基于字典优化的稀疏表示的视频镜头分类*

陈 波[1], 詹永照[1], 成科扬[1,2]

(1. 江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013; 2. 南京航空航天大学 计算机科学与技术学院, 南京 212000)

摘 要: 为了克服稀疏表示中冗余字典分类效果不佳的问题,提出了基于字典优化的稀疏表示算法。该算法制定了新的基于稀疏表示的分类判别规则,采用了基于冗余字典内基元类内平均欧式距离最小以及类间平均欧式距离最大的字典优化方法,形成优化字典进行特征稀疏表示。将该算法应用于视频镜头的稀疏表示特征提取与分类,实验结果表明该方法优化后的字典进行视频镜头的特征提取和分类,其识别率得到了明显的提高。

关键词: 稀疏表示; 字典优化; 视频镜头分类

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)06-2375-04

doi:10.3969/j.issn.1001-3695.2012.06.100

Video shot classification based on sparse representation of dictionary optimized

CHEN Bo[1], ZHAN Yong-zhao[1], CHENG Ke-yang[1,2]

(1. *School of Computer Science & Telecommunication Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*; 2. *School of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 212000, *China*)

Abstract: In order to overcome the ineffective classification results of the redundant dictionary in the sparse representation-based classifier, this paper presented a sparse representation algorithm based on dictionary optimization. The algorithm developed a new classification discriminate rules based on sparse representation. It optimized the dictionary by the method of minimizing the average of the in-class Euclidean distance and maximized the average of the between-class Euclidean distance, formed the optimized dictionary and presented the features based on sparse representation. And the algorithm was applied on video shot to extract feature and classify based on sparse representation. The experimental results show that the recognition rate of feature extraction and classification on video shot based on the dictionary optimized by this method has been significantly improved.

Key words: sparse representation; dictionary optimization; video shot classification

随着多媒体技术的发展,大量的视频数据涌现在日常生活中,如何有效地对其进行组织与分类成为亟待解决且富有挑战的问题。而对视频进行处理,首要解决的就是对视频中一帧图像的处理。而图像处理技术通常需要通过更有效的表示来捕获图像的特征,对分类识别来说,表示需要突出显著特征,对去噪来说,表示需要有效地分离信号和噪声;对压缩来说,表示需要用少量的系数来描述大部分的图像。有趣的是这些应用中看似目的不同,但都有一个共同的目标就是简化图像的表示,即图像的特征提取。

稀疏表示是一种对复杂信号进行压缩表示并能高效重构的方法,近几年在模式识别以及计算机视觉领域也引起了广泛的关注,并且在许多图像目标识别和分类中取得了目前最好的效果。视频图像特征信息复杂且含有噪声,本文拟研究有效的稀疏表示视频信息特征提取并应用于视频镜头分类,以期获得更有效的视频镜头分类效果。

Olshausen 等人[1]提出了稀疏表示模型,指出其模仿人脑 v1 区对自然图像的表示策略。Pati 等人[2]提出了解决稀疏表示问题的 OMP 算法。Wright 等人[3]提出了基于稀疏表示的分类方法。Aharon 等人[4]提出了 K-SVD 字典优化算法,但其迭代过程中包含较为复杂的奇异值分解,且需要额外的训练样本。鉴于此,本文提出一种完全基于冗余字典内基元本身的冗余字典优化算法来形成稀疏表示特征,并将其应用于视频镜头的分类识别。

1 稀疏表示原理

1.1 稀疏表示的数学表示

稀疏表示涉及一个欠定的线性等式 $y = Ax$,基本模型表明自然图像能够被表示成预先定义的原子图像的线性组合,而且这些组合系数是稀疏的,即大部分系数是 0,或接近 0。这对于大数据量的视频分类是至关重要的。数学形式描述如下:设 $y \in \mathrm{R}^m$ 是一个图像的特征向量,一些原子图像特征向量构成字典 $A \in \mathrm{R}^{m \times n}$ $(m < n)$,则稀疏正规化下稀疏表示的就是求解公式

$$\hat{x} = \arg\min \| x \|_0 \text{ subject to } \| y - Ax \|_2^2 \leqslant \varepsilon \tag{1}$$

其中:$\| \cdot \|_0$ 为 L0-范式,即向量中非零元素的个数。

收稿日期: 2011-09-21; **修回日期:** 2011-10-28 **基金项目:** 国家自然科学基金资助项目(61170126);江苏省自然科学基金资助项目(BK2009199);江苏省省属高校自然科学研究资助项目(11KJD520004);江苏省普通高校研究生科研创新计划资助项目(CXZZ11_0216)

作者简介: 陈波(1987-),男,硕士研究生,主要研究方向为视频检索(chenbo19870603@126.com);詹永照(1962-),男,教授,博导,主要研究方向为模式识别、多媒体技术;成科扬(1982-),男,博士研究生,主要研究方向为模式识别、图像处理.

由于解决上述 L0-范式问题是 NP-hard 问题且非常耗时，最新的研究表明，当$\hat{x}$足够稀疏时，求解 L0-范式最优化问题等价于求解下述 L1-范式最优化问题。

$$\hat{x} = \arg\min \| x \|_1 \text{ subject to } \| y - Ax \|_2^2 \leqslant \varepsilon \tag{2}$$

求解式(2)的 L1-范式最优化问题，如今已经有了许多经典算法，如正交匹配算法、OMP Cholesky 算法等。本文使用了 OMP Cholesky 算法来解决此问题。

1.2 OMP Cholesky 算法

OMP Cholesky 算法是稀疏表示的一种经典的也是最简单的算法之一，它试图求得式(2)提出的问题的逼近解。为了简化描述，假定字典 A 的列已经归一化，OMP Cholesky 算法在每一步迭代过程中选择与当前迭代残差最相关的原子，选定原子以后，将信号正交投影到这些原子张成的空间中，重新计算残差，由此循环直到满足约束条件。

OMP Cholesky 算法基本思想基于算法中矩阵是一个正定阵，在每一步迭代更新中仅仅是追加一行或者一列，因此它的 Cholesky 分解仅仅需要计算其最后一行。这很容易验证假定未更新之前矩阵 Cholesky 分解$\tilde{F} = \tilde{L}\tilde{L}^{\mathrm{T}} \in \mathrm{R}^{(n-1)\times(n-1)}$，则在更新追加行之后的 Cholesky 分解为

$$F = \begin{pmatrix} \tilde{F} & v \\ v^{\mathrm{T}} & c \end{pmatrix} \in \mathrm{R}^{n\times n} \tag{3}$$

由 $F = LL^{\mathrm{T}}$ 可知

$$L = \begin{pmatrix} \tilde{L} & o \\ w^{\mathrm{T}} & \sqrt{c - w^{\mathrm{T}}w} \end{pmatrix}, w = \tilde{L}^{-1}v \tag{4}$$

由此每次迭代过程中不再需要显式地知道残差，即不需要显式地计算 r 和它的乘数 A^{T}，取而代之只要计算开销 $A^{\mathrm{T}}r$ 即可。

OMP Cholesky 算法如下：

```
Input: Dictionary A, signal y, target sparsity K or target error ε
Output: Sparse representation x
Init: Set I: = ( ), L: = [1], r: = y, x: = 0, α(0): = A^T y, n: = 1
while( stopping criterion not met) do
  k̂: = arg max_k |a_k^T r|
  If n > 1 then
    w: = solve for w{ Lw = A_I^T a_k^∧ }
    L: = [ L  0 ; w^T  sqrt(1 - w^T w) ]
  end if
I: = (I, k̂)
  x_I: = solve for c{ LL^T c = a_I }
  r: = y - A_I x_I
  n: = n + 1
end while
```

1.3 基于稀疏表示分类规则

假设一共有 c 类样本，A_i 为第 i 类训练样本构成的矩阵，即 $A_i = [y_{i1}, y_{i2}, \cdots, y_{iM_i}] \in \mathrm{R}^{d\times M_i}$，则所有类的训练样本构成冗余字典 $A = [A_1, A_2, \cdots, A_c] \in \mathrm{R}^{d\times M}$，其构成的基元为训练样本。

在使用上述的 OMP Cholesky 算法得到稀疏表示解$\hat{x}$后，可以通过以下方式设计一个基于稀疏表示的分类方法。对每一类 i，向量 $\delta_i(x)$ 中所有非零项为向量$\hat{x}$中与第 i 类相关的项。使用这些只与第 i 类相关的系数，可以将给定测试样本 y 重构为 $v^i = A\delta_i(\hat{x}_i)$，则 v^i 称为测试样本 y 关于第 i 类的原型。样本 y 与其原型 v^i 之间的距离被定义为

$$r_i(y) = \| y - v^i \|_2 = \| y - A\delta_i(\hat{x}_i) \|_2 \tag{5}$$

则基于稀疏表示的分类规则为：如果 $r_l(y) = \min_i r_i(y)$，则 y 分配为第 l 类中。

2 稀疏表示冗余字典的优化

稀疏表示的本质就是将图像用预先定义的冗余字典中的原子图像线性组合。冗余字典的好坏将决定稀疏表示的好坏，以至于决定了最后的分类效果，因此构建一个优异的冗余字典是十分必要的。本文在解决 L1-范式最优化问题的基础上提出了一种优化冗余字典的方法。其基本思想就是使得冗余字典中的每一个基元的类内重构误差尽量小而类间重构误差尽量大。因此，优化后的冗余字典可以得到更优的分类识别结果。

2.1 新的稀疏表示分类规则

上文中基于稀疏表示的分类规则时，冗余字典是由训练样本本身构成的，这样虽可以减少很多计算，但同时由于基元选取时的随机性，可能导致在稀疏表示测试样本时不能达到理想的稀疏效果，进而会影响到最终的分类识别。因此在使用训练样本本身来初始化冗余字典后，对其进行优化，以使得冗余字典可以达到更好的稀疏表示效果。

本文中为了方便描述，假定共有 c 类训练样本，每类都有 M 个训练样本。对于每一个训练样本 y_{ij}，将它从所属类别的训练样本集中取出，并使用剩下的训练样本来线性表示样本 y_{ij}，通过计算 L1-范式最优化问题，得到样本的稀疏表示系数向量 w_{ij}，设 $\delta_s(w_{ij})$ 为样本关于类别 i 的表示系数向量，则样本 y_{ij} 关于类别 s 的原型为 $v_{ij}^s = A\delta_s(w_{ij})$，$s = 1, \cdots, c$。样本 y_{ij} 与类别 i 的距离定义为

$$d_i(y_{ij}) = \| y_{ij} - v_{ij}^s \|^2 \tag{6}$$

根据稀疏表示的分类规则可知，为了使分类效果更好，期望样本 y_{ij} 与所属类别原型 v_{ij}^i 尽可能接近，而与其他类别原型 $v_{ij}^s (s \neq i)$ 则尽可能远离。即样本与所属类别之间的距离 $d_i(y_{ij})$ 尽可能小而与其他类别之间的距离 $d_s(x_{ij}) (s \neq i)$ 则尽可能大。样本 y_{ij} 与其他类别之间的距离定义为

$$d(y_{ij}) = \frac{1}{M-1}\sum_{s\neq i} d_i(y_{ij}) = \frac{1}{M-1}\sum_{s\neq i} \| y_{ij} - v_{ij}^s \|^2 \tag{7}$$

则定义准则

$$j(d_{ij}) = \frac{d_i(y_{ij})}{d(y_{ij})} \tag{8}$$

重新定义稀疏表示分类规则：$r_l(y) = \min_i j(d_{ij})$，则 y 分配为第 l 类中。

2.2 冗余字典的优化

为了使基于稀疏表示的分类效果更好，不能仅仅考虑单个样本在字典上的分类表现，而应该综合考虑这个冗余字典中所有训练样本，即冗余字典可以在所有的训练样本上取得好的分类效果，因此需要计算冗余字典中所有样本类内距以及类间距

的平均值。类内距均值定义为

$$D_i = \frac{1}{M}\sum_{i,j} d_i(y_{ij}) = \frac{1}{M}\sum_{i,j} \| y_{ij} - v_{ij}^i \|^2 = \frac{1}{M}\sum_{i,j} (y_{ij} - v_{ij}^i)^{\mathrm{T}} (y_{ij} - v_{ij}^i) \tag{9}$$

同时类间距均值定义为

$$D_o = \frac{1}{M(c-1)}\sum_{i,j}\sum_{s\neq i} d_s(y_{ij}) = \frac{1}{M(c-1)}\sum_{i,j}\sum_{s\neq i} \| y_{ij} - v_{ij}^s \|^2 \tag{10}$$

根据稀疏表示分类规则,越大的类间距平均值和越小的类内距平均值会得到越好的分类效果。因此可以选择最大化如下判定准则函数来得到好的分类冗余字典。

$$J(D) = \frac{D_o}{D_i} \tag{11}$$

基于以上准则找到最优的冗余字典。因此本文试图将准则转换为与冗余字典 A 相关的函数。

将 $y_{ij} = Ax_{ij}$,$v_{ij}^s = A\delta_s(w_{ij})$ 代入公式 D_o、D_i 中,得到

$$D_i = \frac{1}{M}\sum_{i,j} [x_{ij} - A\delta_i(w_{ij})][x_{ij} - A\delta_i(w_{ij})]^{\mathrm{T}} \tag{12}$$

$$D_o = \frac{1}{M(c-1)}\sum_{i,j}\sum_{s\neq i} [x_{ij} - A\delta_s(w_{ij})][x_{ij} - A\delta_s(w_{ij})]^{\mathrm{T}} \tag{13}$$

将 D_o、D_i 带入以得到准则 $J(D)$,通过最大化准则 $J(D)$ 便可以得到优化的冗余字典。但在应用中时,无法确定一个理想的最大值 $J(D)$ 来终止上述最大化的过程。在实验中,从经验告可知,一个迭代过程中,当前一次的迭代结果与当前迭代结果相差不大时,可以认为此迭代过程已经达到了理想结果。基于以上的经验,定义新的准则为

$$J = |J(D_k) - J(D_{k-1})| / |J(D_k)| \tag{14}$$

当 $J < \varepsilon$ 时,则认为达到了理想结果,迭代结束,即冗余字典优化完成。

2.3　优化的稀疏表示算法描述

输入:初始冗余字典 A,最大迭代次数 K,目标误差参数 ε。

输出:优化的冗余字典 A^*。

初始化:将训练样本本身作为冗余字典基元向量,$k=0$。

a)先对冗余字典中的某一类进行处理,将冗余字典中类 i 中每一个基向量用剩余的基向量进行稀疏表示,通过 OMP Cholesky 算法计算 L1-范式最优化问题获得稀疏表示向量 $\hat{x}$,向量 w_{ij} 中所有非零项为基向量 y_{ij} 的稀疏表示向量 $\hat{x}$ 中与第 i 类相关的项;

b)通过每一个基向量的稀疏表示向量 w_{ij} 计算每个基向量的类内距 $d_i(y_{ij})$ 以及类间距 $d_s(x_{ij})(s\neq i)$,并可由此计算类内所有基向量的类内距均值 D_i 和类间距平均值 D_o。

c)计算每一个样本的 $j(d_{ij})$,并找出类中最小的 $\min j(d_{ij})$。

d)根据其稀疏表示向量 w_{ij} 可得到原型 v_{ij},则将原型 v_{ij} 代替向量 y_{ij} 作为新的基向量。

e)$k = k+1$,在保证 k 小于最大迭代次数 K 的情况下,计算准则 $J(D)$ 以及 $J = |J(D_k) - J(D_{k-1})| / |J(D_k)|$,并判断是否大于 ε。如若大于 ε,则重复步骤 1 ~ 5;若小于 ε 或 k 超出了最大迭代次数,则跳至下一步。

f)对冗余字典中第 i 类基元的优化工作结束。

g)对冗余字典中的其他类重复步骤 a) ~ f),以使字典中的每一类基向量都得到优化。

2.4　优化的稀疏表示算法分析

本文提出的字典优化算法是基于训练样本的特征向量进行操作的,即对字典中的列向量进行优化,每次迭代过程都会更新字典中的一列,因此首先定位需要进行更新的列向量。在上述算法描述中,步骤 a) b)进行了必要的运算以得到相关的量。在步骤 c)中,找到了类 i 中最大的 $j(d_{ij})$,根据基于稀疏表示的分类规则可知,此基向量相对于其他基向量的分类效果最差,因此定位了需要更新的基向量。因为向量 w_{ij} 为基向量相对于其他基向量的稀疏表示,则其原型 v_{ij} 为其他基向量的线性表示,而由分类规则可知,相对于类 i 来说,基向量 w_{ij} 的分类效果最差,其他基向量的分类效果都好于 w_{ij},则其他基向量线性组合的原型 v_{ij} 的分类效果也好于 w_{ij},则可以将原型 v_{ij} 代替样本 y_{ij} 作为新的基向量,即步骤 e)。以此完成了一次迭代过程,冗余字典中稀疏表示效果最差的一列得到了优化。在满足式(14)且未超过最大迭代次数的情况下,继续迭代过程以期得到更好的字典。

由上述的分析可知,经过此算法处理过的冗余字典,其字典中每一类的类内平均欧式距离得以缩小而类间平均欧式距离得以扩大,显而易见,如此情况下,优化后的冗余字典其分类识别率得到改善。实验结果亦表明,经过优化后的字典其分类识别率较优化前字典的识别率有较大提高。

3　视频镜头分类应用

为了克服上文提到的视频图像特征信息复杂且含有噪声的困难,将基于字典优化的稀疏表示用于视频镜头的分类。首先,使用聚类的方法对视频库中不同种类的视频镜头进行关键帧提取,再根据关键帧图像的特点选取特定的特征提取算法进行特征提取,以得到用于稀疏表示的训练样本以及测试样本。在完成上述工作后,使用一定数量的训练样本构成初始的冗余字典后,使用本文描述的字典优化算法对冗余字典进行优化,在得到优化后字典 A 后,对于任一测试样本 y_{ij} 通过 OMP 算法解决式(2)中 L1-范式最优化问题即可得到该测试样本特征的稀疏表示 $\hat{x}$,并由此计算出样本 y_{ij} 与字典中各个类的欧式距离 $d_s(y_{ij})$,然后根据式(8)即可对该样本进行分类。

4　实验结果

本文实验是在安装了 Windows XP 的微机(DELL, Intel® Core™2Duo CPU P735@2.00 GHz)上进行的,采用了 TRECVID 2007 所提供的新闻视频,使用 MATLAB 7.0 编程实现。实验中所使用的视频类别为 TRECVID 2007 中的其中六个类别,分别为 car、person、weather、mountain、sky、road,如图 1 所示。

实验中选取上述六类视频镜头作为冗余字典中的 6 类。为了简单起见,本文使得每一类别的训练样本数是相同的。对上述视频提取关键帧后,提取关键帧的图像视觉特征,本文分别提取了 HSV 颜色特征、纹理特征、形状特征共 24 维特征。

为了保证实验过程中字典的过完备性，也就是字典的列的维数大于行的维数，对每类样本取 8、16、24、32 帧图像的特征值构成原始的冗余字典，则冗余字典的大小分别为 24 × 48、24 × 96、24 × 144、24 × 192，保证了方程 $y = Ax$ 是欠定的。在测试时为了便于比较，将测试样本数目固定为 144。表 1 显示了不同样本数量识别率的情况。从表 1 中可以看到，优化后的字典其识别率明显优于未优化之前字典的识别率。且当训练样本数越少时，其识别率的提升则越明显。且当训练样本达到 144 左右时，冗余字典趋于饱和，识别率的提升不甚明显。

图1 TRECVID 2007新闻视频关键帧

表 1 测试图像在不同字典上的识别率比较

训练图像	测试图像	字典大小	未优化字典识别率/%	优化后字典识别率/%
48	144	24 × 48	76.5	81.0
96	144	24 × 96	89.3	91.2
144	144	24 × 144	95.3	96.9
192	144	24 × 192	95.7	97.0

5 结束语

本文研究了稀疏表示理论并使用 OMP Cholesky 算法解决欠定方程 $y = Ax$ 的 L1-范式最优化问题，在此基础上提出基于类内重构误差以及类间重构误差的冗余字典优化算法，实现了冗余字典的优化，并将该算法优化后的冗余字典应用于 TRECVID 2007 新闻视频库。结果表明，通过该算法优化后的字典识别率得到了明显的提高，证实了该算法的有效性。

参考文献：

[1] OLSHAUSEN B A, FIELD D J. Sparse coding with an over-complete basis set:a strategy employed by v1[J]. Vision Research,1997,37(23):3311-3325.

[2] PATI Y C, REZAHFAR R, KRISHNAPRASAD P S. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition[C]//Proc of the 27th Asilomar Conference on Signals, Systems & Computers. Los Alamitos, CA:IEEE,1993:40-44.

[3] WRIGHT J, YANG A, GANESH A, *et al*. Robust face recognition via sparse representation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2009,31(2):210-227.

[4] AHARON M, ELAD M, BRUCKSTEIN A. K-SVD: an algorithm for designing over-complete dictionaries for sparse representation[J]. IEEE Trans on Signal Processing,2006,11(54):4311-4322.

[5] DONOHO D. For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution[J]. Communications on Pure and Applied Mathematics, 2006, 59(6):797-829.

[6] LEE K C, HO J, KRIEQMAN D. Acquiring Linear Subspaces for Face Recognition under Variable Lighting[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2005,27(5):684-698.

[7] HE Xiao-fei, YAN S, HU Y, *et al*. Face recognition using laplacianfaces[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2005,27(3):328-340.

[8] BELHUMEUR P N, HESPANHA J P, KRIENGMAN D J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997,19(7):711-720.