Vol.2020-CVIM-222 No.26
2020/5/15

# Spatial Transformer Networksを用いた角膜表面反射画像からのシーン識別

前田 響介[1,a)]　大嶋 佑紀[1,b)]　中澤 篤志[1]

**概要**：人の瞳に反射するシーン画像（角膜表面反射画像）は，撮影環境周辺の多くの情報を含んでおり，プライバシ上の大きな問題となりうることが指摘されている．一般に角膜表面反射画像は，通常撮影されたシーンの画像と比較して位置ずれや歪みが大きく，シーン識別性能は低いと考えられてきたが，近年の深層学習技術の発達によりより高度な位置合わせが画像認識が可能になれば，新たな問題になると考えられる．そこで本研究では，角膜表面反射画像に Spatial Transformer Networks を適用して画像の位置合わせを行い，シーン識別の精度を向上させる手法を提案する．従来法との識別精度の比較を行った結果，特に識別手法に Triplet Network を用いたものに対し，Spatial Transformer Networks による精度向上が確認できた．

# Scene Identification from Corneal Surface Reflection Images Using Spatial Transformer Networks

Kyosuke Maeda[1,a)]　Yuki Ohshima[1,b)]　Atsushi Nakazawa[1]

## 1. はじめに

今日，画像分野の研究は様々な応用がなされており，なかでもカメラによって得られた画像からその場所を特定する Visual Odometry の分野では，古くからの研究によって多くの手法が提案されている．しかし，記念写真やセルフィー写真のような，画像中の背景の割合が小さい画像に対しては，一般的な手法でのシーン識別は非常に困難である．

それに対し，画像内の人物の瞳に反射した画像を利用してシーン識別を行う研究の意義は大きい．角膜に反射した情報を悪用して対象人物の個人情報を特定し，凶悪な犯罪に及ぶという実例がある．このような手口の犯罪を未然に防ぐための，新たなプライバシー問題への対策検討のためにも，この研究タスクは重要なものであるといえる．

角膜表面反射画像とは，撮影画像からそこに映った人物の目領域を切り出し，目の幾何構造を考慮して平面展開することで，通常のカメラ画像と同様の透視撮影による画像に変換したものである．しかし，このようにして得た角膜表面反射画像は，

- 眼球内の虹彩パターンが反射したシーン画像と混ざり合う．
- まつ毛が角膜に直接覆いかぶさる．
- 帽子やまぶたが影になりシーン画像を遮って角膜まで到達できなくなる．

といった理由から，うまく識別できない場合も多く識別精度が上がらない [1]．したがって，識別精度向上のためにはこれらのノイズを除去する処理が必要となる．

さらに，一般的に角膜表面に反射したシーンは，元になるシーン画像と比較して傾いていたりするほか，目領域の平面展開が不正確であったために歪みが生じていたりすることが多い．そのため，そのような画像の位置ずれ・歪みを修正することで，識別精度の向上につながると考えられる．

本研究では，深層ニューラルネットの一つである Spatial Transformer Networks を用いて角膜表面反射画像とシーン画像との位置合わせを行い，従来手法よりも高精度なシー

[1] 京都大学 Kyoto University
[a)] maeda@ii.ist.i.kyoto-u.ac.jp
[b)] ohshima@ii.ist.i.kyoto-u.ac.jp

ン識別を行う手法を提案する．画像の変換の手法には，アフィン変換と TPS(薄板スプライン) 補間の 2 種類を用いる．これらの手法で変換された角膜表面反射画像に対し，シーン画像とのマッチングを，1) patch ベースの Triplet Network を用いた手法，2) Vectors of Locally Aggregated Descriptors(VLAD) を用いた手法の 2 種類で行い，その識別精度について検証した．

## 2. 関連研究

### 2.1 距離学習 (Metric Learning) に関する研究

距離学習 (Metric Learning) とは，複数のデータから得た特徴量を，データ間の意味的な距離を考慮して変換する手法であり，画像の分類や情報検索，異常検知といったさまざまなタスクへと応用されている．

Metric Learning の一種である Siamese Network[2] は，重みを共有した 2 つのネットワークに画像のペアを入力し，そのペアが同じラベルを持つならば出力ベクトルの距離が近くなるように，違うラベルを持つならば遠くなるようにネットワークを学習することでクラス識別を行うネットワーク構造である．

Siamese Network の構造を改良したネットワークとして，Triplet Network[3] が提案されている．基準となる画像 (Anchor)，Anchor と近いラベルを持つ画像 (Pos:Positive)，Anchor と遠いラベルを持つ画像 (Neg:Negative) の 3 枚を入力として，Anchor と Pos との距離は近く，Anchor と Neg との距離は遠くなるように学習するネットワーク構造である．入力画像ペアを「近づける」か「遠ざける」かを 2 値で定める必要がある Siamese Network に対して，Triplet Network は「Neg よりも Pos を相対的に近づける」というしかたで学習するため，識別タスクのコンテキストを考慮に入れる必要がない，という長所がある．

### 2.2 画像の特徴量に関する研究

画像からのシーン識別の手法にはさまざまあるが，基本的には，識別対象となる画像の特徴量とあらかじめ得ておいた画像の特徴量とを比較し，最も一致するものからシーンを同定する手法が多い．それに属する研究として，Torii ら [4] は，SIFT 特徴量と Vectors of Locally Aggregated Descriptors(VLAD)[5] という特徴量を用いてシーン識別を行う手法を提案した．任意の局所特徴量を k-means 法でクラスタリングし，各クラスタに属する特徴ベクトルを総和することで算出される．位置ずれに強く次元数の大幅な圧縮が可能な特徴量で，また主成分分析を適用してさらに次元を圧縮することもでき，多数のデータを貯蓄できるため，検索性能が高いことで知られている．

### 2.3 Spatial Transformer Networks に関する研究

Convolutional Neural Network(CNN) を用いた従来の画像識別タスクでは，ネットワークへの入力として画像全体を用いている．そのため，正しい識別には必要ない画像内の箇所が，ネットワークの識別能力に悪影響を及ぼす可能性があると考えられる．Spatial Transformer Networks[6] は，画像の変換を行うパラメータを学習し，識別ネットワークに入力する前に識別に必要な箇所のみを切り取ることで，識別性能を向上させることができるネットワークである．このネットワークは大きく Localization Net, Grid Generator, Image Sampler の 3 つの部分に分かれている．Localization Net では，ニューラルネットワークによって入力画像から変換を行うパラメータ (例えば，アフィン変換を行う $2 \times 3$ の変換行列や，制御点 $n$ 個の TPS を行う $2n^2$ 個の座標パラメータ) を生成する．Grid Generator では，あらかじめ定義したサンプリングカーネルをもとにサンプリンググリッドを計算する．そして，計算されたグリッドをもとに Image Sampler で出力画像を生成して出力する．Image Sampler で出力画像を生成する一般式は，出力空間上の点を $(x_i^t, y_i^t)$，点 $(x_i^t, y_i^t)$ を求めた変換パラメータによって変換した先の点を $(x_i^s, y_i^s)$ とすると以下のようになる．

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y) \qquad (1)$$
$$(\forall i \in [1 \cdots H'W'], \forall c \in [1 \cdots C])$$

ただし，$H', W'$ は出力画像の縦のサイズと横のサイズ，$C$ はチャンネル数，$\Phi_x, \Phi y$ はサンプリングカーネル $k$ のパラメータ，$U_{nm}^c$ は入力空間上のチャンネル $c$ での格子点 (画素)$(n, m)$ の値，$V_i^c$ はチャンネル $c$ での点 $(x_i^t, y_i^t)$ の値である．サンプリングカーネル $k$ には，例えば

$$k(p) = \max\{0, 1 - p\}$$

などが用いられる．このカーネルを利用した場合は，$(x_i^t, y_i^t)$ の値は $(x_i^s, y_i^s)$ の近傍 4 点の線形補間となる．

Localization Net には任意のネットワーク構造を利用することができ，勾配計算可能な変換手法やサンプリングカーネルを用いることで，後に接続した識別ネットワークと同時に学習させることができる，という点が特徴的である．

## 3. 手法

### 3.1 提案手法の概要

本研究では，Spatial Transformer Networks を用いて角膜表面反射画像を変換したのち，変換後の画像のシーン識別を 2 種類の手法で行った．

1 つ目は，学習済みの patch ベースの Triplet Network を用いた手法である．角膜表面反射画像とシーン画像それぞれから小領域 (patch) を切り出し，事前学習した Triplet Network を用いてそれらの特徴ベクトルを計算する．さら

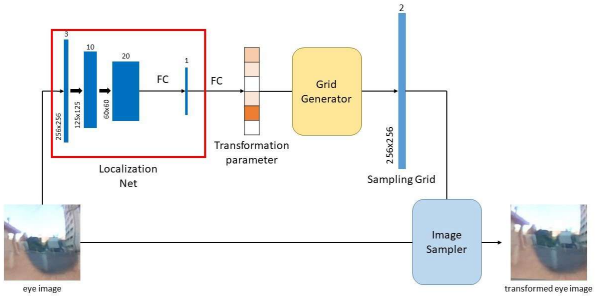


図 1　Spatial Transformer Networks のネットワーク構造

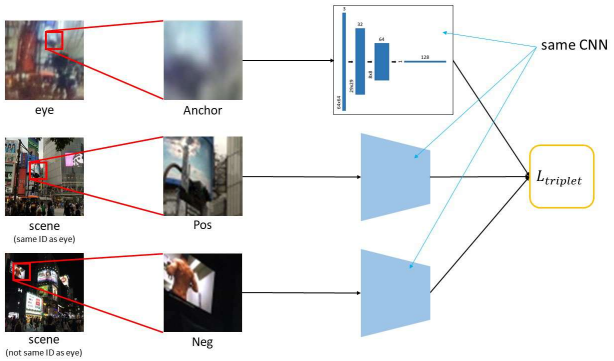


図 2　Triplet Network の学習時のネットワーク構造

に，patch 同士の距離をシグモイド関数に入力した値を求め，その 2 つの特徴ベクトルの L2-loss との積を類似度として識別する．

2 つ目は，Dense SIFT 特徴量を集約した VLAD によってシーン識別をする手法である．

### 3.2　Spatial Transformer Networks による角膜画像の変換

入力画像の変換には，アフィン変換と TPS(薄板スプライン補間) の 2 種類の手法を用いた．コードは PyTorch のチュートリアルコード [7] をもとに実装し，TPS のモジュールには WarBean[8] が公開しているコードの一部も利用した．

Spatial Transformer Networks のネットワーク構造を図 1 に示す．Localization Net は，2 つの畳み込み層とプーリング層，2 つの全結合層からなる CNN で構成した．2 つのプーリング層にはいずれも Max Pooling を用いている．全結合層の出力の次元は，アフィン変換を利用する場合と TPS を利用する場合とで異なり，アフィン変換を利用する場合は 6 次元，TPS を利用する場合は 32 次元である．

Grid Generator は，サンプリングカーネルとして近傍 4 点の線形補間を出力するカーネルを用いている．出力はサイズ $256 \times 256 \times 2$ のテンソルで，出力画像の各画素が変換によってどの座標の点から移ってくるかを示す．

Image Sampler は， Grid Generator で得られたグリッドをもとに実際に出力画像を構成する．入力画像はチャンネル数 3 の RGB 画像であるが，変換時にはその 3 枚の画像すべてに対し，同じサンプリンググリッドを適用することで変換を行う．変換前の点の座標が入力画像の外側にはみ出ている場合は，その点の画素値はゼロとなるよう実装している．

Spatial Transformer Networks は，以下の損失関数 $L_{STN}$：

$$L_{STN} = ||ST(\mathbf{I}_{eye}) - \mathbf{I}_{sc}||_1$$

を最小化するように学習する．ただし，$\mathbf{I}_{eye}, \mathbf{I}_{sc}$ はそれぞれサイズ $256 \times 256 \times 3$ の角膜表面反射画像とシーン画像，$ST(x)$ は画像 $x$ を Spatial Transformer Networks に入力したときの出力画像 (サイズ $256 \times 256 \times 3$) である．

以上述べたネットワークを学習したのち，得られた Localization Net の重みを用いてテストデータの変換を行い，変換後の画像を続くシーン識別の入力に利用する．

### 3.3　Triplet Network によるシーン識別

Bielski[9] の公開しているコードをもとに実装した．Triplet Network の学習時のネットワーク構造を図 2 に示す．

学習時は，ある角膜表面反射画像から得た patch と，その patch に対応するシーン画像の patch，そしてその patch とは対応していないシーン画像の patch の 3 枚組を入力し，重みを共有した CNN によって特徴ベクトルを計算する．ここで patch のサイズはいずれも $64 \times 64$ としている．

CNN のネットワークは，3 つの畳み込み層とプーリング層からなる．全てのプーリング層には Average Pooling を使用している．3 つの畳み込み層には，活性化関数として Tanh 関数を適用した．この CNN の出力は 128 次元の特徴ベクトルとなる．

損失関数には Triplet Loss を用いる．Triplet Network は，以下の $L_{triplet}$ を最小化するように学習する．

$$L_{triplet} = \max\{0, D_{pos} - D_{neg} + m\} \quad (2)$$

$$D_{pos} = ||V_{anchor} - V_{pos}||_2^2 \quad (3)$$

$$D_{neg} = ||V_{anchor} - V_{neg}||_2^2 \quad (4)$$

ただし，$V_{anchor}$，$V_{pos}$，$V_{neg}$ はそれぞれ，角膜表面反射画像から得た patch と，その角膜表面反射画像に対応するシーン画像の patch，その角膜表面反射画像に対応していないシーン画像の patch を Triplet Network の CNN に入力して得られる特徴ベクトルを表す．また $m$ は Triplet Network におけるマージンを表す定数である．この損失関数は，対応する patch 同士の誤差は小さくなるように，同時に対応していない patch 同士の誤差は大きくなるように学習を進めるように構成されている．マージン $m$ が大きいほど，対応する patch 同士の誤差と，対応していない patch 同士の誤差との差を大きくするように要請する．

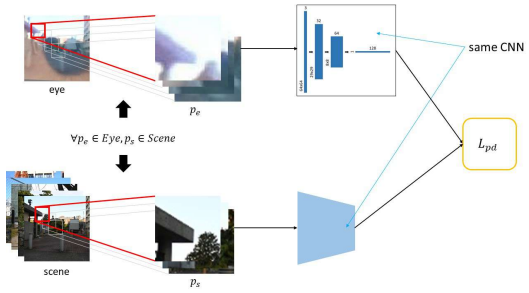


図 3　Triplet Network の評価時のネットワーク構造

以上のネットワークを用いて学習した CNN を用いて，図 3 で示したネットワークで評価を行う．入力された角膜表面反射画像とシーン画像からサイズ $64 \times 64$ の patch を切り出し，学習済みの CNN に入力して特徴ベクトルを得，それらの L2-loss を求める．さらに，2 つの patch の代表点同士の距離を計算し，シグモイド関数に入力して patch distance loss を得る．これらの積を求め，これをこの 2 つの patch の誤差と定義する．この誤差を，入力シーン画像のすべての patch に対して計算し，その中で誤差が最小のものを選択する．以上の操作を，角膜表面反射画像のすべての patch に対して行い，得られた誤差を総和した値を 2 枚の画像の誤差とする．すなわち，角膜表面反射画像中のある patch との類似度が高い patch がシーン画像にあるほど，またその patch 同士の距離が近いほど，2 枚の画像の誤差は小さくなる．

数式で表すと，以下の誤差 $L_{pd}$ が小さいほど，2 枚の画像は類似度が高いとすることになる．

$$L_{pd} = \sum_{p_e \in \mathbf{I}_{eye}} \min_{p_s \in \mathbf{I}_{sc}} pd_{p_e,p_s} ||V_{p_e} - V_{p_s}||_2^2 \quad (5)$$

$$pd_{p_e,p_s} = \sigma(||R_e - R_s||_2^2) \quad (6)$$

$$\sigma(x) = \frac{1}{(1 + \exp(-\alpha x))} \quad (7)$$

ただし，$\alpha = 0.005$, $\mathbf{I}_{eye}$，$\mathbf{I}_{sc}$ はそれぞれ入力された角膜表面反射画像，シーン画像から得られた patch 全体の集合，$R_e, R_s$ はそれぞれ patch $p_e, p_s$ の代表点 (本手法では patch の左上) の座標を表す．

本研究では，patch の切り出しのステップサイズは，角膜表面反射画像では 32，シーン画像では 8 である．

### 3.4　VLAD によるシーン識別

識別に用いる VLAD 特徴量には，Dense SIFT 特徴量を集約して得られた VLAD 特徴量に対し，さらに主成分分析を行って 4096 次元に圧縮し正規化したものを用いた．2 枚の画像の類似度にはコサイン類似度を用いた．

## 4.　データセット

### 4.1　京都大学周辺で撮影した角膜表面反射画像とシーン画像のデータセット対

本研究の実験のデータセットには，京都大学周辺で撮影した各 104 枚の角膜表面反射画像・シーン画像の対を用いた．

撮影には図 4 のような撮影器具を自作した．チンレスト (あご乗せ台) 取り付けた器具に，カラーボードと段ボールを切り出して作った $45 \times 45$cm の正方形のフレーム (枠) をつなげる．フレームにはビニール紐が対角線を描くように張られており，紐の交点によってフレームの中心の位置がわかるようになっている．フレームとチンレストとの距離は固定されている．フレームは左右方向に回転させてチンレストの正面から取り除くことができ，同一のカメラ位置でフレームありの画像とフレームなしの画像の両方を撮影することができる．

図 5 のように，被撮影者にはチンレストにあごを置いてあごの位置を固定し，その正面からカメラで被撮影者を撮影することで目を含む顔の画像を得る．さらに，カメラのレンズをチンレストの上に合わせてシーンを撮影することで，被撮影者が見ていたシーンに対応する，位置ずれの少ないシーンの画像を得られる．

シーン 1 か所あたり a.) フレームありの顔の画像 b.) フレームなしの顔の画像 c.) フレームありのシーン画像 d.) フレームなしのシーン画像 の 4 枚を撮影する．画像 a. の角膜内に映ったフレームの内部には，画像 c. のフレームの内部にあるシーンがほぼそのまま反射して映っていると考えることができる．続いて，顔画像 a. と b. については，目の周辺のみを切り取る作業を行う．切り取った後の画像の黒目の位置が一致するよう，手作業で切り取りを行っている．その後，切り取り作業後の目画像 a. からフレームの四隅と中心，各辺の中点の計 9 点の座標を取得し，その座標情報を使って切り取り作業後の目画像 b. に対し TPS を適用し，図 6 のように切り取る．画像 c. については，フレームの四隅 4 点の座標を取得し，アフィン変換を行って同じく図 6 のように切り取る．このような手順を踏むことで，画像 b，d. から対応した領域のみを抽出し，シーン識別のしやすい角膜表面反射画像を得ることができる．

### 4.2　24/7 Tokyo dataset から作成された角膜表面反射画像データセット

シーン識別に用いた Triplet Network の事前学習には，Torii ら [10] によって作成された 24/7 Tokyo dataset と，それを用いて Edamoto ら [11] によって作成された角膜表面反射画像データセットを用いた．

24/7 Tokyo dataset は，スマートフォンを用いて撮影さ

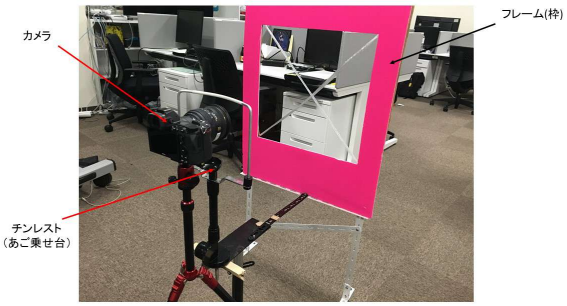


図 4　データセット作成に使用した撮影器具


図 5　撮影風景と得られる画像の一例

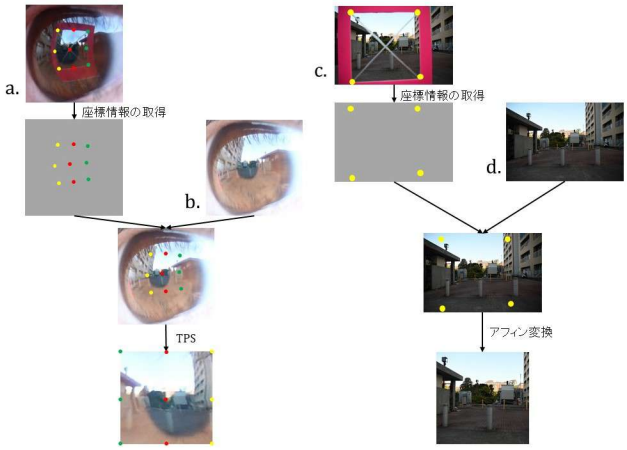


図 6　得られた画像の変換処理

れた東京の離れた 125 地点について，3 つの異なる方向を日中・日没時・夜間の 3 種の時刻に撮影した，計 1125 枚のシーン画像データセットである．図 7 にその一例を示す．

角膜表面反射画像データセットは，以下のような方法で撮影された角膜表面反射画像データセットである．まず，目の位置の固定のためのパターン画像と 24/7 Tokyo dataset の画像とを八面ディスプレイに交互に表示し，それをチンレストによって頭部を固定された被撮影者に見てもらう．その被撮影者の目を正面からカメラで自動撮影することで得られた 2 枚の目画像のセットに対し，明るさの調整とパターン画像に従った切り抜き，平面展開をすることで，1053 枚の角膜表面反射画像を得ている．図 8 にその一例を示す．

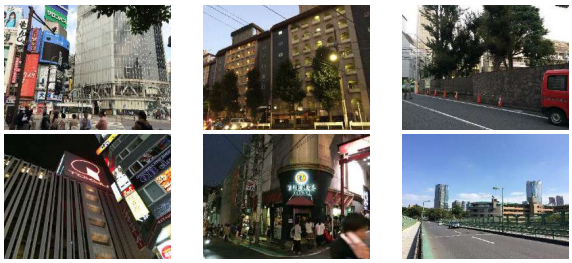

図 7　24/7 Tokyo dataset のシーン画像の一例

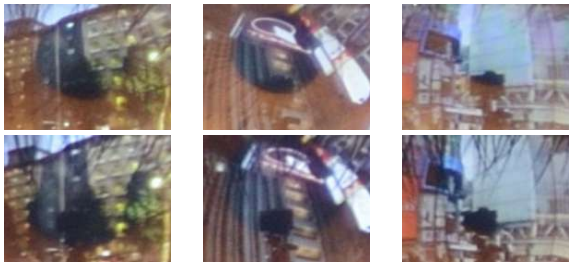

図 8　Edamoto らによる角膜表面反射画像データセットの一例

本実験では，まず角膜表面反射画像とそれに対応したシーン画像とから SURF 特徴量を算出し，それらを利用して対応する patch を求め，1) 角膜表面反射画像 2) シーン画像 3) 角膜表面反射画像の patch の代表点の座標 4) シーン画像の patch の代表点の座標 5) 角膜表面反射画像の patch の，シーン画像の patch に対する拡大率 の 5 つ組のデータを得る．Triplet Network の学習の際には，このデータをもとに対応する patch の切り出しを行っている．

## 5.　実験

本研究の提案手法である，Spatial Transformer Networks を用いて角膜表面反射画像を変換し，1) patch ベースの Triplet Network によってシーン識別を行う手法，2) VLAD によってシーン識別を行う手法の 2 つに対して，Spatial Transformer Networks による変換をせずにそれぞれの手法でシーン識別を行う従来手法と比較し，精度の検証を行った．

### 5.1　実験設定

学習や評価に用いる画像は，「データセット」の章で得られた RGB 画像で，サイズは $256 \times 256$ にリサイズしてある．

シーン識別で用いた VLAD 特徴量は，Torii ら [10] による，24/7 Tokyo dataset から抽出した Dense SIFT 特徴量をもとに学習を行った 4096 次元の VLAD 特徴量である．

Spatial Transformer Networks の学習に用いたデータセットと実際に変換を行ったテストデータは同一のものであり，いずれも京都大学周辺で撮影した 104 枚の角膜表面反射画像である．

Triplet Network の学習に用いたデータは，24/7 Tokyo

表 1　Triplet Network による結果 ($Acc(k)$)

| | 変換なし | affine | TPS |
|---|---|---|---|
| $k=1$ | 0.2788 | 0.3077 | 0.3654 |
| $k=2$ | 0.3462 | 0.3365 | 0.4135 |
| $k=3$ | 0.3654 | 0.3654 | 0.4615 |
| $k=4$ | 0.4231 | 0.4038 | 0.4615 |
| $k=5$ | 0.4327 | 0.4423 | 0.5000 |

表 2　VLAD による結果 ($Acc(k)$)

| | 変換なし | affine | TPS |
|---|---|---|---|
| $k=1$ | 0.8846 | 0.8750 | 0.7788 |
| $k=2$ | 0.8942 | 0.9231 | 0.8462 |
| $k=3$ | 0.9135 | 0.9231 | 0.8750 |
| $k=4$ | 0.9135 | 0.9231 | 0.8942 |
| $k=5$ | 0.9231 | 0.9231 | 0.8942 |

dataset と Edamoto らによる角膜表面反射画像データセットから SURF 特徴量を算出し，それらの対応を求めることで得られた 2249 個の patch のペアデータである．

### 5.2　評価方法

いずれの手法も，各角膜表面反射画像に対し類似度の高いシーン画像を順に出力し，その上位 $k$ 件に正解画像が含まれている割合 ($Acc(k)$ と表す) の高さによって手法の評価を行う．

### 5.3　実験結果

まず，patch ベースの Tripet Network によって評価をした 1 つ目の手法の結果として，Spatial Transformer Networks を用いて角膜表面反射画像を変換した場合の $Acc(k)$ の値を，変換しない場合の $Acc(k)$ の値と比較して $k$ =1〜5 について表 1 に示す．

アフィン変換，TPS のどちらの手法も，画像の変換をしない従来手法の場合と比較して高い精度が得られた．中でも，TPS による手法では，$k=5$ では従来手法より約 6.7% の精度の向上が見られた．

次に，VLAD によって評価をした 2 つ目の手法の結果として，同じく Spatial Transformer Networks によって変換した場合としない場合の $Acc(k)$ を比較して表 2 に示す．

VLAD を用いた評価では，画像の変換をしない場合でも約 92.3% と，Triplet Network による評価よりも高い精度で識別ができている．画像の変換をした場合との比較をすると，アフィン変換を施した手法では識別精度はほとんど変化せず，TPS を施した手法では従来手法よりも識別精度が低下した．

## 6.　考察

Spatial Transformer Networks によって変換された画像の一例を図 9 に示す．

シーン 1 を見ると，大きな建物や空との境界線といったシーンにおける主要なパーツの位置が，Spatial Transformer Networks によって元のシーン画像に近づくように変換されていることがわかる．シーン 2 やシーン 3 は，Triplet Network によって得られた正解のシーン画像との類似度が，アフィン変換を施すことで低下し，TPS を施すことで向上した画像である．変換前の角膜表面反射画像の非線形な歪みが大きい場合は，アフィン変換による識別精度向上はあまり見込めないことがわかった．また，中にはシーン 5 やシーン 6 のように，TPS によって画像の大部分が壊れてしまった画像もあった．これは，Spatial Transformer Networks の学習の際に画像の色彩情報を参照するため，太陽光が眼球内に直接入射していた，などの撮影時の条件に起因する，角膜表面反射画像特有の着色がなされた領域を破壊するように学習が進んだためであると推測される．

評価手法の違いによって TPS による手法の識別精度が大きく変わる理由として，Spatial Transformer Networks の損失関数が色彩情報に強く依存するものであったことが考えられる．前述のシーン 4 のような角膜表面反射画像は TPS によって大きく歪んでしまったが，Triplet Network によって得られた正解のシーン画像との類似度は向上し，逆に VLAD による類似度は低下した．Triplet Network による評価の際にも patch ごとの色彩情報を利用するが，VLAD の構成の際に集約した Dense SIFT 特徴量は色彩情報を利用せず，また特徴量の獲得の過程で画像の極値を利用するため，画像の構造が歪むことで，正解のシーン画像から得られる特徴量とはかけ離れたものが得られてしまうと考えられる．

また，シーン 5 については，いずれの変換手法・評価手法を用いても正しくシーン識別をすることはできなかった．これは，角膜表面反射画像に虹彩パターンが強く出すぎており，色，被写体の輪郭などがほとんど角膜内に確認できないことに起因すると考えられる．

### 6.1　まとめ，今後の課題

本研究では，角膜表面反射画像からのシーン識別の精度向上を目的として，Spatial Transformer Networks を用いて角膜表面反射画像を変換し，シーン画像との位置のずれを補正する手法を提案した．実験には，それぞれの画像から対応した領域を抽出できるように，フレームを利用して撮影した自作のデータセットを用いた．角膜表面反射画像の変換は，アフィン変換と制御点 16 個の TPS の 2 種類の手法で行い，変換された角膜表面反射画像の評価には，patch ベースの Triplet Network による手法と VLAD による手法の 2 つを用いて実験をした．画像の変換をしない従来手法との識別精度の比較をした結果，アフィン変換を施した手法では，どちらの評価手法でもわずかな精度向上が見られた一方，TPS を施した手法では，Triplet Network による評価の精度は大きく向上し，反対に VLAD による

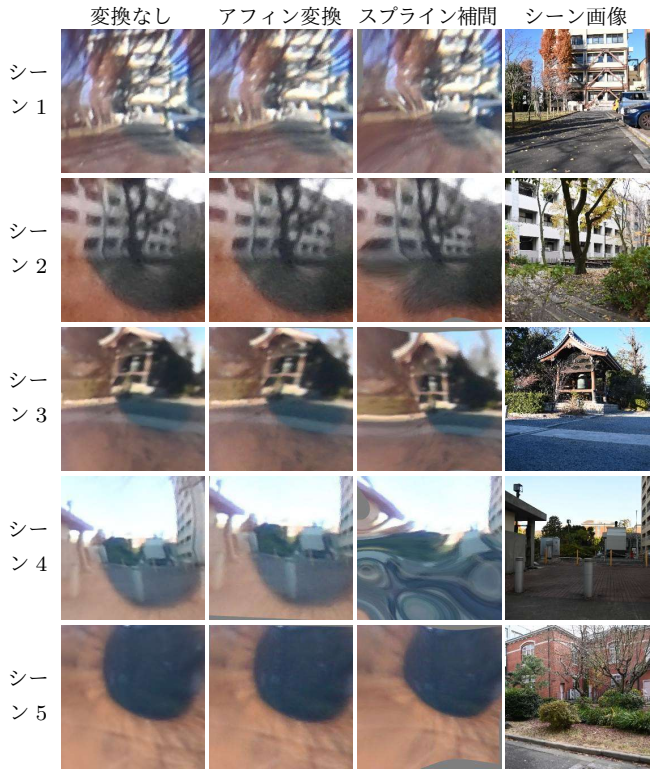


**図 9** Spatial Transformer Networks によって変換された画像の一例

評価の精度は従来手法よりも低下した．

本研究においての問題点として，以下のようなことが挙げられる．

(1) Spatial Transformer Networks の学習データがテストデータと同一のものであるため，学習データを別のデータにした場合，本研究で得られた識別率を維持できない可能性が高い (汎化性能が低い可能性がある)．

(2) Spatial Transformer Networks の損失関数が色彩情報に強く依存するものであったため，VLAD のように色彩情報を参照しない評価手法では，画像変換による精度の向上が期待できない．

(3) 同様の理由で，角膜表面反射画像が暗くシーン画像との明度の差が大きすぎるデータに対しては，(特に TPS による変換では) 期待するような変換ができない．

(4) 強い虹彩パターンの出ている角膜表面反射画像に対しては，画像の変換・シーン識別ともにうまくできない．

問題点 1 に対しては，撮影を重ねてデータセットを増やし，実際に学習データとテストデータを分けた場合について検証をするべきである．しかし，大量の撮影データを獲得するには非常にコストがかかるため，論内で述べた撮影方法より効率的な何らかの手法を考える必要がある．

問題点 2,3 に対する解決策としては，Spatial Transformer Networks の損失関数について再考し，色彩情報以外の情報にも大きく依存するような関数を定義して学習をしてみることが挙げられる．一般に角膜表面反射画像は太陽光の加減や虹彩パターンによって，シーン画像と比べて色が変わっているので，本実験で利用した画素値の L1-loss は，本タスクにおいて最適な損失関数であったとは言い難い．画像中のエッジの位置など，色彩情報を参照しない別の特徴量から算出した誤差を損失関数に加えるなどすることで，画像の変換の精度，評価の精度ともに向上する可能性がある．または，Spatial Transformer Networks の損失関数を，評価手法で用いる誤差関数と同じものにしてしまうことで，その評価手法における識別精度は改善されると推測できる．

**謝辞** 本研究は科研費 17H01779, 26249029, 15H02738, および, JST CREST, JPMJCR17A5 の支援を受けている．

## 参考文献

[1] 江川佳輝，小川太士，中澤篤志ほか：深層学習を用いた自己撮影画像の撮影場所検索，研究報告コンピュータビジョンとイメージメディア（CVIM），Vol. 2018, No. 56, pp. 1–5 (2018).

[2] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. and Shah, R.: Signature verification using a" siamese" time delay neural network, *Advances in neural information processing systems*, pp. 737–744 (1994).

[3] Hoffer, E. and Ailon, N.: Deep metric learning using triplet network, *International Workshop on Similarity-Based Pattern Recognition*, Springer, pp. 84–92 (2015).

[4] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307 (2016).
[5] Jégou, H., Douze, M., Schmid, C. and Pérez, P.: Aggregating local descriptors into a compact image representation, *2010 IEEE computer society conference on computer vision and pattern recognition*, IEEE, pp. 3304–3311 (2010).
[6] Jaderberg, M., Simonyan, K., Zisserman, A. et al.: Spatial transformer networks, *Advances in neural information processing systems*, pp. 2017–2025 (2015).
[7] : Spatial Transformer Networks Tutorial — PyTorch Tutorials 1.4.0 documentation. `https://pytorch.org/tutorials/intermediate/spatial_transformer_tutorial.html`.
[8] : GitHub - WarBean/tps_stn_pytorch: PyTorch implementation of Spatial Transformer Network (STN) with Thin Plate Spline (TPS). `https://github.com/WarBean/tps_stn_pytorch`.
[9] : GitHub - adambielski/siamese-triplet: Siamese and triplet networks with online pair/triplet mining in PyTorch. `https://github.com/adambielski/siamese-triplet`.
[10] Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M. and Pajdla, T.: 24/7 place recognition by view synthesis, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817 (2015).
[11] 枝本祐典，中澤篤志，西田豊明ほか：敵対的生成ネットワークを用いた角膜表面反射画像からのシーン識別，研究報告コンピュータビジョンとイメージメディア (CVIM)，Vol. 2019, No. 12, pp. 1–8 (2019).