自然场景图像与合成图像的快速分类

刘国帅[1,2]，仲伟峰[1]，殷飞[2]，刘成林[2]

[1.]哈尔滨理工大学自动化学院，黑龙江省哈尔滨市，150080；

[2.]中国科学院自动化研究所模式识别国家重点实验室，北京市，100190；

摘　要：本文针对当今互联网中两种主要的图像类型：自然场景图像与合成图像，设计层次化的快速分类算法。该算法包括两层，第一层利用两类图像在颜色，饱和度以及边缘对比度上表现出的差异性提取全局特征，并结合支持向量机 (support vector machine, SVM) 进行初步分类，第一层分类结果中低置信度的图像会被送到第二层中。在第二层中，系统基于词袋模型 (Bag-of-Words) 对图像不同区域的纹理信息进行编码得到局部特征并结合第二个 SVM 分类器完成最终分类。为测试算法的实用性，我们同时收集并发布了一个包含约 30,000 张图像的数据库。实验结果表明，在单核 Intel(R) Xeon(R) (2.50GHz) CPU 上，本文提出的算法分类精度可达到 98.26%，分类速度超过 40FPS。

关键词：　图像类型快速分类；特征提取；词袋模型；层次化分类算法

Fast Classification of Natural Scene and Born-Digital Images

Guo-shuai Liu[1], Wei-feng Zhong[1], Fei Yin[2], Cheng-lin Liu[2]

[1.] School of Automation, Harbin University of Science and Technology, Harbin City, Heilongjiang Province, 150080;

[2.] National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences, Beijing, 100190;

Abstract: In this paper, we propose a hierarchical algorithm for the fast genre classification of natural scene images and born-digital images, which are the most prevalent image types on the Internet. Our algorithm consists of two stages; the first stage extracts global features reflecting distributions of color and saturation and uses a support vector machine (SVM) classifier for classification. The images assigned low confidence by the first-stage classifier are processed by the second stage, which extracts local texture features represented in the Bag-of-Words framework and uses another SVM classifier for final classification. To validate experimentally the effectiveness of our proposed method, we also build a database containing about 30,000 images from different sources. On our test image set, we obtained an overall accuracy of 98.26% and the processing speed is over 40FPS on an Intel(R) Xeon(R) (2.50GHz).

Key words: Fast Genre Classification of Images; Feature Extraction; Bag-of-Words; Hierarchical Classification Algorithm

1　引言

随着互联网、智能手机和通信技术的迅速发展，互联网上的图像数据在快速增长。海量的图像可以在网络信息分析、商业数据挖掘、敏感信息检测等领域中发挥巨大作用，因此，网络图像的处理、分析、搜索和理解得到了学术界和工业界的广泛重视。考虑到网络中图像类型的多样性，以及不同类型的图像需要不同的处理方法，基于图像类型设计的快速分类算法可在实际图像检索与信息挖掘系统中起到前期过滤的作用。本文针对互联网中最常见的两类图像，自然场景图像与网络合成图像设计快速分类算法。该算法基于两类图像在视觉表现形式与图像生成方式上的差异性设计全局与局部特征，并应用于层次化的分类框架中。

第一作者简介：刘国帅，男，哈尔滨理工大学与中国科学院自动化研究所联合培养硕士研究生（1991-），模式识别与智能系统方向，guoshuai.liu@nlpr.ia.ac.cn。

国内外学者针对特定内容与类型的图像提出了很多有效的特征提取与分类方法。文献[1]，[2]分别对城市场景/田园场景图像，室内/室外图像利用边缘方向直方图进行分类，其中[2]在做室内/室外图像分类时，基于室内图像多包含具有水平竖直线条的人工制品这一假设，利用水平与竖直边缘方向直方图作为主要特征进行分类。但在本问题中，互联网中图像数据的混杂性导致即使同一类图像可能在内容上千差万别，或者不同类型的图像却包含相似的图像内容，因此，基于特定图像内容所设计的特征无法用于图像类型的分类。Swain 在[3]中曾利用图像颜色与饱和度信息对拍照图像与早期网络中简单的图形图像（国家旗帜，公司商标，地图等）进行分类，但是目前的网络合成图像中包含了更丰富的内容以及更复杂的版面结构，有些合成图像甚至会包含一部分场景图像，单纯利用颜色等信息的全局统计特征无法再做出正确的分类。Hammoud 等人[4]，[5]在处理扫描油画图像与拍照油画图像的分类问题时提出了颜色边缘与亮度边缘的概念，对解决本问题有一定的意义。在图像真伪检测方面，针对自然拍照图像与计算机合成仿真图像在生成过程中的差异性，Chang 在文献[6]中提出了基于几何特征的分类模型，并取得了一定的分类效果，但是该模型计算复杂度较高，在处理 1280×1024 大小图像时，仅提取分形几何特征就需要128.1s，无法满足实际系统对实时性的要求。此外，近年来，深度神经网络算法，尤其是卷积神经网络[7-10]，在图像特定目标检测与识别领域取得了非常好的性能，但是卷积神经网络需要大量人工标注样本训练网络，同时在训练与应用过程中需要 GPU 加速计算，导致在实际应用中受到了限制。

本文提出的基于层次化的分类模型，充分考虑了系统对于分类精度与速度的要求，第一层采用的全局特征，足以保证对大部分图像做出正确的分类。少量复杂图像（低置信度）被输入到第二层中，在 Bag-of-Words 框架下融合区域纹理信息得到区分性更强的局部特征用于最终分类。实验表明，本文提出的算法能够以超过 40FPS 的分类速度获得 98.26%的分类正确率。

2 层次化分类框架

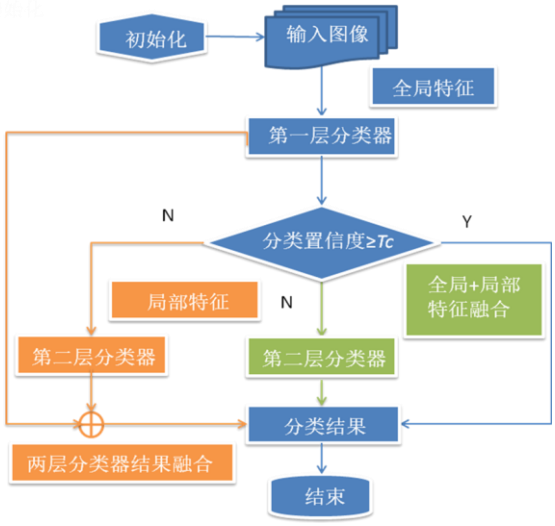


图 1：层次化分类算法整体流程图

2.1 基本思想

本文设计的层次化分类系统如图 1 所示，该系统主要包括两层，第一层对输入图像提取全局特征，并输入到支持向量机（SVM）分类器中进行分类，该层计算出图像后验类别概率并作为分类置信度。分析得到的二维置信度向量，如果该向量中最大值不小于指定阈值 T_c，则直接图像判别为最大值所属的类别并结束分类；反之，则需要进入第二层提取局部特征。局部特征利用 Bag-of-Words 框架编码图像中三种局部区域的纹理特征获得，第二层同样采用 SVM 分类，但是分类过程中需要融合上一层的信息。为此，我们设计了两种融合策略，另外为进一步提高分类速度，在层次化分类框架中，针对第一层我们还分别尝试了线性与非线性的 SVM 分类器。

2.1.1 特征融合

第一种融合策略是两层间的特征融合。在第二层分类器的训练阶段，我们同时提取训练样本的全局与局部特征并将其拼接到一起用来训练 SVM 分类器。在预测阶段，该层分类器的输出直接作为系统最终分类结果。

2.1.2 分类器结果融合

在第二种融合方式中，直接使用局部特征训练第二层分类器。在预测阶段，加权两层分类器的置信度向量并在结果向量中选择具有最大值的类别作为最终分类结果。

2.2 全局特征

考虑到合成图像与自然场景图像在颜色，饱和度以及边缘对比度上的分布具有明显的差异，同时为保证第一层具有较快的处理速度，全局特征设计如下：

2.2.1 高饱和像素聚合度： f_1

合成图像经常由数块单一色彩构成的区域块拼接而成，且设计上常使用饱和度较高的颜色以引人注意。相反，自然拍照图像整体的饱和度较低，同时高饱和度的像素也没有明显的聚合现象。令 I_{bgr} 表示原始图像，I_s 表示饱和度通道，给定饱和度阈值 T_s，计算高饱和度像素的位置模板图像 I_{mask1}，

$$I_{mask1}(x,y)=\begin{cases}1, & if \quad I_s(x,y)\ge T_s\\ 0, & otherwise\end{cases} \tag{1}$$

利用 3×3 的矩形结构元素对 I_{mask1} 做腐蚀操作得到 I_{mask2}，分别统计 I_{mask1} 与 I_{mask2} 上非零元素的个数并记作 N_1, N_2，则 $f_1 = N_2/N_1$。

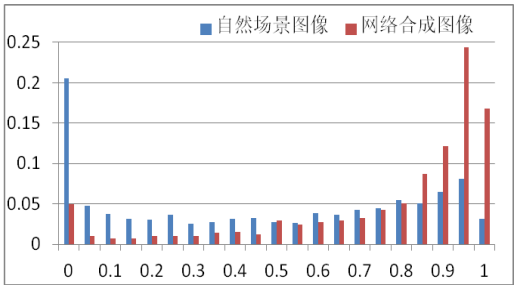
为测试该特征的有效性，我们在数据库中对每一类图像各随机抽取 5,000 张，组成 10,000 张采样数据，并在每一张采样图像上提取高饱和像素聚合度特征。图 2 (a) 展示了两类图像在该特征上的分布直方图，可以看出，合成图像由于存在高聚合度的区域块，在该特征上表现出更高的取值（接近 1）。

2.2.2 边缘像素平均对比度： f_2

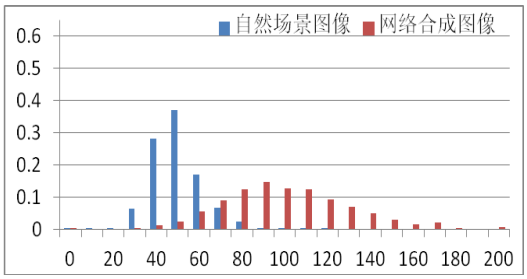
合成图像存在大量由区域块拼接而成的“色彩边缘”，而场景图像中的边缘是由于遮挡，光照与材料反光属性不同形成的“光照边缘”，两类边缘的视觉对比度会存在差异。令 I_c 为原始图像的 *canny* 二值图像（边缘位置图像取值为 1，其他区域取值 0），I_g 为归一化的灰度图像，则对应的局部极大对比度图 M_{ms} 为

$$M_{ms}(x,y)=\begin{cases}max\{|I_g(x,y)-I_g(x',y')|\}, & if \quad M_c(x,y)\ge 0\\ 0, & otherwise\end{cases} \tag{2}$$

其中 $max\{|x-x'|,|y-y'|\}=D$，其中 D 表示像素单位的邻域距离，本文实验中 D 取值为 1。以 M_c 为模板计算 M_{ms} 的均值作为边缘像素的平均对比度 f_2。图 2 (b) 为两类图像在该特征上的分布情况。



(a) 两类图像在高饱和像素聚合度上的分布直方图



(b) 两类图像在边缘像素平均对比度上的分布直方图

图 2：两类图像在全局特征 f_1, f_2 上的分布直方图

2.2.3 颜色直方图： f_3

由于拍照条件、所拍图像内容的任意性，自然场景中的颜色分布具有很大的不确定性，相反，网络合成图像本身颜色数目很少，且经常出现的颜色类别也相对比较固定。由此，可以直接提取原始图像的颜色直方图并以此作为全局特征。本文同时尝试了在两种不同的颜色空间中提取直方图。第一种颜色空间是 *RGB* 空间，为保证计算速度，首先分别将 *RGB* 空间中三个通道的颜色数目从 256 均匀量化到 8，则量化后整个空间内的颜色数目为 512 (8^3)。然后在此颜色空间中提取原始图像的颜色直方图，最终得到 512 维的直方图向量 f_3^{bgr}。

为进一步提高运算速度，第二种方式使用 *HSV* 颜色空间中的色度通道统计图像中的颜色分布信息。具体做法：先将原始图像 I_{bgr} 转换到 *HSV* 颜色空间，并取出色度通道记作 I_h，利用 I_h 的直方图代替 I_{bgr} 直方图，最终得到 180 维的颜色直方图特征 f_3^h，表 1 给出了分别使用两种直方图作为全局特征时分类器的性能（见 1)和 2)），可以看出，在全局特征中使用 f_3^{bgr} 时，分类速度约为 31FPS，精度为 95.34%，而 f_3^h 则能够以更快的速度(超过 40FPS)获取近似的分类精度（94.95%）。

2.3 局部特征

本文基于 Bag-of-Words [11,12] 框架分别对三类局部区域的特征描述子进行近邻约束线性 [12] (Locality-constrained Linear Coding, LLC) 编码，并级联三种编码向量最终得到关于一幅图像局部细节的整体描述。三类局部区域的数目均为 N_{lp}，大小均为 $S_{lp}\times S_{lp}$。

2.3.1 局部区域及对应的特征描述子

- 局部平滑区域

局部平滑区域的确定方法如下：首先利用 *sobel* 算子计算出原始图形的梯度强度图 M_g，然后利用给定阈值 T_g 对 M_g 进行二值化得到 I_{mask3}，最终以 I_{mask3} 作为模板，随机选择 N_{lp} 个大小为 $S_{lp}\times S_{lp}$ 的区域块，保证每一个区域块与 I_{mask3} 的重合面积大于 $S_{lp}\times S_{lp}\times r$,实验中 r 取值 0.7。提取这些区域的局部二值模式（Local Binary Pattern，LBP）作为特征描述子。

- 局部边缘区域

局部边缘区域的选择方法：随机选择 N_{lp} 个以 *canny* 边缘点为中心，大小为 $S_{lp}\times S_{lp}$ 的区域。同样使用 LBP 作为该类区域的特征描述子。

- 局部随机区域

随机区域作为对其他类型区域的补充，通过在原始图像中随机选择 N_{lp} 个大小为 $S_{lp}\times S_{lp}$ 的区域块即可。随机区域采用颜色缩减后的图像颜色直方图进行描述。具体做法：首先将 I_{bgr} $(256\times 256\times 256)$ 的每一个颜色通道从 256 均匀量化到 4，则压缩后的图像具有 $64=4^3$ 中颜色；然后对颜色压缩后的图像中的每一种颜色进行编码，重新得到颜色索引图像 I_{rindex}，统计 I_{rindex} 上随机区域的直方图作为该区域的特征描述，直方图设置为 64 维。假设原始图像汇中某个位置的像素点三通道的颜色值 $(134,201,17)$，量化后则变为 $(2,3,0)$，则 I_{rindex} 上对应的编码为 44 $(2\times 4^2+3\times 4^1+0\times 4^0)$。

2.3.2 局部特征的整体描述

考虑到传统的 Bag-of-Words 方法无法描述局部特征在空间上的分布特性，我们采用了近邻约束的线性编码方式分别对上述每一种局部区域的描述子进行编码，然后将三种类型的编码向量级联到一起，形成局部特征的整体描述，并作为层次化分类系统的局部特征。基于 LLC [12]对图像进行编码，最主要的一步是建立视觉字典，考虑到训练图像样本来自于互联网，同种类型的样本在图像内容上没有太多的关联性，因此本文利用两级聚类的方法分别针对每一类局部区域的描述子生成字典。下面以局部边缘区域为例，介绍字典的层次化生成方式。

设训练集 $S_{train}^{N}=\{I_0,I_1\cdots I_k\cdots,I_{N-1}\}$,第 k 张图像对应的局部边缘区域的特征描述子集合表示为 $S_k^{N_{lp}}=\{x_E^0,x_E^1\cdots x_E^m\cdots,x_E^{N_{lp}-1}\}$，

- 首先对每一张图像的 $S_k^{N_{lp}}$ 进行 $k-means$ 聚类，提取 N_{c1} 个聚类中心，得到单张图像的类中心集 $S_k^{N_{c1}}=\{c_E^0,\cdots c_E^m\cdots,c_E^{N_{c1}-1}\}$。
- 合并训练集中所有图像的类中心集，记作 $S_{center}^{N}=\{S_0^{N_{c1}},S_1^{N_{c1}},\cdots,S_N^{N_{c1}-1}\}$。
- 再次使用 $k-mean$ 方法对 S_{cent}^{N} 中的 $N_{c1}\times N$ 个中心元素进行聚类，提取出 N_{c2} 个类中心并将其作为局部区域描述子的视觉字典。

利用同样的方式针对另外两类区域的特征描述子生成对应的视觉字典，子聚类中心与全局聚类中心均为 N_{c1} 和 N_{c2}，得到视觉字典后，在 Bag-of-Words 框架下对图像中不同区域的纹理信息进行编码，并级联三类编码向量，即可得到 $3\times N_{c2}$ 维的局部特征。

(a) 自然场景图像

(b) 合成图像

图 3：自然场景图像与合成图像

3 实验结果与分析

3.1 自然场景与合成图像数据集

为测试本文提出的层次化分类算法的性能，我们建立并发布了一个包含约 30,000 张图像的数据集。该数据集包含 19,670 张自然场景图像和 10,508 张合成图像。其中，自然场景图像中有 12,654 张直接来自于互联网中，5,175 张来自于 SUN397 Database [14], 剩余 1,841 张通过手动拍照获取。所有的合成图像也均来自互联网。图 3 展示了数据库中一些样本图像。

3.2 实验结果

本文实验采用的分类器为 SVM，非线性核为径向基函数（Radial Basis Function, RBF），为保证分类速度，如果原始图像的高度与宽度大于1000×1000，则直接在原始图像上随机截取1000×1000的图像块输入到分类系统中。参数$T_s=200, T_g=2, S_{lp}=15, N_{lp}=200, N_{c1}=5, N_{c2}=100, T_c=0.95$，从数据集中随机选取70%作为训练样本，剩余图像作为测试样本。

为验证本文中全局与局部特征对分类问题的有效性以及层次化分类框架的优势，我们总共做两组实验。第一组实验用来测试所提取的全局与局部特征的有效性，其中 1)：直接使用全局特征和非线性 SVM 进行分类实验，全局特征中颜色直方图为 f_3^{bgr} 。2)：在 1）的基础上，将 f_3^{bgr} 替换为 f_3^h，其余保持不变。3)：直接使用全局+局部特征和非线性SVM进行分类实验，全局特征中使用 f_3^h 。

第二组实验用来验证层次化分类框架在整体分类性能上的优势，并比较使用线性与非线性分类器对分类性能的影响。其中，4)：采用层次化分类框架以及第一种融合策略（特征融合）进行分类，对于两层分类器均选用非线性 SVM； 5)：采用层次化分类框架，以及第二种融合策略（分类器结果融合），第一层分类器为线性 SVM，第二层分类器为非线性 SVM；6)：将 5）中第一层分类器修改为非线性 SVM，其余保持不变。在第二组实验中，全局特征中颜色直方图均使用 f_3^h，另外，凡采用第二种融合策略的实验，第二层分类器的加权系数设置为 0.9。表 1 给出了所有实验的分类结果。其中 FPS (Frame Per Second) 表示系统的整体分类速度。

表 1：自然场景与合成图像分类结果

实验编号	分类精度（%）	分类速度（FPS）
1）	95.34	31.52
2）	94.95	41.99
3）	**98.39**	**25.75**
4）	**98.26**	**40.18**
5）	97.94	36.69
6）	98.17	39.22

3.3 结果分析

从表 1 中的 1）和 2）可以看出，全局特征本身计算复杂度低，单纯使用全局特征的分类精度为 94.95%，分类速度接近 42FPS；3）表明在全局特征的基础上融合计算复杂度较高的局部特征，一方面可以获得更高的分类精度（98.39%），同时系统要在分类速度上做出牺牲（25.75FPS）。实验 4），5）和 6）均表明，层次化的分类框架能够同时满足系统对于分类精度与速度的双重要求，以接近 2）的分类速度（40.18FPS）获取 98.26%的分类性能，这主要是因为，大部分图像足够“简单”，仅使用全局特征即可获得较高的分类置信度与正确的分类结果，只有极少量“复杂”图像，分类器才需要依靠全局与局部特征共同作用来做出决策。整体上，层次化分类系统通过选择性地处理“复杂”样本来保证以较快的分类速度实现较高的分类精度。

(a) 被错分为合成类的自然场景图像

(b) 被错分为场景类的合成图像

图 4：错分样本

另外，在分类实验中，总有一些图像不能被正确分类，图 4 中展示了这些错分样本。图 4 (a) 中的自然场景图像被错分为合成图像，图 4 (b) 则刚好相反。(a) 中的图像由于所拍照场景的特殊性，含有较少的颜色和大面积相同颜色的区域，在视觉上更像合成图像；(b) 中图像本身包含大面积自然场景（只含有少量合成文字），因而被错分到自然场景类中。接下来我们打算引入场景/合成混合类来解决该问题。

4 结论

在本文中，我们针对自然场景图像与合成图像提出了一种层次化的快速分类算法。通过综合考虑两类图像在颜色，饱和度，边缘对比度以及局部纹理上的差异，我们设计出一组快速有效的特征，结合层次化的分类框架，系统的分类精度为 98.26%，分类速度超过 40FPS。该算法可应用到对实时性要求较高的图像检索与数据信息挖掘等实际项目中。另外，我们还建立并发布了一个包含约 30,000 张自然场景与合成图像的数据库供学术界免费使用。

在接下来的工作中，我们会进一步扩充数据集的规模，细化图像数据的类别，并寻找对图像更加有效的特征描述算法，提高系统的分类速度与精度。

参考文献

[1] Gorkani, M.M., Picard, R.W.: Texture orientation for sorting photos ”at a glance”. In: Proceedings of 12th International Conference on Pattern Recognition. (1994) 459.

[2] Yiu, E.: Image Classification Using Color Cues and Texture Orientation. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science (1996).

[3] Athitsos, V., Swain, M.J., Frankel, C.: Distinguishing photographs and graphics on the world wide web. In: Proceedings of the 1997 Workshop on Content-Based Access of Image and Video Libraries, IEEE Computer Society (1997).

[4] Cutzu, F., Hammoud, R.I., Leykin, A.: Estimating the photorealism of images:Distinguishing paintings from photographs. In: CVPR, IEEE Computer Society (2003) 305–312.

[5] Hammoud, R.I.: Color texture signatures for art-paintings vs. scene-photographs based on human visual system. In: 17th International Conference on Pattern Recognition, IEEE Computer Society (2004) 525–528.

[6] Ng, T.T., Chang, S.F., Hsu, J., Xie, L., Tsui, M.P.: Physics-motivated features for distinguishing photographic images and computer graphics. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, ACM (2005) 239–248.

[7] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012.

[8] Bluche, T., Ney, H., Kermorvant, C.: Feature extraction with convolutional neural networks for handwritten word recognition. In: ICDAR, IEEE Computer Society (2013) 285–289.

[9] Huang, W., Qiao, Y., Tang, X. In: Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees. Springer International Publishing (2014)497–511.

[10] Jaderberg, M., Vedaldi, A., Zisserman, A. In: Deep Features for Text Spotting.Springer International Publishing (2014) 512–528.

[11] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. (2006) 2169–2178.

[12] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vison and Pattern Classification. (2010).

[13] Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: Proceedings of the 9th European Conference on Computer Vision - Volume Part I, Springer-Verlag (2006) 430–443.

[14] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN Database: Largescale Scene Recognition from Abbey to Zoo. In: CVPR. (2010).