基于小波域隐马尔科夫模型的文本图像子带分割方法

侯玉华,宋锦萍,周福娜,文成林,杨晓艺
(河南大学数学与信息科学学院,河南开封 475001)

摘 要: 本文在已有文献的基础上,通过分析不同子带小波系数之间的相关性,提出了一类基于小波域 HMT (Hidden Markov Tree)模型文本图像分割方法. 其基本思想是先在子带分类的基础上,综合考虑不同尺度上的分类,进行多尺度文本图像分割,最后根据后验像素信息对上述方法所得分割结果进行修正,得到优于已有文献的分割效果,而且在一定程度上减少了分割算法的计算量.

关键词: HMT 模型; 二维小波变换; 多尺度文本图像分割
中图分类号: TP391.4 **文献标识码**: A **文章编号**: 0372-2112 (2002) 08-1180-04

A New Document Segmentation Based on Subbands by Wavelet-Domain Hidden Markov Tree Models

HOU Yu-hua, SONG Jin-ping, ZHOU Fu-na, WEN Cheng-lin, YANG Xiao-yi
(College of Mathematics and Information Science, Henan University, Kaifeng, Henan 475001, China)

Abstract: We introduce a kind of new wavelet-domain HMT segmentation method, a finer to coarser HMTseg, which combine with the classification results of the three subbands for the 2-D wavelet transform. We demonstrate that the new method's performance is somewhat better than the raw segmentation in prevenient document. The advantage to our segmentation algorithms is that they can offer improved segmentation accuracy with smaller computational burden. Finally we introduce a new HMTseg method by updating the classification constrainedly, and then we get a better segmentation result for document image.

Key words: HMT model; 2-D wavelet transform; multiscale document image segmentation

1 引言

随着现代印刷技术的发展,排版版式越来越复杂,文本图像的预处理过程越来越重要,因此文本图像的分割已成为当今国内外图像处理领域的热门研究课题. 但真正有效的分割方法并不多,文献[1]中提出了一类小波域 HMT 多尺度分割方法,但为了便于处理,文中对模型做了简化:假定经小波变换后各尺度三个子带之间的小波系数是相互独立的,而实际上,由于小波变换的三个子带对应图像中的同一个数据块,因此三个子带之间就必然存在一定的相关性. 本文通过秩统计量相关性检验证实了这一点. 所以,文献[1]的假设不甚合理,所得原始分割的结果也不十分可靠,且计算量较大. 为了解决这些问题,在此基础上,提出了一类基于小波域 HMT 模型新的文本图像分割方法. 不仅得到了优于文献[1]的分割效果,而且在一定程度上减少了分割算法的计算量.

2 多尺度图像分割

图像分割的基本思想是将文本图像分成不同的区域,每个区域仅含有一种纹理. 而多尺度分割是按照某一原则,在不同的尺度上,把图像分成不同的窗口,随着尺度的增加窗口逐渐变小,再对每个窗口利用某种方法判断其纹理类型.

文献[1]采用的多尺度分割方法是每次将当前窗口的图像进行四等分,如果把整个原始图像视为 0 尺度上的一个数据块,将该数据块四等分,得到 1 尺度上的四个数据块,依此类推,在 j 尺度上原图像被分成 4^j 个数据块,每个数据块与上一尺度的数据块均形成嵌套,最终得到一个由二维数据块组成的四叉树结构(见图 1).

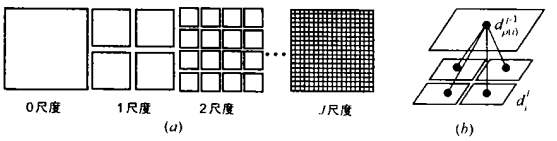

图 1 (a) 二维数据块分割的嵌套结构; (b) 二维数据块的四叉树结构,每个父结点对应下一尺度的四个子结点

文本图像分割的主要困难在于反映图像纹理象素的联合概率分布是未知的,而分割的可能区域又是多种多样且无法事先确定. 因此,人们通常采用某种方法将其转换到变换域中进行讨论. 由于小波变换具有检测局域突变的能力,且可以结合多尺度信息进行边缘检测,故而,通过小波变换利用小波系

收稿日期:2001-07-06;修回日期:2002-03-29
基金项目:河南省教委自然科学基金(No. 1999120003)

数及 HMT 模型进行图像分割是一种很好的图像分割方法. 类似的工作还可在文献[2,3]中见到.

3　二维小波变换[4]

与文献[1]类似,本文采用 Haar 小波变换. 对一个文本图像进行二维 Haar 小波变换可以解释为:利用四个二维小波滤波器,即局部平滑滤波器 $h_{LL}=\frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix}$,水平边界探测器,$g_{LH}=\frac{1}{2}\begin{pmatrix}1&1\\-1&-1\end{pmatrix}$,垂直边界探测器 $g_{HL}=\frac{1}{2}\begin{pmatrix}1&-1\\1&-1\end{pmatrix}$ 和对角边界探测器 $g_{HH}=\frac{1}{2}\begin{pmatrix}1&-1\\-1&1\end{pmatrix}$ 进行滤波.

为了计算一个由 $2^J\times 2^J$ 个像素组成的文本图像 x 的小波变换,首先令平滑图像为 $u_J(k,l):=x(k,l)$,$0\le k,l\le 2^J-1$,让 $u_J(k,l)$ 分别与四个滤波器 h_{LL},g_{LH},g_{HL},g_{HH} 做卷积,得到大小为 $2^{J-1}\times 2^{J-1}$ 的子带图像 u_{J-1},w_{J-1}^{LH},w_{J-1}^{HL},w_{J-1}^{HH},然后再对平滑图像 u_{J-1} 重复以上过程 $J-1$ 次,最终得到 u_0,w_0^{LH},w_0^{HL},w_0^{HH}. 其中的尺度系数矩阵 u_j,$0\le j\le J$,是由原始图片逐步平滑得到的. 小波系数矩阵 w_j^{LH},w_j^{HL},w_j^{HH},$0\le j\le J-1$,分别对应图像在水平、垂直、对角三个方向的小波系数. 分别以三个子带的 w_0^{LH},w_0^{HL},w_0^{HH} 为根,把以上过程所得结果倒提起来得到三棵四叉树,图 2 表示其中 LH 子带小波系数的四叉树结构.

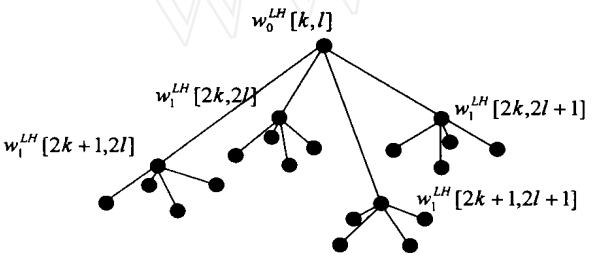

图 2　小波变换的四叉树结构

4　HMT 模型

HMT 模型是最近发展起来并被应用于图像处理的一种参数统计模型,利用此模型可将分布未知的小波系数问题归结为隐状态的确定问题. 当隐状态确定后每个小波系数的分布也随之确定. 由于文本图像边缘处小波系数较大,平滑处小波系数较小,因此本文中的隐状态指控制小波系数大小且不可观测的状态变量,记为 $S_i=m$,$m=S,L$ 分别对应小波系数取大值或小值时的隐状态取值. 同时假设这些状态变量组成的四叉树结构满足一阶马尔科夫树模型,故又称之为隐马尔科夫树(HMT)模型. 图 3 是隐状态的四叉树结构,其中黑点代表小波系数,白点代表隐状态. 每个父结点有四个子结点,且每个子结点的分布都由其父结点的分布确定. 称 $d_{i,m}^{\rho(i),m'}=p(S_i=m'|S_{\rho(i)}=m)$ 为由父结点 $S_{\rho(i)}$ 到子结点 S_i 的状态转移概率. 对应的状态转移矩阵为 $\begin{bmatrix}d_{i,S}^{\rho(i),S}&d_{i,L}^{\rho(i),S}\\d_{i,S}^{\rho(i),L}&d_{i,L}^{\rho(i),L}\end{bmatrix}$,其中 $d_{i,L}^{\rho(i),S}=1-d_{i,S}^{\rho(i),S}$,$d_{i,S}^{\rho(i),L}=1-d_{i,L}^{\rho(i),L}$. 需要说明的是小波系数本身之间的联系并不是一阶马尔科夫的,但每个小波系数的分布由其隐状态唯一确定,并假设在 $S_i=m$ 条件下 w_i 服从正态分布 $N(\mu_{i,m},\sigma_{i,m}^2)$,其中 $N(\mu_{i,m},\sigma_{i,m}^2)$ 是均值为 $\mu_{i,m}$,方差为 $\sigma_{i,m}^2$ 的 Gauss 分布,$m=S,L$. 进一步假设小波系数 w_i 的概率分布为二状态混合 Gauss 分布

$$f(w_i)=\sum_{m=S,L}p_{Si}(m)f(w_i|S_i=m) \tag{1}$$

其中 $p_{Si}(S)+p_{Si}(L)=1$.

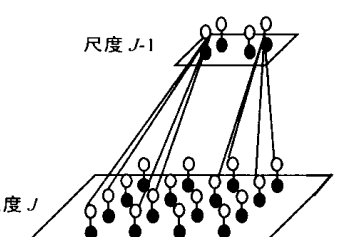

图 3　MHT 模型的四叉树结构,其中黑点代表小波系数,白点代表隐状态. 每个父结点与其四个子结点对应.

5　文献[1]中的原始分割

二维 Haar 小波变换的小波系数组成的四叉树结构与数据块组成的四叉树结构形成一一对应关系. 每个数据块 d_i 与以结点 w_i(有时简记为 i)为根结点的子树 T_i 对应. 文献[1]中假设小波系数的三个子带之间是相互独立的,故对某一种纹理来说,在某一尺度上任一数据块 d_i 的概率分布可以表成

$$f(d_i|M_c)=f(T_i^{LH}|\Theta^{LH})\cdot f(T_i^{HL}|\Theta^{HL})\cdot f(T_i^{HH}|\Theta^{HH}) \tag{2}$$

其中 $M_c=\{\Theta^{LH},\Theta^{HL},\Theta^{HH}\}$ 是纹理 c 对应的训练参数集,而 Θ^{LH},Θ^{HL},Θ^{HH} 分别是各个子带上的参数集,其包括各个尺度上的状态转移矩阵,状态概率矩阵,数学期望及方差阵. 根据文献[1],文本图像 x 在某个尺度上的数据块 d_i 的分类就是使得 $f(d_i|M_c)$ 最大的 c,即

$$c_i^{ML}=\arg\max_c f(d_i|M_c) \tag{3}$$

6　相关性分析

文献[1]在假设三个子带之间小波系数是相互独立的条件下,给出了基于小波域 HMT 模型文本图像的原始分割. 在引言中已经指出三个子带之间存在着一定的相关性,从下面的具体分析中可以看出三个子带之间小波系数确实是相关的,且随着尺度的加粗,其相关性逐渐增强.

下面利用秩统计量计算三个子带之间的相关系数,并进行相关性检验. 具体分析如下:从一个二维连续总体抽得一个容量为 n 的独立同分布样本 (X_i,Y_i),$i=1,2,\cdots,n$.

检验的原假设是 $H_0:X$ 和 Y 独立

备择假设是 $H_1:X$ 和 Y 相关

不妨假设样本 $X_1<X_2<\cdots<X_n$,则其所对应的秩向量[5]就是 $(1,2,\cdots,n)$,而相应 $Y_1,Y_2,\cdots,Y_n$ 的秩记为 $R_1,R_2,\cdots,R_n$. 当 X 和 Y 相关时,样本值 X_i,Y_i 有同时取最大值或最小值的倾向,因而秩统计量对 $(1,R_1),(2,R_1),\cdots,(n,R_n)$ 的相关系数取值较大,定义

$$S=\frac{\sum_{i=1}^{n}(i-\frac{n+1}{2})(R_i-\frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(i-\frac{n+1}{2})^2\times\sum_{i=1}^{n}(R_i-\frac{n+1}{2})^2}}=\frac{\sum_{i=1}^{n}(i-\frac{n+1}{2})R_i}{n(n^2-1)/12}$$

$$= 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (i - R_i)^2 \qquad (4)$$

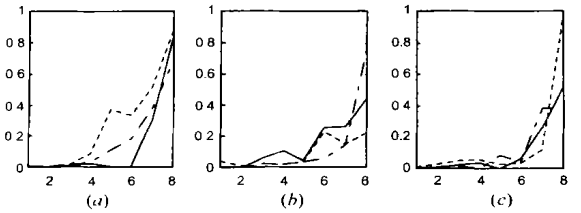

图 4 实线表示 HH 与 HL 子带的相关性,虚线表示 HL 与 LH 子带的相关性,点画线表示 LH 与 HH 子带的相关性,(a),(b),(c) 分别对应背景,文字和图片三种不同纹理.

当 S 较大时利于 H_1,S 较小时利于 H_0. 其中统计量 S 称为 Spear 秩相关系数. 如果把每一层上三个子带之间的小波系数分别看成是三个向量,则可以利用上述方法检验该尺度上三个子带间小波系数的相关性,同时可计算不同尺度上三个子带间的相关系数. 从图 4 中可以看出不同子带之间的相关系数随尺度的加粗而变大.

另一方面,根据以上分析,该检验有否定域$\{|S| \geq s_\alpha\}$,其中 s_α 是显著性水平为 α 的临界值. 由式 (4) 可以看出 S 是一个线性秩统计量,其否定域的临界值可由线性秩统计量的分布求得,当样本容量 n 较大时,$\sqrt{n-1}S$ 有渐近分布 $N(0,1)$,当 n 较小时,可以利用专用的 Spear 秩相关统计量临界值表进行检验. 对大量文本图象作假设检验,得检验 H_0 对 H_1 的显著性水平为 $\alpha = 0.05$ 的临界值如表 1 所示.

表 1 显著性水平 $\alpha = 0.05$ 时各个尺度上的临界值,相应的否定域为$\{|S| \geq s_\alpha\}$.

尺度 j	临界值 s_α
$j = 9$ (2 ×2 数据块)	2.9908×10^{-5}
$j = 8$ (4 ×4 数据块)	1.1964×10^{-4}
$j = 7$ (8 ×8 数据块)	4.7863×10^{-4}
$j = 6$ (16 ×16 数据块)	0.0119
$j = 5$ (32 ×32 数据块)	0.077
$j = 4$ (64 ×64 数据块)	0.2460
$j = 3$ (128 ×128 数据块)	0.5030

利用式 (4) 可以算出各个尺度上的 S 值,它们都落入了检验的否定域当中,根据假设检验理论,我们认为各子带的小波系数相关.

通过以上两个方面的分析,文献[1]中三个子带之间是独立的假设不甚合理,因此我们提出一类改进的 HMT 分割方法.

7 改进的 HMT 分割方法

7.1 基于原始分割的 HMT 分割

本文假设文本图像是由三种不同的纹理组成:背景、文字和图片. 并且在后续的例图中均以白色表示背景,黑色表示文字,灰色表示图片.

首先进行参数训练. 对给定的反映三种纹理的训练图像,用文献[2]中的 EM 算法,分别找出与之相匹配的最佳 HMT 模型参数集 $M = \{\Theta^{LH}, \Theta^{HL}, \Theta^{HH}\}$. 其次再利用小波变换给出待分割文本图像的小波系数. 最后利用上述已求出的 HMT 模型参数集 M,由式 (1) 分别计算该文本图像三个子带上根结点为 i 的三个子树 T_i^{LH},T_i^{HL},T_i^{HH} 的似然函数:

$$f(T_i^{HL} | \Theta^{HL}) = \sum_{m=S,L} \beta_i^{HL}(m)\, p^{HL}(S_i = m | \Theta^{HL}) \qquad (5)$$

$$f(T_i^{LH} | \Theta^{LH}) = \sum_{m=S,L} \beta_i^{LH}(m)\, p^{LH}(S_i = m | \Theta^{LH}) \qquad (6)$$

$$f(T_i^{HH} | \Theta^{HH}) = \sum_{m=S,L} \beta_i^{HH}(m)\, p^{HH}(S_i = m | \Theta^{HH}) \qquad (7)$$

其中 β_i^{LH},β_i^{HL},β_i^{HH} 为条件概率,是由文献[2]中的公式 $\beta_i^{LH}(m) := f(T_i | S_i = m, \Theta^{LH})$,$\beta_i^{HL}(m) := f(T_i | S_i = m, \Theta^{HL})$,$\beta_i^{HH}(m) := f(T_i | S_i = m, \Theta^{HH})$ 计算出来的;而 $p^{LH}(S_i = m | \Theta^{LH})$,$p^{HL}(S_i = m | \Theta^{HL})$,$p^{HH}(S_i = m | \Theta^{HH})$ 为状态概率,可在训练参数过程中得到.

下面利用上述公式分别对三个子带进行分类,并在此基础上对不定类型从细尺度到粗尺度逐级综合,最终确定各尺度上所有数据块的分类.

首先给出文献[1]中的原始分割方法在最细尺度的分割结果. 再由公式 (5) - (7) 对三个子带分别进行分类如下:

$$c_i^{LH} = \arg\max_c f(T_i | M_c^{LH}) \qquad (8)$$

$$c_i^{HL} = \arg\max_c f(T_i | M_c^{HL}) \qquad (9)$$

$$c_i^{HH} = \arg\max_c f(T_i | M_c^{HH}) \qquad (10)$$

其次,根据以上三个子带的分类结果作如下判断:

如果 j 尺度 ($j = 1, 2, \cdots, J-1$) 上的数据块 d_i^j 对应在各子带上的分类中两个或三个一致且都为 c,则该数据块 d_i^j 的分类就定为 c. 如果对应在各子带上的分类均不一致,则考虑该数据块在下一较细尺度 ($j+1$ 尺度) 上的四个子块对应在各子带上的分类情况,如果这四个子块在三个子带上共计十二个分类中,分类一致的个数所占比例大于 1/3,且记一致的分类为纹理 c',则 d_i^j 的分类就是 c';否则,再用更细尺度上的子块的分类来判断,直到最细尺度. 如果该过程进行到最细尺度仍然无法判断数据块 d_i^j 的分类,则采用文[1]中最细尺度上的原始分割结果来判断:即用以 i 为父结点的所有子结点在最细尺度上的原始分割结果中的大多数定为 d_i^j 的分类. 此时,如果仍有无法判断的点(统计结果表明这样的点极少),按以下方式作强制判断,在分类结果中,若背景与文字纹理各占一半,则 d_i^j 判断为文字,若图片纹理占一半,则 d_i^j 判断为图片.

这样,就得到了 j 尺度上各个数据块的分类结果. 对其他尺度采用同样的方法,确定各尺度上所有数据块的分类.

显然,这种从细到粗逐级综合的多尺度分割方法较好的利用了较细尺度上的信息,因此,一方面判断的准确性更高 (见图 5(d)). 另一方面,以上所述方法只需做出最细尺度的原始分割[1],以及其他各尺度每个子带上的似然函数值,而无需计算式 (2) 中的乘积,从而在一定程度上减少了计算量.

7.2 基于后验像素信息的 HMT 分割

本节先引入一种基于像素信息的初始分类:对大量文本图像像素的灰度值所做的统计结果表明,背景纹理的像素灰度值下界均大于 220,文字的像素灰度值上界均小于 129. 因

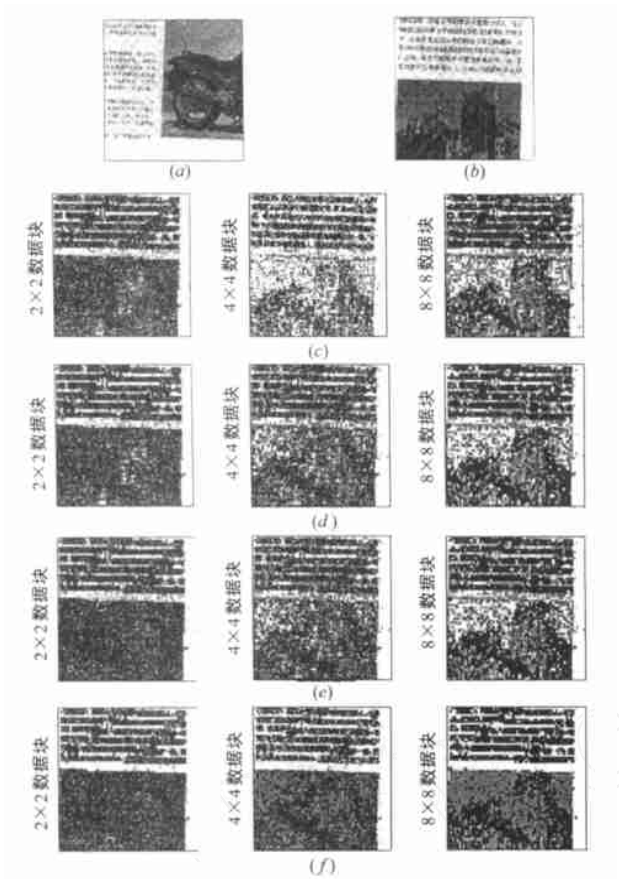

图 5　(a) 用于训练的文本图像; (b) 欲分割的文本图像; (c) 原始的 HMT 分割结果; (d) 利用 7.1 所述方法的分割结果; (e) 利用 7.2 所述方法的分割结果; (f) 利用 7.3 所述方法的分割结果

此,给定一个真实文本图像的灰度值矩阵 x,如果某一点 (i,j) 的像素灰度值 $x(i,j)$ 大于 220,则判断其为背景,如果某一点 (i,j) 的像素灰度值 $x(i,j)$ 小于 129,则判断其为文字,否则,判断为图片. 这样就得到像素级的初始分类. 接下来,给出一种新的基于后验像素信息的文本图像分割方法. 与 7.1 类似,从细到粗逐级综合确定各数据块 d_i 的分类. 所不同的是,当这个过程进行到最细尺度时如果还无法确定 d_i 的分类,则采用本节开始所述子结点对应像素级的初始分类的大多数来确定 d_i 的分类. 实例表明,该方法所得结果(图 5(e))比文献[1]中结果(图 5(c))有所改进,且不需计算文献[1]中的原始分割,从而进一步减少了计算量.

7.3　根据后验像素信息对上述分割结果进行强制更新

因为文本图像中的图片之中存在着大片光滑区域,这样在分割背景和图片时遇到一定的困难. 因此,将在 7.1 和 7.2 的基础上,根据后验像素信息进行各尺度上分类的强制更新,进而得到改进的 HMT 分割结果.

首先,在 7.2 的基础上给出各个尺度上的初始分类,具体过程如下:像素级的初始分类同 7.2;在相邻的粗尺度处,若数据块 d_i 的四个子块分类的大多数为背景,则视 d_i 的分类为背景,否则视其为图片. 重复此过程直到最粗尺度,从而得到各个尺度上的初始分类.

其次,根据这一初始分类结果,对 7.1 和 7.2 的分类结果进行强制更新:如果某个 d_i 的初始分类结果是背景,而子带分割结果不是背景,则把该点在子带分割结果中的分类强制更新为背景;如果 d_i 的初始分类结果是图片,而子带分割的结果是背景,则改变子带分割结果为图片. 与 7.1 和 7.2 中所述的方法相比,此改进的方法可以进一步提高背景和图片判断的准确性(见图 5(f)).

8　结论

本文通过分析子带之间的相关性,提出了一类基于小波域 HMT(Hidden Markov Tree)模型文本图像分割方法,得到了较好的分割效果,也减少了分割的计算量. 如果对上述分割再进行从粗尺度到细尺度的融合,将会得到更好的分割结果.

参考文献:

[1] Hyeokho Choi, Richard Baraniuk. Multiscale document segmentation using wavelet-domain hidden markov models [A]. in Proc. IST/ SPIE's 12th Annual International Symposium Electronic Imaging 2000, Science & Technology [C]. San Jose, CA: 2000.

[2] M S Crouse, R G Baraniuk. Wavelet-based statistical signal processing using Hidden Markov models [J]. IEEE Trans. Signal Proc. 1998, 46: 886 - 902.

[3] H Choi, R G Baraniuk. Image segmentation using wavelet-domain classification [J]. Proceedings of SPIE technical conference on Mathematical Modeling, Bayesian Estimation and Inverse problems, 1999, 3816: 306 - 320.

[4] 杨福生. 小波变换的工程分析与应用 [M]. 北京:科学出版社, 1999.

[5] 孙山泽. 非参数统计讲义 [M]. 北京:北京大学出版社, 2000.

作者简介:

侯玉华　女,1956 年 1 月出生于河南信阳市,教授,1987 年 3 月毕业于西北工业大学应用数学系,理学硕士,现任教于河南大学数学系,近几年主要从事统计图像处理与小波在图像处理中的应用等方面的研究.

宋锦萍　女,1963 年 3 月出生于河南开封市,副教授,1990 年 4 月毕业于中国科技大学数学系,理学硕士,现任教于河南大学数学系,近几年主要从事微分方程,小波与图像处理等方面的研究.

周福娜　女,1977 年 4 月出生,河南鲁山县人,河南大学数学与信息科学学院应用数学专业硕士研究生,研究方面为信号及图像处理.