基于稳健统计的矢量量化器设计算法

桑爱军 陈贺新
(吉林大学通信工程学院仪器系,长春 130022)

摘 要 LBG 算法作为矢量量化的基本算法具有经典意义,但由于在训练图象中,总存在少量的离群矢量,使得在训练码书时,码字的分布受到影响,进而使得压缩性能下降,因而不能充分体现出矢量量化的优越性能.而运用基于稳健统计的方法来设计矢量量化器,由于减少了码书中的离群矢量,同时加强了中心矢量在码书中的权重,因而不仅能够尽量减少码书的冗余,而且能大幅度提高压缩性能.实验结果显示,用基于稳健统计的设计方法设计的码书,其压缩性能比传统的 LBG 算法有了较大的改善,且恢复图象的主观、客观效果都是令人满意的.

关键词 图象处理(510·4050) 图象编码 矢量量化 稳健统计
中图法分类号: TN919.81 **文献标识码**: A **文章编号**: 1006-8961(2003)07-0829-05

Vector Quantizer Designing Based on Robust Statistic

SANG Ai-jun, CHEN He-xin
(*School of Communication Engineering, Jilin University, Changchun* 130022)

Abstract LBG algorithm is classical algorithm in Vector Quantization, which was proposed in 1980 by Y. Linde, A. Buzo, and R. M. Gray. Used this algorithm, an acceptable performance codebook can be get in the acceptable time, it has better performance than Scale Quantization that was proved by Shannon. But few outlines vector of the training image effect the distribution of the codeword in the codebook. To few vectors, it maybe noises, but have large codeword number in the codebook. And decrease the body vectors' codeword in the codebook. That decreases the compress ratio and makes the reconstruct image worse, the advantageous of the Vector Quantization can't be explained adequately. Different people used this algorithm with different image can get different compression ratio. Design the Vector Quantizer based on robust statistic can improve it. Decrease the outlines vectors, improve the center vector effect in the codebook, it can decrease the relativity of the codebook, made the distribution of the codeword of the codebook more economical and bring on the compression ratio. Theoretical analysis and simulation experimental results presented in the paper show that this method can obtain good reconstruction image quality and high compress ratio. It is improved in both subjective and objective. 1

Keywords Image coding, Vector quantization, Robust statistic

0 引 言

矢量量化是用于数据压缩的一种有效方法,目前正受到越来越深入的研究,并被广泛应用于图象压缩中,其中最著名的是 LBG(Linde, Buzo, Gray)算法[1].LBG 算法作为矢量量化的基本算法具有经典意义,目前仍被广泛使用,也是矢量量化与其他压缩技术结合的基础[2],但因为实际存在的一些问题,所以它与图象压缩的结合未能充分地体现出优越性来.人们在发展图象压缩的新算法时[3],通常就与 LBG 算法进行对比.而众多发表的文献中,虽同样使用 LBG 算法对经典图象 Lena 进行压缩,但其效果不尽相同,甚至差别很大.这是因为训练图象中存在少量的离群矢量的缘故,它们在训练矢量中的数量虽然不多,但却对码书中码字的分布有较大影响,由于其使得寻找码书的收敛变得不稳定,从而使得压缩性能下降.本文运用基于稳健统计的方法来设

基金项目:国家自然科学基金项目(60172046)
收稿日期:2002-06-18;**改回日期**:2003-03-03

计矢量量化器,即按照影响函数的度量来抛弃离群矢量,从而减少了离群矢量对码书的影响,由于将中心矢量分裂,可增加中心矢量的权重,并可通过减少码书冗余来尽可能逼近最优的码书分布,从而可大幅度提高压缩性能.实验结果显示,基于稳健统计方法设计的码书,其压缩性能比传统的 LBG 算法有了较大的改善,即训练矢量不再集中于少量码字,且压缩比和峰峰信噪比都得到显著提高,其恢复图象的主观、客观效果都是令人满意的.

1 矢量量化

众所周知,矢量量化过程可以定义为从 k 维欧几里德空间 $\mathbf{R}^k$ 到其一个有限子集 $\boldsymbol{C}$ 的一个映射,即 $\boldsymbol{Q}:\mathbf{R}^k\to\boldsymbol{C}$,其中,$\boldsymbol{C}=\{y_0,y_1,\cdots,y_{N-1}\mid y_i\in\mathbf{R}^k\}$ 称为码书,N 为码书长度.该映射满足:$\boldsymbol{Q}(\boldsymbol{x}\mid\boldsymbol{x}\in\mathbf{R}^k)=\boldsymbol{y}_p$,其中 $\boldsymbol{x}=(x_0,x_1,\cdots,x_{k-1})$,$\boldsymbol{y}_p=(y_{p_0},y_{p_1},\cdots,y_{p_{k-1}})$ 并满足

$$d(\boldsymbol{x},\boldsymbol{y}_p)=\min_{0\le j\le M-1}(d(\boldsymbol{x},\boldsymbol{y}_j)) \tag{1}$$

其中,$d(\boldsymbol{x},\boldsymbol{y}_j)$ 为矢量 $\boldsymbol{x}$ 与码字 $\boldsymbol{y}_j$ 之间的失真测度,常用的失真测度为均方误差准则,其表达式为

$$d(\boldsymbol{x},\boldsymbol{y}_j)=\sum_{l=0}^{k-1}(x_l-y_{j_l})^2 \tag{2}$$

这样信号传输时,只要通过矢量量化编码器在码书中搜索出与输入矢量间失真最小的码字,然后,仅传输该码字的索引,而矢量量化解码器只要根据接受到的索引,用查表操作在码书中查找该码字,并将它作为输入矢量的重建矢量即可.

2 稳健统计

2.1 次序统计[4]

设 $X_1,X_2,\cdots,X_n$ 为样本总体 $\boldsymbol{X}$ 的一个容量为 n 的样本,将其任一实现 $x_1,x_2,\cdots,x_n$ 按大小顺序排成:

$$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)} \tag{3}$$

设 $X_{(m)}(x_1,x_2,\cdots,x_n)=x_{(m)}$,$m=1,2,\cdots,n$,则有

$$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)} \tag{4}$$

$X_{(1)},X_{(2)},\cdots,X_{(n)}$ 称为次序统计量.

(1)某 h 个统计量的联合分布

设样本总体 $\boldsymbol{X}$ 的密度函数为 $f(x)$,其分布函数为 $F(x)$,则次序统计量 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 中某 h 个次序统计量 $X_{(i_1)}\le X_{(i_2)}\le\cdots\le X_{(i_h)}$,$1\le i_1<i_2<\cdots<i_n\le n$,$1<h\le n$ 的联合密度函数为

$$f(x_{i_1},x_{i_2},\cdots,x_{i_h})=\frac{n!\,P_1^{i_1-1}P_2^{i_2-i_1-1}\cdots P_h^{i_h-i_{h-1}-1}P_{h+1}^{n-i_h}f(x_{i_1})\cdots f(x_{i_h})}{(i_1-1)!\,(i_2-i_1-1)!\,\cdots(i_h-i_{h-1}-1)!\,(n-i_h)!}$$

$$x_{i_1}\le x_{i_2}\le\cdots\le x_{i_h} \tag{5}$$

式中 $P_l=\int_{x_{i_{l-1}}}^{x_{i_l}}f(x)\mathrm{d}x=F(x_{i_l})-F(x_{i_{l-1}})$,$l=1,2,\cdots,h+1$,且 $x_{i_0}=-\infty$,$x_{i_{h+1}}=\infty$.

(2)n 个次序统计量 $X_{(1)},\cdots,X_{(n)}$ 的联合分布

在式(5)中,令 $h=n$,则得 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 的联合密度函数为

$$f(x_1,x_2,\cdots,x_n)=n!f(x_1)f(x_2)\cdots f(x_n) \tag{6}$$

$$x_1\le x_2\le\cdots\le x_n$$

(3)前 r 个次序统计量的联合分布

式(5)中,令 $h=r$,$(i_1,i_2,\cdots,i_r)=(1,2,\cdots,r)$,则得前 r 个统计量 $X_{(1)}\le X_{(2)}\le\cdots\le X_{(r)}$ 的联合密度函数为

$$f(x_1,x_2,\cdots,x_r)=\frac{n!}{(n-r)!}f(x_1)f(x_2)\cdots f(x_r)\cdot[1-F(x_r)]^{n-r} \tag{7}$$

$$x_1\le x_2\le\cdots\le x_r$$

(4)某两个次序统计量的联合分布

在式(5)中,设 $h=2$,$(i_1,i_2)=(i,j)$,$i\le j$,则得第 i 个与第 j 个次序统计量

$$X_i\le X_j,\ 1\le i<j\le n$$

的联合密度函数为

$$f(x_i,x_j)=\frac{n!\,f(x_i)f(x_j)}{(i-1)!\,(j-i-1)!\,(n-j)!}\cdot[F(x_i)]^{i-1}[F(x_j)-F(x_i)]^{j-i-1}\times[1-F(x_j)]^{n-j},x_i\le x_j \tag{8}$$

(5)单个次序统计量的分布

在式(5)中,令 $h=1$,$i_1=i$,则第 i 个次序统计量 $X_{(i)}$ 的密度函数为

$$f_i(x)=nC_{n-1}^{i-1}F^{i-1}(x)[1-F(x)]^{n-i}f(x),1\le i\le n \tag{9}$$

式中,$C_m^n=\frac{m!}{n!\,(m-n)!}$.特别地

$$\left.\begin{aligned}f_1(x)&=nf(x)[1-F(x)]^{n-1}\\f_n(x)&=nf(x)F^{n-1}(x)\end{aligned}\right\} \tag{10}$$

(6)样本极差及其分布

通常将 $R=X_{(n)}-X_{(1)}=\max_{1\le i\le n}X_i-\min_{1\le i\le n}X_i$ 称作样本极差,它的分布函数为

$$F_R(x)=n\int_{-\infty}^{\infty}[F(t+x)-F(t)]^{n-1}f(t)\mathrm{d}t \tag{11}$$

式中,$f(x)$ 为样本总体 $\boldsymbol{X}$ 的密度函数;$F(x)$ 为样本

总体 $\boldsymbol{X}$ 的分布函数.

(7)样本中值及其渐进分布

通常将

$$\mathrm{med}(X_1,\cdots,X_n)=\begin{cases} X_{\left(\frac{n+1}{2}\right)} & n\text{为奇数} \\ \frac{1}{2}\left[X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right] & n\text{为偶数} \end{cases} \tag{12}$$

称为样本中值(亦称样本中位数).

设样本总体 $\boldsymbol{X}$ 的密度函数为 $f(x)$,样本总体 $\boldsymbol{X}$ 的中值为 a,若在 $x=a$ 处,函数 $f(x)$ 连续,且大于 0,则样本中值渐进地服从正态分布

$$N\left(a,\frac{1}{4n[f(a)]^2}\right)$$

2.2　秩统计[5]

设 $X_1,X_2,\cdots,X_N$ 为一个样本,它可以是从一个总体中抽出的样本,也可以是从每个样本总体中抽出的样本的合样本. 记其次序统计量为 $X_{(1)}\leq X_{(2)}\leq\cdots\leq X_{(N)}$. 如果 $X_i=X_{(R_j)}$,则称 R_i 为 X_i 的秩.

显然,秩 $R_1,R_2,\cdots,R_N$ 是统计量. 一般地,将只依赖于秩的统计量称为秩统计量.

2.3　影响函数与稳健性度量[6]

在估计分析以及稳健性定义中,由于离群数据对估计运算有重大影响,因此是一个非常重要的问题,但利用影响函数(IF)可以解决离群数据对估计量的影响.

假设 k 维独立同分布的训练矢量 $\boldsymbol{x}^{\mathrm{T}}=(x_1,x_2,\cdots,x_n)$ 属于样本空间 $\boldsymbol{X}$,该空间是实数集 $\mathbf{R}^{\langle k\rangle}$ 的一个子集,则其经验分布函数(直方图)为

$$G_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i} \tag{13}$$

式中,δ_{x_i} 是位于 x_i 的 δ 函数. 当 $n\to\infty$ 时,经验分布函数 G_n 将趋于真实的分布函数 G. 大家知道,观测数据参数模型包括样本空间 $\boldsymbol{X}$ 上一个概率分布函数族 F_θ,它取决于未知参数 θ 的概率分布函数 F_θ 的参数空间,记为 $\boldsymbol{\Theta}$. 在经典参数估计分析中,通常假设观测数据 x 严格服从 F_θ 分布,而且 θ 的估计将基于全部数据 x,而稳健估计则假设 F_θ 仅近似符合实际. 在这种假设下,必须计算 θ 的估计量 $T_n=T_n(x_1,x_2,\cdots,x_n)=T_n(G_n)$.

分布函数 F 的估计量 T 的影响函数 IF 定义为

$$IF(x;T,F)=\lim_{t\to 0}\frac{T((1-t)F+t\delta_x)-T(F)}{t} \tag{14}$$

当 $x\in\boldsymbol{X}$ 时,该极限存在. 如果 G_{n-1} 是 $n-1$ 个观测数据的经验分布,而且 $t=1/n$,那么,影响函数近似于标示在 x 上增加一个观测数据的影响.

3　稳健矢量量化器的设计

在图象中,可将大量分布、距离较近、集中的训练矢量定义为中心矢量,并将这种由中心矢量聚类构成的码字称为中心码字;而将少量远离数据主体(中心部分)的训练矢量定义为离群矢量,并将这种由少量离群矢量聚类构成的码字称为离群码字. 显然,按照码书对图象进行压缩时,大部分的矢量由少量码字来重建,而在码书中占有较大比例的离群码字,却只在很少量的重建矢量中出现. 由此可见,离群码字对寻找码书的收敛性能以及重建图象的恢复质量都有重要的影响.

结合影响函数的思想,本文提出了一种基于稳健统计的矢量量化器设计算法,即把重建图象与原始图象的均方误差作为影响函数,把码书中的码字所拥有的聚类矢量数作为次序统计量. 实验结果表明,用这种算法来设计码书,效果比较理想. 具体计算步骤如下:

(1)确定矢量的维数 k,确定码书中码字的个数 N,误差阈值 $\varepsilon_1>0$,影响误差 $\varepsilon_2>0$. m 表示迭代的次数,赋初始值为零. 给定相对误差 $e^{\langle -1\rangle}$,令 $e^{\langle -1\rangle}=\infty$.

(2)在 0～255 之间,取 $k\times N$ 个随机数作为初始码书 $\boldsymbol{A}^{\langle 0\rangle}$.

(3)对于 $\boldsymbol{A}^{\langle m\rangle}=(\boldsymbol{y}_i;i=1,2,\cdots,N)$,寻找一种对于训练图象的最小误差划分 $\boldsymbol{S}$

$$\boldsymbol{S}(\boldsymbol{A}^{\langle m\rangle})=\{\boldsymbol{S}_i;i=1,2,\cdots,N\}$$

当 $\boldsymbol{x}_j\in\boldsymbol{S}_i$ 时,对任意 L 都有 $d(\boldsymbol{x}_j,\boldsymbol{y}_i)\leq d(\boldsymbol{x}_j,\boldsymbol{y}_L)$,

然后,计算出 $e^{\langle m\rangle}$.

$$e^{\langle m\rangle}=D(\{\boldsymbol{A}^{\langle m\rangle},\boldsymbol{S}(\boldsymbol{A}^{\langle m\rangle})\})=\frac{1}{N}\sum_{j=0}^{N-1}\mathrm{min}d(x_j,y),y\in\boldsymbol{A}^{\langle m\rangle}$$

(4)若 $(e^{\langle m-1\rangle}-e^{\langle m\rangle})/e^{\langle m\rangle}>\varepsilon_1$,则把 $\boldsymbol{S}_i$ 的算术中心作为 $\boldsymbol{A}^{\langle m+1\rangle}$,迭代次数 m 加 1,回到第 3 步.

(5)先计算 $\boldsymbol{S}_i$ 中训练矢量的个数 P_i,再将 $\boldsymbol{A}^{\langle m\rangle}$ 中的码字按照对应的 P_i 排序,即得到次序统计量的一个样本. 定义最后一个码字为离群码字,舍弃它即可得到 $\tilde{\boldsymbol{A}}^{\langle m\rangle}$,重新计算 $e^{\langle m\rangle}$.

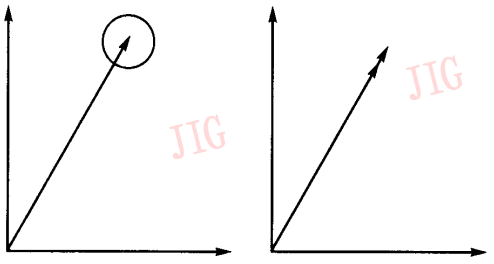
(6)若 $(e^{\langle m-1\rangle}-\tilde{e}^{\langle m\rangle})/\tilde{e}^{\langle m\rangle}>\varepsilon_2$,则将第 1 个码字作为中心码字分裂,首先形成 $\boldsymbol{A}^{\langle m+1\rangle}$,然后将迭代次数 m 加 1,回到第 3 步.

(7)将 $\boldsymbol{A}^{\langle m\rangle}$ 作为最优码书输出. 程序结束.

需要说明的是,步骤 6 中分裂码字的方法也影响着码书设计的收敛性.

4 扰动方式及其大小的选择

在对中心矢量通过施加扰动、进行分裂来产生新的码字时,可以采用对矢量中的每一个分量加上一个小的扰动的方法,也可以采用给矢量乘以一个接近于 1 的系数的方法(见图 1). 显然,图 1(a)产生的新的聚类中心是以原矢量为中心的超空间,图 1(b)产生的新的聚类中心矢量方向未变,仅矢量的模值变化,表现在图象中就是灰度值有所增减. 多次实验发现,使用图 1(a)所示的方法通过施加扰动来产生新的码字的方法较好. 在给聚类中心施加扰动时,从理论上讲,扰动应为这一类聚类样本点之间的平均距离,因为如果扰动太小,则没有什么效果,且产生的依然为无效码字;如果扰动太大,又脱离了这一类样本,则失去了分裂的意义,且同样产生无效码字. 而在高维的矢量空间中,由于无法判断样本点的平均距离,因此只能凭经验逐个进行参数测试. 多次实验结果表明,对矢量中的每一个分量施加在[−2,+2]中随机均匀分布的扰动时,效果最好. 由于每一个分量都是整数,因此扰动实际上就是随机地加了−1,−2,0,1,2,也就是扰动是在以原样本矢量为中心,半径为 2 的球内离散随机分布.



(a) (b)

图 1 扰动方式的选择

在对聚类中心进行分裂时,有几种方法可以选择(为描述方便,不妨设原聚类中心为 c,扰动为 ε):

(1)原聚类中心 c 不变,可通过加一个小扰动来生成新的聚类中心 $c+\varepsilon$.

(2)对原聚类中心,可通过分别加上幅度相同,分布相同的两个小扰动来生成两个新的聚类中心 $c+\varepsilon_1, c+\varepsilon_2$.

(3)对原聚类中心,可通过加减完全相同的小扰动 ε,来生成两个新的聚类中心 $c+\varepsilon, c-\varepsilon$. 这相当于在高维空间中,用一个高维平面把中心一分为二.

从理论上讲,这 3 种方法原理相近,而在实际仿真中,第 1 种方法因为保存了原有的码字,所以要比其他两种方法稍好一些.

5 仿真实验

综上所述,在实验中采用了如图 1(a)所示的施加扰动的方法:首先扰动在区间[−2,+2]上均匀分布;然后用方法 1 产生新的聚类中心,最后用基于稳健统计的设计算法来设计码书. 在实验中,对 256×256 大小的典型图象“Lena”进行压缩,其中每一个像素用 8 比特表示. 为方便起见,由于是取相邻的、方的、互不重叠的 4×4 的小图象块来进行编码,因此码字是 16 维的. 在图 2(b)和图 2(c)的仿真实验中,由于码书的长度取为 256,因此压缩比 $C_R=\frac{8\times16}{\log-2256}=16$;对于图 2(d)的仿真实验,由于码书长度取为 512,因此 $C_R=\frac{8\times16}{\log_2512}=14.22$.

图 3 为运用基于稳健统计的设计方法与传统 LBG 方法获得的码书的效果对比图. 从图 3 中可以看出,对于传统的LBG方法,聚类到第1位中心码

(a) Lena 原图象 (b) LBG 压缩重建图象 (c) 本文算法的重建图象 (d) 本文算法的重建图象

图 2 基于稳健统计的设计方法的重建图象

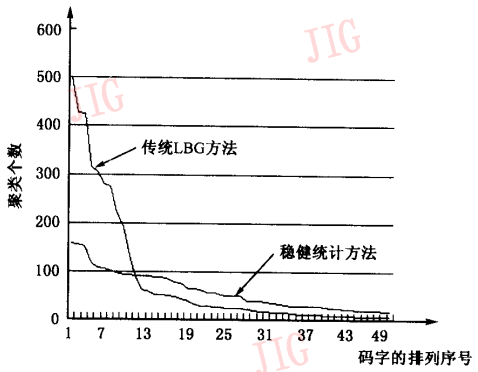


图 3　基于稳健统计的设计算法的码书分布与 LBG 方法的码书比较

字的训练矢量有 500 个，而在基于稳健统计的设计方法中只有 158 个；到了第 20 位时，LBG 方法有 32 个，而在基于稳健统计的设计方法中有 64 个. 也就是说，传统的 LBG 方法中，大量的训练矢量由少数码字重建，而基于稳健统计的方法设计的矢量量化器在抛弃了离群矢量后，不仅将中心矢量分裂，并更多考虑了中心矢量的权重，因此使得码书中的码字更加具有代表性，也就从整体上提高了码书的性能.

从图 2 所示的主观效果和表 1 所示的实验数据可以看出，在压缩比为 16 时，重建图象的均方误差从 119.36 降低到 56.21，而峰峰信噪比则从 27.36dB 提高到 30.63dB；而且在压缩比从 16 稍微降低一点至 14.22 时，均方误差即降低到 36.16，而峰峰信噪比则提高到 32.55dB，这已完全能满足人眼的视觉需要. 由此可见，基于稳健统计的设计算法明显优于单一的 LBG 算法. 实验结果表明，无论是在主观还是在客观质量上都更令人满意. 这也说明，本算法可舍弃码书中的离群码字，并能降低离群矢量对码书的影响，从而使重建图象的质量明显提高，取得较好的效果.

表 1　基于稳健统计的设计算法与 LBG 算法效果对比

	图 2(b)	图 2(c)	图 2(d)
使用的方法	LBG 算法	本文算法	本文算法
压缩比	16	16	14.22
均方误差	119.36	56.21	36.16
PSNR(dB)	27.36	30.63	32.55

6　结　论

本文结合稳健统计的方法，提出了一种基于稳健统计的矢量量化器设计算法，该算法就是按照影响函数的度量标准，将在训练矢量中只有少数矢量聚类的离群矢量抛弃，以增加中心矢量的权重. 实践证明，用这种算法得到的码书，其码书中的码字更加具有代表性. 运用这种算法对传统的图象 Lena 进行压缩，在压缩比为 16 时，重建图象的均方误差只有 57.21，而传统 LBG 算法重建图象的均方误差却高达 119.36；在压缩比为 14.22 时，本文算法重建图象的均方误差又降低到 36.16. 实验结果表明，使用基于稳健统计的方法设计的矢量量化器，其压缩性能比传统的 LBG 算法有了较大的改善，且无论是主观，还是客观效果都有大幅度的提高.

均方误差通常是作为衡量图象质量的客观标准，但与主观评价标准有比较大的差距. 如果能有较好的主观图象质量评价函数作为影响函数，那么压缩效果也许更好.

参 考 文 献

1　Linde Y, Buzo A, Gray R M. An algorithm for vector quantizer design[J]. IEEE Transactions Communication, 1980, 28(1): 84～95.

2　Nasrabadi N M, King R A. Image coding using vector quantization: a review[J]. IEEE Transactions Communication, 1988, 36,(8): 957～969.

3　Ramamurthi B, Gersho A. Classified vector quantization of images[J]. IEEE Transactions Communication, 1986, 34(11): 1105～1115.

4　David H A. Order Statistics[M]. New York: John Wiley, 1980.

5　Huber P S. Rubust Statistics[M]. New York: John Wiley, 1981.

6　陈贺新. 非线性滤波器与数字图象处理[M]. 北京：国防工业出版社，1997.

桑爱军　1973 年生，1994 年获华中理工大学学士学位，1997 年获大连理工大学硕士学位，后分配至吉林大学任教至今，1999 年攻读博士学位. 主要研究方向为多维信号处理、图象编码、变换理论等.

陈贺新　1949 年生，1982 年、1985 年、1989 年先后获得吉林工业大学电子工程系工学学士、工学硕士、工学博士学位，现为吉林大学通信工程学院教授，博士生导师. 主要研究兴趣是多维数字信号处理、人工神经网络、计算机视觉等.