**图片清理预研报告**

吴斌杰

图片是手机上比较占空间的一类文件，如果可以将手机上的相似、模糊图片筛选出来并删除，就能帮用户节省很大一部分空间，提升我们产品的竞争力。

**相似图片**

相似图片筛选的难点在于如何判定两张图片“相似”。通俗的讲，衡量相似程度的维度有两个：一是图片内容的轮廓，二是图片的色彩分布。这两个维度越接近，那么两张图片越相似。

由于现在手机的摄像头像素普遍很高，所以采用对照片原图挨个像素对比的方法肯定是行不通的（内存、时间上都不允许）。市场上各种第三方清理图软件都对原图进行了不同程度的缩小后再进行算法处理。有了照片的缩小就有信息的丢失，那么就会有误差。而这些误差是避免不了的，我们要做的就是尽可能的减少误差。

**感知哈希算法** 这是网上流传比较广泛的算法，计算相似度的步骤是：

****第一步，缩小尺寸（****8\*8大小****）****

****第二步，简化色彩（****转为64级灰度****）****

****第三步，计算平均值****

****第四步，比较像素的灰度。****

****第五步，计算哈希值。****

****第六步，计算两张图片哈希值的汉明距， 距离小相似度越高。****

这种算法比较简单易懂，操作也不复杂。但有个弊端就是忽略了图像的色彩，缩小后的64级灰度图像并不能很好地体现原图的色彩分布。两张轮廓相似但是色彩差异很大的原图，通过这种算法计算后，可能就被判定成相似了。另外，缩成8x8大小极大的丢失了原图的信息，轮廓上相似的判定也不再是那么可靠。

此种算法计算相似度过于简单，误差较大，不适用于商业。但也不是一无是处，可以让我们初步了解相似图片是如何判定的。

**颜色分布法**

有了**感知哈希算法**的先例，我们取长补短，就有了**颜色分布法：**

****第一步，原图缩小尺寸（****100\*100大小****）****

****第二步，转化成灰度图像，计算平均灰度值，保存；****

****第三步，原图缩小尺寸（**6\*6**大小****），保存36个像素点的R、G、B值（36\*3=108长度），暂且叫颜色分布特征值，保存；****

****第四步，同时满足下列条件的两张照片判为相似：****

****1.灰度平均值之差小于4；****

****2.颜色分布特征值相同位置的数值差异之和<32\*108;****

****3.颜色分布特征值颜色相似（相同位置数值差异<40）位数>0.8\*108****

下面简单说明下上述算法原理：

****1.灰度平均值之差小于4；****虽然灰度平均值接近并不代表两张照片相似，但是灰度平均值差异大的照片一定不相似，这一步可以过滤大部分色彩构成差异大的照片。

****2.颜色分布特征值相同位置的数值差异之和<32\*108;****由于颜色分布特征值是按照缩放图从左到右从上到下计算的，大致可以描述原图的位置轮廓。相同位置的数值可以映射到原图相同位置的像素点的R、G、B值，总体差异之和小于32\*108，保证了两张照片整体轮廓的颜色构成上差异不会太大。

****3.颜色分布特征值颜色相似（相同位置数值差异<40）位数>0.8\*108****与第二步有异曲同工之妙。相似位数越多，说明原图相同位置的颜色接近的像素点数越多，轮廓上越接近。

有了颜色、轮廓和整体上（灰度平均值）比较，相似图片的判定就变得相当精准可靠，误差也大大减少。

各种阈值并不固定，可以自己调整以满足不同的需求。减小判定精度，则会扫描出更多的相似照片；增大判定精度，则会扫描较少的相似照片，但更精准。

第三方照片清理软件Facein（以前叫照片梳子）就是采用这种算法。

**模糊照片**

模糊照片的判定难点在于如何定义模糊度。怎么样的照片才算是模糊的？相似照片至少还有对比的参照物，而模糊度只能从图片本身的特征来计算。

既然我们不好定义模糊的标准，那么我们可以换个角度思考，模糊的照片是怎么拍摄出来的？

恩，手抖了没拿稳手机？拍摄的过程中镜头有位移，让照片失去了应有的分明线条，使得某个方向上有大片颜色接近的色块。

于是有了下面的照片模糊度算法：

****第一步，原图缩小尺寸（****按原图长宽比例缩小原图，使得最短边为64像素，至少64\*64=4096个像素点****）；****

****第二步，转化成灰度图像；****

****第三步，遍历缩小后的灰度图像每个像素点，定义一个像素点满足下面条件则为“清晰点”：****

****|该像素点上下左右像素点灰度值之和 - 4\*该像素点灰度值| <１００；****

****第四步，统计所有的“清晰点”数，小于120的为模糊照片。****

简单介绍下原理：

模糊照片的判定不需要色彩这个维度，模糊的彩色照片转化成灰度照片之后，依然是模糊的。所以可以直接缩放转化成灰度图像，减少计算量。

像素点与周围像素点（上下左右、甚至对角线）的差异越小，说明该照片在拍摄过程中镜头晃动的可能性越大；差异越大，说明对焦的越清晰。清晰的像素点越多，说明整体照片的清晰度越高；反之；模糊度越高。

阈值同样不固定。模糊度判定的越严格，筛选出来的模糊照片越少；反之，则筛选出的照片越多，但是可能“误杀”某些不是特别模糊的照片。

该算法有个缺点，就是无法判定某些接近纯色的照片。比如你对着洁白的墙壁拍一张“清晰”的照片，采用该算法，会被判定成模糊。

猎豹清理大师的模糊照片也是采用该算法。

**数据统计**

**相册1531张照片，共扫描出281组共821张相似照片，111张模糊照片。**

**共耗时175.8秒，其中扫描耗时71.2秒，照片归类耗时104.6秒；**

**计算所有相似照片特征值耗时2.9秒，模糊照片特值征耗时1.1秒；**

**相似照片识别准确率大概为85%，模糊照片识别准确率大概为75%。**