Vol.2022-HCI-197 No.46
2022/3/16

視線情報を考慮した機械学習に基づく一人称視点映像の自動要約手法

濱岡啓太[1]　河野恭之[1]

概要：本研究では，長時間にわたる一人称視点映像の高速閲覧を目的とし，視線情報を考慮した機械学習に基づく一人称視点映像の自動要約手法を提案する．一人称視点映像は日常生活やスポーツなどの記録を残す手段の一種であり，ウェアラブルカメラの小型化及び普及に伴い一般に広く浸透した．しかし，ライフログ等のために常時撮影された一人称視点映像は長時間にわたる．そのためユーザにとって有益でないシーンを含むケースが多く，映像の閲覧に時間を要する問題点がある．本研究では視線情報に着目し，視線移動量，瞳孔径，瞼の開き具合を基にした機械学習による予測モデルを構築することで各シーンに対する興味度の推定を行う．興味度が高いシーンは通常速度，低いシーンは高速再生することで入力映像を要約する．評価実験の結果，提案手法により生成した要約映像はユーザの興味や潜在的な意識を反映していることを確認した．本システムを用いて一人称視点映像を要約することでユーザの行動認識や視覚的な日記の作成だけでなく，記憶障害のある患者の支援等多くのアプリケーションに応用可能である．

キーワード：映像要約，視線計測，機械学習

An Automatic Summarization Method for First-Person-View Video Based on Machine Learning Considering Gaze Information

KEITA HAMAOKA[†1]　YASUYUKI KONO[†1]

Abstract: We propose an automatic summarization method for first person view video using interest estimation by machine learning based on gaze information. First-person view video is a method for keeping records in daily life, sports, etc. , and it became widespread as wearable cameras have become smaller. Since a wearable camera can shoot first-person view video with both hands free, it can record the user's natural actions. Many of first-person view video are long, and often include scenes which are not useful to users. There is a problem that it takes a long time to view the movie. Our proposal is the method for summarizing long first-person view video. There are a lot of study focusing on video summarization. For example, Higuchi, et al. summarized first-person view video based on four cues that correspond to the basic user actions of body movement, stillness, hand movement, and human interaction. The user sets the importance of four cues, and the scenes with high importance are reflected in the video after summarization. The contents of the input video are not taken into consideration because the cues are fixed to four. This research proposes the video summarization system employing gaze tracking and machine learning. Because gaze is useful for knowing user's intention and interest, our approach reflects the user's interest and potential consciousness in the video by summarizing the video with the above information. In the previous study, we show first-person-view video to the users using a head mounted display with a gaze measurement function and obtain the amount of change in the user's gaze direction vector, pupil diameter, and eyelid opening condition. The user enters whether or not he/she is interested in each scene by keystroke when viewing the video. We create dataset based on the interest level and gaze information. Preprocessing such as smoothing, removal of missing values, and normalization is performed on the dataset. Based on the pre-processed dataset, we construct a prediction model for estimating the user's level of interest through machine learning. In main process, the user shoot first-person-view video using a head mounted display with a gaze-measuring function, and obtain the user's gaze information and camera images. We estimate the degree of interest for each scene by using the prediction model constructed in the previous study. In the generated summary video, important scenes are played back at normal speed, and other scenes are played back at high speed. As a result of conducting an evaluation experiment, it became clear that our system is useful for fast viewing of videos and videos summarized videos reflect the user's interest. Our system is applicable to many applications such as behavior recognition, to create visual diary, and to support for patients with memory impairment.

Keywords: Video Summary, Gaze Measurement, Machine Learning

1. 序論

1.1 はじめに

本研究では，視線情報を考慮した機械学習に基づく一人称視点映像の自動要約システムを開発する．一人称視点映像は日常生活やスポーツなどの記録を残す手段の一種であり，ウェアラブルカメラの小型化及び普及に伴い一般に広く浸透した．ウェアラブルカメラは一人称視点映像を両手が空いた状態で撮影可能であり，撮影時の負担が小さくユーザの自然な行動を記録可能である．これらの利点から一人称視点映像はアスリートの動きの解析[1]や伝統芸能の継承[2]等，様々な場面で活用されている．しかし，ライフログ等のために常時撮影された一人称視点映像は長時間にわたるため，ユーザにとって有益でないシーンを含むケー

1 関西学院大学
Kwansei Gakuin University

スが多い．そのため映像の閲覧に時間を要する問題点がある．これらの問題点を解決するために，様々な映像要約の研究が行われている. 例えば, 字幕付きの動画を対象とし，字幕のない箇所は高速再生，字幕のある箇所は通常速度で再生することで内容を把握しつつ短時間での鑑賞を可能にする Cinemagazer[3]や入力映像の内容をキーフレームの重要度に応じてサイズを変更し，マンガ形式で表示する Manga[4]が開発された．本研究では一人称視点映像を対象とし，視線計測と機械学習を用いた自動要約システムを開発する．本システムは視線計測機器により抽出したユーザの視線情報を基に機械学習を用いて各シーンの興味度を推定するため，要約後の映像にユーザの興味や潜在的な意識を反映可能である．よって，ユーザは長時間にわたる一人称視点映像を短時間かつ興味度が高いシーンのみを閲覧可能になる．本システムを用いて一人称視点映像を要約することでユーザの行動認識や視覚的な日記の作成だけでなく, 記憶障害のある患者の支援等多くのアプリケーションに応用可能である．

1.2 関連研究

1.2.1 一人称視点映像の要約

粥川ら[5]は, 一人称視点映像中に現れる様々な物体を手がかりに利用し，ユーザの意図に応じて各シーンの再生速度を変化させることで映像を要約するシステムを開発した. ユーザは映像中の様々な物体の重要度を設定する．重要度を高く設定した物体の映るシーンを通常速度，それ以外のシーンを高速で再生する．しかし，この手法ではユーザが各物体に対して重要度を設定する必要があり，一人称視点映像を自動的に要約できない．また，映像に現れた様々な物体から任意の物体を指定することで映像を要約するため, 撮影者の興味や潜在的な意識を要約後の映像に反映できない．Lu ら[6]は，前後のフレームの特徴点を比較しシーンの境界を判断することで映像を要約するシステムを開発した．しかし，この手法では入力された映像の内容から重要シーンを判断するため，ユーザの興味があるシーンを排除する可能性がある．本研究では，視線情報を基に各シーンに対する興味度を推定するため，ユーザの興味や潜在的な意識を要約後の映像に反映可能である．

Higuchi ら[7]は，ユーザの移動，静止，手の動作，人物との対話という基本的な行動に対応した 4 つの手がかりを基に要約映像の生成を行った．ユーザは 4 つの手がかりに対して重要度を設定し，要約後の映像には重要度が高いシーンを反映させる．しかし，手がかりが 4 つに固定されているため，入力映像の内容を考慮していないという問題点が挙げられる．本研究では，一人称視点映像に対するユーザの視線情報を基に興味度を推定するため，映像の内容を考慮した要約映像の生成が可能である．

筆者ら[8]は, 物体に対する注視時間を基にした一人称視点映像の自動要約システムを開発した．物体検出により抽出した物体領域及び視線計測機器により抽出した注視点を基に各物体に対する注視時間を算出する．それらが閾値を超えた場合，重要シーンに設定した．評価実験の結果，一定の有用性が確認できたものの物体が存在しないシーンに対応していないという課題が残った．本研究では，視線情報を基にシーン重要度を判断するため，物体が存在しないシーンにも対応可能である．

1.2.2 視線情報と心的状況

人は五感を通して周囲の様々な情報を取り入れており, 特に視覚は周囲の状況を理解する上で極めて重要な器官である．また，人は興味を持った対象に自然に視線を向けることから，視線はユーザの意図・興味・関心を知る上で有用である[9]．マーケティング等の分野の研究において視線移動を基に消費者行動を理解しようとする取り組みは古くから試みられてきた[10]．近年は視線計測機器の精度の向上や低価格化により様々な分野での研究が拡大しており，視線情報と心的状況を解明するための様々な取り組みが行われている．

視線移動量に関して人間の心理的状態と相関があると考えられている[11].大須賀ら[12]はユーザに画像 2 枚のどちらの方が好みかを質問し，その際の視線の停留時間を計測した．その結果，高く評価した画像に対する視線の停留時間は評価されなかった画像に比べて長いことが明らかとなり，評価の違いを視線の停留時間により計測できる可能性が得られた. 桜栄ら[13]は「人の歩行中のシーンに対する印象」が街路構成要素の何に起因するのかを明らかにするために，アンケートによる主観的印象評価と注視行動情報の分析を行なった．その結果，主観的印象評価と注視行動情報には相関があることを確認した．

瞳孔面積は輝度によって変化し, 輝度が高くなるにしたがって瞳孔面積は小さくなる[14]. しかし, 瞳孔面積は輝度だけでなく精神状況にも相関があると考えられる．Hess ら[15]は物理的な光量が一定の場合でもユーザが興味・関心のある対象を見ると瞳孔面積が広がり，無い対象では狭まると報告している. また, 大山ら[16]は授業評価への視線計測装置の応用を目的とし，授業映像を視聴している際の瞳孔径変化の分析を行なった．その結果，授業への関心が低くなるシーンでは瞳孔径が漸減することを確認した．

山田[17]は瞬目について髄意性瞬目，反射性瞬目及び自発性瞬目の 3 種類が存在し，随意性瞬目はウインクのように意図的に瞼を閉じる行動，反射性瞬目は外的刺激によって誘発される瞬目，自発性瞬目は随意性瞬目や反射性瞬目のように瞬目発生の原因が明白ではないのに生じる瞬目のことであると述べている．被験者が興味の異なる 3 種類の課題を行なっている際の瞬目率を分析した結果，興味・関心の高い課題ほど自発性瞬目率が減少することを確認した．これらの結果から，自発性瞬目は人間の感情が不快のときに促進され，快のときに抑制されると報告している．津田

ら[18]は各被験者の評定結果に基づいて最も興味の高いものと最も低い映像を選出し，主観的興味の程度の異なる２つの条件下で映像を提示した際の主観的興味と瞬目率の関係を計測した．その結果， 主観的興味度が高い映像を視聴した際の瞬目率は主観的興味度が低い映像に比べて有意に低下していることを確認した．

2. 研究概要

2.1 目的とアプローチ

本研究の目的は長時間にわたる一人称視点映像の要約である．ウェアラブルカメラはハンズフリーで撮影可能であり，撮影時の負担が小さくユーザの自然な行動を記録可能である．その反面，長時間にわたる一人称視点映像を閲覧する際はユーザにとって冗長なシーンを含むため時間を要するという課題が挙げられる．本研究ではそれらの課題を解決する一人称視点映像の自動要約システムを開発する．長時間にわたる一人称視点映像を要約するためには映像内のどのシーンがユーザにとって有益であるのかを判断する必要がある．本研究ではユーザの興味・関心に着目し，要約後の映像にはユーザの興味・関心の高いシーンを反映させる自動要約システムを開発する．それらの要件を実現させるために視線情報を考慮した機械学習による予測モデルを構築することで，各シーンに対する興味度の推定を可能にする．ユーザは予め短時間の一人称視点映像を閲覧し，各シーンに対する興味度をラベル付けする．その際の視線情報及び興味度を基に予測モデルを構築する．1.2.2 項で述べたように視線移動量，瞳孔径，瞬目といった視線情報とユーザの興味・関心には相関がある．そのため，それらの視線情報及び興味度を基にした予測モデルによって未知のデータに対する興味度の推定が可能になると考えられる．また，興味・関心は人によって異なるため，個人差を考慮する必要がある．そのために，事前調査で各ユーザの視線情報を及び興味度を取得し，各ユーザに対する予測モデルを構築する．これにより本システムで生成した要約映像は各ユーザの興味や潜在的な意識を反映可能であると考えられる．

2.2 システム概要

図 1 に本システムの処理の流れを示す．事前調査において各ユーザはウェアラブルカメラを用いて一人称視点映像を撮影する．撮影した一人称視点映像を視線計測機能付き HMD によってユーザに提示し，その際の視線情報を取得する.また,ユーザは撮影した一人称視点映像をする際に，各シーンに対する興味度のラベル付けを行う．ユーザが入力した興味度及び視線情報を基にデータセットを作成する．データセットに対して前処理を行い，それらを基にユーザの各シーンに対する興味度を推定する予測モデルを構築する．本処理において各ユーザはウェアラブルカメラを用いて一人称視点映像を撮影し，視線計測機能付き HMD を用いて撮影した一人称視点映像を閲覧する．その際に視線計測機能付き HMD により視線情報を取得し，前処理を行う．前処理を行った視線情報を事前調査で構築した各ユーザに対する予測モデルに入力することで各シーンの興味度を推定する．興味度が高いシーンを通常再生，低いシーンを高速再生した要約映像を生成する．

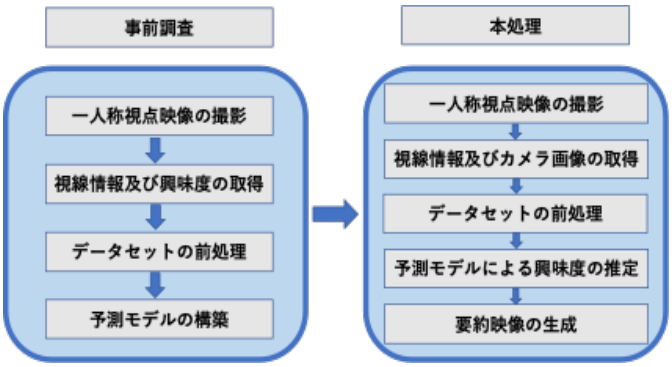

図 1　本システムの処理の流れ

3. 視線情報を考慮した機械学習による予測モデル

本章では,本システムの事前調査にあたる一人称視点映像の撮影から予測モデルの構築を行うまでの過程を述べる．ユーザはウェアラブルカメラにより一人称視点映像を撮影し，視線計測機能付き HMD を装着した状態で撮影した映像を閲覧する．その際にユーザは各シーンに対する興味度のラベル付けを行う．視線計測機能付き HMD により取得した視線情報及び興味度を基にデータセットを作成する．視線計測機能付き HMD により取得した視線情報は欠損値や外れ値を含むため，それらに対して前処理を行う．前処理を行ったデータセットを基に教師あり機械学習による予測モデルを構築する．また，予備実験において分類モデルの代表的な 11 種類の手法を用いて精度評価を行うことで，本システムに適した分類モデルの手法を検討する．

3.1 一人称視点映像の撮影

当初は HTC 社の VIVE Wireless Adapter[21]を Vive Pro Eye に装着し，一人称視点映像を撮影する予定であった．VIVE Wireless Adapter は VIVE Pro Eye のサードパーティ製品であり，Vive Pro Eye 向けの公式無線化キットである．そのため，VIVE Pro Eye を装着した状態においてもユーザの行動範囲を制限することなく，自然なライフログ映像を撮影可能になる．しかし，日本国内では電波法の規制により当デバイスの販売が認可されていない[22]．有線の状態ではユーザの行動範囲が制限されるため，自然なライフログ映像を撮影することができない．それらの課題を解決するために本研究では，予めウェアラブルカメラにより一人称視点映像を撮影し，視線計測機能付き HMD に撮影した映像を提示することで，ユーザの視線情報を取得する．

3.2 視線情報及び興味度の取得

視線計測機能付き HMD を装着した状態のユーザに予め撮影した一人称視点映像を提示することで，視線情報を取得する．視線計測機能付き HMD によりタイムスタンプ，視線の起点，視線の方向，瞳孔径，目の開閉具合を取得する．また，ユーザは一人称視点映像を閲覧する際に各シーンに対して興味度のラベル付けを行う．ラベル付けタスクにおいてラベルの種類が増えるとユーザに対する負荷が大きくなる．そのため，本研究では「興味がある」,「興味がない」の 2 種類のラベルに限定することでユーザに対する負荷を軽減させる．また，本システムではユーザが一人称視点映像を閲覧している際に特定のキーボタンを押すことで各シーンに対するラベル付けを行う．これによりユーザは各シーンをその都度停止させてラベル付けを行う必要がなく，興味があるシーンのみ特定のキーボタンを押し続けるだけでラベル付けを行うことが可能になる．また，視線計測機能付き HMD を装着している際は仮想空間をユーザに提示するため,実空間のデバイスの操作が困難になるが，キーボタンの押下のみに限定することでラベル付けタスクの負荷を軽減することができる．

3.3 データセットの前処理

視線計測機能付き HMD は高精度で視線情報を取得できるが欠損値や外れ値を含んでおり,それらが予測モデルの精度の低下に繋がる可能性が考えられる．そのため，取得した視線情報に対して前処理を行う．また，予測モデルの精度向上のためにデータ整形を行う．

3.3.1 リストワイズ法による欠損値の除去

一般的に欠損値への対処法は大きく分けて「欠損値を補完する」,「欠損値をそのまま扱う」,「欠損値を削除する」の３種類がある．本システムでの欠損値については視線計測機能付き HMD による推定精度に依存しており，特定のパターンが存在せずランダムに発生する．そのため，欠損値を補完する単一代入法や多重代入法等は少なからず誤差のあるデータを代入することになるため，予測モデルの精度が低下する可能性がある．また，欠損値が存在している場合では使用できない機械学習モデルが多く存在する．これらの理由から本システムではリストワイズ法を用いて欠損値を削除する．リストワイズ法とは欠損値が含まれるサンプルデータ自体を削除する方法である．欠損値を削除することによりデータセットのサンプル数は減少するが，視線計測機能付き HMD の視線データ出力周波数は 120Hz かつ長時間にわたる一人称視点映像を想定しているため，膨大なサンプル数を取得可能である．そのため，リストワイズ法によってサンプル数が減少した場合でも予測モデルの精度に影響する可能性は低いと考えられる．

3.3.2 視線移動量の算出及び平滑化による外れ値処理

視線移動量を算出するために視線方向ベクトル及び視線の起点座標を基にユーザの注視点を抽出する．また，視線計測機能付き HMD により取得した視線方向ベクトル及び起点座標は外れ値を含むため，抽出した注視点に対して平滑化を行う．本システムでは Manu ら[23]が提案した加重平均を用いた手法により注視点の平滑化を行う．平滑化を行った注視点を基に視線移動量を算出する．現在のフレームと前のフレームの注視点に対して 2 点間の距離の公式を用いることで視線移動量を算出する．

3.3.3 多重共線性の考慮及び正規化

通常，左右の瞼の開き具合及び瞳孔径はほぼ同じである．予測モデルを構築する際の説明変数に左右それぞれの瞼の開き具合及び瞳孔径を含めた場合，多重共線性が起こることで予測精度が低下する可能性がある．多重共線性とは，多変量解析において説明変数間に強い相関がある場合に解析上の計算が不安定になり，予測精度が低くなる現象である．これらを防ぐために左右の瞼の開き具合及び瞳孔径から平均を算出する．

算出した視線移動量, 瞼の開き具合, 瞳孔径はそれぞれデータスケールが異なる．それらをスケーリングせずに予測モデルを構築した場合, 予測精度が低下する可能性がある．そのため，視線移動量，瞼の開き具合，瞳孔径に対して正規化を行う. 正規化とはデータの最大値が 1, 最小値が 0 になるスケーリング手法であり，各特徴量の持つ重みを平等にすることができる．

3.4 予測モデルの構築

各シーンに対するユーザの興味度を推定するために教師あり学習による予測モデルを構築する．教師あり学習とは学習データに正解を与えた状態で学習させる手法であり，連続する値を予測する「回帰」とクラスを識別する「分類」に分けられる．本システムは各シーンに対して興味があるか，ないかを識別する必要があるため，教師あり学習による分類モデルを用いる．予備実験において教師あり学習による分類モデルの代表的な手法である, Logistic Regression, Nearest Neighbors，Linear SVM，Polynomial SVM，RBF SVM，Sigmoid SVM, Decision Tree, Random Forest, AdaBoost, Naive Bayes，Linear Discriminant Analysis，Quadratic Discriminant Analysis，計 11 種類の手法の精度を比較することで実装するモデルを検討する．目的変数は興味度，説明変数は視線移動量，瞼の開き具合，瞳孔径の計３つである．モデルの精度を向上させるためにグリッドサーチ法を用いてパラメータのチューニングを行う．グリッドサーチとはハイパーパラメータを最適化させる手法であり，学習モデルに用いられるハイパーパラメータの組み合わせを調整することで予測モデルの汎化性能を向上させることができる．また，予測モデルの過学習を防ぎ，汎化性能を向上させるために K-分割交差検証（K=8）を行う．K-分割交差検証は，汎化性能を評価する統計的な手法であり，データを K 個に分割してその内の１つを訓練データ，残りの K -1 個を学習データに設定し，モデルの学習を行う．

3.5 予備実験

本システムに実装する分類モデルの検討を目的とし, 教師あり学習による分類モデルの代表的な手法の精度評価を行う.

3.5.1 実験手法

被験者はウェアラブルカメラにより約 10 分間の一人称視点映像を撮影し，視線計測機能付き HMD を用いて撮影した映像を鑑賞する．その際にユーザは興味があるシーンのみ特定のキーボタンを押下することで各シーンに対するラベル付けを行う．視線計測機能付き HMD によって取得した視線情報と各シーンに対する興味度に対して 3.3 節で述べた前処理を行う．それらを基に教師あり学習による分類モデルの代表的な 11 種類の手法でモデル構築を行う. 被験者は３名であり，22 歳の男性２名と 24 歳の男性１名で行った．データ数は約 54,000 個である．

3.5.2 結果と考察

各分類モデルによる予測結果及びグラフを図 2, 3 に示す. 下記の図では訓練データに対する予測精度が高いモデルから順番に並べている．訓練データにおいて全ての手法で予測精度 80%以上, テストデータにおいて全ての手法で予測精度 70%以上かつ最高スコア 82%を確認した．これらの結果から視線情報による興味度推定は本システムにおいて十分有用であると考えられる．訓練データ及びテストデータに対する予測精度の合計が最も高かった分類モデルは Random Forest であったため, 本システムでは Random Forest を実装する. Random Forest は決定木モデルに対してアンサンブル学習を行った手法である[24]．決定木モデルとは段階的にデータを分割することで目的変数に影響する説明変数を明らかにする手法であり，バイアスは小さくなるがバリアンスは大きくなる特徴を持つ．通常，バイアスとバリアンスはトレードオフの関係になっているが，Random Forest はこれらの特徴を持つ決定木モデルを組み合わせてバリアンスを下げることで，過学習を防ぎつつ，高精度な予測を可能にする．これらのアルゴリズムの仕組みから本予備実験においても Random Forest の予測精度が最も高かったと考えられる.

classifier	train	test
Decision Tree	0.999082	0.763313
Random Forest	0.999082	0.782410
AdaBoost	0.924082	0.818193
Nearest Neighbors	0.921735	0.743855
RBF SVM	0.871020	0.800542
Quadratic Discriminant Analysis	0.834286	0.716325
Linear SVM	0.834082	0.757470
Logistic Regression	0.830918	0.763675
Naive Bayes	0.819694	0.710422
Linear Discriminant Analysis	0.813163	0.754819
Polynomial SVM	0.802653	0.783494

図 2　予測結果

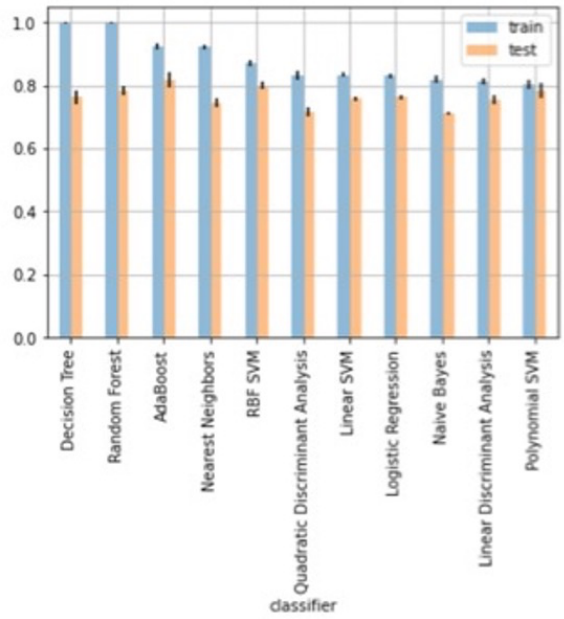

図 3　予測結果のグラフ

4. 要約映像の生成

第 3 章ではユーザが装着した視線計測機能付き HMD により視線情報を抽出し，それらを考慮した機械学習による予測モデルの構築を行った．それらを基に本章では，本処理にあたる一人称視点映像の撮影から要約映像を生成するまでの過程を述べる．ユーザはウェアラブルカメラにより一人称視点映像を撮影し，視線計測機能付き HMD を装着した状態で撮影した映像を閲覧する．視線計測機能付き HMD により取得した視線情報に対して前処理を行い，それらを予測モデルに入力することで，各シーンに対する興味度を推定する．予測モデルによる各シーンに対する興味度を基に要約映像を生成する.

4.1 一人称視点映像の撮影

3.1 節と同様にウェアラブルカメラを用いて一人称視点映像を撮影し，視線計測機能付き HMD に撮影した映像を提示することで，ユーザの視線情報を取得する.

4.2 データセットの前処理

視線計測機能付き HMD は高精度で視線情報を取得できるが欠損値や外れ値を含んでおり,それらが予測モデルの精度の低下に繋がる可能性が考えられる．そのため，取得した視線情報に対して前処理を行った後にデータ整形をすることで予測精度の向上を図る.

4.2.1 線形補完による欠損値処理

第 3 章の予測モデルの構築においては精度の低下を防ぐために，リストワイズ方を用いて欠損値を除去した．本処理では予測モデルを用いて各シーンに対する興味度を推定し要約映像を生成するため, リストワイズ法を用いた場合，ユーザの興味度が高いシーンを排除してしまう可能性がある．そのため，欠損値を補間する必要がある．また，本システムで取得する視線情報は時系列データであるため，前

後のデータに相関関係がある．これらの理由から本システムにおける本処理では,線形補間による欠損値処理を行う．線形補間とは，二点間を直線で結んだときに，その直線上に存在する任意の点を算出する方法であり，計算コストが低いというメリットがある．

4.2.2 視線移動量の算出及び平滑化による外れ値処理

3.3.2 項と同様に，視線移動量を算出するために視線方向ベクトル及び視線の起点座標を基にユーザの注視点を抽出する．また，視線計測機能付き HMD により取得した視線方向ベクトル及び起点座標は外れ値を含むため，抽出した注視点に対して Manu ら[23]による加重平均を用いた平滑化を行う．平滑化を行った現在のフレームと前のフレームの注視点を基に視線移動量を算出する．

4.2.3 多重共線性の考慮及び正規化

3.3.3 項と同様に，多重共線性による予測モデルの精度の低下を防ぐために左右の瞼の開き具合及び瞳孔径から平均を算出する.また,それらはデータスケールが異なるため，正規化を行い各特徴量の持つ重みを平等にすることで，予測精度の向上を図る．

4.3 予測モデルによる興味度の推定

事前調査では,個人によって異なる興味度を要約映像に反映させるために,各ユーザに対する予測モデルを構築した．予測モデルは事前調査においてパラメータチューニングによって高精度で興味を推定できるように最適化されており，本節ではそれらを用いて各シーンに対するユーザの興味度を推定する．前処理を行った視線移動量，瞼の開き具合，瞳孔径を学習済みの予測モデルに入力することで，一人称視点映像の全てのシーンに対する興味度を予測する．

4.4 要約映像の生成

本システムでは興味度が高いシーンを通常速度,低いシーンを高速再生し，一人称視点映像全体の長さを短くすることで，長時間にわたる一人称視点映像の高速閲覧を可能にする．要約映像全体の長さを短くするためには他にも興味度が低いシーンを排除し，興味度が高いシーンのみを繋ぎ合わせる方法が考えられる．しかし，興味度が高いシーンのみを繋ぎ合わせた場合，要約後の映像の持つ情報量が少なくなるという懸念が挙げられる．つまり，興味度が低いシーンを削除した場合，映像の一連の流れを時系列的に解釈することが難しくなる．そのため，本システムでは興味度が高いシーンを通常速度，低いシーンを高速再生した要約映像を生成する．ウェアラブルカメラで撮影した一人称視点映像から各フレームの画像を取得する．各フレームの興味度を基に興味が高いシーンと低いシーンのセグメントに分割する．興味が高いシーンのフレームレートは通常速度，低いシーンは高速に設定し，それらを結合することで要約映像を生成する．

5. 評価実験

本研究では，視線情報を考慮した機械学習による予測モデルを構築し，それらを用いて各シーンに対する興味度を推定することで一人称視点映像を要約するシステムを開発した．本システムの有用性を確認するために，本システムにより生成した要約映像に対する評価実験を行う．また，本システムにより生成した要約映像が各ユーザの興味・関心を反映しているかを検証するために，通常速度で再生するシーンをランダムで設定した要約映像を生成する．それらと本システムにより生成した要約映像の比較評価を行い，アンケートを実施する．

5.1 実験手法

被験者はウェアラブルカメラを用いて約 1 時間の一人称視点映像を撮影する.人によって興味･関心は異なるため，各ユーザに対する予測モデルをそれぞれ構築する必要がある．そのため，各ユーザが撮影した一人称視点映像から約 10 分間を抽出し，それらを基に予測モデルを構築する．ユーザは視線計測機能付き HMD を用いて抽出した映像を閲覧し，各シーンに対する興味度のラベル付けを行う．視線計測機能付き HMD により取得した視線情報及び興味度に対して前処理を行い，それらを用いて Random Forest による予測モデルを構築する．その後，ユーザは視線計測機能付き HMD を装着した状態で予測モデルの構築のために抽出した部分以外（約 50 分）を閲覧する．視線計測機能付き HMD により取得した視線情報に対して前処理を行い，それらを予測モデルに入力することで各シーンに対する興味度を予測し，興味度が高いシーンは通常速度，低いシーンは高速に設定した要約映像を生成する．また，本システムにより生成した要約映像が各ユーザの興味を反映しているかを検証するために通常速度で再生するシーンをランダムで設定した要約映像を生成する．

本システムを用いて生成した要約映像及び通常速度で再生するシーンをランダムに抽出した要約映像を鑑賞した後に，5 段階リッカート尺度を用いたアンケートを集計する．被験者は 20～24 歳の男女計 9 名（男性 7 名，女性 2 名）である．質問内容を図 4 に示す．また，本システムの利点と欠点について自由記述欄を設けた．

	本システムで生成した要約映像について
質問1	要約映像を閲覧する際に疲れなかったか
質問2	映像の高速閲覧に役立つか
質問3	要約映像に興味が反映されていたか
	重要シーンをランダムで抽出した要約映像について
質問4	要約映像に興味が反映されていたか

図 4　質問内容

5.2 結果と考察

評価実験の結果を図 5 に示す．本節では有効数字 3 桁で述べる．質問 1 の疲労度に関するアンケートの結果の平均値は 3.22 であったが，自由記述欄には「本システムにより生成された要約映像は興味があるシーンとないシーンのフレームレートが異なるため，要約後の映像を閲覧する際に疲れた」という回答が寄せられた．本システムは長時間にわたる一人称視点映像を要約するために興味度が高いシーンのフレームレートは通常速度，低いシーンのフレームレートは高速に設定し，映像全体の長さを縮めることで高速閲覧を可能にする．そのため，フレームレートが異なるシーンに切り替わる際，ユーザはストレスを感じる可能性がある．これらを解決するために二つの方法が考えられる．一つ目の方法はフレームレートが異なるシーンを結合する際にフレームレートを段階的に変更する方法である．例えば，フレームレートが通常速度と高速のシーンを結合する際にフレームレートを段階的に変更し，徐々に早くすることでユーザに対する負荷を軽減できると考えられる．しかし，本システムの要約映像に比べて要約後の映像の長さが長くなる欠点が挙げられる．もう一つの方法は興味度が低いシーンを排除し，興味度が高いシーンのみを繋ぎ合わせる方法である．そうすることでフレームレートが異なるシーンの切り替わりにおいてユーザが感じる負荷を無くすことができる．また，本システムに比べて要約映像の全体の長さを短くすることができ，より高速な閲覧を実現可能にする．しかし，興味のないシーンを排除するため，要約後の映像の情報量が少なくなるという懸念が考えられる．これらの理由から今後は本システムの使用用途や目的に適した方法を検討していく必要がある．

質問 2 の有用性に関するアンケートでは平均値 4.11 と高い評価を得ることができた．自由記述欄には「今後，視線計測機能付き HMD の小型化，軽量化によって着用している際のストレスが軽減された場合は，本システムを使用して普段の生活のハイライト映像を記録してみたい」との意見が得られた．本システムで使用した Vive Pro Eye の重量は約 770g であるため，日常的に使用するにはユーザに対する負荷が大きい．しかし，今後，視線計測機能付き HMD の軽量化及び小型化が実現し着用する際の負荷が減ることで，本システムは様々なシーンに活用できると考えられる．本システムにより一人称視点映像を要約することで，ユーザの行動認識や視覚的な日記の作成だけでなく，記憶障害のある患者の支援等多くのアプリケーションに応用可能である．

本システムにより生成した要約映像がユーザの興味を反映しているかを検証するために，通常速度で再生するシーンをランダムに設定した要約映像と比較評価を行う．質問 3,4 の「要約後の映像に興味が反映されていたか」という項目において提案手法は平均値 3.89，通常速度のシーンをランダムで設定した要約映像の平均値は 3.00 であった．また，それらに対して対応のある T 検定を行った結果，有意水準 5%未満で有意差を確認した．これらの結果から，本システムはユーザの興味を反映した要約映像を生成可能であると考えられる．また，自由記述欄では「一人称視点映像の各シーンに対してラベル付けを行う際に興味があるかないかの 2 種類だけでなく，興味度を 5 段階で設定可能にすることでよりシステムの汎用性が上がるのではないか」という意見が得られた．本システムのラベル付けタスクにおいてはユーザに対する負荷を軽減させるために興味があるかないかの 2 種類のラベルで行った．これらのラベリングタスクにおいて各シーンに対する興味度を 5 段階で設定することで，要約後の映像にユーザの興味度をより高精度に反映させることができる．また，興味度の段階に対して各シーンの再生速度を柔軟に変更することで，要約映像の長さをユーザが予め指定した長さに調整可能になる．

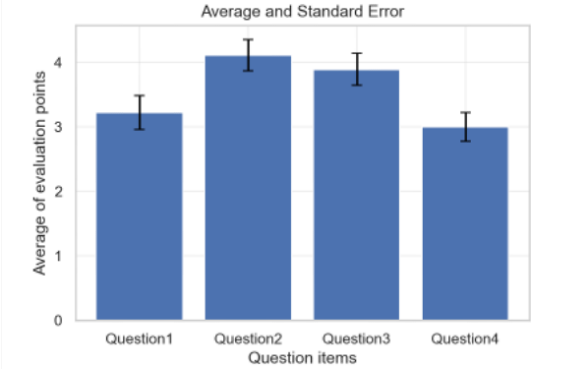

図 5　評価実験の結果

6. 結論と今後の展望

本研究では視線情報を考慮した機械学習に基づく一人称視点映像の自動要約システムを提案した．ユーザはウェアラブルカメラを用いて撮影した一人称視点映像を視線計測機能付き HMD により閲覧し，各シーンに対する興味度のラベル付けを行った．視線計測機能付き HMD により取得した視線情報及び興味度に対して前処理を行い，視線移動量，瞼の開き具合，瞳孔径及び興味度を基に予測モデルを構築した．予測モデルにより各シーンに対する興味度を推定し，興味度が高いシーンは通常速度，低いシーンは高速再生することで一人称視点映像を要約した．

本システムの有用性を検証するために評価実験を行った．その結果，本システムにより生成した要約映像は一人称視点映像の高速閲覧に役立ち，かつユーザの興味を反映していることを確認することができた．しかし，本システムを用いて生成した要約映像が可変フレームレートであるという点に否定的な意見が得られた．本システムは興味度が高いシーンを通常速度，低いシーンを高速再生するため，それらのシーンの切り替わりの際にユーザがストレスを感じる可能性がある．そのため，入力映像から興味度が低いシ

ーンを排除し，興味度が高いシーンのみを繋ぎ合わせるなど，システムの使用目的や用途に適した方法を検討する必要がある．

今後の展望に，ラベリングタスクにおけるラベルの種類の拡張を検討している．本システムではユーザに対する負荷を軽減するために興味があるかないかの２種類のラベルによりラベリングを行った．しかし，興味度を５段階で設定することで，要約後の映像にユーザの興味度をより高精度で反映させることができる．さらに，興味度の段階に対して各シーンの再生速度を柔軟に変更することで，要約映像の長さをユーザが予め指定した長さに調整可能になる．

参考文献

[1] 木村聡樹，三上弾．“打者は打席で何をしているのか？打撃パフォーマンス分析に向けたバーチャルリアリティの活用”，日本神経回路学会誌，vol.24，No.3，pp.109-115，2017.
[2] 檜山淳，土山祐介，宮下真理子，江渕栄貫，関正純，広瀬通考．“一人称視点映像からの多感覚追体験による伝統技能教示支援”，日本バーチャルリアリティ学会論文誌，vol.16，No.4，pp.505-514，2011.
[3] K. Kurihara. “CinemaGazer: a System for Watching Videos at Very High Speed”, Proc. AVI'12, pp.108-115, 2012.
[4] Manga, https://www.fujixerox.com/eng/company/technology/communication/multimedia/manga.html （参照 2022/01/03）.
[5] 粥川青汰，樋口啓太，中村優文，米谷竜，佐藤洋一，森島繁生．“一人称視点映像の高速閲覧に有効なキューの自動生成手法”，Workshop on Interactive Systems and Software 2017, 2017.
[6] Z. Lu and K. Grauman. “Story-driven summarization for egocentric video”, Proc. CVPR’13, pp. 2714–2721, 2013.
[7] K. Higuchi, R. Yonetani, and Y. Sato. “EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines” Proc. CHI’17, pp.6536–6546, 2017.
[8] K. Hamaoka, Y. Kono, “Automatic Summarization Method for First-Person-View Video Based on Object Gaze Time”, Proc. IHSED’20, pp.39-44, 2020.
[9] 沖中大和，満上育久，八木康史．“人の眼球と頭部の協調運動を考慮した視線推定”，情報処理学会研究報告，2016-CVIM-202（18）, pp.1-8, 2016.
[10] 里村卓也．“視線計測による消費者行動の理解”，オペレーションズ・リサーチ，Vol.62, No.12, pp.775-781, 2017.
[11] 金井典彦，小宮一三，百瀬桂子，“マルチメディア通信におけるヒューマンウェアの研究（2）-視線移動と心理要因の関係に関する基礎検討"，神奈川工科大学研究報告 B，理工学編，vol.22, pp.59-62, 1998.
[12] 大須賀智洋，田中元志，新山喜嗣，井上浩，“食品画像を用いた好み評価時の視線停留時間に関する実験的検討”，計測自動制御学会論文集，Vol.49, No.9, pp.880-886, 2013.
[13] 桜栄翔大，中澤篤志．“歩行時におけるシーンの主観評価と注視行動の関係”，情報処理学会研究報告，2020-CVIM-222（29），pp.1-7, 2020.
[14] 浅野樹美，安池一貴，中山実，清水康敬，“輝度変化に対する瞳孔面積変化モデル”，電子情報通信学会論文誌 A, Vol.J77-A, No.5, pp.794-801, 1994.
[15] Hess E. H. “Attitude and Pupil Size” Scientific American, 212（4）, pp.46-54, 1965.
[16] 大山貴紀，金子格，小野文孝，曽根順治，花村剛，“瞳孔径による授業評価”，第 10 回情報科学技術シンポジウム，2011.
[17] 山田富美雄，“瞬目による感性の評価”，心理学評論，Vol.45, No.1, 2002.
[18] 津田兼六，鈴木直人．“主観的興味が瞬目率と体動の生起頻度に及ぼす影響”，Japanese Journal of Physiological Psychology and Psychophysiology, vol.8, no.1, pp.31-37, 1990.
[19] HTC Vive Pro Eye, https://www.vive.com/jp/product/vive-pro-eye/overview/（参照 2022/01/24）
[20] CHINO PC-1, http://chinon.jp/pc-1/ （参照 2022/01/24）
[21] VIVE Wireless Adaptor, https://www.vive.com/us/accessory/wireless-adapter/（参照 2022/01/24）
[22] 総務省, https://www.soumu.go.jp/main_sosiki/joho_tsusin/policyreports/joho_tsusin/02kiban14_04000664.html（参照 2022/01/24）
[23] K. Manu, J. Klingner, R. Puranik, T. Winograd, “Improving the accuracy of gaze input for interaction”, Proc. ETRA’08, p.65, 2008.
[24] L. Breiman, “Random Forests”, Machine Learning, Vol.45, No.1, pp.5-32, 2001.