DEIM Forum 2018 H3-3

一人称画像と位置に基づくライフログセグメンテーション

瀧本　祥章†　　山本　修平†　　西村　拓哉†　　戸田　浩之†

† 日本電信電話株式会社 NTT サービスエボリューション研究所　〒 239–0847 神奈川県横須賀市光の丘 1-1

E-mail: †{yoshiaki.takimoto.ar,shuhei.yamamoto.ea,takuya.nishimura.fk}@hco.ntt.co.jp, hirotoda@acm.org

あらまし　スマートフォンやスマートグラスなどのデバイスの普及により，ユーザの状態や，位置情報，周辺環境を記録したライフログと呼ばれる系列データを得られるようになった．ライフログには多岐にわたる情報が含まれており，ユーザの行動内容理解など，活用しようとする研究が盛んに行われている．しかし，ライフログをセグメントと呼ばれるインデックス可能な単位に分割する標準的なアプローチは未だ存在しない．ここでセグメントとは，料理や買い物をしているなど，それ単独で意味を持ち，利活用のための検索における基本的な単位となるものである．そこで本稿では，ユーザの位置情報と，ユーザの視点から継続的に撮影された画像に注目し，ライフログをセグメントに分割する手法を提案する．評価実験では，NTCIR-13 Lifelog-2 タスクで提供されるユーザ 2 人の延べ 90 日分のライフログと，人手による分割結果を利用して提案手法の検証を行い，提案手法により高精度にセグメントへの分割ができることを示した．

キーワード　ライフログ，DBSCAN，滞留点抽出，Gated CNN，ImageNet，Places，移動軌跡

1. はじめに

スマートフォンやスマートグラスなどのデバイスの普及により，ユーザの状態や，位置情報，周辺環境に関するライフログと呼ばれる系列データを得られるようになった．例えば，スマートデバイスの一種である Fitbit [1] からは心拍数や，歩数，睡眠の質と時間などのユーザの状態を表すデータから構成されるライフログが得られる．また，スマートフォンアプリである Moves [2] からは walking，running，cycling などのユーザの移動状況や，ユーザの移動軌跡，滞在地から構成されるライフログが得られる．

このように得られるライフログは多様な情報を持つため，ユーザの行動内容理解や，ヘルスケアなどへ活用しようとする研究が盛んに行われている [14]．しかし，ライフログのデータ量はその性質上，時間と共に増大し，大量のデータから人手で有用な情報を特定することは困難である．そのため，ライフログをセグメントと呼ばれる索引付け可能な単位に分割すること（ライフログセグメンテーション）が求められている．様々な手法 [8–10, 19, 20, 28, 29] が提案されているが，セグメンテーションされるセグメントの性質や数，長さに制約がある場合や，セグメンテーションを行う上で重要な情報である位置情報を考慮していないなどの課題がある．ここでセグメントとは，料理や買い物をしているなど，それ単独で意味を持ち，利活用のための検索における基本的な単位となるものである．

そこで本稿では，図 1 のように得られるユーザの位置情報と，ユーザの視点から継続的に撮影された画像（一人称画像）に注目し，ライフログセグメンテーションを行い，図 1 中のセグメント 1–6 のようなセグメントを抽出する手法を提案する．セグメントはユーザが同一行動を継続している時間や，環境を表す．例えばセグメント 1 では，ユーザは会話をしており，セグメント 2 ではコンピュータを操作している．これらのセグメントを抽出するため，本稿ではまず，画像類似度アプローチをベースラインとして提案し，技術的な課題を明らかにする．画像類似度アプローチでは系列データ中の一人称画像を前後で比較し，類似度が閾値以下である場合に画像間を分割点として抽出する．この手法はシンプルであるが，ノイズの影響が大きい，図 1 のセグメント 4 やセグメント 5 など移動を伴うセグメントを抽出できないという 2 つの課題がある．この 2 つの課題を解決するため，滞留抽出アプローチでは滞留点抽出技術である D-star [22] を，2 群検定アプローチでは Welch の 2 群検定を活用する．また，Gated CNN アプローチでは前後複数枚の画像を入力として用いることにより，これらの課題を回避する．

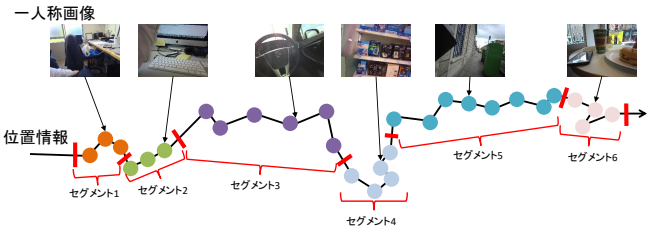


図 1: ライフログとセグメントの例

評価実験では，提案手法を評価するため，NTCIR-13 Lifelog-2 タスク [13] で提供されるユーザ 2 人の延べ 90 日分のライフログと，人手による分割結果を利用して各手法の比較検証を行った．その結果，ベースラインである画像類似度アプローチと比較して，他の 3 つの手法が高い精度を示すことを確認した．

本稿の主な貢献は以下の通りである．

- ベースラインとなる画像類似度アプローチの提案とライフログセグメンテーションにおける課題の明確化
- NTCIR-13 Lifelog-2 タスクの LEST（Lifelog Event

表 1: 関連研究と提案手法との相違点

文献	利用データ	提案手法との相違点
Ellis ら [9,10]	音声	場所，環境のみを考慮
Wang ら [29]	画像	セグメントの単位時間が 5 分
Lin ら [20]	映像	場所のみを考慮
Doherty ら [8]	画像	長時間の移動を考慮できない
Li ら [19]	画像	対象セグメントが限られる
Talavera ら [28]	画像	位置情報が未考慮
Lu ら [21]	映像	位置情報が未考慮
Castro ら [5]	画像	多量の教師データが必要
Poleg ら [23]	画像	対象セグメントが限られる

Segmentation Task）サブタスク [13] の優勝手法を含む，ライフログセグメンテーションの課題を解決する相異なる 3 つの手法の提案

- 複数のユーザおよび長期間日常的に収集したデータを用いた評価実験による提案手法の定性的評価および定量的評価，および有効性の確認

本稿の構成は以下の通りである．まず，2 章 で関連研究として，ライフログセグメンテーションに関する研究を紹介する．3 章 で本稿における基本的な概念の定義や問題設定ついて述べる．その後，4 章 でベースライン手法および提案手法の詳細を論じ，5 章 で提案手法を評価するために行った実験について述べる．最後に，6 章 で本稿のまとめと今後の課題について述べる．

2. 関 連 研 究

ライフログは近年注目され，関連する研究が多く行われている [14]．その活用のため，セグメントと呼ばれる索引付け可能な単位に分割する技術であるライフログセグメンテーションが求められており，表 1 のように様々な手法が提案されている [5,8–10,19–21,23,28,29]．各手法と提案手法との相違点を順に述べる．

Ellis らは 62 時間分の周囲の音声データをライフログとして収集し，スペクトル情報を活用して street や restaurant などの 16 個の場所や環境を表すセグメントに分割した [9,10]．本稿では収集する情報が音声データではなく，位置情報と一人称画像であるという点や，想定するセグメントとして場所や環境だけでなく，家事や食事などユーザの行動に基づくセグメントも想定する点で異なる．Wang らは 6 週間の一人称画像をライフログとして収集し，walking outside や meeting などの 6 種類のセグメントに 5 分単位で分割した [29]．しかし，ライフログに含まれるセグメントは 5 分間単位とは限らない．これに対し提案手法では，このような単位時間を設定しない．Lin らは時間制約付きのクラスタリングにより，映像から構成されるライフログを office など場所を表すセグメントに分割した [20]．本稿では，前述したように家事や食事などユーザの行動を基づくセグメントも想定する点で異なる．Doherty らは Sense-Cam [16] で収集した 1 日当たり 1785 枚の画像からなるデータに対し，前後間の画像について，MPEG-7 のメタデータから得られる色やエッジ情報の類似度を計算し，閾値未満の箇所をセグメントの境界としている [8]．ただし，ユーザが移動中の場合には過剰に境界を検出してしまうことから，後処理によって互いに近いセグメント境界のうち，最初の境界のみを残す処理を行う．しかし，後処理は予め定めた時間間隔（ [8] では 3 分）のみに依存することから，より長時間の移動を伴うセグメントを正しく検出できない．これに対し提案手法では，長時間の移動を伴うセグメントであっても検出可能である．Li らは Sense-Cam により収集した画像系列を時系列データとし，固有値のピークを導出し，セグメントの境界とする手法を提案した [19]．しかしこの手法は，全てのセグメントを検出し，分割することを想定しておらず，ノイズへの対応もしていない．これに対し提案手法では，全てのセグメントを検出し，分割することを想定しており，ノイズへの対応もしている．Talavera らは ImageNet [24] で学習した CNN によりグラフカット技術を用いてセグメントの境界を検出する [28]．また，Lu らは映像に映っている物体に注目し，一人称映像から重要な瞬間を抽出し，一人称映像の要約を行う [21]．しかしこれらの手法では，画像情報のみを考慮しており，ライフログセグメンテーションにおいて重要な情報である位置情報を考慮していない．これに対し，提案手法の滞留抽出アプローチや Gated CNN アプローチでは位置情報を考慮する．Castro らは ImageNet で学習した CNN を再学習することによりセグメントにおけるユーザの行動内容を含めて推定を行う [5]．しかし，想定する行動ごとに教師データが必要となるため，新たな行動を考慮するために数千枚のデータにラベルを付与する必要がある．これに対し，提案手法の滞留抽出アプローチや 2 群検定アプローチは教師データを必要とせず，Gated CNN アプローチにおいても行動別の教師データは必要としない．Poleg らはユーザの頭部の動きに注目し，画像の変位を用いてユーザの行動を認識する [23]．しかし，認識できる行動は sitting や walking などのユーザの移動状態に限られる．これに対し，提案手法は前述のように家事や食事などの行動を想定する点で異なる．

表 2: 本稿で用いる記号一覧

記号	意味
L	ライフログ
l_i	観測データ
n	ライフログのデータ数
$\boldsymbol{S_i}$	観測データ l_i の画像特徴量ベクトル
$s_{i,j}$	画像特徴量ベクトル $\boldsymbol{S_i}$ の次元 j の値
$\boldsymbol{T_i}$	観測データ l_i の潜在トピック分布

3. 準　　備

本章では，本稿における基本的な概念の定義や問題設定を行う．本稿で用いる記号は表 2 の通りである．

3.1 基本的な概念

本稿では，位置情報と一人称画像などの系列データから構成されるライフログをセグメントに分割する．

まず，入力となるライフログを以下のように定義する．

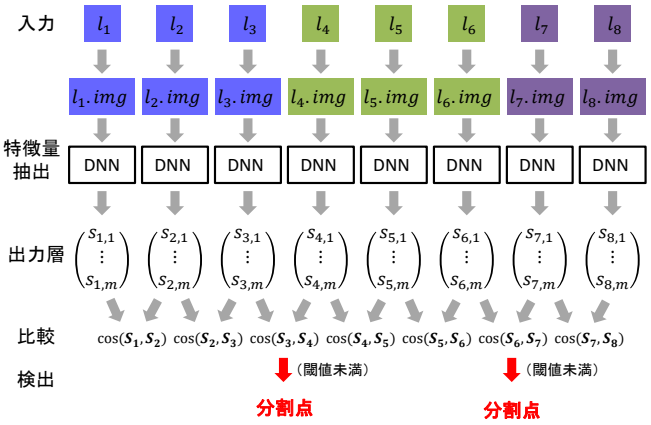
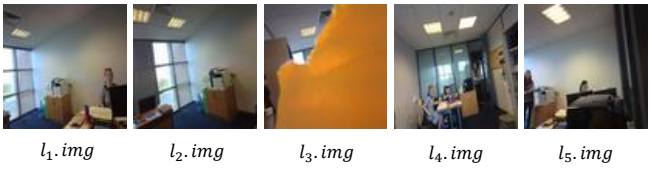


図 2: 画像類似度アプローチの概要



（a） ノイズを含むライフログ例

（b） 移動を伴うライフログ例

図 3: ライフログの例

［定義 1］ **ライフログ** $L = [l_1, \ldots, l_n]$ はユーザの状態や位置情報，周囲の情報の観測データを時系列順に並べた系列である．ここで，ライフログに含まれる各データ $l_i \in L$ は時刻情報 $l_i.time$，位置情報 $l_i.loc$，一人称画像 $l_i.img$ など複数の属性を持つ．位置情報 $l_i.loc$ は位置を表す二次元座標であり，一人称画像 $l_i.img$ は自動的に撮影されたユーザ視点の生画像である．

［定義 2］ **セグメント**は開始時刻と終了時刻の組であり，ユーザや周囲の状態について特定の状態が継続した期間を表す．ここで，特定の状態には料理や買い物，散歩などが例として挙げられる．

3.2 問題設定

本稿では，任意期間のライフログ $L = [l_1, \ldots, l_n]$ を入力して受け取り，セグメントの分割点を抽出し，セグメント集合を出力する．ただし，各データ l_i はいずれかのセグメントに属するものとする．

4. 提案手法

本章では，本稿で扱う 4 つの手法について順に述べる．まず，ベースラインとなる画像類似度アプローチの内容と課題について述べる．その後，課題について別々のアプローチで解決を図った 3 つの提案手法について述べる．

4.1 画像類似度アプローチ

ライフログセグメンテーションの素朴な手法として図 2 に示すような前後の観測データ間を比較する手法が考えられる．この手法では，観測データ中の画像のみを用いてセグメンテーションを行う．図 2 ではまず，周囲の状態を把握するため，観測した一人称画像 $l_i.img$ を GoogLeNet [26] などの，画像のクラス分類が可能なニューラルネットワークに入力する．そして，出力された各クラスの分類確率 $\boldsymbol{S_i} = (s_{i,1}, \ldots, s_{i,m})$ を入力画像の特徴ベクトルとする．その後，前後の観測データ間の画像特徴量 $\boldsymbol{S_i}, \boldsymbol{S_{i+1}}$ をコサイン類似度を用いて比較し，類似度が閾値未満であるデータ間を分割点として検出する．これにより，セマンティックギャップ [6] として知られる画像表現と意味内容の乖離を回避し，観測データの意味内容に基づく比較が可能になると期待できる．

この手法はシンプルである一方で，ノイズに弱い，位置情報を考慮できないという 2 つの課題を持つ．例えば，図 3 のようなライフログを得られたときを考える．図 3（a）は 1 つのセグメントとして抽出すべき部分である．しかし，$l_3.img$ の撮影時にカメラを物体が覆ってしまっていることにより，$l_2.img$ と $l_4.img$ の間の類似度が大きい一方で，$l_2.img$ と $l_3.img$，$l_3.img$ と $l_4.img$ の間の類似度が小さくなってしまい，3 つのセグメントに分割されてしまう．このようにカメラを何らかの物体が覆う，ユーザが体の向きを変えるなどが原因となるノイズにより，セグメントが過剰に分割されてしまう．図 3（b）に移動を伴う単一セグメント中のライフログの例を示す．このライフログはユーザがショッピングをしており，1 つのセグメントとして抽出すべき部分である．しかし，ユーザが移動しているため，撮影される画像が刻一刻と変化しており，必ずしも類似度が高くなるとは限らない．そのため，閾値によっては複数個のセグメントに分割されてしまう．このように，移動を伴うセグメントでは，同一のセグメント中でも画像が変化していくため，正しい分割ができない．

4.2 滞留抽出アプローチ

画像類似度アプローチにおける 2 つの課題を解決するため，滞留抽出アプローチでは画像だけでなく，位置情報にも焦点を当て，分割点を抽出する．本稿では滞留の抽出に D-star [22] を拡張し，用いる．ここで，滞留とはユーザが一定時間以上，一定範囲内に留まることを表す．また，D-star は移動軌跡からユーザが滞留を行った地点である滞留点を抽出する技術であり，ノイズの影響を受けにくい，ストリーム処理に対応しているなどの特徴を持つ．

滞留抽出アプローチの概要を図 4 に示す．まず，滞留抽出アプローチでは画像類似度アプローチと同様に学習済みネットワークに一人称画像 $l_i.img$ を入力し，出力された各クラスの分類確率 S_i を得る．次に，D-star を画像の類似度を考慮するように拡張した simD-star により，ライフログから移動を伴わない（滞留している）セグメントをクラスタとして抽出する．なお，simD-star の詳細は後述する．その後，クラスタの前後を分割点として，移動部分のセグメントおよび滞留部分のセグメントを抽出する．

simD-star のアルゴリズムをアルゴリズム 1 に示す．ここで，アルゴリズム中の関数 $d(l_i.loc, l_j.loc)$ は $l_i.loc$ と $l_j.loc$ の距離

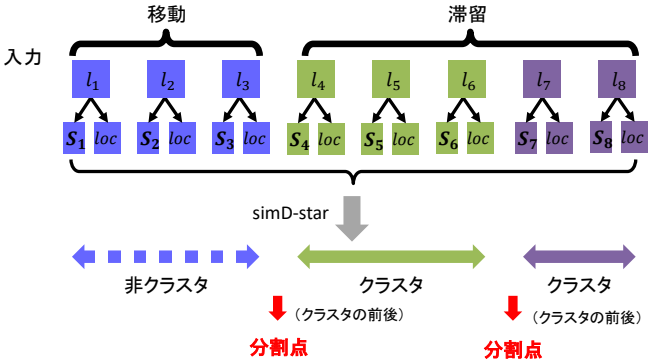


図 4: 滞留抽出アプローチの概要

を表し，$s(\boldsymbol{N}(l_i))$ は近傍データ集合 $\boldsymbol{N}(l_i)$ に含まれるデータの最早観測時刻と最遅観測時刻の差，即ち観測期間を表す．また，$\mathtt{sim}(\boldsymbol{S_i}, \boldsymbol{S_j})$ は画像特徴量間の類似度を表し，画像の意味的な類似度を考慮するため，[27] を参考に以下のように定義する．

$$\mathtt{sim}(\boldsymbol{S_i}, \boldsymbol{S_j}) = \frac{\sum_{k=1}^{m} \min(s_{i,k}, s_{j,k}) \times \mathtt{idf}_k}{\sum_{k=1}^{m} \max(s_{i,k}, s_{j,k}) \times \mathtt{idf}_k} \tag{1}$$

なお，$\mathtt{idf}_k$ は逆文書頻度を表し，

$$\mathtt{idf}_k = \log \frac{n}{\sum_{p=1}^{n} s_{p,k}} \tag{2}$$

と定義され，各次元の重要度を考慮することを可能とする．simD-star は入力として，ライフログデータ L と，ウィンドウサイズ q，距離閾値 ε，最短観測期間閾値 m_{time}（DBSCAN [11] の密度閾値 $MinPts$ に相当），最短滞留時間閾値 m_{stay} を受け取り，ライフログの滞留を伴うセグメントを抽出する．simD-star では時系列順にライフログのデータ l_i の処理を行う．まず，l_i の近傍に存在し，かつ，画像特徴量が類似するデータ l_j をウィンドウ W から抽出し（4–7 行目），各々の近傍データ集合 $\boldsymbol{N}(l_i), \boldsymbol{N}(l_j)$ に加える（8–9 行目）．その後，l_{i-q+1} について，近傍データ集合の観測期間が最短観測期間閾値以上である場合（11 行目）は，近傍データ集合 $\boldsymbol{N}(l_{i-q+1})$ が既存のクラスタと同一のデータを保持する場合には，同一データを保持するクラスタおよび近傍データ集合 $\boldsymbol{N}(l_{i-q+1})$ をすべてマージする（12–13 行目）．そうでない，即ち，近傍データ集合 $\boldsymbol{N}(l_{i-q+1})$ が既存のクラスタと同一のデータを保持しない場合には，新たなクラスタを形成する（14–15 行目）．最後に，観測期間が滞留時間閾値以上であるクラスタを全て，滞留を伴うセグメントとして出力する．

4.3 2 群検定アプローチ

2 群検定アプローチでは，画像類似度アプローチにおける 2 つの課題を解決するため，画像の特徴次元を削減し，対象データデータ前後での画像の分布変化に注目する．

2 群検定アプローチの概要を図 5 に示す．まず，画像類似度アプローチ，滞留抽出アプローチと同様に学習済みネットワークに一人称画像 $l_i.img$ を入力し，出力された各クラスの分類確率 S_i を得る．次に，画像の特徴次元を削減し，ノイズの影響を小さくするため，各画像の画像特徴量 $\boldsymbol{S_i}$ を文書，各特徴量のうち，確率値 $s_{i,j}$ が閾値を超える次元をその画像が含む

アルゴリズム 1: simD-star

Input: $L, q, \varepsilon, m_{time}, m_{stay}, \tau$

1 $W \leftarrow \phi$ // スライディングウィンドウ
2 $\mathcal{C} \leftarrow \phi$ // クラスタ集合
3 **foreach** $l_i \in L$ **do**
4 Push l_i in W
5 $\boldsymbol{N}(l_i) \leftarrow \phi$ // 近傍データ集合
6 **foreach** $l_j \in W$ **do**
7 **if** $d(l_i.loc, l_j.loc) < \varepsilon \wedge \mathtt{sim}(\boldsymbol{S_i}, \boldsymbol{S_j}) > \tau$ **then**
8 Add l_i to $\boldsymbol{N}(l_j)$
9 Add l_j to $\boldsymbol{N}(l_i)$
10 Shift l_{i-q+1} from W
11 **if** $s(\boldsymbol{N}(l_{i-q+1})) \geqq m_{time}$ **then**
12 **if** $\exists \boldsymbol{C} \in \mathcal{C}. \boldsymbol{N}(l_{i-q+1}) \cap \boldsymbol{C} \neq \phi$ **then**
13 同一データを持つクラスタと $\boldsymbol{N}(l_{i-q+1})$ を全てマージ
14 **else**
15 Add $\boldsymbol{N}(l_{i-q+1})$ to $\mathcal{C}$
16 **return** $\{\boldsymbol{C} | s(\boldsymbol{C}) \geqq m_{stay}\}$

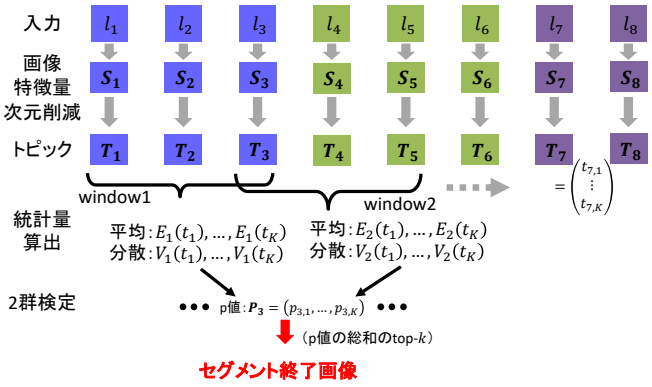


図 5: 2 群検定アプローチの概要

単語とみなし，全てのデータに対して LDA（Latent Dirichlet Allocation）[4] を用いて潜在トピック分布 $\boldsymbol{T_i} = (t_{i,1} \ldots, t_{i,K})$ を抽出した．その後，データ l_i に対して，前後 q 個ずつのデータを対象とする 2 つのスライディングウィンドウ `window1`，`window2` を用意し，各ウィンドウ中の潜在トピック分布集合 $\{\boldsymbol{T_{i-q}}, \ldots, \boldsymbol{T_i}\}, \{\boldsymbol{T_i}, \ldots, \boldsymbol{T_{i+q}}\}$ についてトピック次元ごとに平均 $E_i(t_1), ..., E_i(t_K)$ と分散 $V_i(t_1), ..., V_i(t_K)$ を算出する．その後，Welch の t 検定により，p 値 $\boldsymbol{P_i} = (p_{i,1} \ldots, p_{i,K})$ を算出する．最後に，次元ごとに求めた p 値の総和を計算し，値が大きい予め定めた定数個のデータをセグメント終了データとして抽出する．

4.4 Gated CNN アプローチ

Gated CNN アプローチは近年系列データを扱う上で注目されているネットワークである Gated CNN [7] に，対象時刻のデータとその前後のデータを入力することにより，対象時刻のデータがセグメント終了のデータであるか推定する手法である．なお入力するデータは，画像や位置情報に縛られずに，様々な情報を扱うことができ，入力を変化させてもネットワーク構成

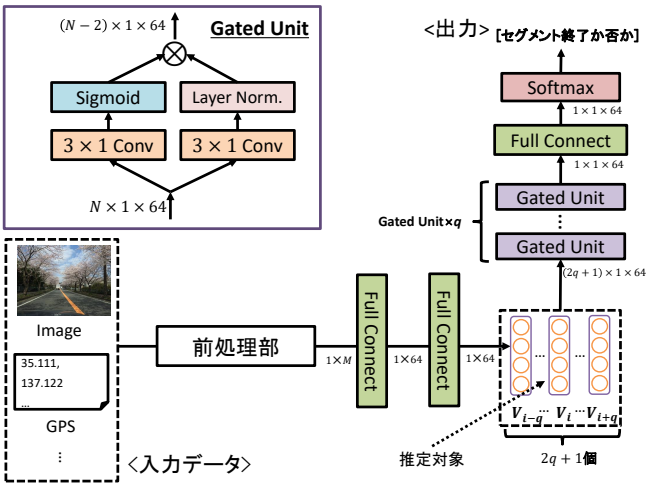


図 6: Gated CNN アプローチで用いたネットワーク構造

の変更を必要としない．図 6 にニューラルネットワークの構成を示す．

まず前処理部では，画像や位置情報などの入力データから複数のデータを生成する．入力データの例として，画像類似度アプローチや 2 群検定アプローチと同様の，画像特徴量 $\boldsymbol{S_i}$ や潜在トピック分布 $\boldsymbol{T_i}$ や，前後の画像特徴量および潜在トピック分布を比較した結果である $\cos(\boldsymbol{S_{i-1}}, \boldsymbol{S_i})$，$\cos(\boldsymbol{T_{i-1}}, \boldsymbol{T_i})$，緯度経度情報に DBSCAN を適用したものが挙げられる．次に，前処理部で生成されたデータは 2 層の Full Connect 層によってベクトル $\boldsymbol{V_i}$ に変換する．その後，ライフログの流れを考慮するため，前後 q 個のベクトル $\boldsymbol{V_{i-q}}, \ldots, \boldsymbol{V_{i+q}}$ と共に Gated Unit を複数回通される．Gated Unit は，入力から 2 つの同次元のベクトルを出力し，一方に Sigmoid 関数を適用してから要素積を取ることにより，必要な情報のみを出力する．本研究では，精度向上のため Sigmoid 関数を適用しないベクトルに対し，Layer Normalization [3] を適用している．最後に，Full Connect 層と Softmax 層によってデータ l_i がセグメント終了のデータであるか判定する．

5. 評価実験

本章では，提案手法の有用性を確認するため，NTCIR-13 Lifelog-2 タスク [13] で提供されたユーザ 2 人の延べ 90 日分のライフログと，人手による分割結果を用いた評価実験について述べる．まず，用いたデータについて説明する．次に，評価指標について述べる．最後に行った実験の結果について述べる．

5.1 データセット

本稿で用いるデータセットは NTCIR-13Lifelog-2 タスクで提供されるユーザ 2 人（ユーザ 1，ユーザ 2）のデータである．データの収集期間はユーザ 1 が 2016 年 8 月 8 日から 10 月 5 日の 59 日間であり，ユーザ 2 が 2016 年 9 月 9 日から 10 月 11 日の 9 月 20 日と 10 月 9 日を除く 31 日間である．ユーザが起きている間，1 分当たり 1 つのユーザ視点の画像および位置情報，Activity（Moves アプリで収集された walking や cycling などユーザの移動状況を表すラベル）を含むライフログデータを収集しており，1 日当たり 1,250 個から 1,500 個のデータが存在する．なお，プライバシー保護の観点から，顔やデバイスの画面が画像に映り込んでいる場合にはぼがしが入れられ，かつ，全ての画像が 1024 × 768 の解像度にリサイズされている．また，位置情報も自宅や職場については，GPS による絶対位置ではなく，それぞれ Home，Work と意味的な位置情報に Moves アプリケーション [2] によって置換されている．

前述のように画像特徴量が複数の手法で必要となるため，深層学習技術により抽出した．抽出に用いたモデルは ImageNet [24] で学習した GoogLeNet [26] および AlexNet [18]（1000 次元），Places365 [30] で学習した GoogLeNet，AlexNet，VGG [25] および ResNet [15] である（365 次元）．これらは *Caffe* [17] 上で動作する学習済みモデルが github(注1)(注2) 上で公開されている．

次に，2 群検定アプローチや Gated CNN アプローチで用いた潜在トピックを LDA によって抽出した．抽出方法は 4.3 節の通りであり，単語とみなす確率値の閾値は 0.1 と設定した．また予備実験の結果から，最も高い性能を示したことからトピック数を 10 に設定した．

5.2 評価指標

本稿では，NTCIR-13 Lifelog-2 タスクの LEST サブタスクによる評価指標に基づいて，Precision，Recall，F1 score を用いて評価を行った．各値の算出方法は以下の通りである．

$$\texttt{Precision} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} And\left(f(l_i,l_j), GT(l_i,l_j)\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} f(l_i,l_j)} \quad (3)$$

$$\texttt{Recall} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} And\left(f(l_i,l_j), GT(l_i,l_j)\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} GT(l_i,l_j)} \quad (4)$$

$$\texttt{F1 score} = \frac{2 \times \texttt{Precision} \times \texttt{Recall}}{\texttt{Precision} + \texttt{Recall}} \quad (5)$$

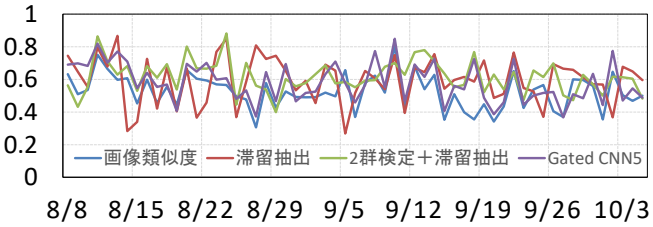
ここで $f(l_i, l_j)$ および $GT(l_i, l_j)$ はデータ l_i と l_j が提案手法による分割または正解データにおいて，同一セグメントに属している場合に 1（$True$），属していない場合に 0（$False$）の二値をとる関数である．また，$And(x, y)$ は論理積を表し，$True$，即ち，$x = y = 1$ のときに 1 となり，$False$ のときに 0 となる．

なお，Precision は各データの組が同一セグメントに含まれると判定した結果のうち，実際に同一セグメントに含まれる組の割合を表し，細かく分割すると，より大きな Precision を得られる可能性が高い．一方，Recall は同一セグメントに含まれるデータ組のうち，同一セグメントに含まれると判定できた組の割合を表し，分割数が小さくなると，より大きな Recall を得られる可能性が高い．また，F1 score は Precision と Recall の調和平均である．

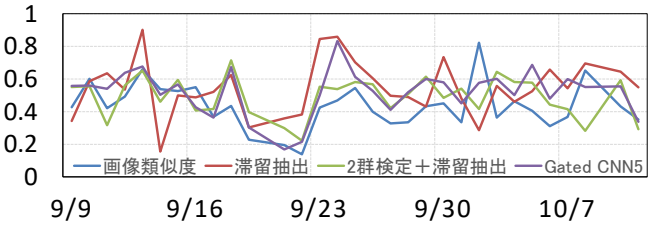
5.3 分割結果

4 章で提案した各手法の比較を行うため，各ユーザのライフログデータに提案手法を適用した．適用した手法と用いた入力の一覧は表 3 の通りであり，パラメータなどは予備実験を元に表 4 のように設定した．なお，「2 群検定 + 滞留抽出」は滞留抽出による分割と 2 群検定による分割の両方を用いて分割し

(注1)：https://github.com/BVLC/caffe/tree/master/models
(注2)：https://github.com/CSAILVision/places365



（a） ユーザ 1



（b） ユーザ 2

図 7: 日ごとの F1 score

た結果である．また，Gated CNN アプローチは入力する特徴量の追加が容易であることから様々な入力を追加したモデルを用意した．例えば,「トピック間類似度」は直前データとの潜在トピック分布のコサイン類似度,「DBSCAN」は緯度経度情報を DBSCAN を適用した結果のクラスタ ID,「Activity」は入力データに付加されていたデータである．ここで，モデルのユーザ独立性は，2 人のユーザに同一のモデル（パラメータや学習データが共通）を用いたことを表す．具体的には，チェックがある場合，ユーザ 2 人に共通の 1 つのモデルを用意し，Gated CNN アプローチにおける学習も 2 人分まとめて行った．チェックがない場合には，ユーザ 2 人別々のモデルを準備し，Gated CNN アプローチにおける学習も個々で行った．

各手法の結果は表 5 および図 7 の通りである．これらの結果から，画像類似度アプローチの Precision が最も大きくなった一方で，Recall の値が最も小さく，F1 score の値も最小となったことがわかる．これは，ノイズや移動を伴うセグメントの影響により過剰な分割を行ってしまったことが原因と考えられる．

その一方で，他のアプローチでは位置情報やデータの流れを考慮することにより，同一セグメントであるデータの判定ができるようになったことから，Recall が大きくなっている．特に，滞留抽出アプローチの F1 score が最大であること，2 群検定アプローチよりも滞留抽出アプローチを組み合わせた方が高精度であることから，位置情報の考慮がセグメント分割の精度向上に大きく寄与することが示唆される．例えば，ユーザ 2 の 9 月 13 日について滞留抽出アプローチの F1 score が他の手法を大きく上回っている．この日，ユーザ 2 は家事やショッピングなど移動を伴うセグメントを主に行っており，これらのセグメント中に含まれる画像の類似度は低くなっている．そのため，画像に基づく手法では過剰に分割を行ってしまい，Recall や F1 score が小さくなる傾向にある．その一方で，滞留抽出アプローチでは，これらを移動を伴うセグメントを移動として捉えることにより，過剰な分割を回避することにより比較的正確な分割が行え，Recall や F1 socre が大きくなる．また，ユーザ 1 の 8 月 27 日についても滞留抽出アプローチの F1 score が他の手法を大きく上回っている．この日，ユーザ 1 は多くの公共交通機関や自動車による移動や長時間のコンピュータの利用を行っていた．これらのセグメント中には周囲の状況変化やユーザの体勢，コンピュータの画面の変化などが原因により，画像の類似度が低くなってしまう．そのため，ユーザ 2 の 9 月 13 日と同様に，滞留抽出アプローチ以外の手法では過剰な分割を行ってしまい，Recall や F1 score が小さくなる傾向にある．これに対し，滞留抽出アプローチでは移動の考慮や，逆文書頻度を考慮した類似度によりこれらの影響を緩和できているため，比較的正確な分割が行え，Recall や F1 socre が大きくなる．

Gated CNN アプローチ内では，Gated CNN5 が最も高精度になり，滞留抽出アプローチ,「2 群検定＋滞留抽出」に次ぐ精度となった．これは，Gated CNN によって，複数の特徴量から有用な特徴を捉えることが可能であることを示す一方で，Gated CNN6 で入力に追加した LDA のトピックの次元数が学習データに対して大きく，過学習してしまうことを示している．

6. ま と め

本稿では，位置情報と一人称画像に注目し，ライフログをセグメントと呼ばれる索引付け可能な単位に分割するライフログセグメンテーション技術を 4 つ提案した．まず，ベースラインとなる一人称画像の類似度に基づく画像類似度アプローチを提案した．その後，画像類似度アプローチにより明らかになった技術的課題を解決するために，異なる 3 つのアプローチを提案した．提案手法の有効性を確認するために，実際に収集したライフログを用いた評価実験では，提案手法により，高精度にセグメントへの分割ができることを示した．また，各アプローチの比較から位置情報の考慮が重要であることが示唆された．

以下に今後の課題について述べる．本研究ではユーザごとに，一律のパラメータで分割を行った．しかし，各アプローチの日ごとの精度が大きく異なっていることから，ライフログは同一ユーザであっても，日によってその性質は大きく異なることが推測される．この日に依存する性質の変化を考慮した適応的なセグメンテーションについては今後の課題としたい．また，他の課題として，より多様なユーザに対する評価実験や，分割結果を活用して，買い物や食事などのユーザの行動内容を推定する技術である人間の行動認識技術 [12] などと組み合わせることが挙げられる．

文 献

[1] Fitbit. https://www.fitbit.com/home.
[2] Moves. https://moves-app.com/.
[3] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
[5] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. Predicting daily activities from egocentric images using deep learning. In *proceedings of the 2015 ACM International symposium on Wearable Computers*, pages 75–82. ACM, 2015.
[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Com-*

表 3: 適用手法と用いた特徴量

モデル	入力						モデルのユーザ独立性
	画像類似度	トピック	トピック間類似度	緯度経度	DBSCAN	Activity	
画像類似度	✓						✓
滞留抽出	✓			✓			
2 群検定		✓					
2 群検定 + 滞留抽出	✓	✓		✓			
Gated CNN1	✓						✓
Gated CNN2	✓					✓	✓
Gated CNN3	✓			✓			✓
Gated CNN4	✓		✓	✓		✓	✓
Gated CNN5	✓		✓	✓		✓	
Gated CNN6	✓	✓	✓	✓	✓	✓	

表 4: 設定パラメータ

画像類似度	画像特徴量：GoogLeNet(ImageNet)，閾値：0.04
滞留抽出	画像特徴量：GoogLeNet(Places365)，ウィンドウサイズ q：5，距離閾値 ε：40 m（ユーザ 2 は 120 m），最短観測期間閾値 m_{time}：3 min，滞留時間閾値 m_{stay}：5 min，類似度閾値 τ：0.4（ユーザ 2 は 0.3）
2 群検定	セグメント終了データ数：50，ウィンドウサイズ q：10
Gated CNN	適応学習アルゴリズム：Adam（$\alpha = 0.00001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$），ミニバッチサイズ：10，エポック数：200，学習データ：人手による分割結果 8 日分（正解データとは異なる），ウィンドウサイズ q：6

表 5: 各手法の分割精度

モデル	Precision	Recall	F1 score
画像類似度	**0.901**	0.352	0.494
滞留抽出	0.559	**0.698**	**0.579**
2 群検定	0.768	0.453	0.550
2 群検定 + 滞留抽出	0.762	0.485	0.573
Gated CNN1	0.848	0.421	0.547
Gated CNN2	0.837	0.421	0.545
Gated CNN3	0.855	0.406	0.535
Gated CNN4	0.860	0.407	0.539
Gated CNN5	0.846	0.436	0.561
Gated CNN6	0.790	0.455	0.551

puting Surveys (Csur), 40(2):5, 2008.

[7] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083*, 2016.

[8] A. R. Doherty and A. F. Smeaton. Automatically segmenting lifelog data into events. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 20–23. IEEE, 2008.

[9] D. P. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 39–47. ACM, 2004.

[10] D. P. Ellis and K. Lee. Accessing minimal-impact personal audio archives. *IEEE MultiMedia*, 13(4):30–38, 2006.

[11] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[12] Y. Guan and T. Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):11:1–11:28, June 2017.

[13] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, D. T. D. Nguyen, R. Gupta, and R. Albatal. Overview of the NTCIR-13 lifelog-2 task. In *The NTCIR-13 Conference*, Tokyo, Japan, 2017.

[14] C. Gurrin, A. F. Smeaton, A. R. Doherty, et al. Lifelogging: Personal big data. *Foundations and Trends® in Information Retrieval*, 8(1):1–125, 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[16] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood. Sensecam: A retrospective memory aid. *UbiComp 2006: Ubiquitous Computing*, pages 177–193, 2006.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105. Curran Associates Inc., USA, 2012.

[19] N. Li, M. Crane, and H. J. Ruskin. Automatically detecting" significant events" on sensecam. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(06):1350050, 2013.

[20] W.-H. Lin and A. Hauptmann. Structuring continuous video recordings of everyday life using time-constrained clustering. SPIE, 2006.

[21] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2714–2721. IEEE, 2013.

[22] K. Nishida, H. Toda, and Y. Koike. Extracting arbitrary-shaped stay regions from geospatial trajectories with outliers and missing points. In *ACM SIGSPATIAL International Workshop on Computational Transportation Science (IWCTS)*, pages 1–6, 2015.

[23] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2537–2544. IEEE, 2014.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[27] Y. Takimoto, K. Sugiura, and Y. Ishikawa. Extraction of frequent patterns based on users' interests from semantic trajectories with photographs. In *Proceedings of the 21st International Database Engineering & Applications Symposium*, pages 219–227. ACM, 2017.

[28] E. Talavera, M. Dimiccoli, M. Bolanos, M. Aghaei, and P. Radeva. R-clustering for egocentric video segmentation. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 327–336. Springer, 2015.

[29] Z. Wang, M. D. Hoffman, P. R. Cook, and K. Li. Vferret: content-based similarity search tool for continuous archived video. In *Proceedings of the 3rd ACM workshop on Continuous archival and retrival of personal experences*, pages 19–26. ACM, 2006.

[30] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.