第 25 回日本バーチャルリアリティ学会大会論文集（2020 年 9 月）

街探索のための Movie Map の自動構築と評価

Automatic Construction and Evaluation of Movie Map for Exploring Cities

杉本直樹 1)，相澤清晴 1)
Naoki SUGIMOTO and Kiyoharu AIZAWA

1) 東京大学 情報理工学系研究科（〒 113-8656 東京都文京区本郷 7-3-1, sugimoto@hal.t.u-tokyo.ac.jp）

概要: 本研究では，全天球映像と地図を連携させたムービーマップの構築と評価について論じる．対象地域を交差点でのターンを含まず道に沿って撮影した全天球映像群と，それらの始点と終点の座標情報のみを入力とし，交差点同士を結ぶ映像群と交差点を曲がる映像群からなる映像データベースを自動的に構築するシステムを提案し，それを活用した探検型インターフェースの実装，ユーザースタディによる実用性の評価を行った．

キーワード： Movie Map，全天球映像，自己位置推定

1. はじめに

自身が実際に訪問することなく特定の地域を体験する，いわば仮想的な探検を行うタスクにおいて動画像データを結び付けた地図アプリケーションは非常に有用である．Google Street View (GSV) [1] はその代表的なインターフェースの 1 つであり，全天球画像が位置的に紐づけられた地図上をユーザーは移動し，自身の場所に対応する画像を閲覧する事であたかも自分がその地域を探検しているかのような体験を得られる．

しかしながら，GSV による体験は没入感の面で完全であるとは言えない．紐づけられたデータが画像である以上，ユーザーは対応する画像の存在する場所を離散的に移動する事しかできず，エリアによっては自由度の高い探検が不可能となってしまう．

この問題を解決する手段として，地図上に紐づけるデータを映像とする手法が考えられる．それを部分的に実現したのが，Mapillary [2] である．Mapillary では，ユーザーが投稿した位置情報付きの動画像データが地図上に集約されており，ユーザーはそれらを再生し切り替えながらメディアの位置情報をもとに地図上を移動する．地域によっては密なデータ量を得ることが出来ており，映像を再生する事でその地域の様子を詳細に知ることが出来る．その一方で，映像に紐づけられた位置情報は GPS による不正確なものであるため，動画間の遷移により移動する方向を切り替えることは困難である．曲がりたい場所で曲がれないはユーザーの体験を著しく損なう．さらに Mapillary には，ユーザー投稿型のアプリケーションであることによるデータの粗密性の課題も見受けられる．

本研究では，全方位映像を利用した，連続的な自由度の高い地図探索インターフェースの実現を目的とする．その際，ユーザーに連続性のある体験を提供するために不可欠な密な映像の空間情報を保持したデータベースを，より少ない撮影コストと労力で構築する技術的課題が発生する．本稿では，特定区画の各道について双方向に移動撮影した全方位映像を用いて，空間情報を紐づけた映像データベースの自動的な構築，そしてそれらを用いた街中の自由な探索を可能にする探索インターフェースの実装について論じる．

2. 関連研究

2.1 Movie Map

全方位映像を再生し，ユーザーの操作によってその切り替えを行う事で仮想的な探検体験を提供するインターフェースは，Movie Map として Lippman [3] に提唱された． [3] では，車両に備え付けられたカメラにより車道全体と交差点のパノラマ画像が 10 フィートごとに撮影された．交差点をそれぞれの方向へと曲がる映像はそれぞれ別々に光学ディスクへと記録され，ユーザーがタッチスクリーンにおいて進行方向を選択する事でディスクの切り替えが行われ連続的な映像の再生を続けることが出来る．これによりユーザーはまるで自分が車両を運転しているかのような体験を得ることが出来る．しかしながら，アスペンを対象として構築されたこのシステムは，物理的な構築コストが大きく，それぞれの映像の結びつきについても手作業で登録を行うため膨大な労力を要した．この例の後にも映像地図を構築する試みはいくつか行われているものの [4] [5]，エリア内の経路が 1 つの建造物を中心に囲むように配置されていることを仮定しているなど条件が課されている場合があり，対象地域について Movie Map を少ないコストで構築するシステムは未だ存在しないと言える．

一方で，全方位画像を地図上に紐づけることで同じようにユーザーに探検の体験を提供するアプリケーションとしては，GSV が 2007 年にサービスを開始した．現在では世界各地にその範囲を拡大し，道案内やバーチャルツアーと

図 1: Movie Map システムフロー

いったタスクにおいて広く利用されている．ただし，[3] が Movie Map において切れ目のない視覚的な情報によりユーザーが実際にその地域を運転しているような体験を得ると主張しているように，GSV は静止画をメディアとして利用していることから離散的な移動や情報の不連続性を含み没入感の高い体験を得られるインターフェースとは言えない．現状，GSV における 1 枚の画像の位置情報は GPS と比較して正確であり屋内でも動作していることから映像も利用した位置情報の特定が行われていると思われる．仮に GSV におけるメディアを映像にそのまま置き換えた場合，その位置情報の付加には多大なコストがかかり，それらを GPS の利用などにより自動化したとしても交差点における別経路への連続的な遷移を実現する方法については自明でない．

2.2 Visual SLAM

Visual Simultaneously Localization and Mapping (vSLAM) は，単眼カメラ映像を入力として 3 次元復元とカメラ位置の推定を同時に行う技術である [6] [7] [8]．vSLAM は，フレーム内に検出された画像特徴点を基に，連続したフレームの相対的なカメラの向きと位置の変化の最適化を行う．その際，検出された特徴点の 3 次元的位置情報は保存されるため，3 次元空間上に映像から推定される地図を復元する事が出来る．そのため，一度 3 次元地図を構築すれば，その地図内に含まれる画像や映像を入力した際にその位置が地図内のどこに対応しているかを求めることが出来，この処理はローカライゼーションと呼ばれる．ただし，この処理は広く疎な空間について行う事は計算コストの面で現実的ではない．数十，数百という映像が存在し，それらが全く別の場所を通りほんの一部のみが交差するという一般的な Movie Map の構築において現れる状況を考えると，広い 3 次元空間上に多量に検出される特徴点全てとの成功することの無いマッピングを続けることは，無駄な演算の多さと並列化が出来ない事から現代での実現は難しいと言える．

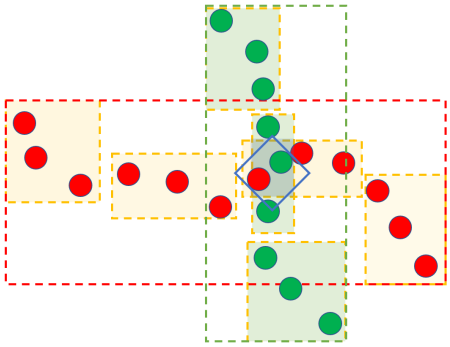

図 2: 交差点の検出

本研究では，OpenVSLAM [9] を用いて，収集した各道の全方位映像についてその相対的なカメラ姿勢の変化の推定を映像別に独立に行った．さらに，1 つの映像に対して 2 つずつ付与した基準座標情報を基に共通の空間にマッピングを行う．これは並列性の担保とともに，比較的短い経路について vSLAM を適用する事で vSLAM の課題であるスケールドリフト [10] の影響を小さく抑える事にも貢献する．

3. 提案手法

本研究の提案は，道ごとに撮影を行った全方位映像を用いて，最低限の付加情報を伴い適用する事で自動的に空間情報を保持した映像のデータベースを構築する Movie Map システムと，構築したデータベースを用いたユーザーが自由度の高い連続した地域の探検を可能にする探索インターフェースの実装の 2 つの要素から成る．その全体像を図 1 に示す．

3.1 映像データベース構築

まず収集した全方位映像それぞれについて，共通する空間上における始点と終点の座標情報を付与する．その上で，各映像について独立に vSLAM を適用する．その後，付与した座標情報を基に推定された相対的なカメラ位置と向きの推移をスケーリングし，全ての映像の座標情報を共通の座標空間上にマッピングする．

次に，マッピングした座標情報に基づいて映像同士が構成する交差点の検出を行う．検出の様子を図 2 に示す．1 つの映像ペアについて，それぞれの映像のカメラ位置の軌跡全体を覆う矩形を求め，矩形同士の重なりを判定する．重なりが判定された場合，より細かく分割した軌跡について同様に矩形の設定と重なりの検出を行っていき，最終的に閾値を超える近さを持つ映像間のフレームのペアを見つけ出す．

以上の検出手段は vSLAM の推定結果を付与した基準座標をもとにマッピングした結果を用いており，様々な要因によりある程度の誤差を含むため検出された交差点フレームが本当に 2 つの映像間で最も画像上近しいものである可能性は低い．そこで本手法では，画像特徴点を利用した校正を行う．対象となる 2 映像について，検出された交差点フレームの前後のいくつかのキーフレーム画像について，そ

図 3: 探索インターフェースの全体像（本郷キャンパス）

図 4: インターフェース上での交差点のターンシークエンス

の特徴点を検出し，全てのフレーム画像のペアについて特徴点のマッチングを行いより多くのマッチングが成功した画像のペアを正しい交差点フレームのペアとして決定する．特徴点の検出には SPHORB [11] を利用し，画像の回転の向きに関わらず同じ構造の場所からは同様の特徴量の抽出が可能としている．全ての映像のペアについて交差点の検出と校正を終えた後，交差点を基準とした各映像の分割を行い，どの交差点のはざまの映像であるかという情報とともにデータベースに保存する．

最後に，交差点における各経路の映像同士の遷移映像の合成を行う．インターフェース上でのユーザーの連続的な移動体験の実現には，交差点における自然な映像の移り変わりによるターンが不可欠である．本手法では，交差点が検出された映像のペアについて，その検出された交差点フレームを利用した合成映像を遷移映像として利用する．具体的には，ターン前後の映像の交差点フレーム画像を，vSLAM のカメラの向きの推定結果を用いて向きを揃えた上で重み付きの重ね合わせを行い，同時に回転させていくことで切り替えを補完するようなターンの映像を合成する．詳細を式 1 に示す．

$$F(i) = (1 - \frac{i}{N}) \cdot Rotate(I_{before}, \frac{i}{N}q) + \frac{i}{N} \cdot Rotate(I_{after}, (1 - \frac{i}{N})q^{-1}) \quad (1)$$

式 1 において，$F(i)$ が合成映像の i フレームを，I が切り替え前後の交差点フレーム画像を表し，q は 2 画像間の相対的な回転角である．N は合成映像の総フレーム数であり，これにより切り替え前の画像が徐々に回転しながら切り替え後の画像に切り替わっていく映像が合成される．

3.2 Movie Map 探索インターフェース

前述した位置関係を保持する映像データベースを利用した，自由度の高い連続的なメディア体験を提供する探索インターフェースを提案する．図 3 に東京大学本郷キャンパスを対象として構築したインターフェースの例を示す．

ユーザーはインターフェース上で交差点を基準として映像の再生を行う．どの交差点から交差点に向けて移動するかを指定する形で移動が開始され，交差点においては次なる進行方向を映像内にポップアップする矢印を選択する形で決定する．再生された全方位映像は映像をドラッグする事で周囲を見回しながら進行させられる他，再生箇所の変更や速度変更といった一般的な映像再生インターフェースとしての機能を備えている．交差点においてターンを行う場合には，図 4 に示すように前節で述べた合成映像を遷移映像として挿入し，異なる経路映像の再生へとつなげる．これにより，ユーザーは途切れの無い移動をエリア内で行うことが出来る．

また拡張的な機能として，再生されている映像の一部分を指定して画像を埋め込むことで，図 3 の再生画面右に表示されているような仮想的な看板や目印を追加する事も可能である．

4. 実験

京都駅周辺のエリアについて Movie Map の構築を行い，与えられた区画の中で特定のランドマークを発見するタスクを GSV と我々の提案するインターフェースを用いて行った．図 5 に探索範囲を赤枠で示した京都駅周辺の地図とランドマークを表す画像を示す．被験者は狭い探索範囲を示す地図とランドマークの手掛かりとなる画像を与えられ探索を行い，各インターフェースの操作性と没入感について，

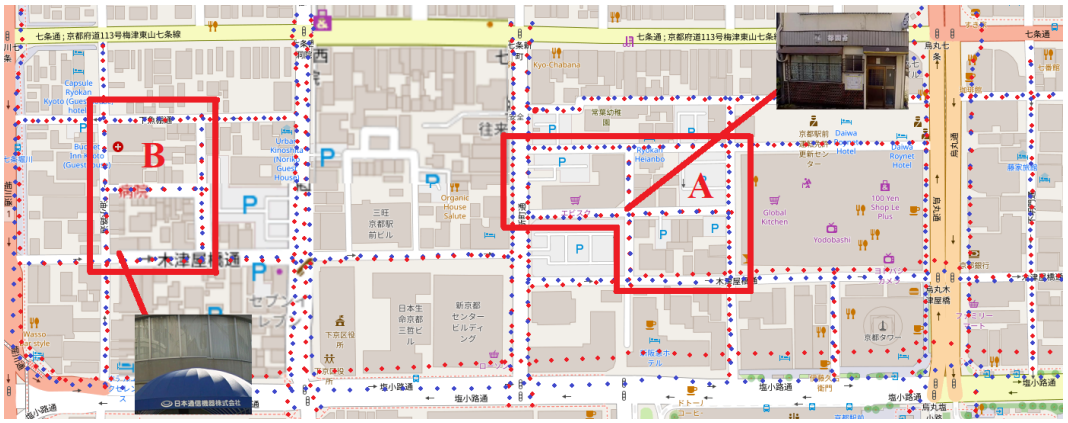

図 5: 探索タスクを行う範囲と対象となるランドマーク画像

Question	Average	
	GSV	提案手法
ユーザビリティ	3.81	3.44
探検の没入感	3.19	**4.38**

** : $p < 0.01$

表 1: インターフェースの操作性，没入感に関する評価

5 を最大，1 を最低とする主観評価を行う．

表 1 に結果を示す．タスクは簡単であるため，どちらのインターフェースによっても達成されるものであったが，その没入感には有意な差が表れ，提案するインターフェースが GSV と比較して操作性を損なうことなくより高い没入感を得られることが示された．

5. まとめ

本研究では，全方位映像を分析しデータベースを自動的に構築するシステム，そしてそれを利用した探索インターフェースからなる Movie Map システムを設計，構築した．実験においては，交差点フレームの回転とブレンディングによる合成された遷移映像の挿入により交差点での連続的な移動体験が実現したこと，そしてそれを利用した提案する探索インターフェースが GSV と比較して操作性を損なうことなく高い没入感を提供できることをユーザースタディによって示した．

謝辞 本研究の一部は，(株)VTEC 研究所の支援を受けた．

参考文献

[1] Google. Google street view. `https://www.google.co.jp/intl/ja/streetview/`, 2005.

[2] Mapillary AB. Mapillary. `https://www.mapillary.com/`, 2014.

[3] Andrew Lippman. Movie-maps: An application of the optical videodisc to computer graphics. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '80, pp. 32–42, New York, NY, USA, 1980. ACM.

[4] M. Naimark. *Field Recording Studies*. MIT Press, Cambredge, 1996.

[5] Michael Naimark. A 3d moviemap and a 3d panorama. *Proc. SPIE*, Vol. 3012, , 01 2004.

[6] P. Lothe, S. Bourgeois, F. Dekeyser, E. Royer, and M. Dhome. Towards geographical referencing of monocular slam reconstruction using 3d city models: Application to real-time accurate vision-based localization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2882–2889, June 2009.

[7] M. Tamaazousti, V. Gay-Bellile, S. N. Collette, S. Bourgeois, and M. Dhome. Nonlinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment. In *CVPR 2011*, pp. 3073–3080, June 2011.

[8] G. Balamurugan, J. Valarmathi, and V. P. S. Naidu. Survey on uav navigation in gps denied environments. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 198–204, Oct 2016.

[9] Shinya Sumikura, Mikiya Shibuya, and Ken Sakurada. Openvslam: a versatile visual slam framework. `https://github.com/xdspacelab/openvslam`, 2019.

[10] Jakob Engel, Thomas Schoeps, and Daniel Cremers. Lsd-slam: large-scale direct monocular slam. Vol. 8690, pp. 1–16, 09 2014.

[11] Qiang Zhao, Wei Feng, Liang Wan, and Jiawan Zhang. Sphorb: A fast and robust binary feature on the sphere. *International Journal of Computer Vision*, Vol. 113, No. 2, pp. 143–159, 2015.