中图法分类号：TP37

视频局部特征描述子的紧凑表示方法

张翔，王诗淇，张新峰，马思伟，高文

北京大学信息科学与技术学院数字媒体研究所，北京市 中国 100871

摘　要：目的：随着手持移动设备的迅猛发展和大数据时代的到来，以多媒体数据为核心的视觉搜索等研究和应用得到了广泛关注。其中局部特征描述子的压缩、存储和传输起到了举足轻重的作用。本文拟在传统图像/视频压缩框架中，提出一种高效的视觉局部特征的紧凑表示方法，使得传统内容编码可以适应广泛的检索分析等需求。**方法：**为了得到紧凑、有区分度、同时高效的局部特征表示，我们首先引入了多参考的预测机制，在消除了时空冗余的同时，通过充分利用视频纹理编码的信息，消除了来自纹理-特征之间的冗余。此外，我们还提出了一种新的率失真优化方法——码率-准确率最优化方法，使得基于匹配/检索应用的性能达到最优。**结果：**通过在不同数据集上进行的验证实验，和最新的视频局部描述子压缩框架进行比较，表明提出的方法能够在保证匹配和检索性能的基础上，显著地减少特征带来的比特消耗。**结论：**本文提出的方法适用于传统图像/视频编码框架，通过在码流中嵌入少量表示特征的信息，即可实现高效的检索性能，是一种面向检索等智能设备应用的新型多媒体内容编码框架。

关键词 ：计算机科学与技术，视觉搜索，视频局部特征描述子，尺度不变特征变换，高效视频压缩

Compact Representation of Video Local Feature Descriptors

Zhang Xiang, Wang Shiqi, Zhang Xinfeng, Ma Siwei, Gao Wen

Institute of Digital Media, Peking University, Beijing 100871, China

Abstract: Objective: Compression, storage and transmission of local feature descriptors have shown remarkable importance in a wide range of image and video applications. In this paper, we aim at developing a hybrid framework of conventional image/video compression and compact feature representation to make the multimedia smarter for retrieval/analysis driven applications. **Method:** Towards compact, discriminative, and meanwhile efficient representation of feature descriptor, the multiple-reference prediction technique is introduced to remove both the redundancy from both spatial-temporal domain and texture-feature, by efficiently leveraging the information in video coding. Furthermore, the rate-accuracy optimization (RAO) technique is introduced, targeting at optimizing the performance in matching/retrieval based applications. **Result:** Based on the extensive simulations on different test databases, it is shown that the bitrate of visual features can be significantly reduced according to the proposed scheme, while keeping state-of-the-art matching/retrieval performance. **Conclusion:** The proposed system has demonstrated its efficiency and effectiveness towards novel multimedia compression framework for the smart device applications.

Key words: Computer science and technology, visual search, video local feature descriptor, scale-invariant feature transform, high efficiency video coding

基金项目:973 项目(2015CB351800)，国家自然科学基金(61322106)
收稿日期:2015-07-22；　**改回日期**: 2015-08-24
第一作者简介：张翔(1991 年),男,目前就读于北京大学,计算机应用技术专业博士研究生,本科毕业于哈尔滨工业大学,主要研究方向为视频编码与处理。E-mail: x_zhang@pku.edu.cn

0 引 言

随着大数据时代的到来，图像和视频数据量爆发式地增长。同时手持智能设备性能快速发展，以图像和视频视觉特征为核心的移动搜索、增强现实等应用得到了学术界和工业界的广泛关注。通常视觉描述子包含一个全局描述子和若干局部描述子。其中全局描述子刻画了图像的基本特征，一般在大规模视觉检索等应用中进行快速有效的初始匹配。但是全局描述子丢失了重要的图像细节和位置信息，因此引入了局部描述子对图像中重要的兴趣点位置进行细致的描绘，使得在目标匹配、定位等应用中得到更加鲁棒的性能。局部描述子一般保持了对常见变换的不变性，如平移、旋转、相机运动、光照变化、拍摄角度等等。常用的局部描述子包括：Scale-Invariant Feature Transform (SIFT)[1]，Speeded Up Robust Features (SURF)[2]，这些局部描述子在实际应用中被广泛采用，如图像/视频视觉检索、拷贝检测、目标识别、图像压缩等等[3-5]。

以视觉检索应用为例，常用的两种框架包括：1）传输视觉内容，即图像或者视频到服务器端，服务器端提取特征并检索；2）客户端提取局部视觉描述子，并将视觉描述子传输到服务器端进行检索。上述两种方式各有优劣，首先在方案 1 中，为了减少传输代价，图像或者视频都是经过压缩处理的，在压缩后的视觉内容上提取特征会降低特征描述的准确性，进而降低检索等的性能。并且服务器端需要执行完整的特征提取过程，带来较大的复杂度提升。方案 2 较好地解决了传输代价和检索性能的问题，并且在实际中得到了广泛的应用，MPEG 基于此提出了基于图像的紧凑描述子标准 CDVS[6-7]。但是该方案完全丢弃了视觉内容，使得其很难适用于其它基于内容的应用场景。

因此本文提出了一种新的针对视频局部特征的压缩框架，主要思想是将少量的特征流嵌入视频流同时编码并传输。通过重用视频内容编码中的有用信息，使得特征编码达到更高的压缩效率，同时保证了检索效率。我们首先引入了多参考的预测机制，消除了时空冗余以及纹理-特征之间的冗余。此外，还提出了一种新的率失真优化方法——码率-准确率最优化方法（RAO），使得基于匹配/检索应用的性能达到最优。大量实验表明提出的方法能够在保持现有匹配和检索性能的基础上，显著地减少特征压缩带来的比特消耗。

1 总体系统框架

系统总体框架如图 1 所示。视频编码框架保持不变，对视频特征经过特征提取和特征选择后进行高效编码。多参考预测机制从多种预测模式，包括帧内、帧间和重构帧预测中选择，极大地提高了预测的准确性。预测后得到预测残差信号，经过变换、量化、熵编码输出码流。同时为了提高压缩效率，在预测和变换过程中都引入了跳过模式。在预测模式决策以及跳过模式决策的过程中引入了新的率失真模型，即码率-准确率最优化方法，使得在编码器能够根据检索性能自适应地选择最优模式，最大化编码效率。最后编码形成的少量特征码流可以嵌入视频流共同传输，以支持不同应用场景的需求。

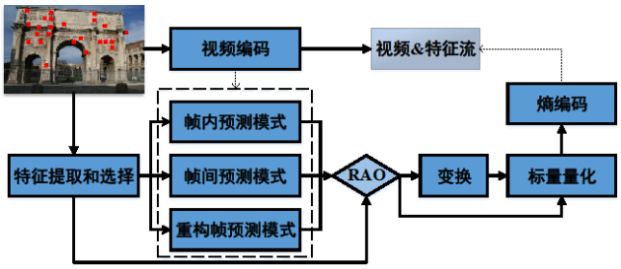

图 1 特征压缩系统总体框架示例图

Fig.1 The overview of the proposed feature compression system

不失一般性地，本文选择 SIFT 特征压缩作为示例，但本文不限于 SIFT 特征，任何局部描述子均可以采用本文提出的方法进行高效压缩。具体的算法设计将在后续章节中详细介绍。必要的数学符号说明见表 1。

表 1 数学符号说明

Table 1 Definition of mathematical symbols

数学符号	含义
$\boldsymbol{S}^i$	第 i 帧中所有特征的集合
$S_j^i \in \boldsymbol{S}^i$	第 i 帧中第 j 个局部描述子
(x_j^i, y_j^i)	第 i 帧中第 j 个局部描述子的横纵坐标
$\boldsymbol{v}_j^i$	第 i 帧中第 j 个局部描述子的特征向量
$\tilde{\boldsymbol{v}}_j^i$	第 i 帧中第 j 个局部描述子的重构特征向量
$\hat{\boldsymbol{v}}_j^i$	第 i 帧重构图像中第 j 个局部描述子的特征向量

2. 多参考预测机制

类似于视频压缩，我们使用多参考的预测机制最大限度地提升特征预测效率。多参考预测模式包

括：帧内预测模式、帧间预测模式以及重构帧预测模式。

2.1 帧内预测模式

帧内预测模式为了去除因图像非局部相似性带来的冗余，在当前帧已编码特征描述子中选择最近邻的作为当前待编码描述子的预测。假设当前待编码局部描述子为$\boldsymbol{v}_j^i$，最优的帧内参考描述子可以通过最小化率失真代价函数得到，

$$(\boldsymbol{v}_{intra}, k) = \underset{\boldsymbol{v}_k^i, k\in[0,j)}{\operatorname{argmin}} \|\boldsymbol{v}_j^i - \boldsymbol{v}_k^i\|_1 + \lambda \cdot R(j-k), \tag{1}$$

式中第一项$\|\boldsymbol{v}_j^i - \boldsymbol{v}_k^i\|_1$表示预测误差，第二项$R(j-k)$代表编码索引的比特数。$\lambda$是拉格朗日乘数，用来控制误差和码率的相对重要性。最优λ的计算将在第3章中详细介绍。

2.2 帧间预测模式

帧间预测的目标是去除时域相邻帧之间的冗余。为了实现高效的帧间运动搜索，该模式中重用了视频编码中的运动矢量信息作为指导。在现有视频编码标准中，每一个帧间预测单元都需要编码运动矢量（MV），包含了水平和垂直位置的偏移（d_x, d_y）以及参考帧索引（d_i）。

假设当前待编码特征为$S_j^i = (x_j^i, y_j^i, \boldsymbol{v}_j^i)$，该特征对应的帧间搜索原点定义为第$i-d_i$帧的（$x_j^i + d_x, y_j^i + d_y$）位置，搜索范围$\Psi$包含$K_\Psi$个距离搜索原点最近的重构特征描述子。类似地，最优的帧间参考描述子可以通过最小化率失真代价函数得到，

$$(\boldsymbol{v}_{inter}, t) = \underset{\boldsymbol{v}_t^{i-d_i}\in\Psi, t\in[0,K_\Psi)}{\operatorname{argmin}} \|\boldsymbol{v}_j^i - \boldsymbol{v}_t^{i-d_i}\|_1 + \lambda \cdot R(t), \tag{2}$$

其中t表示最优参考描述子的索引。值得注意的是当K_Ψ无限大时，该算法退化为蛮力搜索算法。通过实验验证，采用$K_\Psi = 2$是在压缩性能和效率之间较好的权衡。

2.3 重构帧预测模式

第三种预测模式为重构帧预测，即通过在重构视频帧上提取特征作为当前描述子的预测。在这种模式中存在的主要问题是提高了解码复杂度，因为在解码端也需要进行完整的特征提取的步骤。为了解决这个问题，我们提出了通过编码部分描述子参数，使得解码端只需要执行小部分特征提取步骤。编码的参数包括SIFT特征的层数、尺度以及主方向θ，这样解码端可以跳过兴趣点检测过程，直接在对应位置提取特征向量即可。

需要注意的是层数和尺度都是整数，可以采用定长码直接编码。但是主方向θ是浮点数，需要量化到整数范围。设N_θ为量化后主方向的数量，最优的$\widetilde{N_\theta}$可以通过求解最优化函数得到，

$$\widetilde{N_\theta} = \underset{N_\theta}{\operatorname{argmin}}(D(N_\theta) + \lambda \cdot R(N_\theta)), \tag{3}$$

其中$D(N_\theta)$和$R(N_\theta)$分别表示预测误差和码率，两者都可以看作是N_θ的函数。假设使用定长码编码N_θ，那么可以得到$R(N_\theta) = \log_2(N_\theta)$。为了进一步探究预测误差$D(N_\theta)$和$N_\theta$之间的关系，我们通过实验的方式得到如图2所示的散点图，图中横坐标N_θ从4、8、16、32、64中取值，纵坐标表示每一个N_θ对应的预测误差，不同颜色形状的数据点表示不同的测试序列。进而将这些数据拟合成如下指数函数，

$$D(N_\theta) = a{N_\theta}^b + c, \tag{4}$$

其中a、b、c均为模型参数，拟合函数在图2中用黑线画出。将$D(N_\theta)$和$R(N_\theta)$带入式(3)得到，

$$\widetilde{N_\theta} = \underset{N_\theta}{\operatorname{argmin}}(a{N_\theta}^b + c + \lambda \cdot \log_2(N_\theta)). \tag{5}$$

令导数为零，得到最优的$\widetilde{N_\theta}$为λ的函数，

$$\widetilde{N_\theta} = (\frac{-\lambda}{ab\ln 2})^{1/b}. \tag{6}$$

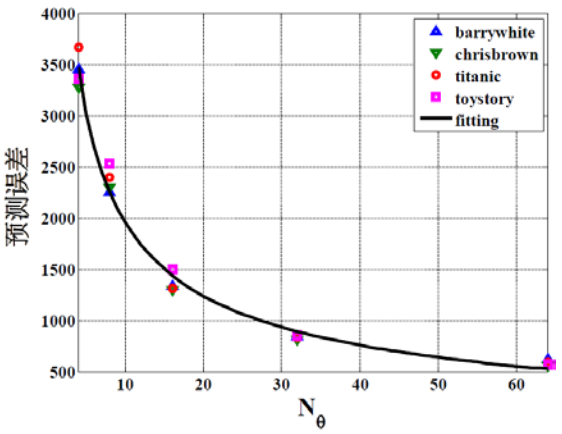

图 2 量化后的方向数量N_θ与预测误差之间的散点图，黑线代表拟合的曲线。

Fig.2 Scatter plot of the prediction error in terms of N_θ, the black solid line is fitted by training data.

3 率失真和码率-准确率最优化方法

率失真优化方法（RDO）被广泛应用在视频编码技术中，目标是在限定码率的条件下最小化编码失真[8-9]，对提高整体编码性能起着至关重要的作用。在本文提出的框架中，传统的率失真最优化方

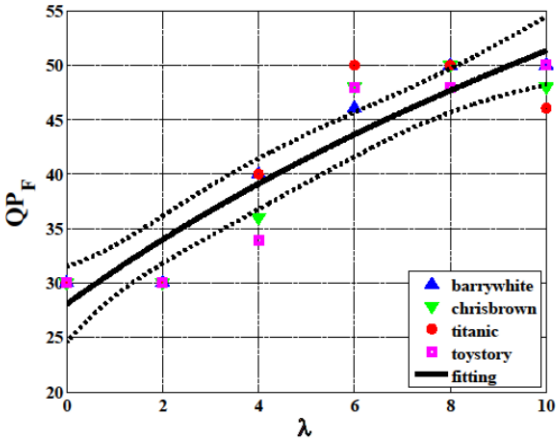

图 3 λ和最优QP_F散点图。黑线实线表示拟合的曲线$QP_F(\lambda) = a \times \ln(\lambda + b) + c$，其中$a = 38.25, b = 11.93, c = -66.83$。黑色虚线表示 95%置信区间。

Fig.3 The relationship between the optimal QP_F and λ. The fitted function $QP_F(\lambda) = a \times \ln(\lambda + b) + c$ is plotted with 95% confidence bounds, where $a = 38.25, b = 11.93, c = -66.83$.

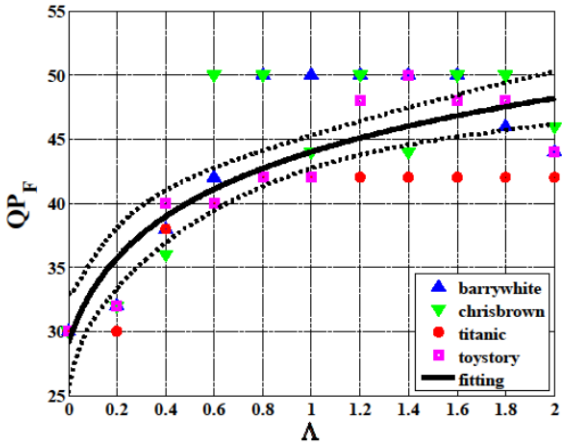

图 4 Λ和最优QP_F散点图。黑线实线表示拟合的曲线$QP_F(\Lambda) = a \times \ln(\Lambda + b) + c$，其中$a = 6.539, b = 0.114, c = 43.28$。黑色虚线表示 95%置信区间。

Fig.4 The relationship between the optimal QP_F and Λ. The fitted function $QP_F(\Lambda) = a \times \ln(\Lambda + b) + c$ is plotted with 95% confidence bounds, where $a = 6.539, b = 0.114, c = 43.28$.

法应用在各个预测模式中，用来选择最优的参考描述子，该过程可以通过下式表示，

$$\min(J), where\ J = D + \lambda R, \qquad (7)$$

其中R和D分别表示码率和失真，D使用绝对值差和（即 SAD 指标）来计算。λ是拉格朗日乘子，用来控制R和D之间的相对重要性。

众所周知，λ的选择对编码效率的影响极大。为了得到最优的λ，我们通过训练的方式来探究λ和QP_F之间的关系，其中QP_F表示编码特征时的量化参数。对于一个固定的λ，QP_F从 30 变化到 50，在特征压缩过程中，使用使得率失真代价函数最小的模式进行编码。于是可以得到一个最优的QP_F，使得编码后根据式(7)计算得到的总代价最小。最后通过改变λ的取值范围，得到每一个λ下对应的最优QP_F，画出散点图如图 3 所示，图中不同颜色形状的数据点表示不同的测试序列。然后使用对数函数$QP_F(\lambda) = a \times \ln(\lambda + b) + c$拟合训练数据点。在实际编码过程中，即可使用训练得到的对数模型在给定QP_F的条件下估计最优的λ值。

然而特征描述子是用来做匹配和检索的，传统率失真优化方式中使用 SAD 来衡量失真的方式并不能很好地反映其在具体应用中的真实失真。因此我们提出了一种新的码率-准确率最优化方法（RAO），使用特征描述子在匹配中的准确率指标代

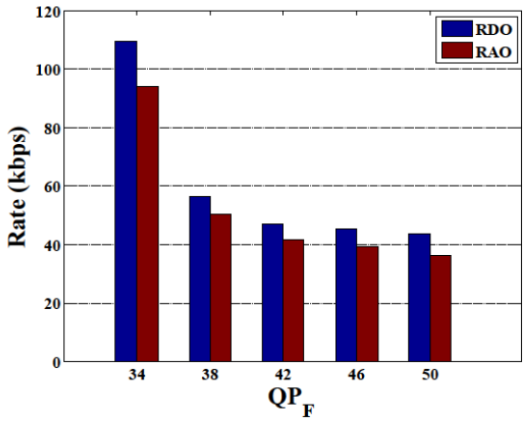

图 5 率失真优化方法和码率-准确率优化方法对比结果。

Fig. 5 Rate comparison between RDO and RAO.

替传统率失真方法中的失真项。类似地，码率-准确率最优化方法可以形式化地表达为最小化以下代价函数，

$$\min(J_A), where\ J_A = D_A + \Lambda R, \qquad (8)$$

其中R和Λ表示码率和拉格朗日乘子。D_A量化了特征压缩在匹配或者检索性能上的失真。本文中使用基于成对比较的排序之间的差异来近似估计D_A。假设$\boldsymbol{v}$和$\tilde{\boldsymbol{v}}$分别代表原始和重构的特征描述子。$\mathbb{F}$表示一个描述子的集合，作为匹配的目标集合。对于$\mathbb{F}$中的任意一个描述子$\boldsymbol{d}^i \in \mathbb{F}$，根据其与原始特征描述子之间的距离可以得到一个排序$\boldsymbol{R_o}$，使其满足下式：

$$r_o^i < r_o^j, s.t. \|\boldsymbol{d}^i - \boldsymbol{\nu}\|_1 < \|\boldsymbol{d}^j - \boldsymbol{\nu}\|_1, \quad (9)$$
$$i \neq j \in [1,K]$$

其中$r_o^i, r_o^j \in \boldsymbol{R_o}$，$K$表示集合$\mathbb{F}$的大小。同样地，可以得到关于重构描述子$\tilde{\boldsymbol{\nu}}$的排序$\boldsymbol{R_r}$，

$$r_r^i < r_r^j, s.t. \|\boldsymbol{d}^i - \tilde{\boldsymbol{\nu}}\|_1 < \|\boldsymbol{d}^j - \tilde{\boldsymbol{\nu}}\|_1, \quad (10)$$
$$i \neq j \in [1,K]$$

其中$r_r^i, r_r^j \in \boldsymbol{R_r}$。然后使用斯皮尔曼相关性系数来计算两个排序之间的距离作为失真项的度量，即

$$D_A \triangleq 1 - SROCC(\boldsymbol{R_o}, \boldsymbol{R_r}) = \frac{6\sum(r_o^i - r_r^i)^2}{K(K^2-1)}. \quad (11)$$

然而码率-准确率模型中的率失真关系已经完全不同，因此需要重新训练得到Λ和最优QP_F之间的关系。训练的方式和上述率失真模型中的一样，不再赘述，得到的散点图和拟合曲线如图 4 所示。从图中可以看出码率-准确率模型中曲线的变化趋势较率失真模型收敛得更快。本文将码率-准确率最优化方法应用在了预测以及跳过的模式决策过程中。图 5 给出了率失真优化方法和码率-准确率优化方法的对比实验结果，可以看出在不同QP_F测试点下，码率-准确率方法都能显著减少特征编码的比特消耗。同时匹配和检索性能几乎没有改变。

4. 实验结果与分析

为了验证提出方法的有效性，我们在斯坦福数据库[10]上进行了大量实验，该数据库中所有的视频均由手持设备拍摄，视频序列格式为 VGA（640 × 480），帧率为 30 帧每秒，拍摄对象包括 CD、DVD、书本等常见兴趣对象，拍摄过程涉及相机运动以及物体运动，是研究视频特征压缩的权威数据库之一。实验中，视频编码器采用的是最新标准 HEVC[11]提供的参考软件 HM-14.0，编码配置采用的是低延时 P 帧，且 GoP 大小设置为 4。为了有效地评估压缩后特征在匹配检索上的性能，采用经过几何校验后成对匹配的个数$N_{inliers}$作为其评价指标。

将本文提出的方案和最新的视频局部特征压缩方法[12]进行了比较，该方法主要去除了特征的帧间冗余，并且在特征编码模式决策过程中使用了传统的率失真优化方法。图 6 中展示了部分对比实验结果，两者采用了相同的码率和检索性能对比指标，需要注意的是纹理图像编码的码率没有参与计算。从中可以看出提出的方法在检索性能几乎保持不变的基础上，极大地降低了特征的比特消耗。使得可以通过在视频码流中加入少量的特征信息达到高效

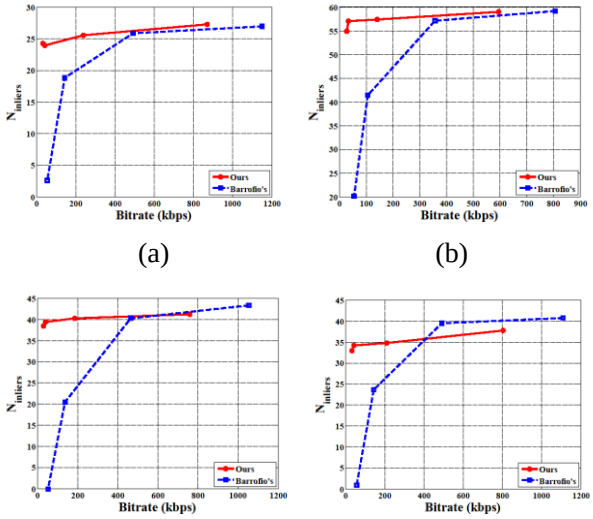

图 6 本文提出方案（红实线）与方法[12]（蓝虚线）对比实验结果。横坐标表示码率，纵坐标表示匹配性能$N_{inliers}$。测试序列为：(a) barrywhite, (b) chrisbrown, (c) toystory, (d) titanic。

Fig. 6 Performance comparison between the proposed scheme (red solid line) and the Barrofio's framework [12] (blue dashed line). The horizontal axis indicates the rate and the vertical axis is the number of matched pairs $N_{inliers}$.

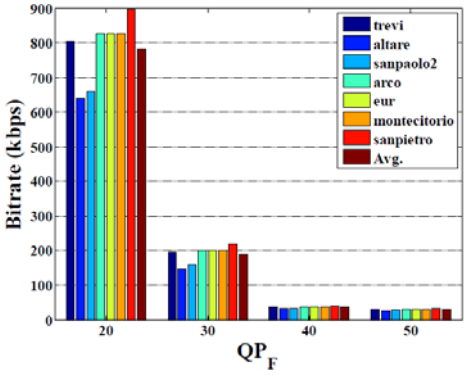

图 7 算法在地标检索数据库[13]上的实验结果，其中不同QP_F下的第一检索正确率都为 100%。

Fig. 7 Simulation results of bitrate under landmark retrieval database [13], where the average precise of rank 1 for each operation point is always 100%.

的检索性能。

为了进一步检验该方法对检索有效性，我们针对特定的应用场景，即地标检测进行了实验。数据库采用的是罗马地标数据库[13]。其中包含了一万张图像，十个查询视频在数据库中分别对应九幅图像。我们使用提出的方法，在不同QP_F测试点下进行了实验，发现在第一检索正确率保持 100%的基础下，特征可以达到很高的压缩比。

5 结　论

本文创新性地提出了一种面向多媒体检索分析等新型智能应用的混合压缩框架。通过在传统多媒体码流中加入少量表示特征的信息，使得其可以达到高效的检索性能。本文解决的主要问题是如何利用图像/视频编码中的信息更加有效地表示视觉特征。首先我们引入了多参考的预测机制，消除了时空冗余以及纹理-特征之间的冗余。此外，我们还提出了一种新的率失真优化方法，即码率-准确率最优化方法，进一步了提高特征表示的效率。大量实验表明提出的方法能够在保持现有匹配和检索性能的基础上，显著地减少特征带来的比特消耗。并且允许用户根据不同应用场景和实时性要求，自适应配置编码方法。今后我们将在该框架下进一步研究更加紧凑高效的特征压缩算法，同时研究如何利用紧凑特征提高图像/视频内容编码效率的方法，使得特征和视觉内容的压缩能够达到互相促进的作用。

参考文献(References)

[1] D. Lowe. Object recognition from local scale-invariant features [C]. Proceedings of IEEE International Conference on Computer Vision, Kerkyra: IEEE, 1999, 2: 1150–1157. [DOI：10.1109/ICCV.1999.790410]

[2] H. Bay, T. Tuytelaars, L. Van Gool. Surf: Speeded up robust features [C]. Proceedings of Computer Vision–ECCV, Berlin Heidelberg: Springer, 2006, 404–417. [DOI:10.1016/j.cviu.2007.09.014]

[3] Li Y N, Liu T, Jiang S Q, Huang Q M. Video matching method based on “bag of words” [J]. Journal on Communications, 2007, 28(12). [李远宁，刘汀，蒋树强，黄庆明. 基于“Bag of Words”的视频匹配方法 [J]. 通信学报, 2007, 28(12).] [DOI: 10.3321/j.issn:1000-436x.2007.12.025]

[4] SU Y, Shan S G, Chen X L, Gao W. Integration of Global and Local Feature for Face Recognition [J]. Journal of Software, 2010, 21(8): 1849-1862. [苏煜，山世光，陈熙霖，高文. 基于全局和局部特征集成的人脸识别方法 [J]. 软件学报, 2010, 21(8): 1849-1862.]

[5] Qin L, Hu Q, Huang Q M, Tian Q. Action Recognition Using Trajectories of Spatio-Tempral Feature Points [J]. Chinese Journal of Computers, 2014, 37(6): 1281-1288. [秦磊，胡琼，黄庆明，田奇. 基于特征点轨迹的动作识别 [J]. 计算机学报, 2014, 37(6): 1281-1288.] [DOI: 10.3724/SP.J.1016.2014.01281]

[6] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y. A. Reznik. Mobile visual search: Architectures, technologies, and the emerging mpeg standard [J]. IEEE Trans. Multimedia, 2011, 18(3): 86–94. [DOI: 10.1109/MMUL.2011.48]

[7] Duan L Y, Huang T J, Gao W. Overview of the MPEG CDVS standard [C]. Proceedings of Data Compression Conference (DCC), Snowbird, UT: IEEE, 2015. [DOI: 10.1109/DCC.2015.72]

[8] G. Sullivan and T. Wiegand. Rate-distortion optimization for video compression [J]. IEEE Signal Processing Magazine, 1998, 15(6): 74–90. [DOI: 10.1109/79.733497]

[9] Ma S W, Gao W. Rate Distortion Optimization Based Video Coding [J]. Journal of the Graduate School of the Chinese Academy of Sciences, 24(1):137-143, 2007. [马思伟，高文. 基于率失真优化的视频编码研究[J]. 中国科学院研究生院学报, 24(1):137-143, 2007.]

[10] M. Makar, V. Chandrasekhar, S. Tsai, D. Chen, B. Girod. Interframe coding of feature descriptors for mobile augmented reality [J]. IEEE Trans. Image Processing, 2014, 23(8): 3352–3367. [DOI: 10.1109/TIP.2014.2331136]

[11] G. Sullivan, J. Ohm, W.-J. Han, T. Wiegand. Overview of the high efficiency video coding (HEVC) standard [J]. IEEE Trans. Circuits System Video Technology, 2012, 22(12): 1649–1668. [DOI: 10.1109/TCSVT.2012.2221191]

[12] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, S. Tubaro. Coding visual features extracted from video sequences [J]. IEEE Trans. Image Processing, 2014, 23(5): 2262–2276. [DOI: 10.1109/TIP.2014.2312617]

[13] BAROFFIO L, Rome landmark dataset [EB/OL]. [2015-03-22]. https://sites.google.com/site/greeneyesprojectpolimi/downloads/datasets/rome-landmark-dataset.