J-033

Web 閲覧中の視線と動画を活用した情報推薦装置の提案

Proposal for information recommendation system using eye gaze and video analysis on web browsing

李　昊聞[1]　山田光穂[1]　石井英里子[2]　星野　祐子[1]
Li Haowen　Yamada Mitsuho　Ishii Eriko　Hoshino Yuko

東海大学情報通信学研究科情報通信学専攻[1]　鹿児島県立短期大学[2]
Department of information and Telecommunication Engineering,
Graduate School of information and Telecommunication Engineering, Tokai University[1]
Kagoshima Prefectural College[2]

1. はじめに

近年，情報通信技術の発達により，インターネットを利用する人数や回数が増加する傾向がある．事前調査[1]によると，2024 年におけるインターネットの普及率は 84.9%であり，「情報探し」を目的としてインターネットを利用する人は 69.4%に達している．これは，日本人の若者から高齢者までの全年齢層にわたるインターネットの普及を示していると同時に，Web 上での情報探索の重要性が高まっていることも示している．しかし，膨大な情報の中から自分に必要な情報を見つけ出すことは困難である．この問題を解決するために，我々の研究室では観光情報推薦システムに無意識の情報選択の一つである視線情報を導入している．我々の研究室が開発した既存システムでは Web サイト閲覧中のユーザー視線情報をもとに注視文章と静止画像を取得し，ユーザーの嗜好を抽出し，観光関連の検索キーワードとして推薦している[2]．

しかし情報通信技術の発展により Web 観光サイトは文章と静止画だけではなく，観光スポットの「説明用動画」も載せていることも多い．既存システムでは静止画像と文章による嗜好抽出には対応しているが，動画も含めている観光サイトでは対応できていない．この問題を解決するために既存システムへの動画情報の追加を提案する．

2. 既存システム推薦手法

2.1 既存システムの処理

既存システムでは，視線検出装置に Tobii 社のアイトラッカーTobii pro spark[3]を使用し，検索キーワードの推薦には閲覧した Web ページ内の文章と画像を使用している.先ず Web ページを閲覧中のユーザーの視線情報を取得し，熟読している部分の文章と画像を取得し，解析してユーザーの嗜好を抽出する.抽出した内容に基づいてユーザーが興味を示した内容と類似性の高い新たな情報を得るための検索キーワードを推薦する．

2.2 文章・画像からのユーザー嗜好抽出

ユーザーが注視している部分は主に①文章②静止画像の二つのパターンに分けることができる．①文章に注目している場合は視線が停留している箇所の文字を抽出する．②静止画像に注目している場合は画像を GPT - 4 - vision - preview に入力し，画像のキャプション(説明文)を生成する．抽出したテキストとキャプションを GPT-3.5 Turbo(今後の動作検証では 4.0 を使用)に入力し，要約文を作成する．その要約文をもとに，本文関連情報，共起関連情報，類似観光スポットを三つの情報を推薦する．図 1 に文章および画像の抽出処理の流れを示す．

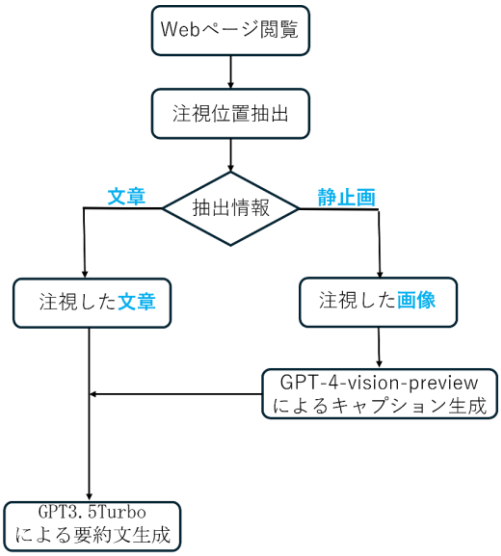

図 1 ：文章および画像の抽出処理の流れ

3. 提案手法

3.1 動画の切り抜き

本研究の提案は図 2 で示すように既存システムの中に動画情報を組み込むことである．Runway ML[4]のような既存動画 AI では，一本の動画に対して，ある程度の動画内容を解析し，その内容を基づいて動画のタイトルを生成することができる．逆に文章から新しい動画を生成することも可能である．しかし，これはあくまで動画全体に対しての要約である．本研究のポイントは動画の中で，ユーザーが特に注視していた部分に着目し，その部分を切り抜くことである．

3.2 キーフレーム抽出

まず動画を静止画像に変換する必要がある．最近の動画は画素数 2k～4k，フレームレートは 30fps，60fps が主流となっている．今回は画素数 2k，フレームレート 60fps の 1 分間の動画に対して静止画像を生成する．1 分間の動画では 60＊60＊1＝3600 枚のフレーム画像が得られる．これらのフレーム全てを静止画として処理するには時間がかかりすぎるため，キーフレームの抽出を行う．

動画中でユーザーの視線が最も長く停留している部分をユーザーが最も興味を示した部分と見なし，フレーム抽出を行う．具体的には，動画内で視線が最も長く停留していた区間の最初のフレーム（第１）と次のフレーム（第２）を抽出し，画素値を比較する．あらかじめ決めた大きな変化がなければ更に次のフレーム（第３）を抽出し，第１フレームと画素値を比較する．大きな違いがあらわれるまで，次のフレームの抽出と比較を繰り返す．大きな違いが出たフレームをキーフレームとする．画素値が大きく変化する要因には次の二つが考えられる．一つはユーザー視線の移動，二つ目は動画シーンの切り替えである．

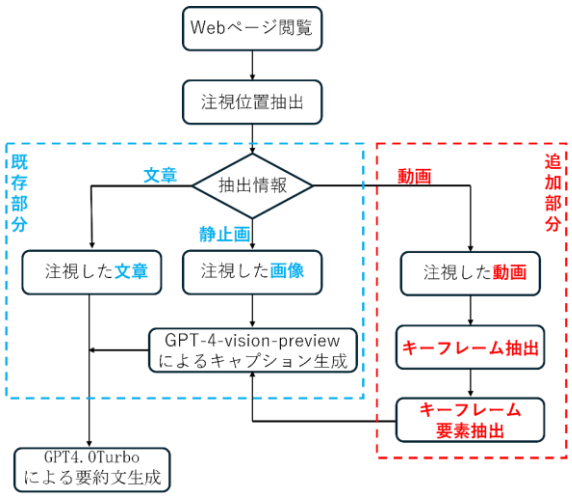

図 2：追加システムの流れ

一つ目はキーフレームを得た後に，さらにキーフレームと比較したフレームが 120 枚（2 秒）を超えていれば，比較されたキーフレームは興味を持っている部分と判断する．なお閾値 120 枚（2 秒）は赤松らの研究[6]をもとに決定した．次にキーフレームを静止画として切り抜きキャプションを生成する．以上の処理を 1 分の動画内で繰り返し行う．このルールを当てはめると図 3 のようなシーンの切り替えがある動画において東京タワーを見ていた時間が 2 秒未満，富士山を見ていた時間が 2 秒以上であれば，キーフレームは 1 枚で，キーワードは「富士山」となる．逆に，東京タワーが 2 秒以上，富士山が 2 秒未満の場合は，キーフレームは 1 枚でキーワードは「東京タワー」となる．また，両方とも 2 秒以上見ていた場合は，キーフレームは 2 枚となり，それぞれのキーワードは「東京タワー」と「富士山」となる．反対に，両方とも 2 秒未満であれば，キーフレームは抽出されない．

しかし，一つの問題がある．それは二つ目のように画面のシーンの切り替わりがある時，全てのシーンの流れる時間が 2 秒未満の場合である．図 3 のように都市の風景が 1.5 秒ほど続いた後，映像のテーマが急に変わり，「森の中の猿」のシーンへ切り替わり，「森の中の猿」が 1.5 秒ほど流れたとする．この場合，全てのシーンで 2 秒以上の注目が集まることはない．そこで，同じ風景の中で最も注目された部分のフレームを抽出する．例えば，都市の風景の中で東京タワーを 1 秒，富士山を 0.5 秒見た後にシーンが切り替わった場合，より長く注目された「東京タワー」をキーワードとするキーフレームを抽出する．

3.3 キーフレーム要素の抽出

既存システムでは静止画像を GPT - 4 - vision - preview に入力し，静止画像に関する説明文を生成している．しかし動画の中には図 3 の「東京タワー」と「富士山」のように複数のオブジェクトが存在していることが多い．そのため，キーフレームを抽出するだけではなく，静止画像中のユーザーが注目したオブジェクトを抽出することが必要である．そこで，抽出したキーフレームを OCR や Google Lens に入力し，オブジェクトを抽出行い，更に視線の位置に合わせて視線が停留しているオブジェクトだけを出力することを検討している．

図 3：シーンの切り替え

4. まとめ

本研究では，既存の観光情報推薦システムを改良し，動画情報の追加し観光情報推薦に活用することを提案した．現在はまだシステム全体を通した稼働ができていない段階であり，以下 3 つの課題がある．すなわち①各シーンの長さがが 2 秒未満であった場合，キーフレーム取得方法の再検討②キーフレーム中の注視しているオブジェクトだけの抽出③システム精度検証の方法である．

本研究により，ユーザーの嗜好がより正確に判別できることが期待される．

謝辞

本研究の一部は日本学術振興会科研費 JP23K11635 助成を受けております．

参考文献

[1] 総務省．総務省|令和 6 年版情報通信白書|インターネット .https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nd21b120.html
[2] 森大河，山田光穂，石井英里子，星野祐子：視線情報活用した Web 検索システムの開発，パーソナルコンピューター利用技術学会論文誌，JJPCATS，Vol16，No.2
[3] Tobii . 明瞳孔法と暗瞳孔法 : https:connect.tobii.com/s/article/What-is-dark-and-bright-pupil-tracking?language=en_US.
[4] Runway ML ：https://platform.vidu.cn/
[5] 片岡裕雄,”コンピュータビジョンによる動画認識”, VISION Vol. 31, No. 1, 1–4, 2019
[6] Kazuaki Akamatsu, Tomohiro Nishino, Yoichi Miyawaki, "Spatiotemporal bias of the human gaze toward hierarchical visual features during natural scene viewing," Scientific Reports, 13, 8104