車載動画キュレーションのための観光地らしさ推定モデルの構築

河中 昌樹[1,a)]　松田 裕貴[1,2]　諏訪 博彦[1,2]　安本 慶一[1,2]

概要：近年，ドライブレコーダで撮影した動画に対してキュレーションを行うことで，ドライブ観光のメモリアル経路動画の自動生成法が提案されている．しかし，従来手法では “沖縄らしさ” といった特定の尺度に基づいてキュレーションを行うため，各ユーザが想定する多様な “沖縄らしさ” に基づくメモリアル経路動画を生成できないことや他の観光地での適用が難しいことなどの問題がある．この問題を解決するために我々は，汎用的な尺度の組み合わせによって “観光地らしさ” を表現することを考え，感性工学分野で用いられる手順に基づき，“観光地らしさ” を構成する代表語の抽出している．本研究では，これらの代表語を用いて “観光地らしさ” を構成する尺度を明らかにするとともに，ドライブレコーダから得られる車載動画データを入力として各尺度のスコアを推定するモデルの構築を行う．実験結果より，車載動画キュレーションに必要な 6 つの尺度を明らかにし，それぞれの尺度スコアは推定可能であることを示した．

1. はじめに

近年，観光終了後に観光地での画像やメモリアル動画をソーシャルメディアに投稿するユーザが増加している．観光後のメモリアル動画の作成には，動画編集ソフトを用いる方法に加えて，RealNetworks 社の提供する RealTimes [1] のような写真に対応する位置情報に基づいて画像からショートムービーを自動的に作成・提供するサービスが提案されている．RealTimes [1] のような自動的にメモリアル動画を作成するシステムは，動画編集のスキルがないユーザでも簡単に使えることから，需要が高まっている．

しかし，従来のシステムによって生成されるメモリアル動画の多くは，複数の観光スポットの写真をまとめるものであり，観光移動経路に着目したメモリアル動画の作成支援を行うサービスは，みあたらない観光における移動の割合は多いため，印象に残る場面も多く存在する．また，移動経路に関する動画をキュレーションして提供することは，観光の流れを直感的に想起・共有する手助けとなる．

観光における移動方法として，バスや電車などの公共交通機関のほかに，歩行や車での移動が考えられる．バスや電車などの公共交通機関や歩行の場合には，観光者自身が移動中の経路動画を撮影することは困難である．しかし，車の場合には，事故や煽り運転の被害を受けた際の，自身に過失がないことの映像記録を残すためのドライブレコーダが搭載されていることが多く，また走行中に時経路の動画がは随撮影されていることから，そのデータを用いることができる．近年では，ドライブレコーダの画質などの性能が向上したことによって，運転技術向上のためのシステム [2]，観光支援システム [3], [4], [5], [6]，路上駐車の検出 [7] にも活用されている．

片山らは，ドライブレコーダの思い出深いシーンをメモリアル動画に残すための動画キュレーションアルゴリズム [6] を開発している．このアルゴリズムでは，代わり映えのしない道のり，信号での停車や渋滞などの冗長な部分を削除し，見どころとなる思い出深いシーンの抽出を行なっている．具体的には，沖縄をドライブ観光した際のドライブレコーダの動画の “沖縄らしさ” をクラウドソーシングを用いて収集・尺度化するとともに，沖縄らしさを予測する機械学習モデルを構築し，メモリアル動画キュレーションを行っている．しかし，“沖縄らしさ” などの曖昧な表現は人によって感じ方が異なるため，各ユーザが想定する多様な “沖縄らしさ” に基づくメモリアル経路動画を生成できないことや他の観光地での適用が難しいことなどの問題がある．本研究では，この問題の解決策として，先行研究で用いている解釈に差が生じやすい尺度の代わりに，人による解釈の差が生じにくい汎用的な尺度の組み合わせによって “観光地らしさ” を表現することで，メモリアル動画キュレーションの品質を向上することを目指す．

その実現には，“観光地らしさ” が，人が観光地から得る

1　奈良先端科学技術大学院大学，
Nara Institute of Science and Technology
2　理化学研究所 革新知能統合研究センター（AIP），RIKEN AIP
a)　kawanaka.masaki.kj1@is.naist.jp

印象のどのような組み合わせによって表現できるのかを明らかにする必要がある．我々はこれまでに，沖縄県をドライブ観光した際のドライブレコーダの動画を用い，メモリアル経路動画キュレーションに用いるための尺度を導くため，“沖縄らしさ” を構成する印象語のクラスタ抽出および代表語の選定を，感性工学分野で用いられている手順 [8] に基づき実施している [9].

本稿では，これらの代表語を用いて，“沖縄らしさ” を構成する尺度を明らかにし，ドライブレコーダから得られる車載動画データを入力として各尺度のスコアを推定する CNN モデルの構築を行う．その結果，車載動画キュレーションに必要な 6 つの尺度を明らかとし，また各尺度についてスコアを推定できる可能性を示した．

本稿の構成は以下の通りである．2 章にて関連研究を紹介し，本研究の立ち位置を整理する．次に 3 章でドライブレコーダの映像を用いたメモリアル動画キュレーション手法について述べる．4 章にて “観光地らしさ” を構成する代表語の抽出結果を用いて，尺度を明らかする方法について述べるとともに，5 章にて各尺度のスコアの推定を行う CNN モデルの構築と評価を行う．最後に 6 章にて本稿をまとめる．

2. 関連研究

2.1 経路案内のための動画キュレーション

観光支援システムでは，歩行経路や自動車の運転経路を正しく伝えることが必要である．単純な経路案内を行うシステムとして地図が存在するが，ユーザによっては地図だけでは，経路を理解することが困難なこともある．そこで，経路案内にキュレーション動画を活用することで，経路を理解することを容易にする手法が提案されている [3], [10]. これらの手法では，直進などの経路案内に必要のない部分は再生速度を速くし，右左折などの経路案内において重要な部分では通常の速度で再生を行うようにキュレーションを行う．歩行経路案内を目的とした研究 [10] では，1 人称映像を用いてヒストグラム差分から風景の切り替わりを検出し，可変フレームレート方式でキュレーションを行っている．また，運転経路案内を目的とした研究 [3] では，ドライブレコーダの映像から LucasKanade アルゴリズムに基づいたオプティカルフローの計算を行い，右左折を検出することで動画のキュレーションを行う．

2.2 メモリアル動画作成のための動画キュレーション

近年，観光終了後に思い出の写真やメモリアル動画をソーシャルメディアに投稿するユーザが増加している．動画編集ソフトを用いたメモリアル動画作成は，動画編集スキルを有したユーザ以外は作成が難しい問題がある．そのため，動画編集スキルがない人でも容易にメモリアル動画を自動的に作成するシステム [1] が提案されている．しかし，従来のシステムの多くは観光において大部分を占める移動中の様子が欠落しているため，移動中に発生した思い出深いシーンなどを反映させることができない．そこで，ドライブレコーダ映像を用いて移動を含めた観光全体のメモリアル動画自動キュレーション手法 [6] が提案されている．

片山らの手法 [6] では，観光経路を撮影したドライブレコーダの動画を 3 秒ごとのセグメントに分割を行い，セグメントごとに観光メモリアル動画として必要であるか評価する重要度推定し，メモリアル動画キュレーションを行う．重要度推定モデルとして，3 秒ごとのセグメント動画から 3 枚の画像フレームを抽出し，CityScapes [11] と BDD100k [12] を用いて学習された DeepLabv3 [13] を用いてカテゴリ別占有率を算出した結果と YOLOv3 [14] を用いて得られるランドマーク情報などを特徴量として “沖縄らしさ” を推定するモデルを構築する．この際，“沖縄らしさ” などのスコアについては，Yahoo!クラウドソーシングを用いて，セグメント動画を見た際の “沖縄らしさ” を評価してもらい，正解ラベルとして利用している．しかし，“沖縄らしさ” という曖昧な表現は人によって感じ方が異なるため，各ユーザが想定した “沖縄らしさ” に基づくメモリアル経路動画を生成できない可能性がある．

2.3 本研究の立ち位置

本研究では，2.2 節で提案されている重要度スコアに基づく動画キュレーション手法をベースとしつつ，解釈に差が生じやすい尺度の代わりに，人による解釈の差が生じにくい尺度の組み合わせによって “観光地らしさ” を表現することでメモリアル動画キュレーションの品質向上を目指す．

我々はこれまでに，沖縄県をドライブ観光した際のドライブレコーダの動画を用いて，メモリアル経路動画キュレーションに用いるための尺度を導くため，“沖縄らしさ” を構成する印象語のクラスタ抽出および代表語の選定を，感性工学分野で用いられてい手順 [8] に基づき実施している [9].

本稿では，これらの代表語と因子分析を用いて，“沖縄らしさ” を構成する尺度を明らかにする．その後，それぞれの尺度のスコアを推定する CNN モデルの構築を行い，それぞれの尺度スコアは推定可能かどうかについて，評価・考察する．

3. 提案システム

提案手法の概要を図 1 に示す．提案手法は，まずドライブレコーダの動画に対してセグメント分割を行う[*1]．その

[*1] ドライブレコーダの動画をセグメント動画に分割する際には，多くの人々の視覚情報が 3 秒間で安定期に入る特性を考慮し [15], 3 秒間ごとに分割した．なお，3 秒は時速 60 km の車から見て 50 m 先に存在する物体がフレームアウトするまでのおおよその時間であるため，車外の風景の変化を確認するのに十分な時間で

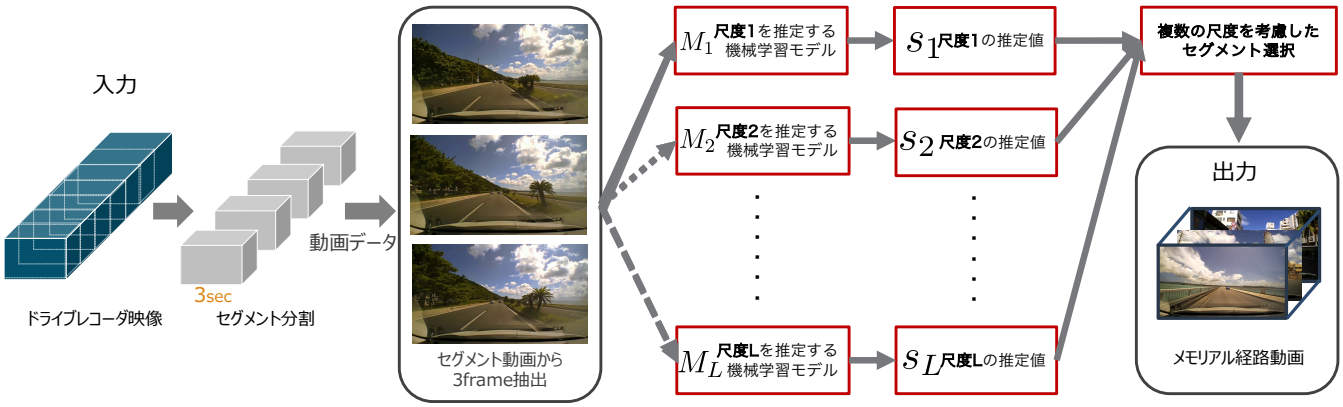

図 1　システムの概略図

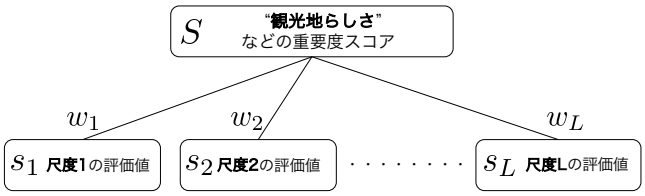

図 2　"観光地らしさ" を構成する尺度

後，セグメント分割した 3 秒の動画の最初，中間，最後のフレームを抽出する．その後，"観光地らしさ" 推定モデルにこれらの画像を入力し，各尺度の推定値を取得する．得られた各尺度の評価値を組み合わせ，セグメント選択することで "観光地らしさ" に基づくメモリアル経路動画を生成する．

このセグメント選択に用いる重要度スコア（すなわち，"観光地らしさ"）は，**図 2** に示すように，構成する複数の尺度の組み合わせから計算される．本研究では，"観光地らしさ"S を，尺度の評価値 s_l を組み合わせることで以下のように定義することとする．

$$S = \frac{1}{L}\sum_{l=1}^{L} s_l w_l \qquad (1)$$

ここで，L は S を構成する尺度の総数，w_l は尺度の評価値の重みを示す．

4.　"観光地らしさ" を構成する尺度の選定

提案システムを実現するためには，"観光地らしさ" が，人が観光地から得る印象のどのような組み合わせによって表現できるのかを明らかにする必要がある．そこで本章では，"観光地らしさ" を構成する尺度の導出方法について述べる．

4.1　導出方法の概要

メモリアル経路動画キュレーションに用いるための尺度を導くため，"観光地らしさ" を構成する尺度の導出を，感性工学分野で用いられている次の手順 [8] に基づき検討する：(1) ドライブレコーダの動画を見た際の印象語の収集する，(2) それぞれの印象語に対して，印象語間の意味空間上の距離を算出する心理実験を行い，得られた距離を基に階層クラスター分析を行うことで，"観光地らしさ" を表現する印象語のクラスタを抽出し，代表語を選定する，(3) 得られた代表語と因子分析を用いて "観光地らしさ" を構成する尺度を明らかにする．これまでに我々は，沖縄のドライブレコーダの動画を用いて (1) と (2) を実施し，クラスタ抽出によって得られた各クラスタの代表語を**表 1** に示す [9]．

4.2　"観光地らしさ" を構成する尺度の導出

"観光地らしさ" を構成する尺度の導出するために，沖縄のドライブレコーダの動画を見た際の表 1 に示す代表語にいついてどのように感じるか 7 段階（1. 非常に当てはまる，2. 当てはまる，3. やや当てはまる，4. どちらとも言えない，5. やや当てはまらない，6. 当てはまらない，7. 非常に当てはまらない）で評価させるタスクを Yahoo!クラウドソーシングを用いて実施した．

本実験で使用した動画は，沖縄県で撮影されたドライブレコーダの動画から生成した 50 個の 3 秒のセグメント動画を用いた．セグメント動画では，様々な印象語が収集できるように**図 3** に示すように様々な場面のドライブレコーダの動画を使用した．1 つの動画に対して，10 人が動画を見た際に提示された 3 個の代表語についてどのように感じるか評価を行なった．また，チェック設問を用意し，チェック設問が正しく回答されていない場合には，回答者としての信頼がないとして，その人のデータは全て破棄した．

その後，クラウドソーシングによって得られた 10 名分の各動画に対する評価項目ごとの平均に対して，最尤法とプロマックス回転による因子分析を行なった．因子数については，固有値が 1 以上であることを基準とした結果，因子数は 7 が適切であると判断された．しかし，尺度数が増加するとキュレーションシステムをユーザが利用する際に，

ある．

表 1　“沖縄らしさ” を構成する代表語

クラスタ 1	クラスタ 2	クラスタ 3	クラスタ 4	クラスタ 5	クラスタ 6	クラスタ 7	クラスタ 8
晴れやかな 明るい 清々しい	気持ちいい 心地良い	のんびりした ゆったりとした 穏やかな	海沿いの 海岸沿いの 優雅な	郊外の 閑静な 日常の	平穏な 落ち着いた 静かな	開放的な 開けた 広大な	暖かい 生き生きした 南国な
クラスタ 9	クラスタ 10	クラスタ 11	クラスタ 12	クラスタ 13	クラスタ 14	クラスタ 15	クラスタ 16
眩しい 常夏の 新しい	高速道路の 夜の 道路沿いの	異国情緒な 異国の アジア風の	賑やかな 暑い	ごみごみした 渋滞した 息苦しい	渋滞の 雑踏の 窮屈な	都会の 街中の	暗い 狭い 暗がりの

図 3　沖縄で撮影されたドライブレコーダの動画から作成したセグメント動画

どの尺度を重視するか選択する項目が増えることでシステムのユーザビリティが低下する可能性が考えられる．そのため，本実験では因子数を 6 個とした．

因子分析によって得られた各因子の正の相関と負の相関を示す代表語の因子負荷量の絶対値の大きさの上位 3 つを**表 2** に示す．表 2 より，第 1 因子は “閑静な” や “静かな” などの人が少ないことを示す語に正の相関があり，“渋滞した” や “ごみごみした” などの人が多い様子を示す語に負の相関があることから第 1 因子は “都会感” を尺度とし，第 2 因子は，“明るい” や “晴れやかな” などの朝や昼の時間帯の様子を示す語に正の相関があり，“暗がりの” や “夜の” などの夜の時間帯の様子を示す語に負の相関があることから第 2 因子は “明るさ” を尺度とした．次に，第 3 因子では，“異国情緒な” や “南国な” などの日常にはない異国間な様子を示す語に強い相関があることから第 3 因子は “異国感” を尺度とし，第 4 因子では，“海沿いの” や “海岸沿いの” などの景色の様子を示す語に強い相関があることから第 3 因子は “美景感” を尺度とした．第 5 因子は “高速道路の” などの爽快感のある語に正の相関があり，“街中の” や “のんびりした” などの爽快感のない語に負の相関があることから第 5 因子は “爽快感” を尺度とし，第 6 因子は，“開けた” や “広大な” などの景色が開けた様子を示す語に正の相関があり，“狭い” や “窮屈な” などの景色が開けていない様子を示す語に負の相関があることから第 6 因子は “開放感” を尺度とした．

因子分析の結果より，“観光地らしさ” を構成する 6 つの尺度として “都会感”，“明るさ”，“異国感”，“美景感”，“爽快感”，“開放感” が導出された．

5.　“観光地らしさ” 推定モデルの構築

本章では，4 章にて導出された “観光地らしさ” を構成する 6 つの尺度について，ドライブレコーダから得られる車載動画データを入力としてスコアを推定するモデルの構築と評価を行う．

5.1　データセット

本実験では，2020 年の 7 月 11 日から 14 日の 4 日間に撮影されたドライブレコーダの映像を用いて実験を行う．ドライブレコーダの映像を 3 秒のセグメント動画に分割すると 1 日目は，459 個，2 日目は，831 個，3 日目は，1244 個，4 日目は，569 個生成され計 3103 個の動画をデータセットとした．正解ラベルは Yahoo！クラウドソーシングを用いて動画を見た際に “観光地らしさ” を構成する尺度がどのように感じるか評価するタスクを実施することで収集した．また，チェック設問を用意し，チェック設問が正しく回答されていない場合には，回答者としての信頼がないとして，その人のデータは全て廃棄した．“都会感”，“明るさ”，“爽快感”，“開放感” は，4 章において尺度を決定する際に正の相関と負の相関の両方を考慮して尺度を決定したため 7 段階で評価した．“都会感” については 1 に近づくほど都会感があること，7 に近づくほど田舎感があることを示す．“明るさ” については 1 に近づくほど明るいこと，7 に近づくほど暗いことを示す．“爽快感” については 1 に近づくほど爽快感があること，7 に近づくほど混雑感があることを示す．“開放感” については 1 に近づくほど開放感があること，7 に近づくほど窮屈感があることを示す．一方，“異国感” と “美景感” は，4 章において尺度を決定する際に正の相関のみを考慮して尺度を決定したため 4 段階で評価したため，1 に近づくほど “異国感”，“美景感” が**ある**こと，4 に近づくほどそれらが**ない**ことを示す．1 つの動画に対して，5 人が動画を見た際に感じた尺度の評価値の平均を正解ラベルとした．

表 2　各因子の正の相関と負の相関を示す代表語の因子負荷量の絶対値の大きさの上位 3 つ

	第 1 因子	負荷量	第 2 因子	負荷量	第 3 因子	負荷量	第 4 因子	負荷量	第 5 因子	負荷量	第 6 因子	負荷量
正の相関	閑静な	0.98	明るい	0.87	異国情緒な	0.91	海沿いの	0.92	高速道路の	0.86	開けた	0.47
	静かな	0.91	晴れやかな	0.82	異国の	0.88	海岸沿いの	0.91	新しい	0.50	道路沿いの	0.43
	のんびりした	0.87	暖かい	0.51	南国な	0.74	広大な	0.35	眩しい	0.37	広大な	0.42
負の相関	渋滞した	-0.99	暗がりの	-0.97	高速道路の	-0.43	日常の	-0.35	日常の	-0.42	狭い	-0.81
	渋滞の	-0.91	暗い	-0.74	郊外の	-0.27	賑やかな	-0.24	街中の	-0.36	窮屈な	-0.29
	ごみごみした	-0.88	夜の	-0.72	渋滞した	-0.14	狭い	-0.16	のんびりした	-0.33	アジア風の	-0.26

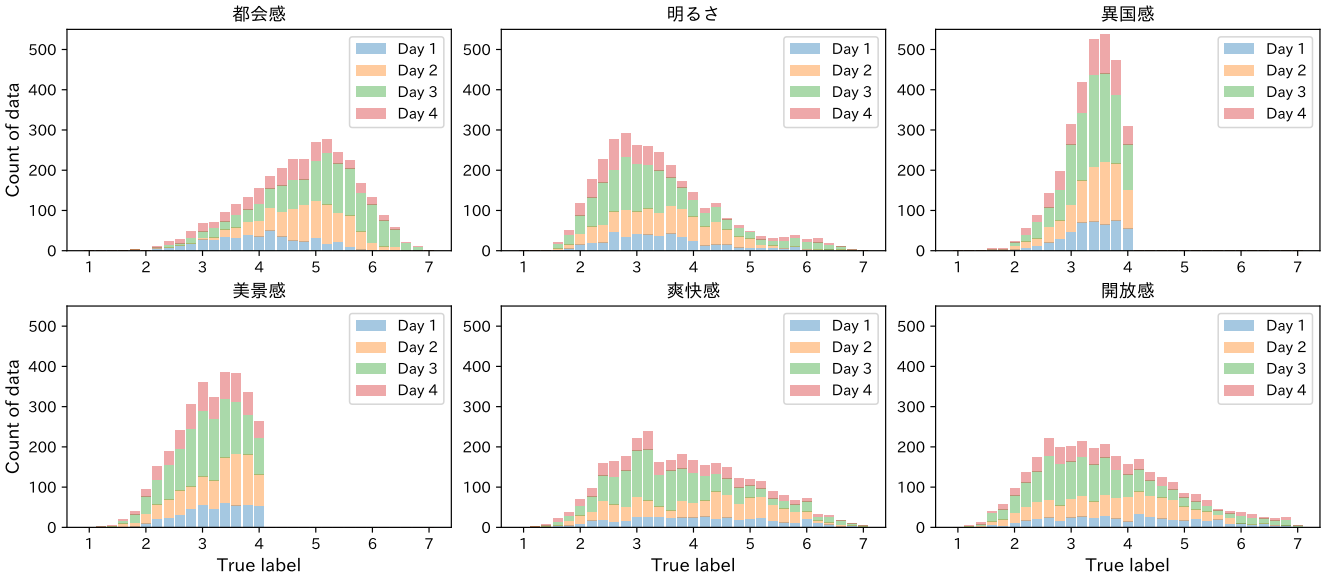

図 4　各セッションにおける正解ラベルとデータ数の関係

表 3　6 つの尺度スコア推定モデルの評価結果（MAE）

都会感	明るさ	異国感	美景感	爽快感	開放感	平均
0.73	0.52	0.36	0.40	0.59	0.60	0.53

5.2　実験条件

本実験では，画像認識の分野で有効性が示されている ResNet [16] を用いて，観光地らしさを構成する尺度のスコアを推定する．モデルは，PyTorch が提供している ResNet50 の学習済みモデルを用いる．入力には，**図 1** に示すように，ドライブレコーダの映像を 3 秒ごとにセグメント動画に分割し，セグメント動画から抽出した 3 枚の画像を用いた．入力に用いる 3 枚の画像は 3 秒のセグメント動画の最初，中間，最後のフレームの画像を利用する．機械学習モデルに入力する際には，RGB が格納されている次元方向に 3 枚の画像データを結合した．出力として，各尺度のスコアを得るために出力クラス数を 1 に変更し，回帰モデルとして扱った．汎化性能を評価するために，1 日分のデータを単位として Leave-one-day-out 交差検証を実施した．各セッションにおける正解ラベルとデータ数の関係は**図 4** に示す．学習時の損失関数には，Mean Absolute Error（MAE），最適化手法には，学習率 0.01 の Stochastic Gradient Descen（SGD）を利用し，ミニバッチ数は 32，エポック数は 20 で学習した．

5.3　実験結果

MAE の評価結果を**表 3**，CNN を用いた尺度スコアを推定した結果を**図 5** に示す．表 3 より，全ての尺度に関する MAE の平均は約 0.53 であることから，比較的に良好な結果が得られたといえる．さらに図 5 より，“明るさ”，“美景感”，“爽快感”，“開放感” においては，正解ラベルのスコアの増加に伴って推定スコアも増加していることから，傾向を捉えられていることが分かる．一方で，“都会感”，“異国感” については，推定値が中央（3〜4）付近に集中するような結果となっており，正しく推定できていないといえる．この結果の原因として，Leave-one-day-out 交差検証を行う際のテストデータの分割の際に生じたデータの偏りの影響が考えられる．図 4 に示すように，“都会感” は他のデータと比較すると正解ラベルの 6 付近のデータは Day 3 に多く含まれており，Day 1，2，4 を学習データに用いたモデルでは，正解ラベルの 6 付近のデータを正しく学習できていないことから，推定できていないと考えられる．また，“異国感” においては，他の尺度と比較して 3.5 付近に極端にデータが集まっており，低いスコアを学習できなったと考えられる．この問題を解決するためには，学習データが少ない部分については，今後より多くのデータを収集することや，Data Augmentation を適用することでデータ不均衡問題を解決する必要があると考えられる．

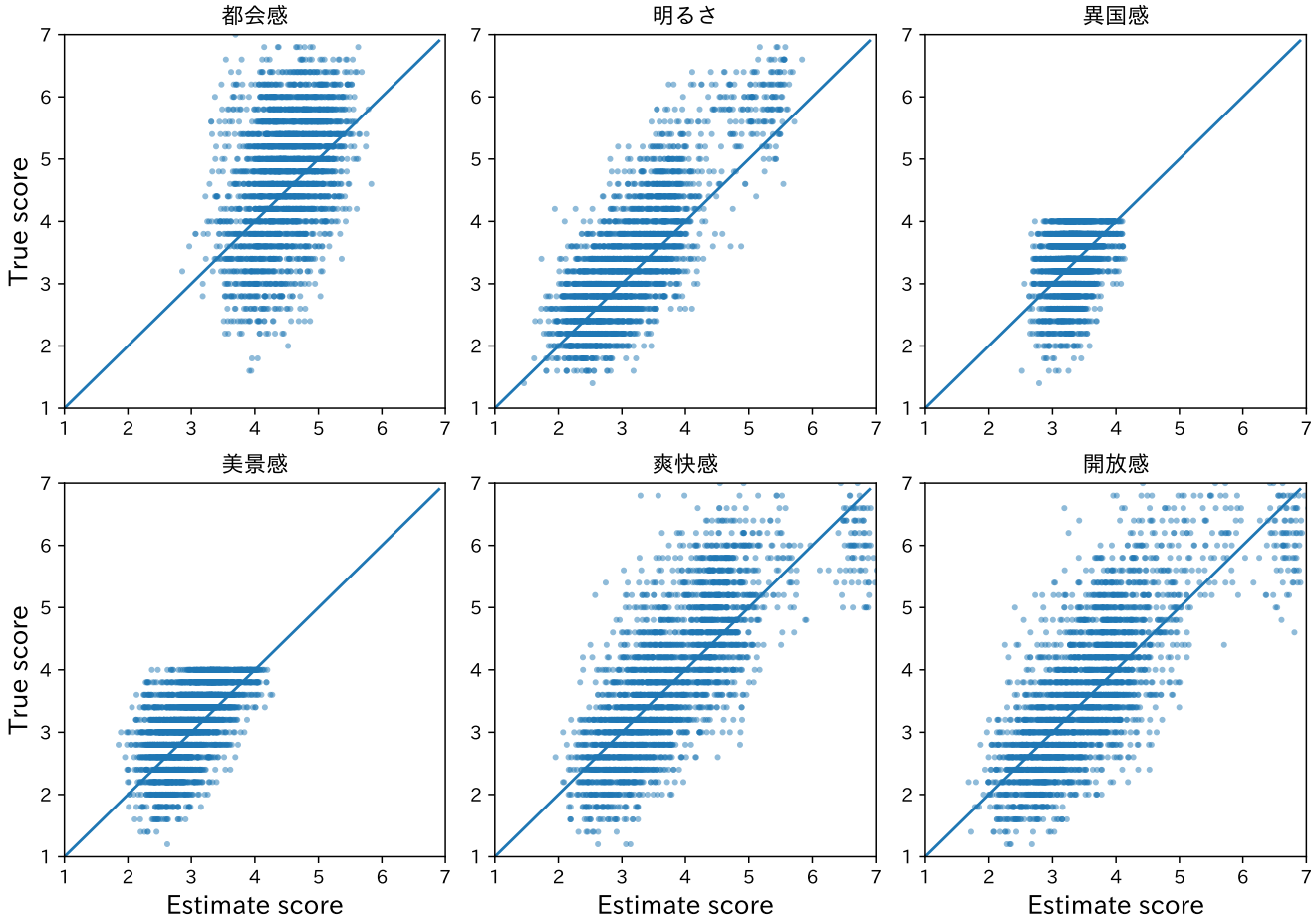

図 5　6 つの尺度スコア推定モデルの評価結果と真値との関係

6.　まとめ

本研究では，既存研究で用いられてきた解釈に差が生じやすい尺度の代わりに，人による解釈の差が生じにくい汎用的な尺度の組み合わせによって “観光地らしさ” を表現することで，メモリアル動画キュレーションの品質を向上させることを目的としている．本稿では，沖縄を題材として，“観光地らしさ” を構成する 6 つの尺度を明らかにするとともに，各尺度を推定する CNN モデルの構築を行った．実験結果より，全ての尺度の MAE の平均は約 0.53 であり，比較的に良好な結果が得られた．一方で，データ不均衡問題により，一部の尺度においては十分な性能を示すモデルを構築できていないことも明らかとなったため，今後より多くのデータを収集することや，Data Augmentation などを活用することで性能向上を図るとともに汎用的なモデルの実現を目指す．

謝辞　本研究の一部は，株式会社デンソーテンの協力，JST さきがけ（JPMJPR2039），JSPS 科研費（JP22H03648）の助成を受けて行われたものです．

参考文献

[1] Real Networks. https://jp.real.com/realtimes/.

[2] Kazuhito Takenaka, Takashi Bando, Shogo Nagasaka, and Tadahiro Taniguchi. Drive video summarization based on double articulation structure of driving behavior. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM ’12, p. 1169–1172, New York, NY, USA, 2012.

[3] 佐藤享憲, 成沢淳史, 柳井啓司. シーン文字認識と自己動作分類を用いた車載動画の要約. 画像の認識・理解シンポジウム (MIRU), 2015.

[4] Shigeya Morishita, Shogo Maenaka, Daichi Nagata, Morihiko Tamai, Keiichi Yasumoto, Toshinobu Fukukura, and Keita Sato. Sakurasensor: Quasi-realtime cherry-lined roads detection through participatory video sensing by cars. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’15, p. 695–705, New York, NY, USA, 2015.

[5] 片山洋平, 平野陽大, 諏訪博彦, 伍洋 and 安本慶一. 観光メモリアル動画のための車載動画キュレーションアルゴリズムの検討. 第 28 回マルチメディア通信と分散処理ワークショップ論文集（DPSWS2020）, 2020.

[6] 片山洋平, 諏訪博彦, 安本慶一. dash-cum: ドライブレコーダを用いたメモリアル経路動画キュレーション. 第 27 回社会情報システム学シンポジウム (ISS 2021), 2021.

[7] Akihiro Matsuda, Tomokazu Matsui, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. A Method for Detecting Street Parking Using Dashboard Camera Videos.

Sensors and Materials, Vol. 33, No. 1, pp. 17–34, 2021.

[8] 飛谷謙介, 松本達也, 谿雄祐, 藤井宏樹, 長田典子. 素肌の質感表現における印象と物理特性の関係性のモデル化. 映像情報メディア学会誌, Vol. 71, No. 11, pp. J259–J268, 2017.

[9] 河中昌樹, 松田裕貴, 諏訪博彦, 安本慶一. ドライブレコーダを用いたメモリアル経路動画キュレーションのための”観光地らしさ” の再考. 第 30 回マルチメディア通信と分散処理ワークショップ論文集（DPSWS2022）, 2022.

[10] Yuki Kanaya, Shogo Kawanaka, Hirohiko Suwa, Yutaka Arakawa, and Keiichi Yasumoto. Automatic route video summarization based on image analysis for intuitive touristic experience. *Sensors and Materials*, Vol. 32, No. 2, pp. 599–610, 2020.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2633–2642, 2020.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[14] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[15] Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, Vol. 105, No. 38, pp. 14325–14329, 2008.

[16] 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)*, pp. 770–778, Los Alamitos, CA, USA, jun 2016.