文本页面图像的图文分割与分类算法

王加俊 黄贤武 郭玮玮 仲兴荣
(苏州大学电子信息学院,苏州 215021)

摘 要 为了能对包含不规则图片区和表格的倾斜文本页面图像进行图文分割与分类,提出了一种新的图文分割和分类算法。该算法先采用数学形态学和分级霍夫变换来进行文本倾斜的检测和校正;然后为了使算法能够对包含不规则图片区的文本页面图像进行处理,提出在传统的投影轮廓切割算法中,引入中点切割的过程,以便利用一系列的矩形来近似地逼近不规则的图片区。对于分割后的图像,则提出利用黑白像素比(R_{bw})和近邻像素间的交叉相关性(R_{cc})两个特征来作为分类的判据。实验结果证明,算法速度快、可靠性高。该算法只适用于二值图像。

关键词 文本图像 形态学 图像分割 霍夫变换
中图法分类号: TP391.4 O4 **文献标识码**: A **文章编号**: 1006-8961(2004)05-0571-07

Page Segmentation and Classification Algorithm for Document Images

WANG Jia-jun, HUANG Xian-wu, Guo Wei-wei, ZHONG Xing-rong
(*School of Electronics and Information Engineering, Soochow University, Suzhou* 215021)

Abstract In this paper, a system valid of the segmentation and classification of skewed document images with irregular graph regions and form regions is proposed. In this system, the skew angle of the document images is detected with a novel algorithm based on the morphological operation of Hit-or-Miss and the hierarchical Hough transform. The former (Hit-or-Miss operation) is for the detection of the baseline points while the latter (Hough transform) is for the detection of the skew angle of the baseline which is also of the page image. To make the system valid for the document images with irregular graph regions involved, we proposed to introduce a middle point cut process to the traditional projection profile cut algorithm so that the irregular graph regions can be approximated with a lot of small rectangles. The segmented regions are classified with two features of the black to white ratio and the cross correlation between adjacent pixels of the sub-blocks. Experimental results have proved the fastness and the reliability of the system proposed in this paper.

Keywords document image, morphological operation, image segmentation, hough transform

1 引 言

由于电子文件相对于传统的纸张文件来说不仅存储安全,而且检索方便,传输快捷,因此,将现有的纸张文件转换成电子文件具有非常重要的实际意义,但要完成这种转换,则必须对文本页面进行理解。众所周知,光字符识别系统(OCR)只能对文字部分进行识别,然而,由于文本图像的页面形式可以多种多样,可能包含图片、图表等非文字区域,因此为了提高文字识别的效果,有必要在进行字符识别前,将文字区和非文字区加以分割和区别。对于文本页面图像分割与分类算法的研究最早可以追溯到20 世纪 80 年代[1],已有的算法大体上可以分成自顶向下和自底向上两类。其中自顶向下算法的优点是速度快,缺点是要对页面有一定的先验知识,其代表性的算法有投影轮廓切割法(PPC)[2]等;自底向上的算法的优点是不需要对页面有先验知识,缺点是耗时较多,其代表算法有行程平滑法[3]、近邻线密度法[4]、连通分量分析法[5]等。由于文本图像中的不同区域(文字、图片、图表等)在纹理特征上存在显著的差别,因此对文本图像的分割还可以建立在对

基金项目:江苏省教育厅自然科学基金项目(L0112419925);江苏省自然科学基金项目(BK2001137)
收到日期:2002-12-17;**改回日期**:2003-12-19

文本的纹理特征进行分析的基础之上[6~8]。该类方法的优点是它与图文区域的形状无关，其缺点是算法较为复杂和需要较长的处理时间。

普通的 PPC 算法是建立在印刷区域主要由矩形块组成这样的假设基础之上的。该方法的主要思路是，首先得到页面图像在 X 轴和 Y 轴上的投影轮廓(Projection Profile)；然后在这些投影轮廓上寻找较深的谷点，再在这些谷点的位置对页面图像进行切割；最后对由此得到的每一个图像块重复以上过程，直到不能分割为止。但是，当页面图像中含有非矩形的图片区时，由于投影轮廓没有明显的谷点，致使算法失去作用，同时，该算法对文本的倾斜也很敏感，为此，本文首先提出了一种基于形态学运算和霍夫变换的倾斜检测算法来对文本图像进行倾斜校正；然后，发展了一种新的分割算法，即通过在普通的 PPC 算法中引入中点切割，使得该算法可以对含有非矩形图片区域的页面图像进行分割。本文选用了黑白像素数目比 R_{bw} 以及相邻像素间的交叉相关性(Corss-Correlition)R_{cc} 两个特征来对分割后的图像区域进行分类，同时将分割和分类过程加以结合，从而提高了处理速度和分类准确度。实验证明，该算法对中英文文本页面图像具有较好的分割和分类效果。

2 倾斜检测与校正

文本页面图像的倾斜检测与校正，对于以投影轮廓切割为基础的页面分割和分类算法来说是至关重要的，因为这类算法不能对倾斜的文本进行处理。为了能够使该类算法适用于倾斜文本的情形，有必要发展快速而可靠的文本倾斜检测和校正算法。通过分析发现，文本页面具有一定的特殊性，由于其中的文本行都是按照一定的规则有规律地排列的，所以文本行具有特别的意义，一方面，文本行的走向表示了文本页面图像的大体走向；另一方面，文本行最后一行黑像素所在直线(本文称之为文本行的基线，如图 1 所示)的走向代表了文本行的走向，因而也代表了文本页面的走向。基于以上事实，本文提出一种快速而新颖的倾斜检测和校正算法。该算法分成 3 步：第 1 步通过形态学的运算找出文本行的最后一行黑像素(称之为基线点)；第 2 步通过将提取出的基线点作为分级霍夫变换的输入来找出文本行的基线及其倾斜角；第 3 步利用上述倾斜角来对文本页面图像进行倾斜校正。

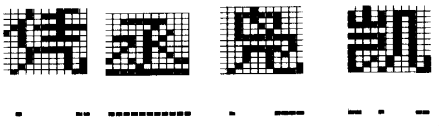


图 1 点阵汉字组成的文本行及其基线点

2.1 基于形态学的文本基线点的检测

这里的基线点是指文本行的最后一行黑像素的集合。图 1 显示了一张由 4 个字组成的文本行和它的基线点，(为了说明方便，选用了点阵型的汉字)。

本文提出应用形态学中的击中与击不中变换来检测文本图像中的基线点。设用集合 $\boldsymbol{X}$ 来表示文本图像中所有前景像素(黑像素)的集合，而用 $\boldsymbol{X}^c$ 来表示所有背景像素的集合，结构元素对用 $\boldsymbol{Y}=(\boldsymbol{E},\boldsymbol{F})$ 来表示，则击中与击不中变换(HMT)定义为

$$\boldsymbol{X} * \boldsymbol{Y} = (\boldsymbol{X} \ominus \boldsymbol{E}) \cap (\boldsymbol{X}^c \ominus \boldsymbol{F}) \tag{1}$$

其中，$*$ 表示 HMT 运算符，$\ominus$ 表示形态学的腐蚀运算，$\boldsymbol{E}$ 表示击中结构元素，$\boldsymbol{F}$ 表示击不中结构元素，且

$$\boldsymbol{E} \cap \boldsymbol{F} = \varnothing \tag{2}$$

其中，$\varnothing$ 表示空集。从击中击不中变换的定义可以看出，当且仅当 $\boldsymbol{E}$ 平移到可以填入前景集 $\boldsymbol{X}$ 的内部，$\boldsymbol{F}$ 平移到可以填入背景集 $\boldsymbol{X}^c$ 的内部时，该变换才可能有输出。同时，由图 1 可以看出，由于文本的基线点实际上是图像中每一文本行最下面前景与背景交界处的前景像素点，因此，如果选择适当的结构元素对，并将击中结构元素与击不中结构元素相邻排列，则可以对文本行的基线点进行检测。经过反复实验，利用图 2 所示的结构元素对来对文本行的基线点进行检测，可以取得较好的效果，图中，● 表示击中结构元素，○ 表示击不中结构元素，△ 表示坐标原点所在位置。

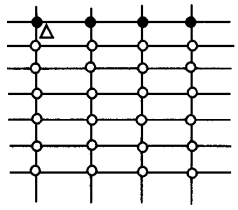


图 2 结构元素对示意图

然而，这种基于击中与击不中变换的文本行基线点的检测方法却受到如下因素的干扰：如在诸如汉语文字“二”及大写英文字母“E”，“F”等中存在这样的点，这些点满足击中击不中变换的输出条件，但是它们却不是基线点。为了消除这种干扰，本文提出

在利用击中击不中变换提取基线点之前,先用一个 5×5 的正方形结构元素来对原文本页面图像进行膨胀运算,以达到填补上述情形下文字和子母中的空洞的目的。

2.2　基于霍夫变换的文本行基线的检测

众所周知,霍夫变换是对图像进行线条检测的有效方法,它不仅具有很强的抗干扰能力,而且能对不连续的线条进行检测,但是由于在变换中需要进行大量的浮点运算,因此这种方法需要较大的运算量。若霍夫变换为从直角坐标(x,y)到极坐标(γ,θ)的变换(如图 3 所示),则其算法可总结如下:

(1) 生成判断数组 $A(\gamma,\theta)$,并初始化;

(2) 进行坐标变换;

for every pixel in the line

for$(\theta=\theta_{\min};\theta<\theta_{\max};\theta+=\theta_{\text{step}})$

$\{\gamma=x\cos\theta+y\sin\theta;$

$A(\gamma,\theta)++;\}$

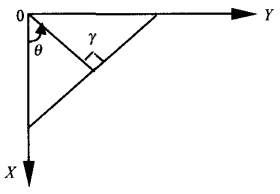


图 3　霍夫变换及其坐标系

(3) 对 $A(\gamma,\theta)$进行峰值检测。

从以上的算法可以看出,霍夫变换的运算量与参与变换的像素点的数目、变换中所取角度间隔 θ_{step},以及检测范围$(\theta_{\max}-\theta_{\min})$有关,也就是说,当参与变换的点数越多时,则计算量就越大;当变换中所使用的角度间隔 θ_{step}越小,即精度越高时,则算法所需要的运算量就越大。本文利用霍夫变换来由基线点检测出文本行的基线。由于此时参与霍夫变换的只有基线点,其像素点的数目远小于整个图像中像素点的数目,因此就大大减少了算法的运算量。为了进一步减小运算量,同时不致于降低对基线检测的精度,本文提出分级霍夫变换的方法,即首先采用低分辨率的霍夫变换,检测出 θ 可能的范围,然后在这个范围之内逐步减小 θ_{step},以达到所要求的精度。

2.3　文本图像的倾斜校正

利用霍夫变换不但可以检测出文本行的所有基线,而且可以得到这些基线的倾斜角度。校正时,可以先求其平均值和方差,然后去除掉那些与均值的偏差大于方差的角度,再重新计算基线倾斜角度的均值,并以此作为文本图像的倾斜角度,最后将文本图像旋转相应的角度,即可完成对文本图像的倾斜校正。

3　文本图像页面分割算法

为了能对包含不规则图片区的文本页面图像进行处理,本文提出一种改进的递归投影轮廓切割算法。为了叙述方便,定义沿某一条直线的投影轮廓为位于该直线上的所有像素值的和。基于投影轮廓,Nagy 等提出一种分割算法[2],这种算法首先寻找投影轮廓的谷点,然后在谷点所在的位置对文本图像进行切割,但是这种方法容易受文本图像的倾斜和页面图像中的非矩形图片区的影响,因为在这两种情况下投影轮廓中不再存在明显谷点。由于文本图像的倾斜问题已经通过发展快速的倾斜校正算法得到解决,因此下面仅考虑包含非矩形图片区的文本图像页面分割问题。

利用几个预设的参数,本文所提出的分割算法步骤如下:

(1)对每一个图像块,分别计算 X 轴上(沿 Y 轴)和 Y 轴上(沿 X 轴)的投影轮廓。

(2)如果 X 轴上或 Y 轴上的投影轮廓有谷点,则在这些谷点所在的位置将文本图像切割成子块。否则转步骤 5。

(3)对于每一个子图像块,如果其高度 h 满足 $h<T_{\text{line}}$或其宽度 $w<T_{\text{line}}$,则该图像子块判为线条不再进行进一步的处理,其中 T_{line}为表征线条宽度的阈值。

(4)对于每一个子图像块(已判为线条的除外),如果其高度 h 满足

$$|h-m_{\text{height}}|<\sigma_{\text{height}} \tag{3}$$

且

$$h<T \tag{4}$$

则相应的子图像块被视为文字区,且对该子图像块将不再进行进一步的处理。这里,m_{height}和 σ_{height}分别为该子图像块所在的图像块中各子图像块(包括该子图像块在内)高度的均值和方差,T 为一个阈值,其可以由有关页面图像的先验知识(如页面的高度等)确定。如果所有的子图像块都可以看成文字区,则对该图像块的处理结束;否则转步骤 1。

(5)如果图像块的高度 h 满足

$$h > T_{\text{height}} \tag{5}$$

或者图像块宽度 w 满足

$$w > T_{\text{width}} \tag{6}$$

则分别在图像块长、宽的中点进行切割。这里 T_{width}、T_{height} 是设定的切分阈值。由此可以看出，这种方法实际上是用许多小的矩形来逼近一个不规则的非矩形区域，T_{width}、T_{height} 就是用来逼近不规则区域小矩形的最大宽度和高度。

值得注意的是，在算法的第 5 步对图像块进行了预判，这将为后面的分类算法提供可供利用的先验知识。由于投影轮廓算法运算量小，所以这种处理方法的速度非常快。

从上面对分割算法的讨论可以看出，本文对传统的 PPC 算法的改进主要是在分割过程中引入了中点切分的过程，即在本文算法中有谷点切分和中点切分两种切分的过程。一般地，如果文本页面图像含有多栏(不同的栏中的文本行在编排上可能会发生交错)或者含有非矩形的图形区，则投影轮廓中将不存在明显的谷点，而通过引入中点切分，即使投影轮廓中没有明显的谷点，文本页面图像也可以分割成不同的子图像块，因此，位于单一栏中的纯文字区可以通过连续中点切分来得到，此时该区域的投影轮廓中将存在谷点；而非矩形的图片区则最终可以用长宽不超过 T_{width}、T_{height} 的矩形来近似。显然，T_{width}、T_{height} 越小，对不规则区域的逼近结果越精确，其处理时间也越长。实验证明，以上方法快速且可靠。

4 文本页面的分类

当文本页面被切分成各个小区域以后，需要对各个区域进行分类，以便将其分成文字区和非文字区。这样就需要找到一种快速、有效的分类方法。由于各种变换域的分类方法，例如傅立叶变换、Gabor 变换，虽然非常有效，但是运算时间太长，因此，本文直接在空间域中进行特征提取并以此对页面图像进行分类。考虑到文字区和非文字区的一些特性明显不同，因此，它们可以作为进行分类的判决条件。这些特性选取如下：

(1)文本页面图像的文字区和图片区黑像素和白像素的比值 R_{bw} 是不同的，一般图片区的 R_{bw} 值要大于文字区；(2)文字区黑白像素的交叉相关性 R_{cc}，即黑像素到白像素和白像素到黑像素的变化往往大于图片区。以上两个特性均可作为分类的根据，其定义为

$$R_{\text{bw}} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w} f(i,j) \oplus (0)}{\sum_{i=1}^{h}\sum_{j=1}^{w} f(i,j) \oplus (1)} \tag{7}$$

$$R_{\text{cc}} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w} f(i,j) \oplus f(i,j+1)}{\sum_{i=1}^{h}\sum_{j=1}^{w} f(i,j)} \tag{8}$$

其中，h 和 w 分别表示图像块的高度和宽度，$f(i,j)$ 表示图像中 (i,j) 位置的像素值，$\oplus$ 表示异或运算。合理地选择 R_{bw} 和 R_{cc} 的阈值 T_{bw}，T_{cc}，就可以把文字区和图片区分开。但是因为文本页面的复杂性，固定不变的阈值显然是不合理的，因此，找到一个合理的 T_{bw}，T_{cc} 是问题的关键。本文提出一种动态确定阈值 T_{bw}，T_{cc} 的方法，该方法利用了分割部分对子图像块进行预分的信息。在分割阶段，如果某一子图像块，其高度 h 满足 $|h - m_{\text{height}}| < \sigma_{\text{height}}$，则该子图像块将被预分为文字区。因为上述图像块的特征量 R_{bw} 的均值及 R_{cc} 的均值反映了文字区的特性，所以本文将区分文字区与非文字区的判决阈值 T_{bw}，T_{cc} 分别取成 R_{bw} 的均值和 R_{cc} 的均值。实验结果表明，这样处理是合理有效的。

利用以上两个阈值，分类算法的步骤如下：

(1)计算子图像块的特征量 R_{bw} 和 R_{cc}。

(2)如果 $R_{\text{bw}} > T_{\text{bw}}$ 则该子图像块被划分为图片区，该子图像块的分类终止；否则，转步骤 3。

(3)如果 $R_{\text{cc}} > T_{\text{cc}}$，则该子图像块被划分为文字区，该子图像块的分类终止；否则，转步骤 4。

(4)子图像块被划分为图片区。

5 实验结果和讨论

为了验证所提算法的效果，本文用它来对一组 A4 幅面的杂志页面经过扫描后得到的页面图像(其分辨率为 300dpi)进行了处理实验，实验平台是 P Ⅲ 450 计算机。其对不同的页面图像处理结果见图 4、图 5、图 6、图 7 所示。图 4(a)给出的是原始中文文本页面图像，为了验证倾斜检测与校正算法，在扫描时，有意识地将文本页面倾斜了 45°；图 4(b)给出的是利用一个 5×5 的结构元素对原始图像进行膨胀以后的结果；图 4(c)给出的是利用击中击不中变换提取出的文本行的基线点；图 4(d)给出的是该汉语页面图像经过倾斜校正后的结果。大量的实验证明，

(a) 原始汉语页面图像

(b) 膨胀后的汉语页面图像

(c) 提取出的汉语文本行基线点

(d) 经过倾斜校正的页面图像

(e) 分割后的页面图像

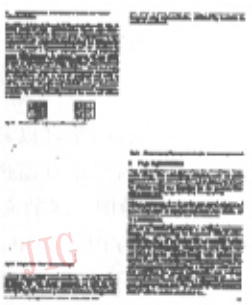
(f) 去除图片区后的页面图像

图 4　本文算法处理的中文页面图像

本文所提出的倾斜检测算法速度快、精度高,对于一幅 2 917×2 813 像素的文本页面图像,在 P Ⅲ 450 平台上倾斜检测和校正算法的运行时间大约 4s,检测精度为 1°。图 4(e)表示的是分割后的结果,显然,这里用了许多小矩形对不规则的图片区进行了近似。经过反复试验,几个阈值的取值如下:T 和 T_{height} 分别取页面图像高度的 1/15 和 1/40;而 T_{width} 取页面宽度的 1/15,T_{line} 取 6 个像素宽度。图 4(f)给出的是去除图片区后的页面图像。显然利用本文所提出的算法,已经正确地提取出了文本图像的文字区。

(a) 倾斜的英文页面图像

(b) 分割后的英文页面图像



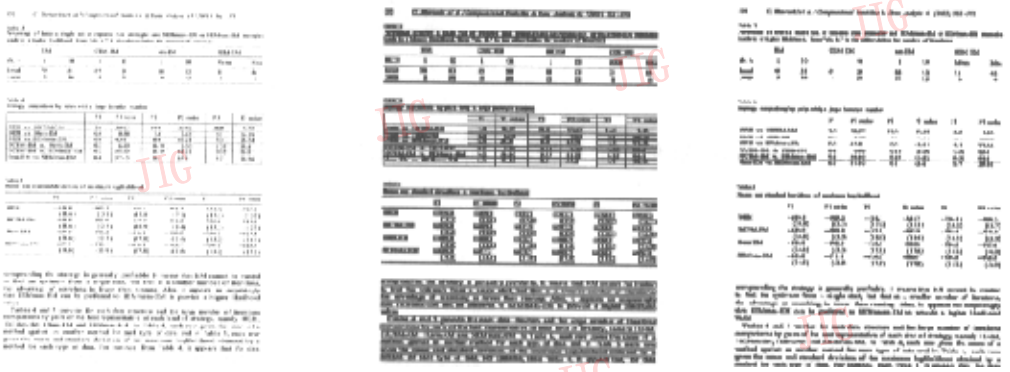
(c) 去除图片后的英文页面图像

图 5　本文算法处理的英文页面图像

本文给出的算法不仅适用于对汉语页面图像进行分割和分类，而且还能对英文页面图像进行正确的分割和分类。图 5(a)是一幅倾斜的英文页面图像。图 5(b)和图 5(c)分别给出了其分割和分类后的结果。和中文页面图像的处理结果一样，本算法也可以正确地取出其中的文字区。作为对比，图 6(a)和图 6(b)给出了利用传统的 PPC 算法对图 4(d)所示的页面图像进行分割和分类的结果。显然，由于在 PPC 分割算法中没有中点切割这一步骤，因此不能用许多小的矩形来近似不规则的图片区域，这就造成了与图片区域相邻的文字区域的误分，而这种现象在本文的算法中则得到了很好的避免。图 7(a)是一幅含有表格的页面图像，图 7(b)和图 7(c)给出的分别是对其进行分割和去除线条以后的结果。从这一结果可以看出，本文的算法基本上可以正确地取出表格中的文字，特别是对于没有竖线的表格仍可以完全正确地提取出文字。

(a) PPC 的分割结果　　(b) 去除图片后的结果

图 6　PPC 算法处理结果



(a) 含有表格的页面图像　　(b) 表格页面的分割结果　　(c) 去除线条后的结果

图 7　本文算法对含有表格的页面图像处理的结果

从上面的实验结果可以看出，本文所提出的算法不仅能对倾斜的文本页面进行处理，而且还能对含有不规则形状的图片区的文本页面进行正确的分割和分类。例如，在图 4(a)中，由于图片为非规则形状，且其周围布满了文字，因而不能用一个外接矩形将文字区和图片区分开。对于这种情况，传统的 PPC 算法无能为力，而本文通过引进中点切割的过程则能很好地实现图文的分割。同时，该算法也能对含有表格的文本页面图像进行基本正确的分割和分类。本研究的主要贡献在于，提出了一种快速而稳健的倾斜检测和校正算法以及一种改进的基于投影轮廓切割的分割算法。该算法的优点是速度快、可以实现对较为复杂的倾斜文本页面进行分割和处理；但该算法也存在一些缺点，例如它对页面中表格不能进行完全正确的分割和分类，如图 7(b)和图 7(c)所示，算法未能将含有竖线的表格的水平线条完全去

除,其原因是表格的线条距离文字太近,以及表格的竖线及英文字母向上突出和向下突出的部分对竖直投影在水平线条附近的谷点的准确定位产生了干扰,为了避免将文字部分作为线条去除,算法在执行过程中,由于忽略了此类谷点,因此导致结果中部分线条的残留。这一问题将在今后的研究中加以解决。

参 考 文 献

1 Abele L, Wahl F, Scherl W. Procedures for an automated segmentation of text, graphic and halftone regions in documents [A]. In: Proceedings of the 2nd Seandinavian Conference on Image analysis[C], Hellsinkii, Finland,1981: 177～182.

2 Nagy G, Seth S C. Document analysis with an expert system [A]. In: E. S. Gelsema and L. N. Kanal (Editors), Pattern Recognition Practice , Elsevier Science Publishers B. V. (North-Holland), 1986:149～159.

3 Strouthopoulos, C Papamarkos, N. PLA using RLSA and a neural network [J]. Engineering Applications of Artificial Intelligence, 1999,**12**(2):119～138.

4 Kubota K, Iwaki O, Arakawa H. Document understanding system[A]. In: Proceedings of the 7th International Conference On Pattern Recogniton[C], Montreal, Canada, 1984:612～614.

5 Fletcher L A, Kasturi R A. A robust algorithm for text string separation from mixed text/graphic images[J]. IEEE Trans On Pattern Recognition and Machine Intelligence, 1998,**10**(6): 910～918.

6 Jain A K, Bhattacharjee S. Text segmentation using Gabor filters for automatic document processing[J]. Machine Vision and Applications, 1992,**5**(3):169～184.

7 Jain A K, Zhong Y. Page segmentation using texture analysis [J]. Pattern Recognition, 1996,**29**(5):743～770.

8 Deng S, Latifi S, Regentova E. Document segmentation using polynomial spline wavelet[J]. Pattern Recognition, 2001, **34**(12):2533～2545.

王加俊　1969 年生,1999 年获浙江大学工学博士学位,副教授。现主要从事图像处理、图像重建、模式识别等领域的研究工作,主持国家自然科学基金、省自然科学基金、以及省教育厅自然科学基金项目各一项,获江苏省科技进步二等奖一项。

黄贤武　1941 年生,1966 年毕业于南京大学,教授,博士生导师。现主要从事图像处理与模式识别等领域的研究工作。

郭玮玮　1980 年生,2001 年获苏州大学工学硕士学位。现主要从事模式识别的研究工作。

仲兴荣　1970 年生,2000 年获西安建筑科技大学工学硕士学位,讲师。现主要从事图像处理与模式识别等领域的研究工作。