基于 logistic 回归辨别图像的真伪

摘要

随着图像处理及分析技术的迅猛发展，图像编辑越来越广泛，同时也产生了一些不法分子对图像进行恶意操作来牟取利益。因此建立识别图像真假的模型，并能够准确判断图片是否被改动过，具有较强的现实意义与应用价值。

对于问题一：基于给出的真假图片，人为找出部分分辨图像真假的指标并说明理由。因为统计类的相关方法很难找出与图片进行相关性分析的指标，所以我们通过人为查看比对大量的真假图片，发现大量图片存在物体颜色受污染的情况，使得颜色渐变程度变小，色彩变单一，图片偏向暖色调，以及存在图片颜色加深等情况。经查询一些有关图像的指标，将虚假图片变化情况进行统计，一一对应到图片各个颜色指标上去。观察真假图片直方图、饱和度、亮度、图片 RGB 占比、平衡梯度、对比度、信息熵、联合熵，明度等指标，发现真假图片的指标均有不同，故初步筛选出以上 8 个指标为影响指标，说明理由详见文中，人为观察图片可以更加直观快速的找出真假图片的差异，节省时间。

对于问题二：根据问题一找出的指标，给出识别标准，建立分辨图像真伪的识别模型，并检验模型的实用性。该问题输出结果为图像真伪，属于典型的二分类预测类型问题。考虑到问题一中选取指标过多，可能需要删去部分无效指标，故我们取出真假前 4700 张图片作为已知样本，使用这些数据进行逐步回归，构建 logistic 回归模型，然后在删除亮度、对比度、RGB 占比指标后，达到拟合效果良好，为 89.41%。故最终选取了色调，饱和度，明度，梯度，三种颜色的直方图这些指标，建立了回归模型。

通过指标建立的回归方程，我们将剩下的 2376 张照片的数据输入，使用回归方程计算阈值 y。然后我们确定一个阈值作为判断标准，经过检验得到阈值为 0.64 时效果好。当 y 大于阈值时待判图片为真，否则待判图片为假。通过 2376 张图片预测出的值与真实值进行比较，得到准确率为 97.76%，预测效果较优，模型具有一定的实用性。

对于问题三：基于问题二的识别模型，辨别给出的数据集中各图像的真伪，并写入表格中。由于问题二中得到了较为优化预测模型，我们再次提取待判的 1000 张照片中我们所需要的数据，将这些数据输入到模型中进行求解，最后分别得到真伪的图片判断，并将判断为假的图片编号输入到 fake 表格中，该模型输出结果快速。

对于问题四：评价识别模型的优缺点。根据问题二建立的模型，对比真实值与预测值之间的差距，对得出 logistic 图像识别模型的拟合优度进行分析，并对模型的准确率，运行效率和模型的应用性进行评价。

最后，我们探讨了模型的推广前景和改进方向。

关键词：图像识别　评判指标　logistic 回归　图像处理

一、问题重述

1.1 问题的背景

随着图像处理的技术发展，图像编辑使得图像变得更加易用、美观，同时也增强了人们的娱乐效果。但同时也带来一些问题，诸如图像拼接、裁剪、添加水印、增强对比度等技术不容易被人眼所察觉，这使得不法分子利用图像从中牟取利益。因此，研究者们针对图像编辑问题进行了大量研究并用以检测恶意行为。

图像处理，图像识别是有关图像的热点问题，同时，检测图像的颜色是否改变也是技术难点问题，解决图像修改问题可用于水印识别，人脸识别等方面的应用。因此，建立图像真假的识别模型，准确定位图片是否被改动，有着广泛的应用和实际价值。

1.2 图像处理的研究现状

图像检索功能是建立在对图像的特征提取和相似度特征分析基础上的，常用的传统图像处理的方法有边缘提取法和特征点提取法。现深度学习图像的检索方法有深度置信网络（DBN）、判别特征网络（DFN）、卷积神经网络（CNN）、循环神经网络（RNN）等。

其中边缘提取法包括微分算子检测和 Canny 边缘检测，其中二阶算子中拉普拉斯算子对噪声非常敏感。特征点提取主要指的是 Hattis 角点检测，角点即沿任意方向移动，均产生明显变化的点，但采用支持向量机检索器进行图像分析，该方法计算开销较大，图像检索的实时性不好。深度学习减少了人为设计特征造成的不完备性，在满足特定条件下的应用场景，已超越了现有算法的识别或分类功能。

1.3 数据说明

数据中文件 1-True 为真实图像，文件 2-Fake 与 3-Fake 为对应真实图像改变颜色后的虚假图像，validation.zip 数据集中为待判真假性的图片，fake.xlsx 为待判图片真假的求解结果。

1.4 问题的重述

检测图像的颜色是否被改变是难题。我们查阅相关文献，学习了深度学习的知识，建立了建立数学模型，解决了如下问题：

问题一：选出一些分辨图像的相关指标，在其中找出能分辨图片真假的判断指标，并说明理由。

问题二：运用问题一找出的指标，给出识别标准，建立分辨图像真假的识别模型，并检验模型的实用性。

问题三：根据问题二的模型，辨别给出数据集中各图像的真假，并用 *EXCEL* 表格列出虚假图像的名称。

问题四：评价识别模型的优缺点。

二、问题的假设

假设 1：图片文件在下载过程中没有出现损坏，丢失情况。
原因：若出现上述情况，模型的拟合程度将会降低。
假设 2：原始数据集不存在误判情况。
原因：若出现上述情况，模型判断结果的可信度将会下降。

三、主要符号说明

符号	符号说明
X,Y	表示两幅图像相交叠的各自部分
x，y	表示 X,Y 的灰度值
β_i	回归系数
$\ln L(\hat{\beta})$	模型最大似然值的对数值
q	模型中的自变量个数
s	图像饱和度的大小
light	图像亮度的高低
p	图像真假的概率
v	图像明度的大小
gradient	梯度

注：其他符号见文中说明。

四、问题的分析

4.1 总体分析

对于题目中给出的四个问题，我们查阅有关图像处理，图像识别的文献，并对深度学习进行了学习。图像的真假代表着图片进过修改，比如添加水印，增强对比度等等。我们研究的主要是图像颜色的改变，首先需要找出能够分辨图像真假的判断指标，然后根据指标建立分辨图像真假的识别模型。该问题属于二分类预测类型的问题，且自变量与因变量没有线性关系，建立 logistic 回归识别模型，去判断给出数据集中图像的真假。

4.2 问题一的分析

对于问题一，基于数据中图片的真假，找到分辨图片真假的判断指标，并说明理由。首先我们查询一些有关图像的指标，通过人为观察比对大量的真假图片，发现大量图片存在物体颜色受污染、颜色渐变程度变小，色彩变单一，图片偏向暖色调的情况，以及存在图片颜色加深的情况。将这些虚假图片变化情况进行统计一一对应到图片各个指标上去，即观察真假图片的直方图，饱和度，对比度，亮度，图片 RGB 占比，横向梯度，纵向梯度，色相，饱和度，明度等指标，对这些指标进行筛选，选出分辨图像真假的影响指标。

4.3 问题二的分析

对于问题二，运用问题一中的指标，建立分辨图像真假的识别模型并给出识别标准。真实图片和虚假图片有许多可以分辨的指标，经过查阅资料，我们选取了图片中的颜色 rgb，色调 h，饱和度 s，明度 v，亮度 light 对比度 dbd，梯度 gradient，三种颜色的直方图 hist 作为选取指标。考虑到选取指标过多，可能需要删去部分无效指标，故我们取出真假前 4700 张图片作为已知样本，使用这些数据进行逐步回归，构建多元回归方程，然后在删去 rgb 亮度、对比度指标后达到拟合效果良好，为 89.41%，故最终选取了色调 h，饱和度 s，明度 v，梯度 gradient，三种颜色的直方图 hist 这些指标。

通过指标建立的回归方程，我们将剩下的 2376 张照片的数据输入，使用方程进行计算 y。然后我们需要确定一个阈值，作为识别标准，当 y 大于阈值时待判图片为真，y 小于阈值时待判图片为假，经过检验得到阈值为 0.64 时效果好，故将此定为检验的阈值。然后将剩下的 2376 张图片预测出的值与真实值进行比较，检验模型的实用性。

4.4 问题三的分析

对于问题三，根据问题二建立的模型，判断所给数据集中图像的真假。第二问得到了较为优化预测模型，我们再次提取待判的 1000 张照片中我们所需要的数据，将这些数据输入到模型中进行求解，最后分别得到真假的图片判断，并将错误的图片编号输入到 excel 表格中。

4.5 问题四的分析

对于问题四，评价建立的识别模型的优缺点。根据问题二建立的模型，对比真实值与预测值之间的差距，对得出 logistic 图像识别模型的拟合优度进行分析，并对模型的准确率，运行效率和模型的应用性进行评价。

五、模型的建立与求解

5.1 问题一：分辨真假图片指标

5.1.1 相关指标的定义

对比度：图像颜色之间的差别，对比度越大，颜色的反差越大，图像越刺眼。

亮度与灰度：图片的明暗程度，主要指光线的明暗程度，亮度越高图片越耀眼，亮度越低越灰暗。

明度：亮度与灰度的总和成为明度，是颜色的固有属性，明度分为高、中、低调，不同的明度对比，决定了图片的光感，清晰感。

饱和度：图片颜色的浓度，浓度越高，图片显得鲜艳，饱和度越低，颜色显得陈旧，暗淡。

色相：即颜色。光波长的长短决定了颜色，红色波长最长，紫色波长最短，色相，指的是这些不同波长的色的情况。色调表示三种原色的明暗程度，在三种原色的基础上，改变其灰度，得到不同颜色，比如红色可改为暗红、深红、浅红等等色调。

色调：图像的整体明暗程度，当偏红色，橙色，对于人体感觉来说属于暖色调，有白色，银色属于冷色调，图像大多数有多种色彩，我们可以调整颜色的通道，来对图片的色调进行细微调整。

信息熵：熵指的是整体的混乱程度，衡量不确定性的指标，而信息熵指的是图像信息量的大小，当信息熵越大时，整体的色彩效果鲜明，图像轮廓越清楚，反之图像色彩比较单调。

联合熵：当随机变量 X，Y 的联合熵为：

$$P(X,Y)=-\sum_{x,y}p_{AB}(a,b)\log p_{AB}(a,b) \tag{1}$$

联合熵指 X,Y 的总信息，X,Y 分别指两幅图像相交叠的各自部分，x，y 分别表示 X,Y 的灰度值，p 表示概率。X 和 Y 之间的互信息为：

$$S(X,Y)=P(X)+P(Y)-P(X,Y) \tag{2}$$

此联合熵与概率的求解十分相似。

平均梯度：图像的边界或影线两侧灰度的变化程度，可用来表示图像清晰度。平均梯度可分为横向、纵向梯度，可表示图像在多为方向上密度的变化率，它反映了图像微小细节反差变化的速率，象征图像的相对清晰程度。

RGB：是一种颜色标准，通过 R 红色、G 绿色、B 蓝色三种颜色通道的变化来得到各式各样的颜色，是目前运用最广的色彩模式之一。

5.1.2 指标选取的理由

对比度：对比度越高表示图像越生动、色彩越丰富，通过对比查看多对真假的图片，发现有许多数量的图片，产生了背景图案污染原本物体图案的情况。即图片色彩变得单一，图像的对比度减少，图像明显的变得暗淡，色彩冲突性降低，故选此指标。例如编号为 00037096 的两组图片。

图 1：真实图片编号 00037096　　图 2：虚假图片编号 00037096

亮度：通过对比观察大量图片，发现有很多数量的图片的颜色变暗，显得图片变得陈旧后者图片变深，失去了原有的质感，故选此作为图像的评价指标。

图片 R 的总量：通过对比大量图片，发现很多图片偏暖色调，物体图案偏红。如图 00031912 的两组图片。

图片 G 的总量：通过对比大量含有植物的图片，发现含有植物的绝大多数图片中植物的颜色由黄变绿由翠绿变成墨绿深绿色。例如 00034224。

图片 B 的总量：通过对比物品图片，发现有较多图片的物品被该图片蓝色天空的颜色所污染，RGB 是影响颜色的关键因素，判断图像的影响因素，图像中 R、G、B 的总量作为评判指标必不可少。如：图片 00033306。

图 3：编号为 00033306 的真实图片　　图 4：编号为 00033306 的虚假图片

由上面两幅图对比可知，天空的蓝色影响到火箭外壳的颜色，受到了周围环境的影响，也可以说蓝色的含量增加。而且还可以看出虚假图片的色彩变得暗淡，色彩对比度下降，色彩信息量减少，说明一幅图片的改变不仅仅时一个指标的改变，而是多种指标

的共同影响的结果。

信息熵，联合熵：大量图片色彩变得单一，且偏暗，图片信息减少，信息熵减少，例如图 00033306，此指标较好的表示了图像信息的内容，此评价指标较好。

平均梯度（横向梯度绝对值+纵向梯度绝对值）：由于大量图片物体的颜色被背景颜色，如图 00033306，污染，例如人体，屋顶，车辆，长椅，导弹等等，对比较度之前减少，梯度下降，故选此指标。

色相：颜色普遍变深变暗，色相减少，如图 00033306。

饱和度：饱和度指色彩的纯洁性，也叫饱和度或彩度，是“色彩三属性”之一。如大红就比玫红更红，这就是说大红的色度要高。它是 HSV 色彩属性模式，孟塞尔颜色系统等的描述色彩变量。各种单色光是最饱和的色彩，物体的色饱和度与物体表面反色光谱的选择性程度有关，越窄波段的光发射率越高，也就越饱和。对于人的视觉，每种色彩的饱和度可分为 20 个可分辨等级。由于大量的虚假图片是在原来颜色的基础上添加不同颜色，造成色彩纯度下降，从而图片饱和度下降，例如图 00043095。

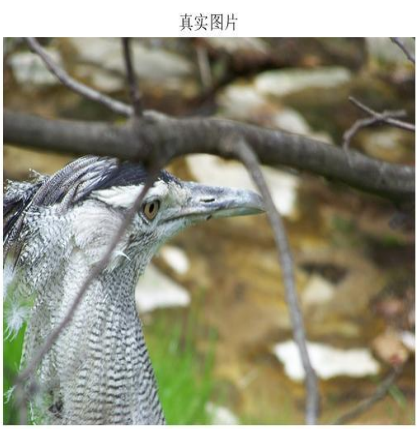

图 5：编号为 00043095 的真实图片　　图 6：编号为 00043095 的虚假图片

明度：明度（brightness）是眼睛对光源和物体表面的明暗程度的感觉，主要是由光线强弱决定的一种视觉经验。一般来说，光线越强，看上去越亮；光线越弱，看上去越暗。对比大量图片发现普遍图片色调变暖，颜色变深，真假图片对比有种”清晨与黄昏“之感，如 00035132 图。

总而言之，根据观察大量真假的图片，并且结合影响图像改变的重要因素，我们给出了共 8 种判断指标，这 8 种指标多数影响着图像颜色的变化，具有一定的选取依据。

5.1.3 结果分析

基于以上选取指标的分析，经真假图片的对比，我们找到了 8 个判断图像真假的指标，分别是对比度、亮度、信息熵以及联合熵、平均梯度（横向梯度，纵向梯度）、饱和度、明度。

表 1：判断指标的选取结果

判断指标
对比度、亮度、信息熵，联合熵、平均梯度、饱和度、明度、图片 RGB 占比

5.2 问题二：分辨图像真假的识别模型

5.2.1 logistic 回归模型的介绍

Logistic 回归是常被采用的统计方法，与线性回归相似，基本原理为利用一组数据，拟合回归模型，求出自变量与一个因变量取某个值的概率之间的关系。结合问题一的指标，我们选取 个指标作为影响因素，

5.2.2 logistic 回归模型的建立与求解

（1）回归模型的确立

logistic 函数：

$$f(x)=\frac{e^{x}}{1+e^{x}}=\frac{1}{1+e^{-x}} \tag{3}$$

自变量$x\in(-\infty,+\infty)$，函数值 f(y)在[0,1]取值，为单调递增的 S 型曲线。

logistic 回归模型表达式如下：

$$\log it(p)=\ln\frac{p}{1-p}=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_rx_r \tag{4}$$

其中，ln 是以 e 为底的自然对数；$\beta_0,\beta_1,\cdots,\beta_r$成为回归系数。

（2）回归参数估计

logistic 回归系数的估计通常采用最大似然法。最大似然法的基本思想为，建立似然函数和对数似然函数，使对数似然函数最大时求解相应的参数值，所得的估计值为参数的最大似然估计值。

假设有 n 个观测值$y_1,y_2,\cdots,y_n$，设$p_i=P\{y_i=1|X=x_i\}\ \ i=(1,2,\cdots,n)$为在定$X=x_i$的条件下$y_i=1$的条件概率，而在同样条件下得到$y_i=0$的概率为$1-p_i$，所以得到观测概率为$P\{y=y_i|X=x_i\}=p_i^{y_i}(1-p_i)^{1-y_i}$，其中，$y_i$取值为 0 或 1，相应的似然函数为：

$$L(\beta)=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i} \tag{5}$$

两边取对数并将 logistic 回归方程代入得：

$$\ln L(\beta)=\sum_{i=1}^{n}y_i\left(\beta_0+\sum_{k=1}^{r}\beta_kx_{ik}\right)+\sum_{i=1}^{n}\ln\left(1+\exp\left(\beta_0+\sum_{k=1}^{r}\beta_kx_{ik}\right)\right) \tag{6}$$

采用迭代算法即可求得参数估计值$\hat{\beta}_j(j=1,2,\cdots,r)$

5.2.3 logistic 回归模型的检验

判断模型的拟合程度的好坏，需要看评价模型有效匹配观测数据的程度，若模型的预测值与对应的数据有较高的一致性，则该回归模型拟合数据较好，拟合效果较优。

1.信息测量指标评估模型的拟合优度

（1）AIC 指标

AIC 的计算公式为：

$$AIC=-2\ln L(\hat{\beta})+2(q+s) \tag{7}$$

其中，$\ln L(\hat{\beta})$：模型最大似然值的对数值；q：模型中的自变量个数；s：因变量的类别数减 1；较小的 AIC 值表示模型拟合较好。

（2）SC 指标

SC 的计算公式为：

$$SC=-2\ln L(\hat{\beta})+(q+s)\ln n \tag{8}$$

其中，$\ln n$：样本量的自然对数，其余含义与 AIC 指标一样，SC 值越小模型拟合越好。

（3）R-square 拟合优度

2.回归系数的显著性检验

（1）似然比检验

设原假设为$H_0:\beta_k=0$，似然比检验是通过分析对数似然值来分析变量是否有统计学意义的，检验公式为：

$$LR=2\ln\frac{L(\hat{\beta})}{L(0)}\overset{H_0}{\sim}\chi^2(1) \tag{9}$$

其中，$L(\hat{\beta})$：包含所有自变量的最大似然函数的对数；L(0)：忽略了自变量的最大似然函数的对数；给定显著水平α，拒绝域为$W=\{LR\chi^2>\chi_\alpha^2(1)\}$,若$\alpha=0.05$，满足拒绝域条件即可拒接原假设，认为该变量对模型有显著性影响。

（2）T 值、P 值检验

当 P 值小于 0.05 时，说明模型拟合较好，自变量对因变量有影响效果，保留其指标，否则将该指标删除，重新进行拟合，并检验该方程的拟合优度。

5.2.4 logistic 回归的预测

给定预测点$(x_{01},x_{02},\cdots,x_{0r})$，代入回归模型中，计算出因变量发生概率为：

$$p=\frac{\exp\left(\hat{\beta}_0+\sum_{k=1}^{r}\hat{\beta}_k x_{0k}\right)}{1+\exp\left(\hat{\beta}_0+\sum_{k=1}^{r}\hat{\beta}_k x_{0k}\right)} \tag{10}$$

若 p>0.5，则认为为$\hat{y}=1$，否则$\hat{y}=0$。

5.2.5 模型识别标准和实用性检验

5.2.5.1 真假图片的识别标准

根据上述建立的模型，对图像的真假发生的概率进行预测，当 p>0.5 时，输出值为 1，该图片为原始图像，否则输出值为 0，为假图像。

5.2.5.2 模型实用性检验情况

对各项指标带入 logistic 回归模型中进行预测，在 matlab 中运行相应的函数，得出验证集的输出结果如下表所示：

表 2：问题二检验的部分结果

图片编号	IL00044580	IL00044581	……	IL00045815	rich00044580	……	rich00044581
真实值	1	1	……	1	0	……	0
预测值	1	1	……	1	0	……	0

预测结果为 97.76%为准确预测，在 2376 张验证集图片中共有 53 张判别失败。验证失真率为 2.24%。模型拟合较好。

5.3 问题三：判断图像的真假性

5.3.1 结果的输出

（1）参数估计的输出

考虑到选取指标过多，可能需要删去部分无效指标，我们取出真假前 4700 张图片

作为已知样本，构建 logistic 回归方程，然后在删去 rgb 亮度、对比度指标后达到拟合效果良好，作为图像真假的影响因素。各指标系数的结果如下：

表 3：各参数的结果

	系数	t	p
色调 H	0.0285	3.8149	0.0001
饱和度 S	0.0359	5.4239	0.0000
明度 V	-0.0397	-2.9147	0.0036
红色 hist1	0.7101	268.2149	0.0000
红色 hist2	0.3795	47.5228	0.0000
绿色 hist1	0.1814	15.6861	0.0000
绿色 hist2	0.1494	11.9902	0.0000
蓝色 hist1	-0.0529	-6.3119	0.0000
红色梯度 gradient	-3.92E-08	-10.5789	0.0000
绿色梯度 gradient	1.56E-08	3.2645	0.0011
蓝色梯度 gradient	2.55E-08	8.4619	0.0000

由上表可知，各指标的 p 值均小于 0.05，与图像的真假具有显著的影响效果。得到相应的 logistic 回归方程为：

$$\log it(p)=\ln\frac{p}{1-p}=0.0285\times H+0.0359\times S-0.0397\times V+0.7101\times Rhist1+0.3795\times Rhist2+0.1814\times Ghist1+0.1494\times Ghist2-0.0529\times Bhist1-3.92\times10^{-8}\times Rgradient+1.56\times10^{-8}\times Ggradient+2.55\times10^{-8}\times Bgradient$$

我们对模型优度进行拟合，真实值与预测值之间是有差异的，对模型的进行了 F 值检验和 P 值检验，当 P 值小于 0.05，说明模型整体拟合效果较好。

表 4：拟合优度的结果

R-square	拟合优度	89.41%
F	F 检验	7933.8
P	P 检验	0

由上表可知，logistic 模型的拟合优度为 89.41%，拟合优度较高，说明该模型的拟合效果较好。P 值为 0，远小于 0.05，所以模型的整体拟合效果较好。

5.3.2 图像真假的部分结果

我们将问题二建立的 logistic 模型得出了真假图片的判断结果，我们将判断出来的虚假图像的序号放入 fake.xlxs 表格中，图像真假的部分结果展示如下。

表 5：图像真假的部分结果

图片编号	00048982	00048983	……	00049600	……	00049999	00050000
预测值	1	1	……	0	……	1	1

5.4 问题四：评价图像识别模型

5.4.1 模型的优点

（1）我们建立的模型在针对同种类型图片的问题处理上拥有较强的判断能力，对于第二问中的自测准确率达到了 97%，足以满足一般的工作需求任务。同样如果想进一步的增强准确率的判断，只需对输入的数据种类进行合理的增加即可。（2）建立的 logistic 回归模型对因变量的分布没有要求，采用此回归可以巧妙的避开分类型变量的分布的问题，可以解决许多实际应用问题。Logistic 训练起来非常的高效，不需要运行太多的计算量，输出校准较好的预测值。

（3）若想检验其他类型图片的真假，只需替换本文程序中的数据即可，框架可以方便其他程序的建立，同时具有广泛的应用性。

（4）模型易于建立，简单，运行时间较少，模型识别正确率较高。

5.4.2 模型的缺点

模型在选取数据后进行了逐步回归删去了部分无影响数据，但是这部分无影响数据只是针对该问题中这 5888 张图片无影响，对于其他的图片是否有影响暂且未知，因此我们的模型不具备强推广性，若强行使用该模型检测不符合模型的图片，则会造成结果偏差较大。

六、模型的改进与推广

6.1 模型的改进

（1）可以通过用遗传算法对神经网络模型进行优化改进，在较小的空间中搜索出全局最优解。

（2）可以对图片进行压缩，用卷积神经网络进行图像真伪分类。

（3）可对图像的特征值进行提取，使用 BP 神经网络进行训练，以此辨别真伪。

6.2 模型的推广

该模型若有庞大的数据集进行参数拟合，可以大规模快速的进行真伪判别，且正确率较高。图像识别包括生物识别，人脸识别等等，图像识别和处理可用于人脸关键性息的标注，为人脸识别技术提升提供数据保障；电子商务数据，判断商品的销售金额，数量以及是否为用户最真实的反应；各种新闻，论坛对其进行娱乐监控，该图像判别的方法应用领域十分广泛。

七、参考文献

[1] 陈娜,李向伟.基于特征提取和回归分类建立卷积神经网络图像识别双重优化模型[J].电脑迷,2018(06):80.

[2] 高玉龙.基于图像识别的造假美术作品检测[J/OL].新乡学院学报,2019(06):39-42[2019-09-01].http://kns.cnki.net/kcms/detail/41.1430.Z.20190712.1932.016.html.

[3] 文政颖,卫欣.多分辨批量古典建筑图像深度学习检索算法[J].河南工程学院学报(自然科学版),2019,31(02):66-71.

[4] 王海燕.基于 Logistic 回归和 BPNN 的二值人脸图像识别[J].计算机应用与软件,2019,36(02):240-244+268.

[5] 扈华,付学良.基于逻辑回归模型的木片和树皮的图像识别[J].计算机应用与软件,2015,32(05):189-192+215.

[6] 孙娜,管一弘,罗亚桃,崔云月.基于颜色特征判别的纸病图像分割研究[J].信息技术,2019(03):87-90.
[7] 樊庆楠. 基于深度学习的图像处理算法研究[D].山东大学,2019.
[8] 周鹏飞. 自然场景图像中的文本检测与识别技术研究[D].西安理工大学,2019.
[9] 王丹妹. 基于卷积变换的图像匹配方法研究[D].重庆交通大学,2015.

八、附录

附录 1：问题一画出图像直方图所用 matlab 程序

```
%%%
%运行时间约为 5 分钟
%程序目的:求出 t_hist1.mat t_hist2.mat f_hist1.mat f_hist2.mat tg_hist1.mat
%tg_hist2.mat fg_hist1.mat fg_hist2.mat tb_hist1.mat tb_hist2.mat
%fb_hist1.mat fb_hist2.mat 数据文件
%以上文件的含义为红绿蓝三种颜色的直方图比例其中 hist1 指直方图前 150 个元素之和/所有元素之和
%，hist2 指直方图 151-256 个元素的和/所有元素之和

%%%
clc,clear
Path1 = 'D:\1-True\';                    % 设置数据存放的文件夹路径
File1 = dir(fullfile(Path1,'*.JPEG'));  % 显示文件夹下所有符合后缀名为.txt 文件的完整信息
FileNames1 = {File1.name}';              % 提取符合后缀名为.txt 的所有文件的文件名，转换为 n 行 1 列
Path2 = 'D:\2-Fake\';                    % 设置数据存放的文件夹路径
File2 = dir(fullfile(Path2,'*.png'));  % 显示文件夹下所有符合后缀名为.txt 文件的完整信息
FileNames2 = {File2.name}';              % 提取符合后缀名为.txt 的所有文件的文件名，转换为 n 行 1 列
Length_Names = size(FileNames2,1);     % 获取所提取数据文件的个数
n=5888;
t_hist1=zeros(1,5888);
t_hist2=zeros(1,5888);
f_hist1=zeros(1,5888);
f_hist2=zeros(1,5888);
tg_hist1=zeros(1,5888);
tg_hist2=zeros(1,5888);
fg_hist1=zeros(1,5888);
fg_hist2=zeros(1,5888);
tb_hist1=zeros(1,5888);
tb_hist2=zeros(1,5888);
fb_hist1=zeros(1,5888);
fb_hist2=zeros(1,5888);
for i=1:n
    i
    K_Trace1= strcat(Path1, FileNames1(i));
    K_Trace2= strcat(Path2, FileNames2(i));
    K_Trace1=K_Trace1{1};
    K_Trace2=K_Trace2{1};
    I1=imread(K_Trace1);
```

```
    I2=imread(K_Trace2);
    [t_r,~]=imhist(I1(:,:,1));
    [f_r,~]=imhist(I2(:,:,1));
    [t_g,~]=imhist(I1(:,:,2));
    [f_g,~]=imhist(I2(:,:,2));
    [t_b,~]=imhist(I1(:,:,3));
    [f_b,~]=imhist(I2(:,:,3));
    t_hist1(i)=sum(t_r(1:150))/sum(t_r);%1-150
    t_hist2(i)=sum(t_r(151:256))/sum(t_r);%151-256
    f_hist2(i)=sum(f_r(1:150))/sum(t_r);%1-150
    f_hist2(i)=sum(f_r(151:256))/sum(t_r);%151-256

    tg_hist1(i)=sum(t_g(1:150))/sum(t_g);%1-150
    tg_hist2(i)=sum(t_g(151:256))/sum(t_g);%151-256
    fg_hist1(i)=sum(f_g(1:150))/sum(f_g);%1-150
    fg_hist2(i)=sum(f_g(151:256))/sum(f_g);%151-256

    tb_hist1(i)=sum(t_b(1:150))/sum(t_b);%1-150
    tb_hist2(i)=sum(t_b(151:256))/sum(t_b);%151-256
    fb_hist1(i)=sum(f_b(1:150))/sum(f_b);%1-150
    fb_hist2(i)=sum(f_b(151:256))/sum(f_b);%151-256
end
save t_hist1 t_hist1
save t_hist2 t_hist2
save f_hist1 f_hist1
save f_hist2 f_hist2
save tg_hist1 tg_hist1
save tg_hist2 tg_hist2
save fg_hist1 fg_hist1
save fg_hist2 fg_hist2
save tb_hist1 tb_hist1
save tb_hist2 tb_hist2
save fb_hist1 fb_hist1
save fb_hist2 fb_hist2
```

附录 2：问题二 logistic 模型所用 matlab 程序

```
%%建立预测模型
%运行时间约为半分钟
%对于弹出的窗口需点击 export 导出数据
%%%
%程序目的：根据第一问所得指标，做 logistic 回归。输出验证集合的正确率与模型拟合参数
%！！！！！！！注意！！！！！！！运行本程序时中间会弹出逐步回归窗口，此时要求您点击窗口中 export 按钮，点击确定输出模型参数 beta
%若您在规定时间（弹出窗口到点击确定之间的时间>16s）没有完成上述操作则程序会
```

报错，此时请您将 68 行 pause（16）内的数字改大一些，以
%便您在规定时间内能完成操作。然后重新运行一遍程序。感谢！！！！！！！！！！！。
%%%
```
clear
clc
prefix=('D:\支撑材料\2t\');
d=dir(fullfile(prefix,'*.mat'));
filename={d.name}';
for i=1:length(filename)%导入数据
    trace=strcat(prefix,filename(i));
    Trace=trace{1};
    load (Trace)
end
%%整理数据
rhist1=[t_hist1,f_hist1];
rhist2=[t_hist2,f_hist2];
ghist1=[tg_hist1,fg_hist1];
ghist2=[tg_hist2,fg_hist2];
bhist1=[tb_hist1,fb_hist1];
bhist2=[tb_hist2,fb_hist2];

h=[ave_true_h,ave_fake_h];
s=[ave_true_s,ave_fake_s];
v=[ave_true_v,ave_fake_v];
dbd=[picture_contrast_true,picture_contrast_fake];
b=[sum_true_b,sum_fake_b];
g=[sum_true_g,sum_fake_g];
r=[sum_true_r,sum_fake_r];
gradient=[true_gradient_data,fake_gradient_data];
Light=[Y_true,Y_fake];
%%删除不用数据
clear trace filename d i prefix Trace ave_fake_h ave_fake_s ave_fake_v
ave_true_h Y_true t_hist2 f_hist2...
    ave_true_s  ave_true_v  picture_contrast_true  picture_contrast_fake
sum_fake_r Y_fake t_hist1 f_hist1...
    sum_fake_b sum_fake_g sum_true_r sum_true_g sum_true_b fake_gradient_data
true_gradient_data...
    tg_hist1 fg_hist1 tg_hist2 fg_hist2 tb_hist1 fb_hist1 tb_hist2 fb_hist2%
删去不必要变量
%%logistic 回归
leibie=[ones(1,5888),zeros(1,5888)];%真假图片标志
%p=[h;s;v;r;g;b;rhist1;rhist2;ghist1;ghist2;bhist1;bhist2;Light;dbd;gradien
t;leibie];%输入数据 色调 饱和度 ...
p=[h;s;v;rhist1;rhist2;ghist1;ghist2;bhist1;gradient;leibie];%回归后剩余指
```

```
标
p=p';
format long

PP=[p(1:4700,:);p(end-4699:end,:)];%真假各选取 4700 张作为已知数据输入
pp=p;pp(end-4699:end,:)=[];pp(1:4700,:)=[];%剩余 2376 张作为验证数据
X0=PP(:,1:end-1);%9400 张已知种类的数据
XE=[PP(:,1:end-1);pp(:,1:end-1)];%11776 张图片的数据 前面 9400 为已知种类 后面为未知种类
Y0=[p(1:4700,end);p(end-4699:end,end)];%前 9700 张图片真假
%X0=[X0;dx];Y0=[Y0;dy];
n=size(Y0,1);%求出 Y0 的大小
for i=1:n
    if Y0(i)==0
        Y1(i,1)=0.25;
    else
        Y1(i,1)=0.75;
    end
end

[~,nn]=size(p);
imodel=1:nn-1;stepwise(X0,Y1,imodel)%先对原输入进行逐步回归 找出删去的值再次运行
disp('请点击 Export 输出数据')
%需要点击 export 导出数据
pause(16)
b=beta;t=0;
for j=1:-0.02 :0%对阈值遍历
    for i=1:size(XE,1)
        pai0(i)=exp(XE(i,:)*b(1:end))/(1+exp(XE(i,:)*b(1:end)));
        if(pai0(i)<=j)
            P(i)=0;
        else
            P(i)=1;
        end
    end
    p0=[PP(:,end);pp(:,end)];
    pval=length(find(p0==P'))/11776;
    t=t+1;
    if t==1
        Pval=pval;
        yuce=P;
        J=j;
        Pai=pai0;
```

```
    else
        if pval>Pval
            J=j;
            Pval=pval;
            yuce=P;
            Pai=pai0;
        end
    end
    if Pval>0.9
        break
    end
end
Pval;%全部准确率
con=[p0,yuce'];
length(find(con(end-2375:end,1)==con(end-2375:end,2)))/2376%验证集的准确率
disp(['回归系数' num2str(b') '  ']);
save b b
```

附录 3：判断待判图像真假所用 matlab 程序

```
clear
clc
%%%
%运行时间约为 10 秒
%程序为：将待判图片的指标带入第二问建立的分类模型中。输出结果，将分类结果保存到 39 行所视目录中的 fake.xlsx 文件中，即本目录下
%%%
load b.mat
prefix=('D:\支撑材料\3t\');
d=dir(fullfile(prefix,'*.mat'));
filename={d.name}';
for i=1:length(filename)%导入数据
    trace=strcat(prefix,filename(i));
    Trace=trace{1};
    load (Trace)
end
%%整理数据
XE=[ave_dp_h;ave_dp_s;ave_dp_v;dp_hist1;dp_hist2;dp_hist1g;dp_hist2g;dp_hist1b;dp_gradient_data]';%输入数据 色调 饱和度 ...%修改后的输入数据
t=0;
for i=1:size(XE,1)
    pai0(i)=exp(XE(i,:)*b(1:end))/(1+exp(XE(i,:)*b(1:end)));
    if(pai0(i)<=0.6400)
        P(i)=0;
    else
        P(i)=1;
```

```
    end
end
true=find(P==1);fake=find(P==0);
length(find(P==1))
length(find(P==0))

%写入虚假图片标号
prefix=('D:\validation');
d=dir(fullfile(prefix,'*.JPEG'));
filename={d.name}';
for i=1:length(fake)
    Fake{i}=filename{fake(i)};
end
delete fake.xlsx;
xlswrite('fake.xlsx',Fake);
```

此处为问题的主程序，其余程序见支撑材料。