画像類似度判定ソフト By Sukiyakitti Ver1.0.0.0

想定した用途は、大量に画像を貯めている人が、重複画像を炙り出すのに使うこと。

特長は、負荷の少なさ、速さ。
　　計算方法は平均値算出だけで、重い処理していないです。

重複かどうかの確認はユーザーが行う必要がある。
　　判定結果には重複そうな組み合わせ順に並んでいて、上の方から確認し、適当な所で作業をやめる。

　　※小さな差分や色合い相違の有る組み合わせを重複とするかは人間の主観のため、
　　　人工知能（笑）では重複判定しきれない。

権利関連

使用料は、筆者のサイトの宣伝に協力ということで。
　　ツイッターやブログで「画像類似度判定ソフト By Sukiyakitti 」という語句を付けて何か書くか、
　　筆者のサイトのアドレス（http://theendou.adam.ne.jp/）を貼っておいて。

オープンソースですが、改造物の公開の際は「コピー元が筆者である旨」と自身の名前・サイトの明記を。

注意

利用は自己責任で。
サポートやる気ないです。

処理概要

・各画像毎にR（赤）・G（緑）・B（青）成分の濃さの平均を計算する。

・平均を計算する際、設定次第では画像を縦横数区画に分割し、それぞれの区画毎で平均を計算する。
　　例：「２」に設定すると、左上・右上・左下・右下の４区画に分割。

・各画像間の差異を計算する。
　　各色素（R・G・B）・区画毎に値の差（単純に引き算して絶対値出すだけ）を計算し、
　　それらの平均を差異度とする。※値が低いほど類似画像ということになる。

・差異度の低い順に結果グリッドに記入される。つまり、重複そうな組み合わせほど上に記入される。

・結果を表示しグリッドの何処かの行を選択すると、その行の比較組み合わせがサムネイル表示されるので、
　　各個人の感覚で重複かどうか判定する。

・「↓」キーを押す次の行のが表示される。

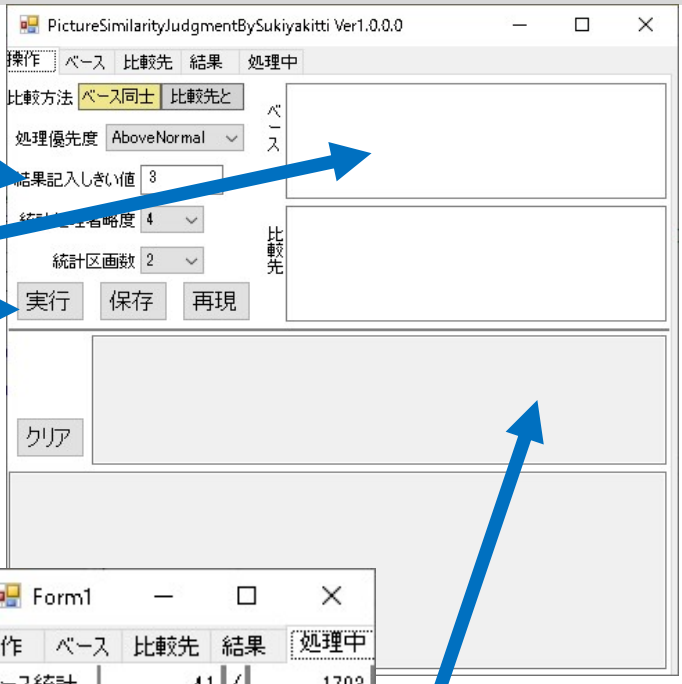
インストール方法

本体と設定ファイルを同じフォルダに置くだけ。
　　PictureSimilarityJudgmentBySukiyakitti.exe
　　PictureSimilarityJudgmentBySukiyakitti.txt

↓↓↓操作方法は次頁から↓↓↓

とりあえず使ってみる

画像が沢山入っているフォルダを用意。
ファイル数は、試すだけなら30～300がいい。
結果記入しきい値を調整する。重複・類似画像が無い場合、
既定の設定では何も出てこない。
無いかもと思ったら6まで上げておく。



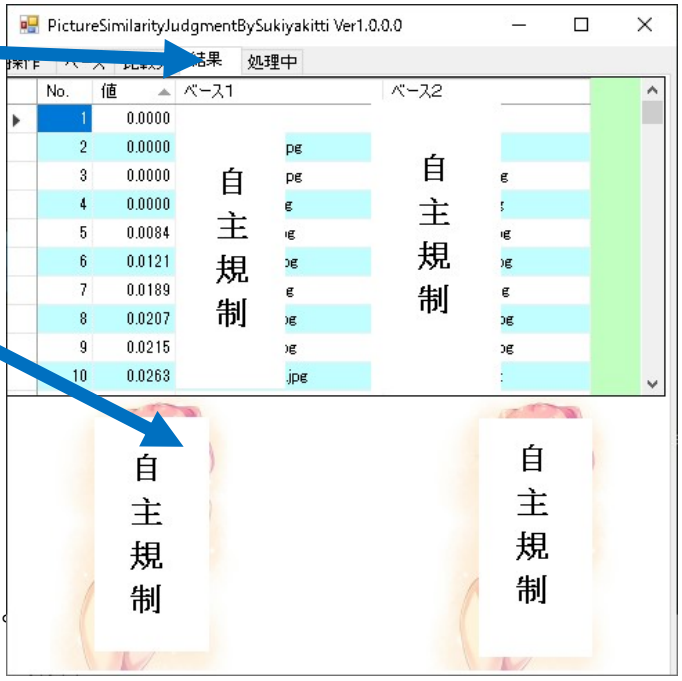
用意したフォルダを「ベース」テキストにドラッグ＆ドロップ
（または、そのフォルダのパスを、そのテキストに貼る）

実行ボタンを押す
※各設定の初期値は、要するに早さ重視
（その分、粗い）

処理中は、右記のように画面が小さくなり、
進捗が表示される。待つのみ。

処理が終わると画面サイズが元に戻り、元の表示に戻る。
画面下部のエラー表示枠（矢印の先）に異常そうな表示が
有るか確認、無ければ多分成功。

「結果」タブを押す。
↓
そのタブに移動時に、上部のデータグリッドの
一番左上のセルが選択状態のはず。
↓
下ボタンを押すと、次の組み合わせに移動する。
各組み合わせを1個ずつ確認する。
重複画像だと判断したら手作業で消す。
（削除機能は、バグると危険なため、無しにした）



どんどん下に行くと、そのうち重複画像が出てきにくくなる。
適当なところで作業をやめる。
結果の値について：
・完全に同じファイル→0
・ファイルサイズや拡張子による差→大抵0．2以下。
・ちょっとした差分→0．5位
・2超える→人間の感覚的に類似と思われるものは稀。
※上記はデフォルト設定時、計算方法次第で差が出る。

画面の各部分の説明－操作タブ

比較方法：
　ベース同士→同じフォルダ内の画像同士の比較
　比較先と→異なるフォルダのファイル間での比較
処理優先度：
　利用は自己責任で。自信のない人 はAboveNormal以上に上げてはならない。
　処理性能をどれくらい優先的に当ソフトに充てるかの設定。
　大抵の状況ではおまじない程度、活きるケースは処理性能の余裕が少ない場合か、
　　一時的に空気読まずに処理性能喰いまくるソフトを入れている場合。
結果記入しきい値
　10以上にあげるのは、どれ位の値でどれ位沢山比較結果が出るか大体分かってからにすべき。
　差異度がこの値以下のものが比較結果グリッドに記入される。
　ざっくりと重複を無くす程度でいいのなら1で十分、3で殆どの重複を炙り出せる。
統計処理省略度
　この値を上げるほど処理が早くなる反面、精度は落ちる。
　画像の各ドットを読む際、この値おきに読み飛ばす（1＝読み飛ばさずに全て読む）。縦横共に。
　例：3にすると、縦横3ドットおき→全部読む場合の9分の1しか読まない、その分処理が早くなる。
統計区画数
　例：「2」に設定すると、左上・右上・左下・右下の4区画に分割し、各々のR・G・B平均を算出。
　画像ファイルを縦横何区画に分割して平均算出するか。（1＝分割しない）
　原則、上げるほど精度は上がるが、統計処理省略度と共に上げ過ぎると精度が落ちる。
　（事情は、その場合は各区画毎の読込ドット数が減りすぎ、読む位置による誤差が大きくなるため）

「ベース」テキストボックス
　処理対象にする画像が入ったフォルダのパスを設定する場所。（ファイルでも可）
　パスの記入方法は、フォルダのドラッグ＆ドロップでも可。
　単一ファイル指定も可能。
「比較先」テキストボックス
　異なるフォルダの画像同士の比較を行う際に、もう片方のフォルダを設定する場所。
　（比較方法を「比較先と」にした場合に処理される）

実行ボタン
　処理実行用。
保存ボタン
　実行した結果をファイルに保存。後で再現できる。
再現ボタン
　保存しておいた処理結果を再現する時に押す。
　予め、再現したいファイルを「ベース」フォルダに設定しておく。

「クリア」ボタンの右の枠
　処理成功時は処理時間が表示される。失敗時にはその内容が表示される。
「クリア」ボタンの下の枠
　プログラムの事分かる人しか分からない。エラー発生時に、いわゆる「スタックトレース」が出る。
「クリア」ボタン。
　このボタンの下・右のテキストをクリアする。

画面の各部分の説明－ベースタブ・比較先タブ

画面の上部の灰色の表示枠
　処理対象のフォルダが表示される。
　※単一ファイル指定の場合、ルートフォルダ。
グリッド
　No：1から連番。目的は、ソート順を元に戻すのと、他人にどのデータか指す時用。
　Name：ファイル名
　3列目以降：各区画・色素の平均値
　　アンダーバー区切りで3種類の値を記述。
　　1個目：区画のY座標、上から1
　　2個目：区画のX座標、左から1
　　3個目：色素名
画面下の画像表示枠
　上記グリッドで選択した画像を表示

画面の各部分の説明－結果タブ

グリッド

No：1から連番。

値：差異値

【統計結果の各区画の差異の合計】÷区画の総数

3列目：

比較方法が「ベース同士」の場合、比較の片方のファイル名。

〃　「比較先と」の場合、「ベース」テキストで設定したフォルダ内のファイル名。

4列目：

比較方法が「ベース同士」の場合、比較のもう片方のファイル名。

〃　「比較先と」の場合、「比較先」テキストで設定したフォルダ内のファイル名。

画面の各部分の説明－処理中タブ

一番左の日本語表記：現在何の処理しているか。

数値：【　現在の進捗数　／　処理すべき数　】

処理すべき数：

統計の場合、ファイル数

比較（ベース同士）の場合、（ファイル数×（ファイル数－1））÷2

比較（比較先と）の場合、ベースファイル数×比較先ファイル数

設定ファイルについて。

便利にしたいなら（念のためコピーを確保したうえで）変更して。

1行目：どの拡張子のファイルを画像ファイルと見なすか。カンマ区切りで記述。

2行目：結果記入しきい値の初期値

3行目：統計処理省略度の初期値

4行目：処理優先度の初期値（コンボボックスの選択肢のどれかを書く）

5行目：比較結果グリッドの表示上限数

6行目：統計区画数の初期値

コードについて。

・開発環境持っていなくても読める。

「Form1.cs」を、何らかのテキスト編集ソフトで開けばいい。

・中間結果は極力持たないようにしてある。

例：各処理結果は、データグリッドに直接突っ込めるものは直接突っ込む。

※中間結果を増やす程、メンテの際に踏まえるべき箇所が増える、バグ率も増える。

・役に立たないコメントやサマリーは付けてない。

類似記事：http://labs.timedia.co.jp/2012/08/indescribable-code-featuring-comments.html

・高度な機能（笑）は出来るだけ使わない。

そういう機能は、凝った処理を開発システムに丸投げできる代わりに、自身のその辺の能力を捨てるもの。

例1：using宣言を使うと、後処理をusingに丸投げできるが、自信のエラー対処能力を捨てることになる。

例2：設定ファイル読み込みは、XML用の高度な機能（笑）を使わないどころか、

旧式の設定ファイル読み込み機能（INIファイル）すら使っていない。

・ステップ数削減

より多くの判断要因を同時に見渡すためには、PC画面に同時に表示できる量を多くすべき。

上記に伴い、処理種別毎にコードを整頓するのも済ましてある。

・必要以上の精度よりも低負荷や速さ優先のため、数値型はdecimalよりfloat（桁数少ない方）がメイン。

さらに、表示時は小数点以下は4桁にカット。それ以上有っても読み辛いだけ。

・オブジェクト名の先頭文字

c：クラス変数

a：引数（argument）

l：ローカル変数

f：関数（function）

btnやtxt等のコントロール種略称：イベント

連絡先

theendou@adam.ne.jp