
    Endif
    }
Endfor
```

以上算法中取小波系数幅值进行处理,是为了降低类内变化性,提高对光照的适应性. 将向量 $\boldsymbol{W}$ 归一化到[0,1]之间,一方面是为了提高数据值比较小的属性对分类的贡献;另一方面可以减少计算量,节约计算资源. 对 $\boldsymbol{W}_{\text{norm}}$ 进行阈值化处理是因为比较小的系数幅值一般代表可能的噪音或表面亮度比较均匀的细节信息,而这部分细节信息对分类的贡献较小. 此外,经阈值化处理后得到的小波特征向量 $\boldsymbol{W}_{\text{feature}}$ 中包含着大量的零元素,这样的向量被称为稀疏向量,它可以大大节约存储空间.

#### 1.1.2 训练样本重构

以上对摩托车训练样本和非摩托车训练样本经特征提取得到的数据集分别记为 $P$ 和 $N$,$P_{\text{num}}$ 和 $N_{\text{num}}$ 分别表示摩托车样本和非摩托车样本的数目,记 $k=\left[\frac{N_{\text{num}}}{P_{\text{num}}}\right]$. 对非摩托车样本数据进行间隔抽样处理,可用伪码描述如下:

For $i=1$ To $k$

{

Set($i$) = null; //初始化第 $i$ 个非摩托车样本子集为空集

For $j=i$ Step $k$ To $N_{num}$

{ //从第 $i$ 个数据开始,每隔 $k$ 个数据抽样一次放入到第 $i$ 个非摩托车样本子集 $N\_i$ 中

Set($i$) = Add(Set($i$), $N(j)$);

}

Endfor

}

Endfor

//数组 Set 中即为经抽样所得到的 $k$ 个非摩托车样本子集

间隔抽样的好处是在不丢失非摩托车样本信息的前提下,能较好地维持原始非摩托车样本的分布规律,且操作便捷. 将以上方法得到的 $k$ 个非摩托车样本子集分别与摩托车样本集 $P$ 组成 $k$ 个子训练集,用于训练相应的 $k$ 个 SVM 分类器.

### 1.1.3 训练 SVM 分类器

支持向量机(SVM) 是基于统计学习理论(Statistical Learning Theory, SLT)的机器学习算法,最初是针对两类模式分类问题而提出来的,它改变了传统的经验风险最小化原则,是根据结构风险最小化原则提出的,因此具有很好的泛化能力. 另外,支持向量机还通过引入核方法,将分类问题归结为解一个二次规划(Quadratic Programming, QP)问题,从而有效地克服了高维以及局部极小问题,并很好地解决了非线性分类问题. 因此,被广泛应用于模式识别及其他领域[8-9].

针对两类分类问题,设有两类共 $l$ 个样本:

$$(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_l,\boldsymbol{y}_l),$$
$$\boldsymbol{x}_i\in\mathbf{R}^N,\boldsymbol{y}_i\in\{-1,+1\},i\in[1,l]. \quad (1)$$

SVM 寻找一个最优分类面,使它不但能将两类无错误地分开,而且要使两类的分类间隔最大. 前者是保证经验风险最小;后者实际上就是使推广性的界中的置信范围最小,从而使真实风险最小. 这等同于使结构风险最小化,从而达到较好的泛化性能. SVM 分类超平面定义为

$$f(\boldsymbol{x}) = \left(\sum_{i=1}^{l}\boldsymbol{y}_i\,\alpha_i\,k(\boldsymbol{x},\boldsymbol{x}_i)+b\right). \quad (2)$$

式(2)中计算所得 $f(\boldsymbol{x})$ 的值为广义置信度,$\alpha_i$ 可以通过求解二次规划问题得到,$k(\boldsymbol{x},\boldsymbol{x}_i)$ 为核函数,包括线性核函数、多项式核函数以及径向基核函数(Radial Basis Function, RBF)等. 构造最优分类面等价于找出所有非零 $\alpha_i$,与其对应的 $\boldsymbol{x}_i$ 即为最优分类面的一个支持向量.

对以上训练样本重构得到的各训练子集,分别训练相应的 SVM 分类器,本文各 SVM 分类器所选用的核函数均为径向基核函数(Radial Basis Function, RBF). SVM 分类器所需选择参数为惩罚系数 $C$ 和 RBF 核函数宽度 $\gamma$. 实验中采用 5-fold 交叉验证法来进行参数选择,即将训练样本子集分成 5 份,任选其中 4 份作为训练集,取剩下 1 份作为测试集,设定参数 $C$ 和 $\gamma$ 的取值范围. 选取最高平均识别精度对应的参数对 $C$ 和 $\gamma$,若最高平均识别精度对应几个不同的参数对,取最少平均支持向量个数对应的参数对 $C$ 和 $\gamma$,这是因为支持向量机的计算复杂度为 $O(L\times D)$,其中 $L$ 代表支持向量个数,$D$ 代表特征向量维数,$L$ 越小所需计算量越少,泛化性愈好. 利用训练得到的各个子 SVM 分类器,就可用于在线识别阶段的摩托车识别了.

## 1.2 在线识别阶段

在线识别阶段如图 2 所示,用于判定 ROI 中摩托车的存在性.

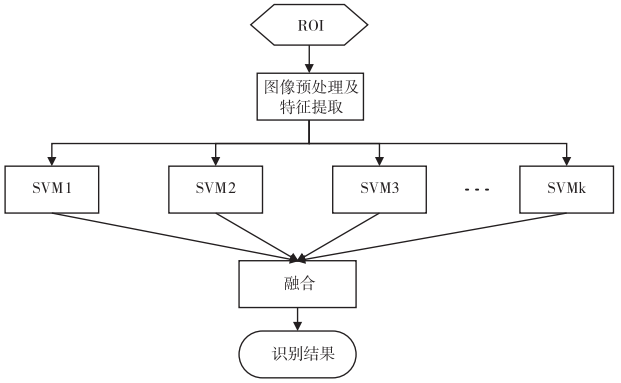


图 2 在线识别

Fig. 2 On-line recognition

首先对 ROI 应用离线训练阶段所采用的图像预处理及特征提取方法得到特征向量,将其分别输入到训练阶段得到的各个 SVM 分类器;然后对各分类器的广义置信度融合,即可得到最终的识别结果. 目前两类分类中融合方法比较常用的有投票法、取最大值法和求和的方法,文中按公式(3)采用求和的方法来进行融合.

$$R = \operatorname{sign}\left(\sum_{i}^{k} f_i(x)\right). \quad (3)$$

这是因为:第一,投票的方法虽然综合了所有SVM分类器的分类结果,但融合时对所有结果是等权值对待的,从而导致易受分类能力较差的分类器的干扰;第二,求最大值方法意味着一旦最大值对应的分类器识别出现错误,无论其他分类器分类结果正确与否,均无法纠正此项错误;最后,求和的方法综合了所有SVM分类器的分类结果,同时各个分类器输出结果对最终分类结果的贡献依赖于其广义置信度的绝对值的大小,绝对值大的分类贡献大,反之对分类的贡献则小,相当于进行了加权处理,因而不易受分类能力较差分类器的干扰.

还有一个很重要的环节是系统的再学习,这对于一个初步设计好的算法识别系统来说,是提高检测率和降低误识别率的必经步骤,因此需要逐步地、有耐心地进行. 通常的方法是将分割算法得到的ROI经图像预处理和特征提取后送入系统的在线识别部分,将最终识别结果中那些误报为包含非摩托车的ROI作为新的摩托车样本送到对其未正确识别的SVM分类器所对应的摩托车训练样本子集中去,将那些误报为包含摩托车的ROI作为新的非摩托车样本送到对其未正确识别的SVM分类器所对应的非摩托车训练样本子集中去;然后,根据新的训练样本子集,对相应的支持向量机重新进行训练. 系统的再学习可以不断地进行下去,直到其性能满意为止. 这样做的好处是在摩托车和非摩托车样本的选择上,在提高分类器性能的同时,可以很好地控制训练样本子集的规模.

## 2 实验结果及分析
Experimental results and analysis

为了验证文中方法对白天基于单目视觉静态图像中的自车后方摩托车的识别性能,将文献中常用特征提取方法与本文构造的特征提取方法结合SVM分类器应用到本文的路面对象识别系统中. 文中共收集了7 218个训练样本,其中包括1 150个摩托车样本,6 068个非摩托车样本;4 819个测试样本,其中包括1 349个摩托车样本,3 470张非摩托车样本. 图3为摩托车和非摩托车训练样本样例.

F-Measure(记为$F_m$)能较好地评价分类器基于不平衡样本集的识别性能,当分类器对摩托车识别性能较差时,相应的$F_m$值就低. 实验中采用$F_m$来评价不同算法的摩托车识别性能,其定义如式(4)所示.

a. 摩托车样本

b. 非摩托车样本

图3 摩托车和非摩托车训练样本样例
Fig. 3 Examples of motorcycle and non-motorcycle training samples

$$F_m=\frac{2P_r\times R_e}{P_r+R_e}. \tag{4}$$

其中,$P_r$表示查准率,$R_e$表示查全率,定义如下:

$$P_r=\frac{N_{TP}}{N_{TP}+N_{FP}};\quad R_e=\frac{N_{TP}}{N_{TP}+N_{FN}}. \tag{5}$$

式(5)中的$N_{TP}$、$N_{FP}$以及$N_{FN}$分别表示测试样本中算法正确检测的摩托车数目、将非摩托车误识别为摩托车的个数以及将摩托车误识别为非摩托车的个数. 本文共组织了以下4组实验.

实验1. 基于原始训练样本集数据训练SVM来进行识别,评估结果如表1所示.

实验2. 利用Under-Sampling[10]抽样方法从非摩托车训练样本数据中随机选取一些样本与摩托车样本数据组成训练集,其中选取的非摩托车样本数量与摩托车样本数量相等,然后训练SVM来进行识别,评估结果如表2所示.

实验3. 利用Hybrid-Sampling方法重构训练集,即先用SMOTE[11](是一种常用的上取样方法(Over-Sampling),它通过在小类和其同类近邻间插值生成

人工样本的方式扩大小类)方法将摩托车样本数量增加 1 倍,然后用 Under-Sampling 方法选取与摩托车样本数量相等的非摩托车样本,最后将所有的摩托车样本和选取的非摩托车样本组成新的训练集训练 SVM 来进行识别,评估结果见表 3.

实验 4. 应用文中所提出的样本重构方法得到的若干训练子集训练相应的 SVM 分类器来进行识别,评估结果如表 4 所示.

表 1 ~4 中的黑体为每个评估指标的最好结果.

**表 1 单一 SVM 分类器的评估结果**

Table 1 Results of single SVM imbalanced set

| 特征提取方法 | $P_r$ | $R_e$ | $F_m$ |
|---|---|---|---|
| PCA[1-2] | 0.883 | 0.401 | 0.551 |
| Gabor[6] | 0.925 | 0.731 | 0.817 |
| Wavelet[5] | 0.879 | 0.491 | 0.630 |
| Wavelet + Gabor[7] | 0.934 | **0.732** | **0.820** |
| 本文特征提取方法 | **0.938** | 0.677 | 0.786 |

**表 2 Under-Sampling 情形下单一 SVM 分类器评估结果**

Table 2 Results of single SVM by under-sampling method

| 特征提取方法 | $P_r$ | $R_e$ | $F_m$ |
|---|---|---|---|
| PCA[1-2] | 0.726 | 0.694 | 0.710 |
| Gabor[6] | 0.800 | 0.860 | 0.829 |
| Wavelet[5] | 0.708 | 0.749 | 0.728 |
| Wavelet + Gabor[7] | 0.806 | 0.903 | 0.852 |
| 文中特征提取方法 | **0.831** | **0.915** | **0.871** |

**表 3 Hybrid-Sampling 情形下单一 SVM 分类器评估结果**

Tab. 3 Results of single SVM by hybrid-sampling method

| 特征提取方法 | $P_r$ | $R_e$ | $F_m$ |
|---|---|---|---|
| PCA[1-2] | 0.803 | 0.597 | 0.685 |
| Gabor[6] | 0.841 | 0.830 | 0.835 |
| Wavelet[5] | 0.783 | 0.638 | 0.703 |
| Wavelet + Gabor[7] | 0.869 | **0.854** | 0.861 |
| 文中特征提取方法 | **0.906** | 0.846 | **0.875** |

**表 4 基于多 SVM 分类器的评估结果**

Tab. 4 Results of SVM ensembles

| 特征提取方法 | $P_r$ | $R_e$ | $F_m$ |
|---|---|---|---|
| PCA[1-2] | 0.749 | 0.683 | 0.714 |
| Gabor[6] | 0.808 | 0.898 | 0.851 |
| Wavelet[5] | 0.735 | 0.763 | 0.749 |
| Wavelet + Gabor[7] | 0.824 | 0.901 | 0.861 |
| 本文特征提取方法 | **0.857** | **0.916** | **0.886** |

由表 1 中的查全率 $R_e$ 值可以看出,受样本不平衡的影响,使得所有方法对摩托车的识别效果受到严重影响;由表 2 和表 3 的 $R_e$ 值可以看出,采用Under-Sampling 和 Hybrid-Sampling 方法对训练数据重构以后,摩托车的识别效果均得到明显改善,且由 $F_m$ 值可以看出,应用 Hybrid-Sampling 方法的识别性能整体上要好于 Under-Sampling 方法,这是因为 Under-Sampling 方法丢失的潜在对分类有贡献的样本数据信息要多于 Hybrid-Sampling 方法. 由表 2 ~4 中的查全率 $R_e$ 和查准率 $P_r$ 对比效果可以看出:本文所提出的对不平衡样本数据的解决方法对摩托车的识别性能整体上要好于 Under-Sampling 和 Hybrid-Sampling 方法;本文所提出的特征方法分类识别时的对应的 $F_m$ 值在表 2 ~4 中均为最高,对应的查全率 $R_e$ 值在表 2 和表 4 中为最高,由此可以看出本文提出的特征提取方法要比其它特征提取方法更有助于提高整个算法的识别性能.

此外,基于支持向量机的系统具有再学习的功能,通过不断地增加训练样本,还可以进一步提高系统的识别性能. 图 4 是本文方法对后视摩托车成功识别的例子.

图 4 不同数量摩托车被成功检测的例子
(白色矩形框表示识别的摩托车)

Fig. 4 Examples of successful detection by the proposed method (white bounding-boxes referring to the detected motorcycles)

## 3 结论
## Conclusion

在道路障碍物识别中,涉及对摩托车识别的文

献算法所述不多,针对摩托车识别中所存在的样本数据不平衡以及如何较好地描述摩托车图像特征的问题,本文提出了一种基于 HSV 颜色空间的 Haar 小波特征和多 SVM 分类器的摩托车图像识别方法,所构造的特征提取方法能有效增加类间距离、增强类内聚类效果以及提高分类器的抗噪能力;所提出的样本重构方法较好地解决了样本不平衡对摩托车识别造成的影响. 实验结果表明:本文识别算法不仅具有较好的抗噪能力,而且较好地解决了样本数据不平衡情形下对摩托车的识别问题,具有较好的应用和推广价值.

## 参考文献

References

[1] Matthews N D, An P E, Charnley D, et al. Vehicle detection and recognition in greyscale imagery[J]. Control Eng Practice, 1996, 4(4):473-479

[2] Sidla O, Paletta L, Lypetskyy Y, et , al. Vehicle recognition for highway lane survey[C]//The 7th International IEEE Conference on Intelligent Transportation Systems, 2004:531-536

[3] Goerick C, Detlev N, Werner M. Artificial neural networks in real-time car detection and tracking application[J]. Pattern Recognition Letters, 1996, 17(4):335-343

[4] Papageorgiou C, Poggio T. A trainable system for object detection [J]. International Journal of Computer Vision, 2000, 4(4):15-33

[5] Schneiderman H. A statistical approach to 3D object detection applied to faces and cars[C]// Proceedings IEEE Conference on Computer Vision and Pattern Recognition, CMU-RI-TR-00-06, 2000:746-751

[6] Sun Z, Bebis G, Miller R. On-road vehicle detection using Gabor filters and support vector machines[C]//IEEE 14th International Conference on Digital Signal Processing, 2002:1019-1022

[7] Sun Z, Bebis G, Miller R. Improving the performance of on-road vehicle detection by combining Gabor and wavelet features[C]// The IEEE 5th International Conference on Intelligent Transportation Systems, 2002:130-135

[8] Seyedi V A, Haji S S, Masoud R A. Enhancing automatic speed estimation systems performance using support vector machines[C] //ICCP IEEE International Conference on Intelligent Computer Communication and Processing, 2009:185-188

[9] Cui B, Xue T, Yang K. Vehicle recognition based on support vector machine[C]// Intelligent Information Technology Application Workshops, IITAW, 2008:443-446

[10] Weiss G M, Provost F. The effect of class distribution on classifier learning[R]. Tech Rep Department of Computer Science, Rutgers University, 2001

[11] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16:321-357

# An algorithm based on imbalanced data sets for motorcycle recognition

WEN Xuezhi[1] ZHENG Yuhui[1] ZHAO Yingnan[1] WU Yi[1]

1 School of Compuer and Software, Nanjing University of Information Science and Technology, Nanjing 210044

**Abstract** A motorcycle recognition algorithm based on Haar wavelet features of HSV space and SVM ensembles is proposed to solve the poor performance of motorcycle recognition generated by imbalanced data. At first, a feature extraction algorithm based on unsigned wavelet coefficients of HSV space is presented. Then a reconstruction approach is applied to training data so as to generate several subsets, and several SVM classifiers are trained based on all subsets respectively; the final recognition result is obtained by aggregating the outputs of all SVM classifiers. Experimental results demonstrate that the proposed recognition approach has better performance and robustness than the current methods and shows promising value and prospect for popularization especially on occasions where classification performance is badly affected by imbalanced data in practical application.

**Key words** motorcycle recognition; feature extraction; imbalanced data; SVM(Support Vector Machine)