GSLetterNeo vol.131

2019年6月

データの特徴を2軸にしてインタラクティブにデータセットの分布を見る

松原 伸人 matubara@sra.co.jp

はじめに

今回はデータセットをデータの特徴に基づいて2軸の散布図のようにプロットして見る方法を紹介します。冒頭の図1は、試作中の画像群の分布を見るツールのプロトタイプのスクリーンショットです。撮影して撮りためてきた写真群や描きためた絵をスキャンした画像群など、大量にある画像群にどのような傾向があるか見る時に用いています。JavaScriptとCSSとHTMLで実装している、Webブラウザ上で実行するWebアプリケーションです。

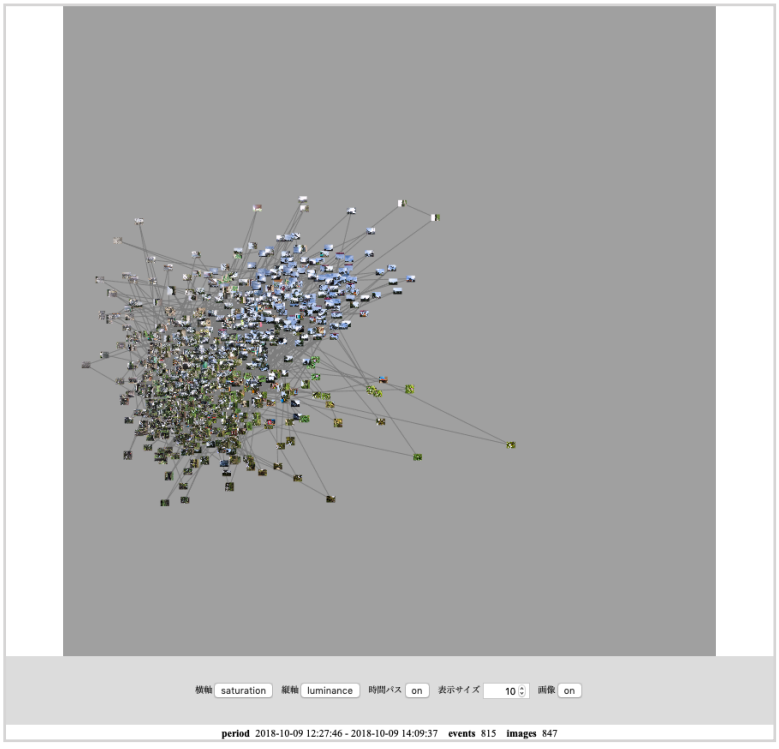

図1 画像の分布を見るツールの画面

本内容は、科学技術振興機構(JST)戦略的創造研究推進事業(CREST)「データ粒子化による高速高精度な次世代マイニング技術の創出」プロジェクト(研究代表者:宇野毅明 教授 (国立情報学研究所))で行なっている研究開発の一部です。今回使用している画像は、筆者自身と、共同研究者の山本恭裕 特任教授 (公立はこだて未来大学)、 および先端技術研究所所長でもある中小路久美代 教授 (公立はこだて未来大学)の協力により得られた研究データの一部です。両者の同意のもと掲載させていただいています。

ツールの概要

この試作ツールは、画像群の特徴のうち2つの特徴を選び、それぞれ横軸と縦軸に指定すると、特徴量に応じた横方向の位置および縦方向の位置に画像群を表示します。あらかじめ各画像は、画像の縦幅と横幅が 16:16 や 256:256 など正方形の矩形におさまるように画像の縦横比を維持して縮小画像を作成しておきます。試作ツールは、各縮小画像からピクセルのRGB値の平均値やHSB値やグレースケール換算の平均値など画像のピクセルから得られる数値を特徴として用います。ピクセルから得られる特徴の他に、各画像の撮影日時と撮影位置を書いたデータセットファイルを作成しています。データセットファイルは、「時間情報を持つテキストの年表化スクリプト(1)(2)」(GSLetterNeo Vol.103, 104) などで紹介した MC データ記述形式のテキストファイルです。この例のデータセットファイルには timeline の各行に画像の4つの属性、撮影年月日 と 撮影時刻、画像ファイルパス、縮小画像ファイルパス、撮影位置の緯度および経度を書いています。特徴量もあらかじめ計算しておいて、データセットファイルに書いておけば毎回特徴を計算しなくて済みます。次の図2は、データセットファイルの記述例です。2018年10月9日に北海道函館市にある函館公園内で3名があらかじめ決められた経路を、ほぼ同じ時間帯に散策してiPhoneで撮影したデータの一部です。このデータセットは「画像群の可視化方法のプロトタイピング」(GSLetterNeo Vol.123)と「距離が近いデータをグループにして表示する」(GSLetterNeo Vol.126)で掲載させていただいたのと同じデータセットです。紙面の都合で実際のファイルパスを簡略化し経緯度の数値の桁を減らして掲載しています。7行目以降に撮影日時順に1つの撮影を1行に、撮影年月日(yyyy-mm-dd)、撮影時刻(hh:mm:ss)、画像ファイルパス(image)、縮小画像ファイルパス(image256x256)、撮影位置の緯度(latitude)、経度(longitude)をタブ区切りで書いてあります。

試作ツールでは縦軸と横軸に指定する特徴の種類をあらかじめプログラム内に書いてあります。縦軸と横軸で指定できる特徴の種類は同じで、画像のピクセルのRGBの赤色(red)成分、青色(blue)成分、緑色(green)成分、HSLの色相(hue)成分、彩度(saturation)成分、輝度(luminance)成分、グレースケール換算値(grayscale)、撮影時刻(time)、撮影緯度(latitude)および撮影経度(longitude)の10種類です。図1は横軸に彩度、縦軸に輝度を指定した様子です。

軸を変えて見る

画面下の「横軸」や「縦軸」と書いてある右となりのボタンを押すと、10種類の軸が順に切り替わるようになっています。「時間パス」を「on」にすると、画面上の画像を撮影時刻順に直線で結んだ線を表示します。[図3]「表示サイズ」は画面上にプロットする画像の大きさの縦横幅です。「画像」を「off」にするとプロットしている画像を非表示にします。撮影時刻による経路を見やすくする際に「画像」を「off」にして「時間パス」を「on」にします。

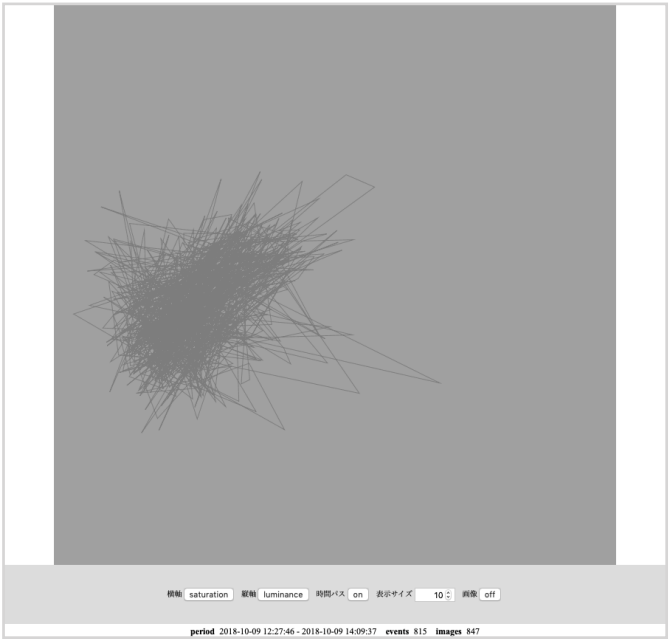

図3 画像off時間パスonにした様子

画面下部の period は、データセットファイルに書かれている撮影日時の最小値と最大値により得られた期間です。events は、データセットファイルに書かれている撮影日時の数です。同時刻の撮影が複数ある場合1つとして数えています。images は、データセットファイルに書かれている撮影した写真の枚数です。例えば地図のように表示する場合、横軸に

```
---
timeline
---
yyyy-mm-dd hh:mm:ss    image    image256x256    latitude    longitude
2018-10-09 12:27:52    IMG_1241.JPG    image256x256/IMG_1241.JPG    41.75726    140.71675
2018-10-09 12:32:45    IMG_1242.JPG    image256x256/IMG_1242.JPG    41.75705    140.71656
2018-10-09 12:35:39    IMG_1243.JPG    image256x256/IMG_1243.JPG    41.75667    140.71592
```

図2 データセットファイルの記述例

longitude を指定し、縦軸に latitude を指定します。[図4] 上を北方向、右を東方向として函館公園とその周辺を歩いて撮影した位置に画像群が描かれます。

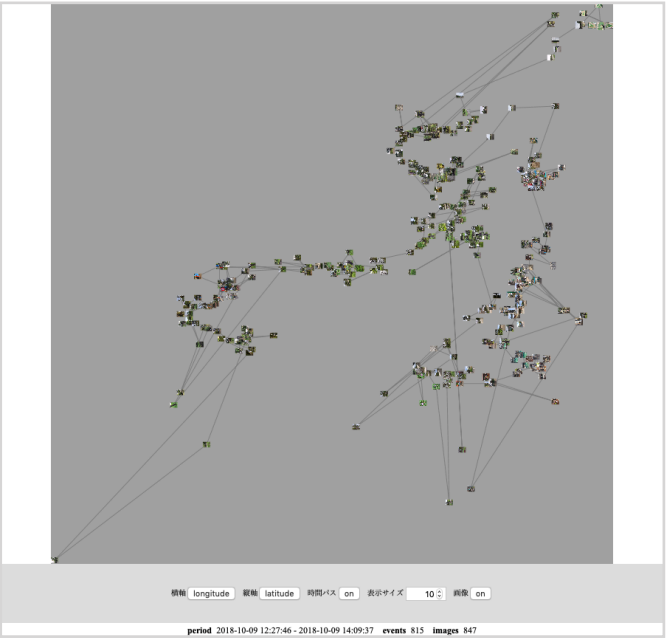

図4 横軸をlongitude、縦軸をlatitudeにした様子

図5のように横軸に time を指定し、縦軸に latitude を指定すると、時間の経過とともに南下して中盤過ぎからまた少し北上し終盤にまた南下しているのが見えます。

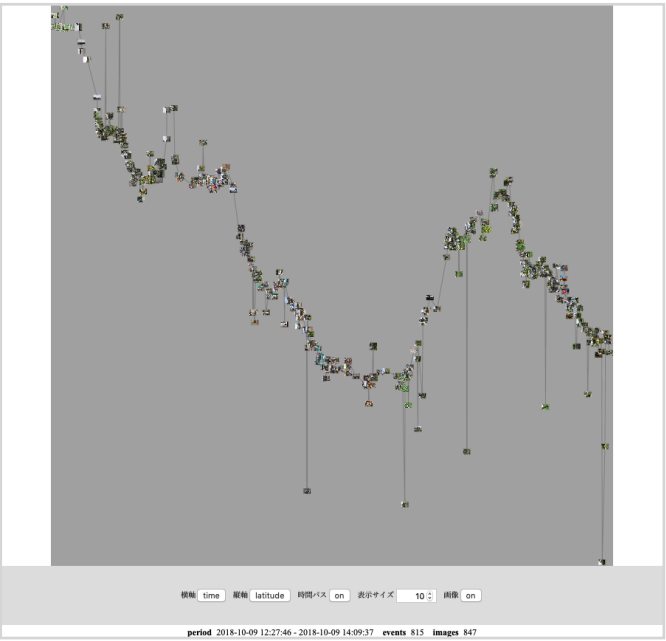

図5 横軸をtime、縦軸をlatitudeにした様子

図6のように横軸に longitude を指定し、縦軸に time を指定すると、序盤に西にちょっと進んで戻り、終盤にかけてなだらかに西に進み、終盤一気に西に移動してるように見えます。

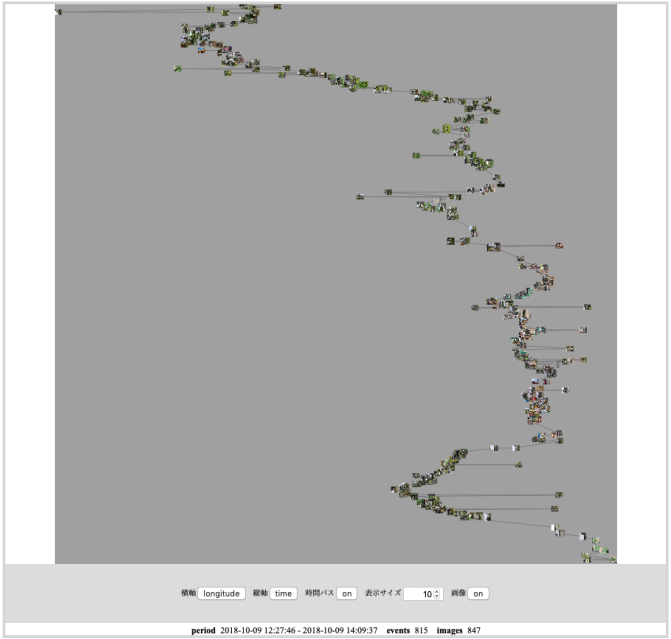

図6 横軸をlongitude、縦軸をtimeにした様子

図5、図6のように撮影時刻と撮影位置を軸にして画像の並びを見ると、多くは連続して連なっていますが、ところどころ撮影位置が外れている箇所が目立って見えます。該当する写真や前後の写真を見てみないと判断できないですが、多くの場合は写真撮影時に記録される経緯度による位置情報が、極端に実際とは異なる位置として計測しています。データ分析をする前にこのようなデータがあることがわかれば、フィルタリングするなどの対応方法を考えられます。撮影した写真に現れる色の傾向が、どれくらいの期間にどのように分布しているか、正確に分かるわけではないですが、なんとなく多い範囲や少ない範囲が見えます。詳細な分析に入る前に、手短におおまかに見ておきたいような時に効果があると考えています。

このつづきでは、今回紹介したツールを題材にデータの可視化と指定を行うプログラミングを紹介します。

GSLetterNeo vol.131
発行日 2019年6月20日
発行者 株式会社 ＳＲＡ 先端技術研究所
編集者 土屋 正人
バックナンバー https://www.sra.co.jp/gsletter/

お問い合わせ
gsneo@sra.co.jp
〒171-8513 東京都豊島区南池袋2-32-8
株式会社SRA