修士論文

視線特徴と画像特徴に基づくマルチメディアコンテンツの選好推定

東京大学大学院
情報理工学系研究科
電子情報学専攻

学籍番号 48126410

尾崎安範

指導教員　佐藤 洋一 教授

2014 年 2 月 6 日

内容梗概

本論文の目的はマルチメディアコンテンツの選好を推定する方法を実現することである．この目的を達成するアプローチとして，視線と画像の情報とそれに関連付けられた選好の情報から機械学習を用いて機械的に推定する手法を用いた．

本論文の構成は以下のとおりである．まず、本論文の目的とアプローチについてまとめた．その後、本論文に関連する研究について調査した結果を報告し，本研究の意義について議論した．本論文は３つの研究で構成される．１つは並列提示した場合における画像の選好を推定した結果を分析し議論した．ここで選好とは対象物が好きか好きでないかを表す情報とする．実験の結果，画像の情報よりも視線の情報が推定に有効である場合が有意にあると認められた．次に順次提示した場合における画像の選好を推定した結果を分析し議論した．実験の結果，画像のみで推定した時と比べ画像と視線情報を組み合わせて推定したときのほうが精度が平均的に向上した．最後に順次提示した場合における動画の選好を推定した結果を分析し議論した．実験の結果，画像のみで推定した時と比べ画像と視線情報を組み合わせて推定したときのほうが精度が平均的に向上した．本論の末尾に得られた知見から結論を出し，今後の課題について議論した．

以上の内容について，本文では詳細に述べる．

目次

内容梗概

図表目次

図目次

表目次

第1章

序論

人は，言葉だけで他人とコミュニケーションで行なっているわけではなく，身振り手振り，あるいは視線なども利用している．たとえば，衣料品店で服をじっと眺めている人はおそらくその服に興味をもっていることが店員にわかり，これがコミュニケーションのきっかけになることがある．我々は，このような言葉でない情報を積極的に読み取り，コミュニケーションを円滑に進めている．このような人間と人間の間でのコミュニケーションの豊かさに比べ，人と機械とでこのようなコミュニケーションを行うことはほぼ実現されていない．人と機械とでより豊かなコミュニケーションを行うためには，言葉でない情報から人の気持ちを推定することが重要であるといえる．

「目は心の鏡」という言葉があるように，目の様子や視線の変化，すなわち視線情報は人間の目は情動などの内部状態が知る手がかりとなる．そこで本研究では目の様子と見ているものの対象を考慮して選好の推定を行うことを考える。

1.1 選好と審美的品質

選好 (preference) を，本研究では対象を好むかそうでないかという情報と定義する．この選好と関連する概念として審美的品質 (Aesthetics quality) がある．審美的品質は，ノイズや圧縮ひずみなどの品質とは違い，人が画像を感覚的に好むかどうかを示す量として定義される [4]．審美的品質と選好とが異なる点として，審美的品質は画像の意味に着目せず，色や構図などの感覚で好みを調べる点にあり，選好は審美的品質や画像の意味を含め好むかどうかである．したがって，審美的品質が必ずしも選好と一致するわけではないが，深く関係はしている．選好をさらに細かく分類すると，特定ユーザの選好 (personal preference) と一般的な選好 (general preferences) が考えられる．本研究では主に特定ユーザの選好について議論し，推定を行う．

1.2 研究目的とアプローチ

本論文の目的はマルチメディアコンテンツの選好を推定する方法を機械的に実現することである．本論でのマルチメディアコンテンツとは画像と音声のない動画のことを指す．具体的な目的として，3 つの条件下での選好を推定の対象とする．まず，2 枚の画像が並列に

提示された時，いずれを好んでいるかを推定する実験である．次は１枚の画像が提示された時，その画像を好んでいるか推定する実験である．最後に動画が提示された時，その動画を好んでいるか推定する実験である．この目的を達成するアプローチとして，視線情報と画像，それらに関連付けられた選好の情報から機械学習を用いて推定する手法を用いた．視線情報は統計量を計算することにより，画像は審美的品質を計算することにより推定の性能を高めた．また，視線情報と画像を組み合わせた手法についても検討した．

1.3　論文構成

本論文の構成は以下のとおりである．2 章では本論文に関連する研究について調査した結果を報告し，本研究の意義について議論する．3 章においては画像を並列提示した場合における画像の選好を推定した結果を分析し議論する．4 章においては画像を順次提示した場合における画像の選好を推定した結果を分析し議論する．5 章においては動画を順次提示した場合における動画の選好を推定した結果を分析し議論する．6 章にではこれまでの章で得られた知見から結論を出し，今後の課題について議論する．以上の内容について，以降では詳細に述べる．

第2章

関連研究

これまでにユーザの行動からのユーザの内部状態推定は盛んに行われてきた．内部状態とは，感情や情動，思考などの人間の内部で起こっている直接観測不能な状態を差す．ユーザの行動からのユーザの内部状態推定で最もわかりやすい例は表情である．たとえば，笑っている人は喜んでいると推定でき，泣いている人は悲しんでいると推定できる．このような内部状態推定のさきがけとして，Ekman らが行った表情の研究 [5] がある．この研究では Action Unit と呼ばれる表情筋を基に作られた単位に表情を分解することで，表情を定量的に扱えることを示した．このように表情に対し，視線情報からの内部状態推定のほとんどは現在研究している段階にある．ただし，研究されながらも実用化されている，数少ない例として，高度交通システムの分野で利用されている居眠り防止技術 [6] がある．

2.1 視線の計測と計測結果の区分

視線計測器 (Eye tracker) は種々あるが，ここでは本論文と関係する非接触型視線計測器のみ説明する．非接触型視線計測器は人間とある程度離れた距離から見ている位置を測定できる機器である．通常，図 2.1 のように人間とディスプレイが向きあうようになっている状態でディスプレイの近くに設置される．この場合，測定器によりディスプレイ上で見ている場所，すなわち注視点などを測定できる．実際に非接触型視線測定器を用いてディスプレイ上の注視点を測定し重畳した結果を図 2.2 に示す．図 2.2 のような測定結果を得ることでコンピュータは目の様子や視線の変化に関係する情報を得ることができる．非接触型視線計測器に利用される測定方法については参考文献 [7] を参照してほしい．

視線測定器から得られたデータは特定の時間における注視した座標が単に並べられているなど，このままでは人間には理解しにくい形になっている．そこで，これらの情報を人間の眼球運動と，眼球内または眼球外部の運動から情報を区分する．本発表ではこの区分した情報を視線情報と呼ぶことにする．視線情報は以下のような区分がある [8, 7, 9]．

- 固視
- 衝動性眼球運動（サッケード）
- 固視微動
- 瞬目
- 瞳孔の大きさ

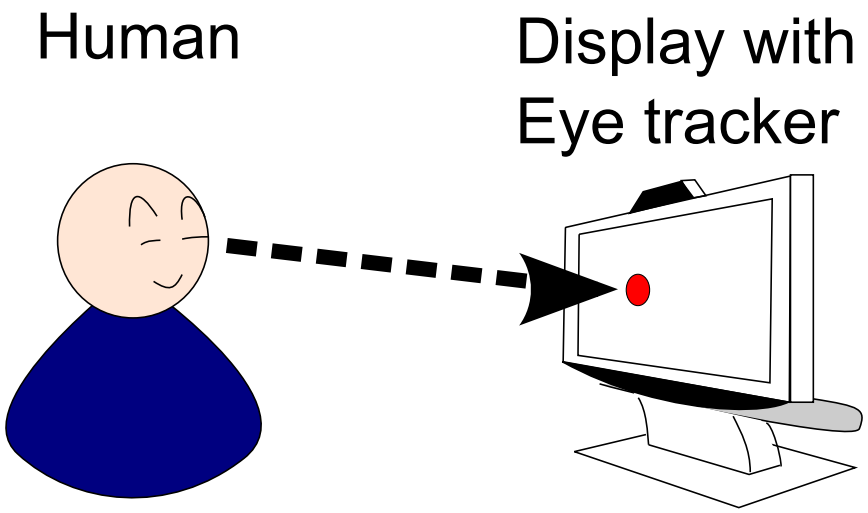

図 2.1: 非接触型視線測定器による視線計測時の模式図

- ピント調節
- スキャンパス

以上の視線情報について，以下では，それぞれの特性と視線情報に関連するユーザの内部状態を説明する．

固視

固視 (fixation) とは，一点を注視している状態のことである．人は興味がある対象を無意識的に注視する傾向がある．注視している対象に興味がある場合，固視の継続時間 (注視時間) がふえるといわれている [10]．しかし，静的な対象であれば，注視時間を簡単に取り扱うことができるが，動画のような動的な対象の注視時間を取り扱うことは難しい．

衝動性眼球運動 (サッケード)

衝動性眼球運動 (saccade) とは，非常に高速度な眼球運動である．衝動性眼球運動は読みづらいため，以下ではサッケードと呼ぶことにする．注視する領域を切り替える際に起きるとされている．そこで逆にサッケードとサッケードの間を固視と通常定義される．固視が興味と関係が有ることから，サッケードは興味の移り変わりを意味している．

固視微動

人間は固視中にも無意識に絶えず眼球を動かしている．これを固視微動という．ある点を固視したまま，周辺のある部分に注意を向けると，固視微動の方向が注意を向けた方向へシフトするという研究 [11] があるため，固視微動と内部状態に関係があるとされる．

瞬き

瞬きは眼球が乾燥しないよう眼球を定期的に涙液で被うなどの役割を持つ．瞬きの

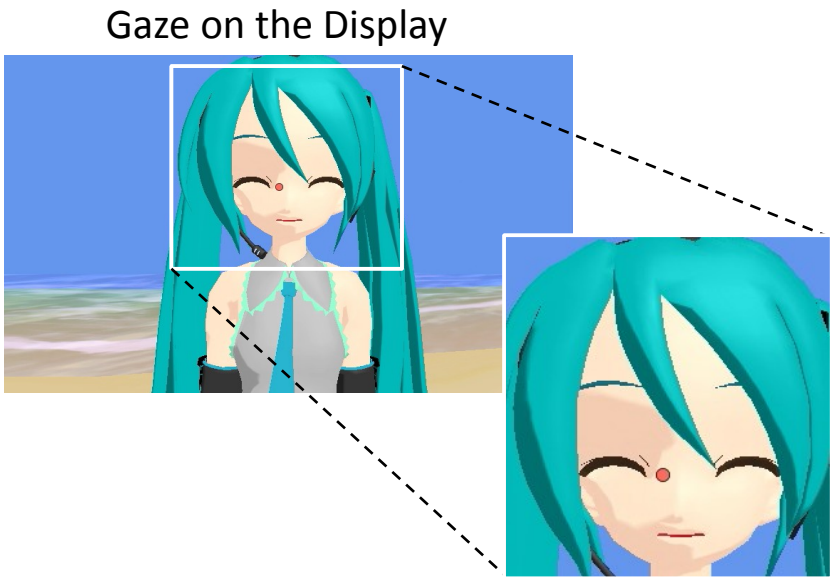

図 2.2: 視線を重畳したディスプレイのスクリーンショット (赤丸が視線先，キャラクターの権利はクリプトン・フューチャー・メディア株式会社が保有)

周期は環境 (湿度，明るさ) に依存するが，性別や年齢によっても変わる．その他に瞬きは注視している対象に興味がある場合，興味がない場合に比べ，瞬目の回数が減ると言われている [10]．また，覚醒度の周期によっても変動する．

瞳孔の大きさ

瞳孔 (pupil) は眼球内に入る光量を調節するための眼球内にある器官である．人は異性のヌード写真を見ると，瞳孔が大きくなることがわかっている [12]．また，一般に人は興味のある対象を見ると，瞳孔が大きくなるとされている [9]．ただし，瞳孔が小さくなることとユーザの内部状態との関係はまだ特にわかっていない．

ピント調節

ピント調節 (accomodation reflex) は眼球の網膜に結ぶ像のピントを水晶体により調整することである．眼精疲労がたまるとピント調節が遅くなることが知られており，ここから疲労について調べることができる [9]．

スキャンパス

スキャンパスとは固視の軌跡である．顔が写っているヌード写真に性的魅力度を感じている場合，体と顔の交互に見る傾向があり [13]，脳外科手術における習熟度に応じてスキャンパスの傾向が変わる [14] など，スキャンパスとユーザの内部状態と関係があると言える．しかし，スキャンパスの類似度などを始めとするスキャンパスの解析は困難である．

人間の眼球の性質上，人間の視線は絶え間なく動いており，視線データから固視微動とサッケードを区分することは難しい．そこで通常，固視とサッケードを区分するために研

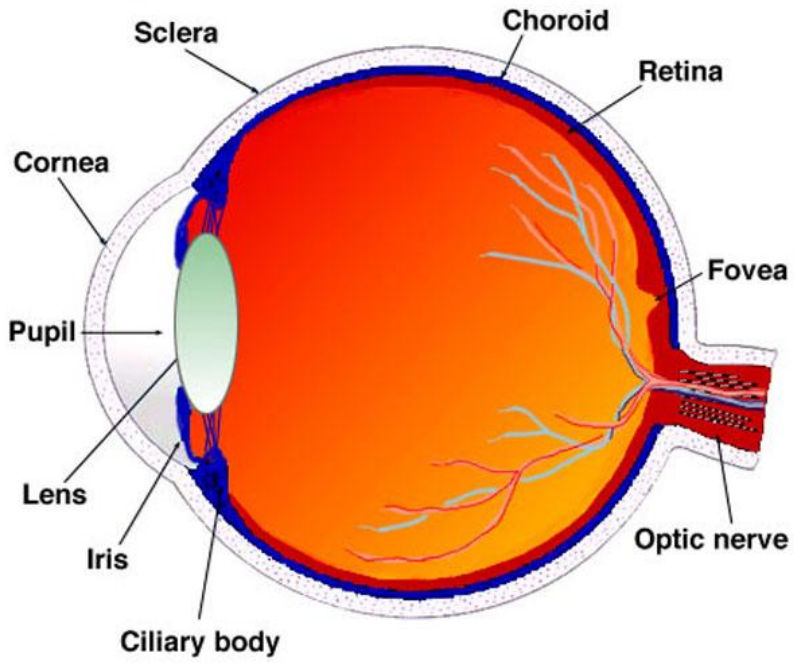

Fig. 6. Vertical sagittal section of the adult human eye.

図 2.3: 人間の目の構造を表した図 ([1] より引用)

究ごとで一定の基準を設けて区別することにしている．

2.2　視野の特徴

人間は目でとらえた光を視線の方向によらず一様に感じ取っているわけではない．このことを説明するために人間の目の構造を図 2.3 に示す．人間の目は水晶体 (Lens) や網膜 (Retina)，中心窩 (Fovea) などで構成されている．このうち，中心窩の周辺には光受容体の密度が周辺に比べ高いことが Curio らの研究 [15] でわかっている．水晶体中心と中心窩を結ぶ線 (以下，光軸) と視線はおおよそ一致し，光受容体の密度のため，視線先は光を感じやすいようになっている．また光を感じる錐体細胞には方位性があり，光の入る角度が光軸と一致する方向であるほど光を強く感じる．この現象をスタイルズ・クロフォード効果 [16] という．たとえば，対向車のヘッドライトが対向車の位置によって眩しい時と眩しくない時があるのは，スタイルズ・クロフォード効果のためである．

この視線先の範囲を中心視野 (central vision) とよび，視線先の周りの範囲を周辺視野 (peripheral vision) という．中心視野と周辺視野が光を受容するという低レベルな違いだけではなく，認識という高レベルでも差がでていることを示す研究がある．

Larson らは被験者にシーン認識を行わせたところ，中心視野の情報よりも周辺視野の情報よりが認識に役に立つ場合があると結論づけた [17]．したがって，視線が関わる画像認

識問題は視野を考慮しなければならない．

2.3　視線情報による内部状態推定

視線情報やその他の情報を基に，ユーザの内部状態を機械的に推定する研究が近年盛んに行われている [18, 19, 20]．fMRI によるデータや脳波 (EEG) からユーザの内部状態推定を行う研究もある [21, 22, 23] が，これらに比べ，視線情報から行うものはユーザにとって測定の手間が少なく，幅広い環境で利用できる利点をもつ．視線情報からの内部状態推定の中で盛んに研究されている内部状態は集中度である．この例として，Takahashi は，適切な映像コンテンツを推薦するための情報を得るために情動計測を用いた映像に対する集中度推定を行なっている [24]．この研究では，映像視聴中におけるユーザの身体動作量や瞬目間隔，視線変動量といったユーザに関する情報のほか，字幕や映像中の顔の数などのコンテンツに関する情報から SVM 回帰により集中度を導出している．また，Yonetani らは，一般動画を視聴している際の集中度を予測するため，視線と顕著性マップを用いた集中度推定の手法 [18] を提案している．

また，集中度ほどではないが，内部状態のうち，画像の選好を推定する研究もある．このような研究のさきがけとして，顔画像や抽象的な人工画像と視線の関係性を示した Shimojo らの研究 [25] がある．Shimojo らは [25]2 枚並べられた画像の選好を判断させる実験において，選好を入力する直前で選択する画像へ視線がかたよることを示した．このような研究を受け，工学的な応用を行った Sugano らは [26] 自然画像の選好を視線情報を基に機械的に推定した．この結果，Shimojo らと同様に視線の注視時間が選好の推定に役立つことが工学的にも判明した．

以上のようにユーザの行動から内部状態を推定する研究がある一方で，ユーザに与えられるコンテンツそのものから内部状態を推定する研究もある．この研究の分野の 1 つに画像から審美的品質を推定する研究がある．審美的品質の推定は美学 (Aesthetics) という哲学から始まり，今日では良い写真を取るための技術として使われている．今日までに画像からの審美的品質の推定は数多く行われている [27, 28, 2]．この中で，Marchesotti らの研究 [27] は一般的な画像認識の手法が審美的品質を推定することに有効であることを示した．Bhattacharya らは，審美的品質を推定を応用し，画像の審美的品質を向上させる研究 [2] を行った．審美的品質を向上させた例を図 2.4 に示す．

このように視線情報と画像はユーザの内部状態に推定する手がかりとなるが，それらを組み合わせて議論した研究がある．Yanulevskaya らの研究 [3] は，抽象的な絵画からポジティブな感情を引き起こす部分を計算し，その場所が視線に集まることを仮説として示している．この仮説についての視覚的情報を図 2.5 に示す．

2.4　本研究の位置づけ

これまでに紹介した研究は視線情報による選好推定や画像による審美的品質の推定であった．一方で，視線情報と画像を組み合わせて選好を予測した研究は未だ存在しない．1.1 節

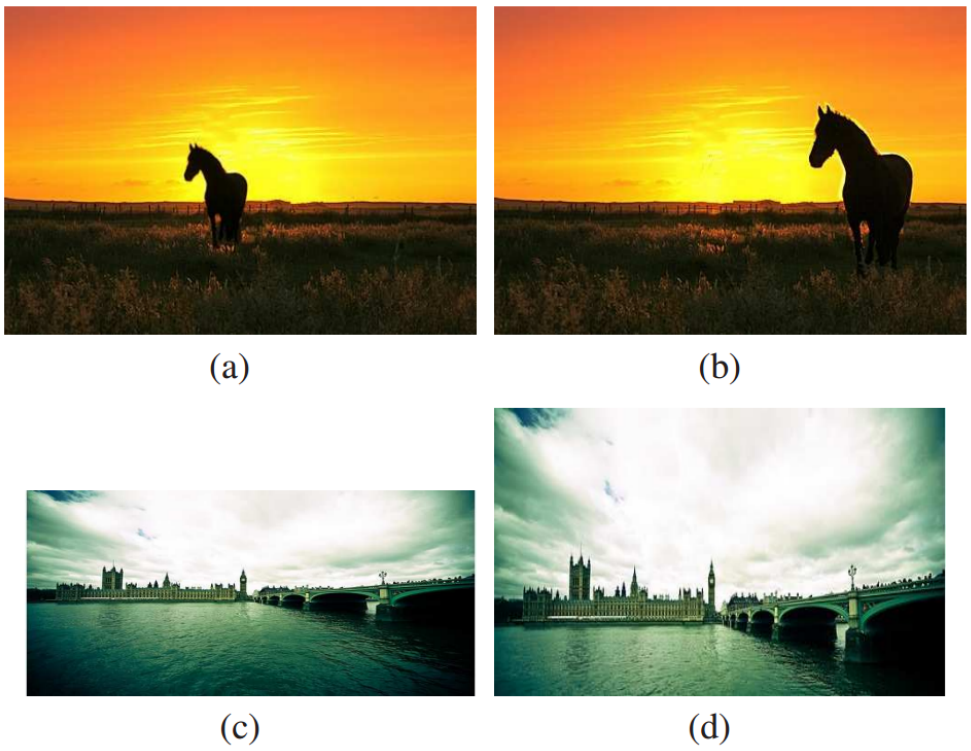

図 2.4: 審美的品質の推定を応用して生成された画像の例．(a) は原画像，(b) は審美的品質が向上するように馬を再配置した (a)．(c) は原画像，(d) は審美的品質が向上するようにクロッピングした (d) ([2] より引用)

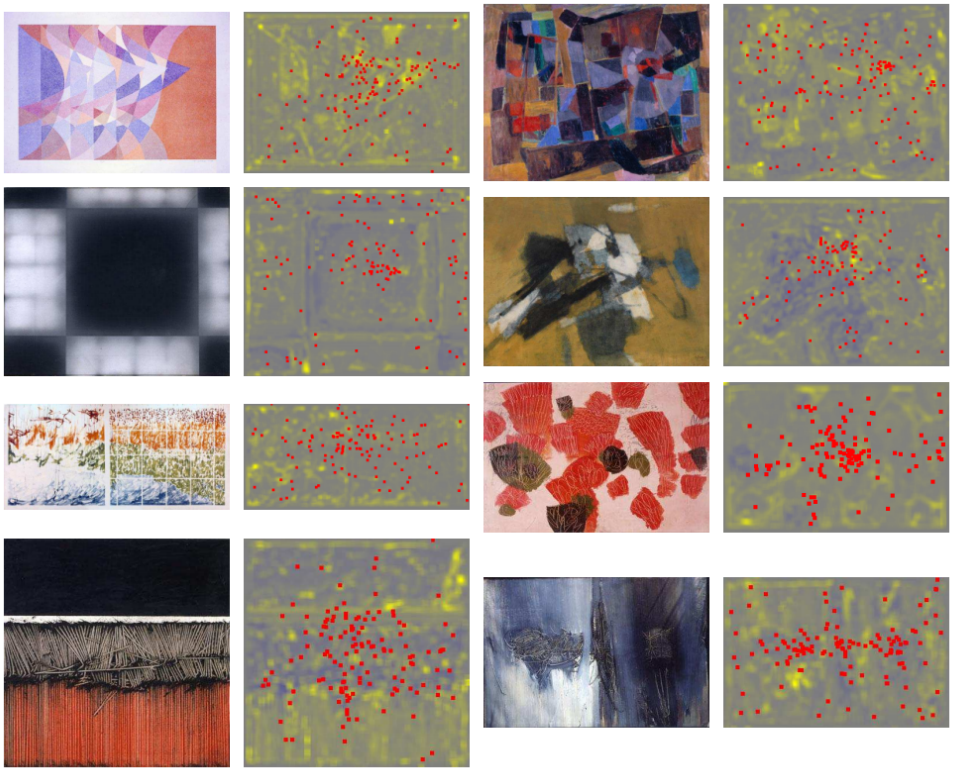

図 2.5:　抽象的な絵画と視線傾向の関係性を示した図．ペアとなる画像の左側が原画像であり，右側が画像に対応するポジティブな感情を引き起こす部分 (黄色) と視線 (赤点) を重畳したものである．確かに黄色い部分に視線が比較的集中していることがわかる．([3] より引用)

にで述べたとおり，画像の選好と興味は相関を持つことがわかっており，マルチメディアコンテンツの選好を推定することは，コンテンツへの興味を推定することにつながる．コンテンツへの興味を推定することは人間の情報処理を分析することができ，デジタルサイネージの広告効果や美術館の展示品の注目度，あるいは動画の推薦に応用することもできると考えられる．したがって，視線情報と画像を用いたマルチメディアコンテンツの選好を推定することは基礎研究だけでなく応用研究においても意義のあることといえる．

第3章

並列提示した場合における画像の選好推定

本実験では視線特徴量と画像特徴量を用いて，図 3.1 のように 2 枚の画像が並列提示されている場合におけるユーザの画像選好を推定することを目的とする．以下の節では本実験とその結果，および考察について詳細に記述する．

3.1 データ収集

まず，本実験を行うために必要なものとして，並列提示するための画像とそれに対する選好の情報が必要となる．本実験では Sugano らが行った実験 [26] で得られた正解データを用いる．本節では Sugano らのデータの収集方法について説明する．

画像収集では被験者実験を行うための並列画像を作成するために画像投稿サイトから画像を収集した．今回使用した画像収集サイトは flickr[1] で，効率よく収集するために flickr API[2] を用いた．これらの画像はディスプレイの両側に二枚並べて使用される．ただし，個々の画像のアスペクト比は必ずしもディスプレイの半分に一致するとは限らないため，レターボックスによるクロッピングを行った．レターボックスによるクロッピングとは画像のアスペクト比を変えず，リサイズし，中央に配置した時に，開いた領域を黒で埋めることで画像全体のアスペクト比を一定にする手法である．この並列画像を画像ペアと呼ぶことにする．

次にこれらの画像を用いて，選好の情報を集めるための実験を行った．以下ではこの実験のことを被験者実験と呼ぶことにする．まず，被験者実験の準備としてディスプレイと視線計測器，顎台を用意した．ディスプレイと視線計測器の仕様を表 3.1 に示す．ディスプレイと顎台の距離を約 60cm に固定し，被験者の顔を顎台で固定して実験を行った．このとき，表 3.1 から計算すると，中心視野から 1 度の範囲はディスプレイ上で直径約 40[px/deg] の円に相当する．被験者実験の流れは図 3.2 のとおりである．被験者実験は大きく自由閲覧と選好タスクに分かれる．この 2 つは図 3.3 のような流れになる。自由閲覧では実験の意図を伝えず，80 組の画像をただ視聴してもらった．視聴時には各組の画像を 10 秒間画像を視聴してもらい，それらの間には視線を固定させるための画像を 3 秒間提示する．こ

[1] http://www.flickr.com/
[2] http://www.flickr.com/services/api/

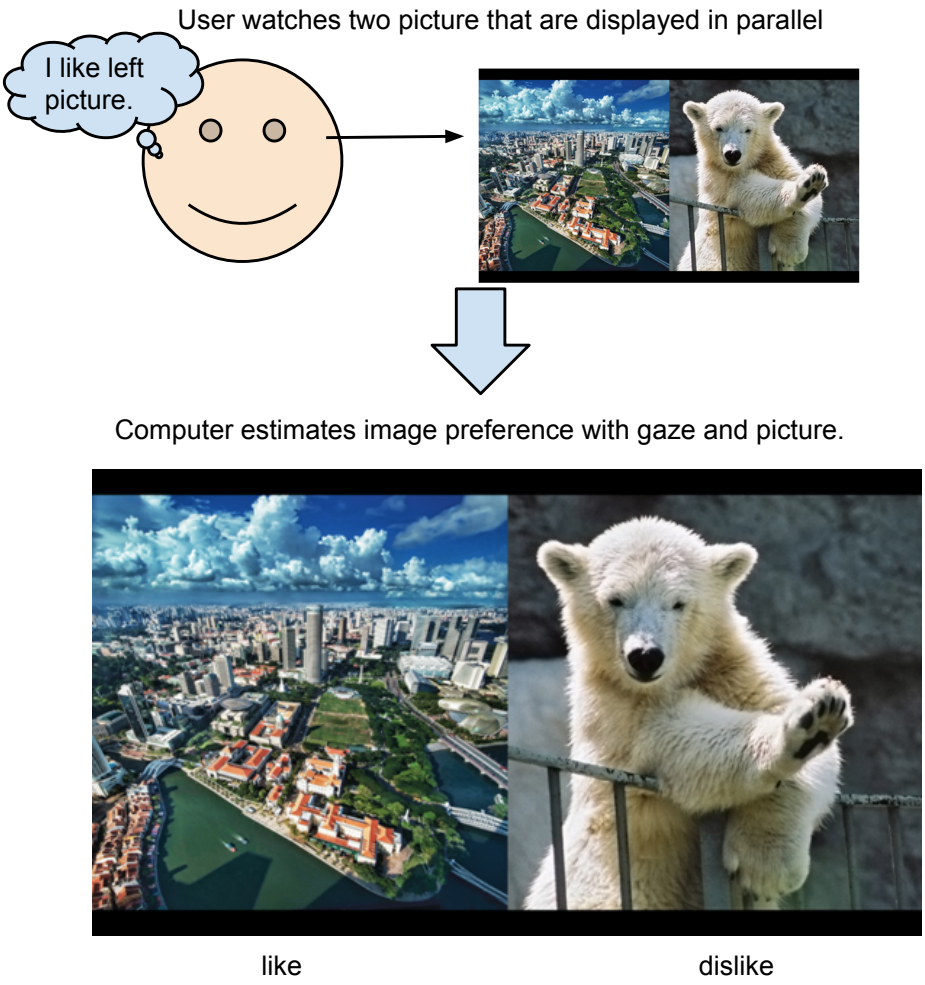

図 3.1: 並列提示した場合における画像の選好推定

こでは選好情報を入力しない．その後，選好タスクでは，画像に選好をつけることを伝えた上で，自由閲覧と異なる画像の組を 400 組視聴してもらった．各組の画像を 10 秒間画像を視聴してもらうが，視聴したあとどちら側が好きかをキーボードで入力してもらう．この点において，自由閲覧と異なる．自由閲覧と同様にそれらの間には視線を固定させるための画像を 3 秒間提示する．最後に自由閲覧で見た画像の 80 組の画像の選好をキーボードで入力し，実験終了となる．

表 3.1: 並列画像提示条件下での実験で使用した視線計測器とディスプレイの仕様

Eye tracker	type	Tobii TX300
Display	sampling rate	60[Hz]
	size	23”
	resolution	1920[px]×1080[px]

図 3.2 の被験者実験を経て，最終的に得られる情報は，以下のとおりである．

画像ペア I　画像ペアは I は左画像 I_L と右画像 I_R で構成される．

視線情報 $\{(g,t)\}$　視線情報は画像提示開始からの時間 t とディスプレイ上における視線先 g とで構成される時系列データである．g はディスプレイ上の座標 (x,y) の成分を持つ．

選好情報 y　選好情報は y はある画像ペアが与えられた時，左が好き (1) か右 (-1) が好きかで二値を取る値である．

3.2　提案手法

本実験では画像の選好を機械的に推定するために，視線情報と画像から特徴を計算し，その計算値からコンピュータを学習させ，学習結果を基に選好をコンピュータに推定させる手法を提案する．

まず，視線情報と画像から特徴を計算することを説明する．計算の流れを図 3.4 に示す．視線情報は認識に影響をおよぼす雑多な問題が混じっている．このような雑多な問題を回避するため，本手法では視線情報の特徴として全体傾向をみることにする．全体傾向を見る方法として 1 つは統計量の計算がある．統計量とは，平均や分散を一般化した概念であり，データに統計学的なアルゴリズムを適用して得られた結果である．本手法ではこの統計量を計算することを視線の特徴を抽出する方法とする．一方，画像の特徴を抽出する方法として，審美的品質を評価するための特徴抽出方法を用いる．以上のステップを経て，視線情報から得られた計算値を視線特徴量と呼ぶ．また同様に画像からの計算値を画像特徴量と呼ぶことにする．これらの計算値を総じて特徴量と以下では呼ぶ．

次に先のステップで得られた特徴量にもとづいて行うコンピュータの学習について説明する．学習の流れを図 3.5 に示す．先のステップで得られた特徴量は元の画像と対応して生成される．したがって，元の画像に対応する特徴量は元の画像の選好情報と対応付けられる．

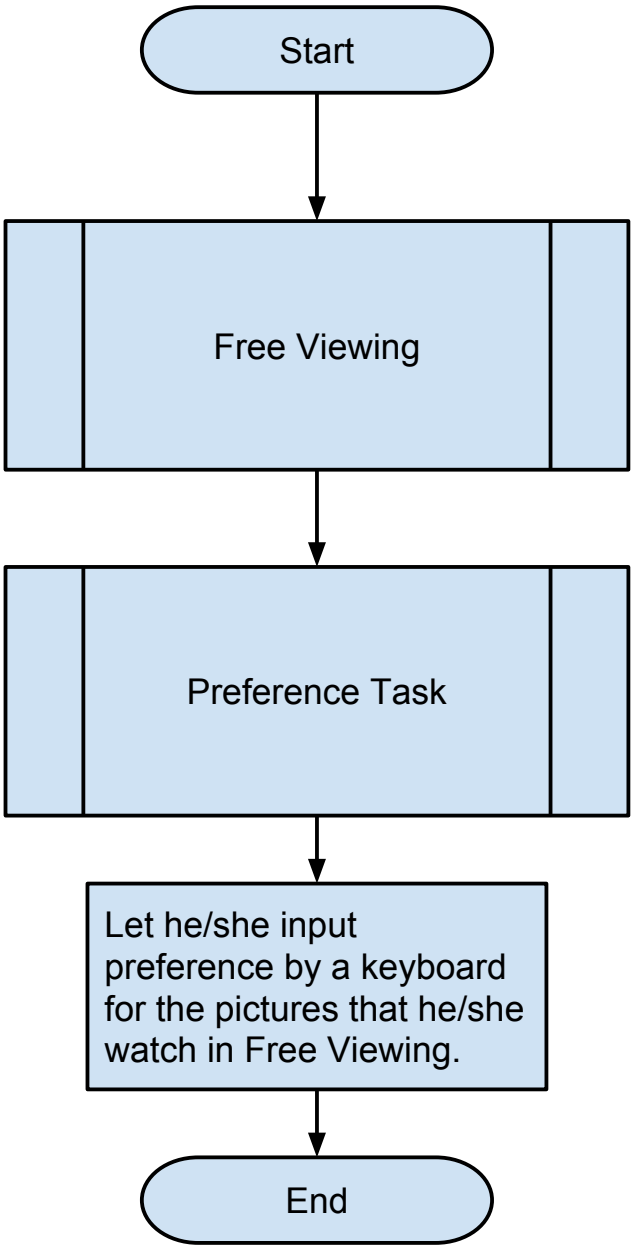

図 3.2: 並列提示条件下における選好情報の収集方法

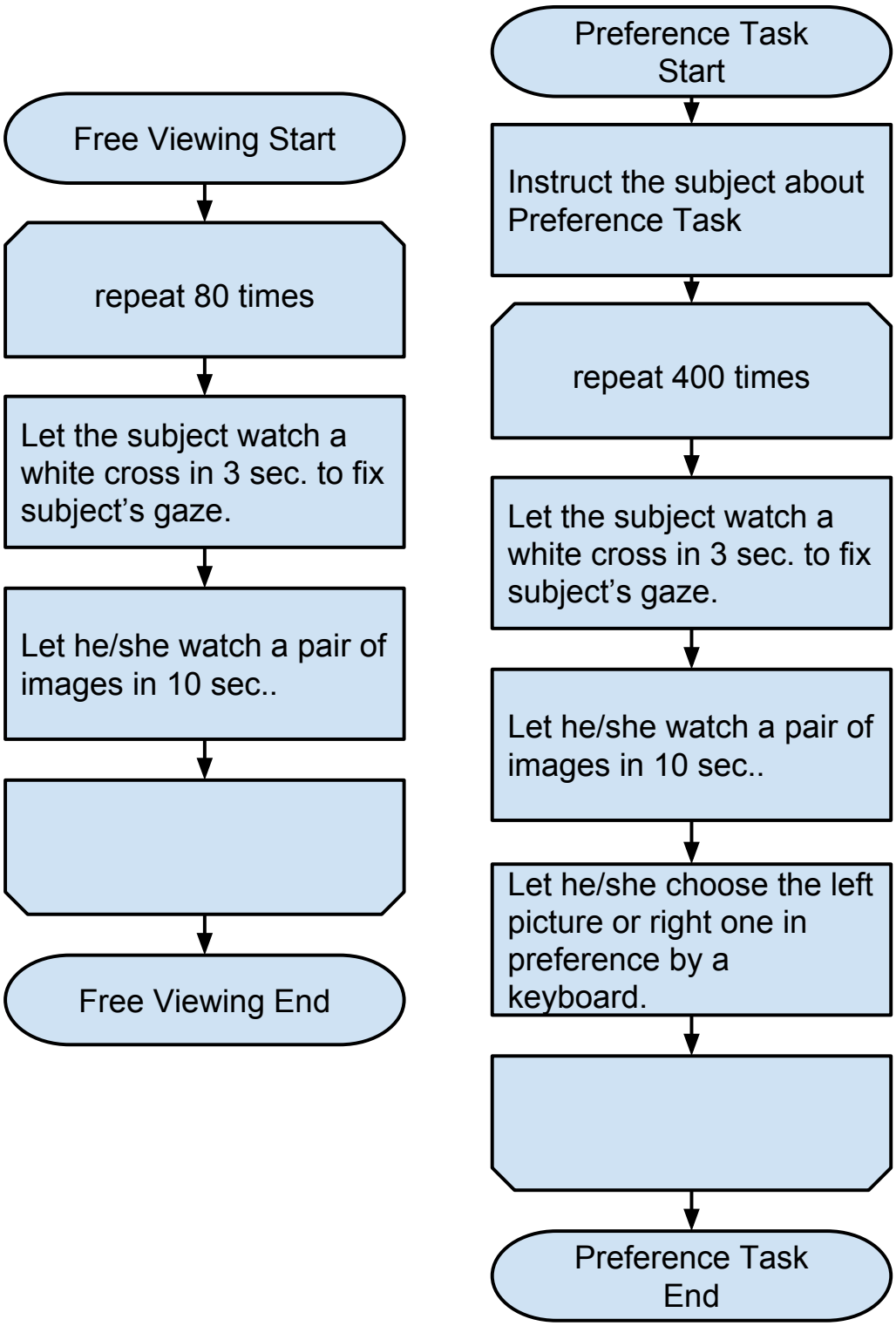

図 3.3: 並列提示条件下における選好情報の収集方法の詳細

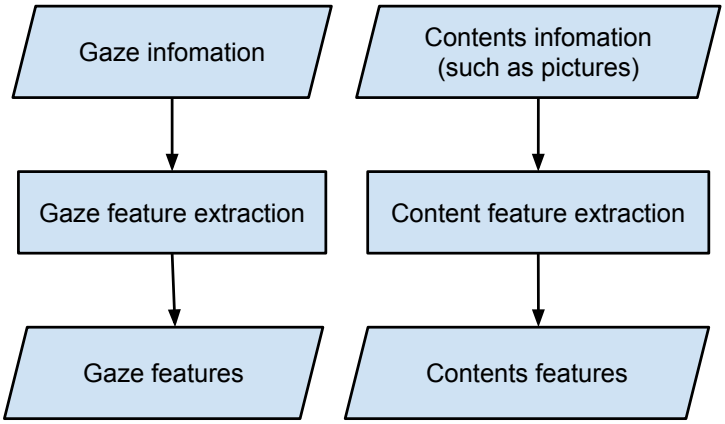

図 3.4: 特徴量抽出の流れ

この選好情報とそれと関連付けられた特徴量から識別器により選好の推定に有効なパターンを発見する．推定性能と原因分析を行いやすくするため，識別器には Random Forest[29] を用いた．

推定の流れを図 3.6 に示す．これまでのステップで得られた推定器を用いて，特徴量から選好を推定する．推定器に特徴量を入力すると，学習した結果として得られた推定に有効なパターンから，自動的に選好を推定する．

以上の流れにより，選好の情報が与えられていない画像に対しても，選好を推定することができる．

3.2.1　特徴量の抽出方法

選好を推定することにとって特徴量を抽出することは，機械に対し選考に関する事前知識を与えることに相当する．そこで本節では今回用いた特徴量の抽出方法について説明する．

i)　視線特徴量

今回，視線全体の傾向を特徴として選好に活用するため，視線情報から抽出した統計量を特徴量とした．今回使用した視線特徴量は表 3.2 のとおりである．まず，実験によって得られた視線情報 $\{(g,t)\}$ を視線の移動速度により Fixation と Saccade に区分した．その後，表 3.2 に表されるような統計量を計算し，これらを特徴量とした．画像 1 枚あたり 25 次元の特徴量が計算されるため，1 ペアから 2 枚分抽出し，その特徴量を連結すると，1 ペアあたり 50 次元の特徴量が得られた．

ii)　画像特徴量

今回使用した画像特徴量は，審美的品質を測定するため，Marchesotti らの研究 [27] や Yanulevskaya らの研究 [3] を参考にした．Marchesotti らの研究の中で今回は Gist と，BoF

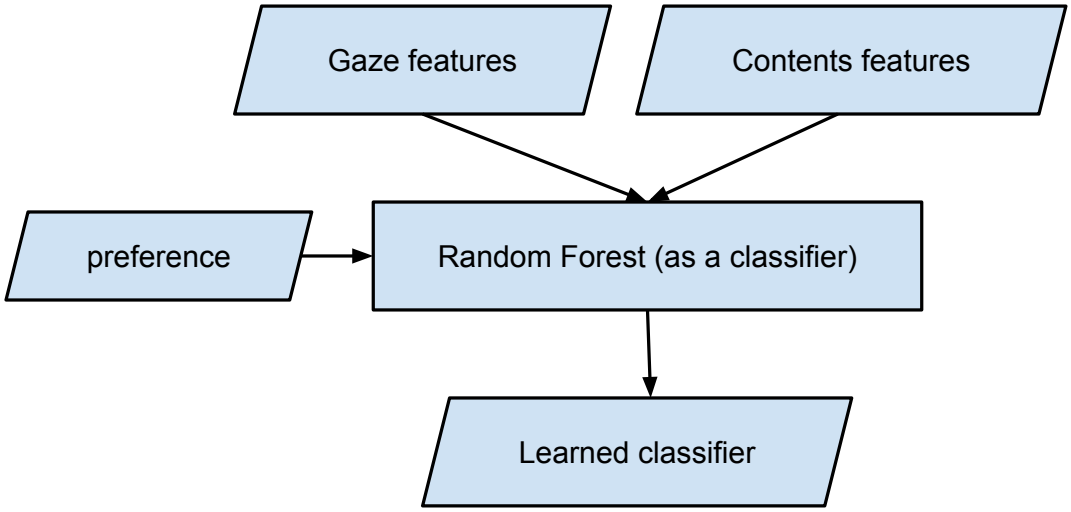

図 3.5: 学習の流れ

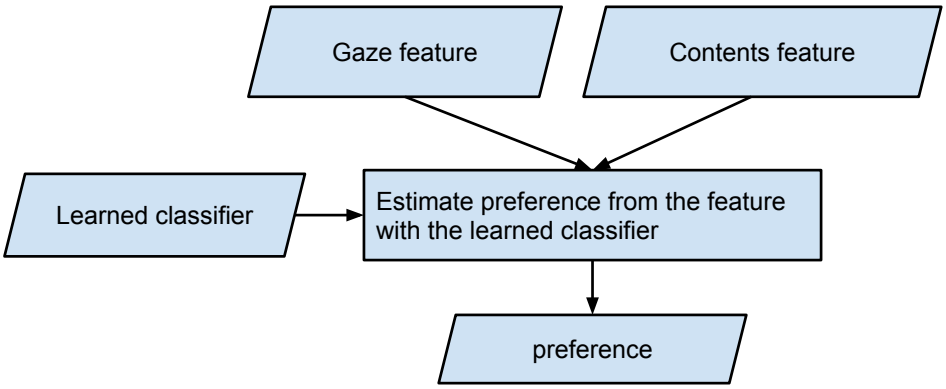

図 3.6: 推定の流れ

表 3.2: 視線特徴量の計算に用いた視線情報と統計量

Fixation	Position	Mean(×2)
		Variance(×2)
		Covariance
	Duration	Mean
		Variance
	Time	Sum
		Mean
		Variance
	Count	
Saccade	Direction	Mean(×2)
		Variance(×2)
		Covariance
	Length	Mean
		Variance
	Duration	Sum
		Mean
		Variance
	Time	Sum
		Mean
		Variance
	Count	

(SIFT) と BoF (Color) 組み合わせた特徴量 (以下，BoF (SIFT)+BoF (Color)，または BoF) を使用した．Gist[30] は Oliva らが提案したシーン認識用の画像特徴量である．Gist は画像のサイズを変化させつつ，Gabor フィルタを各領域で行って得られた結果を結合した特徴量である． 人間の視覚特性を利用しており，内部状態の推定に役立つと考えられる．Marchesotti らの研究では Gist 特徴量の次元は 960 次元だったが，今回はサンプル数が少なく，この場合での性能を上げるため，192 次元になるようパラメータを変更した．具体的には画像ピラミッドを 2，ガボールフィルタの方向を各画像ピラミッドに対し 6 方向，画像のグリッド数を 4×4 とした．

BoF (Bag-of-Features) は，画像を局所特徴量のヒストグラムとして表現する手法 [31] である．局所特徴量は，画像中に存在する複数の特徴的な局所領域から特徴量を抽出する特徴量抽出手法である．局所特徴量の抽出は，コーナーなどの特徴的な点を検出する検出器 (detector) と，特徴点まわりの領域 (局所領域) から特徴ベクトルを抽出する記述子 (descriptor) を用いて行われる．BoF (SIFT) + BoF (Color) は SIFT 記述子と色記述子を BoF で表現し連結したものである．特徴抽出の概要を図 3.8 に示す．今回用いた検出器は dense sampling である．dense sampling は図 3.7 の通り，局所領域のサイズを変えながら，空間的に一定間隔に画像をサンプルする検出器である．この画像を以下ではパッチと呼ぶ．パッチサイズは 64[px]×64[px]，96[px]×96[px]，128[px]×128[px]，サンプリング間隔は 64[px]×64[px] とした．これらの各特徴点の周辺領域から SIFT 記述子と色記述子を計算した．SIFT 記述子 [32] は局所領域を等間隔に分割し，それぞれから輪郭線などの輝度勾配を計算することで得られる特徴量のことである (128 次元)． 色記述子とは色の情報を利用した特徴量のことである．色記述子を実現するために，今回は局所領域を 4 × 4 のグリッドに切り分けてから，RGB チャネルごとで画素値の平均と標準偏差を求めた (96 次元)．性能を高めるため，これらを次元圧縮アルゴリズムである PCA に通し，それぞれ次元を 64 次元にまで圧縮した．なお，PCA で用いた記述子はそれぞれ約 70 万本のうち，メモリの都合により，それぞれランダムサンプリングされた 10000 本分である．圧縮を行ったあとの記述子を BoF により量子化したベクトルをビンとするヒストグラムにする．

BoF の処理はコードブック作成とヒストグラム作成の 2 つの段階に分かれる．コードブック作成では，まず，学習用画像から抽出された局所特徴量をクラスタリングする．次に，各クラスタの中心を示す特徴量の集合 (コードブック) を作成し，コードブックをプロトタイプとした最近傍識別器を構築する．ヒストグラム作成では，まずテスト用画像から抽出された局所特徴量を最近傍識別器にかける．次に各クラスタに属する特徴量を計数し，ヒストグラムを作成する．BoF によって，画像を位置の変化に頑健なヒストグラム特徴として表現することができる．このため一般物体認識の分野で多く使用される．クラスタリングアルゴリズムは Marchesotti らと同じように EM アルゴリズムで GMM にフィッティングした分布を用いたシンプルなクラスタリングを用いた．今回，BoF のコードブックサイズ (ビン数) は 100 にした．以上のステップから，SIFT 記述子と色記述子から得られた BoF のヒストグラムを連結し，これを画像特徴量として推定に用いる．

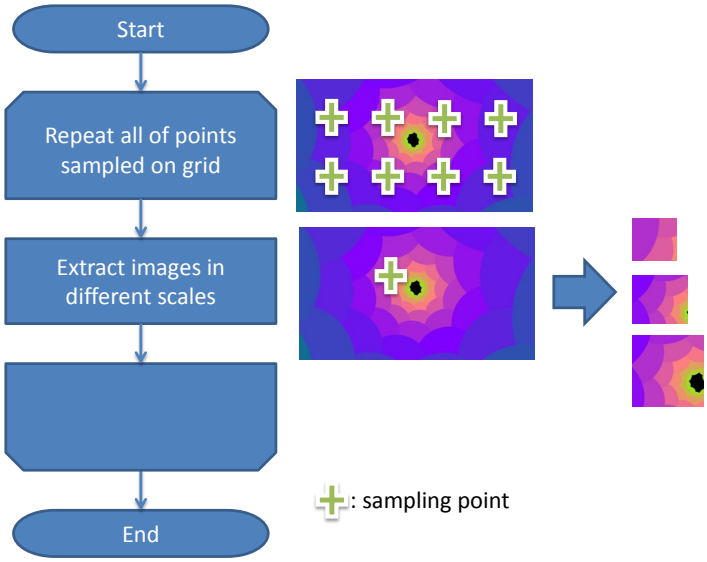

図 3.7: dense sampling による画像のサンプリングの流れ

3.3　実験方法と実験結果

本実験では，提案手法の有効性を分析するために，3 種類の実験を行った．実装には，画像特徴量抽出は OpenCV[3] で用いて，認識は機械学習ライブラリである scikit-learn[4] で行った．なお，Random Forest の決定木は 1000 本とし，各決定木の最大の深さを 3 とした．

3.3.1　選好タスクにおける被験者内の推定精度

被験者実験のうち，選好タスクによって得られた選好を同一の被験者だけのデータを使用して推定する実験を行った．評価方法は Leave one out で評価指標は認識率である．画像特徴量 (Gist，BoF) や視線特徴量 (Gaze)，それらを連結した特徴量 (Gaze + BoF) にて比較した結果を図 3.9 に示す．また図 3.9 の詳細な結果を表 3.3 に示す．各手法の有効性を調べるため，被験者ごとで算出した精度を用いて Wilcoxon の順位和検定を各手法のすべての組み合わせで行った．その結果を表 3.4 に示す．表 3.4 より Gaze での精度と BoF での精度で，危険率 α =0.01 において有意差が認められた ($p = 0.003 < 0.01$)．Gaze での精度と Gaze+BoF での精度で，正確な値は出なかったが，有意差はなかった ($p = 0.563$)．

[3]http://opencv.org/
[4]http://scikit-learn.org/stable/

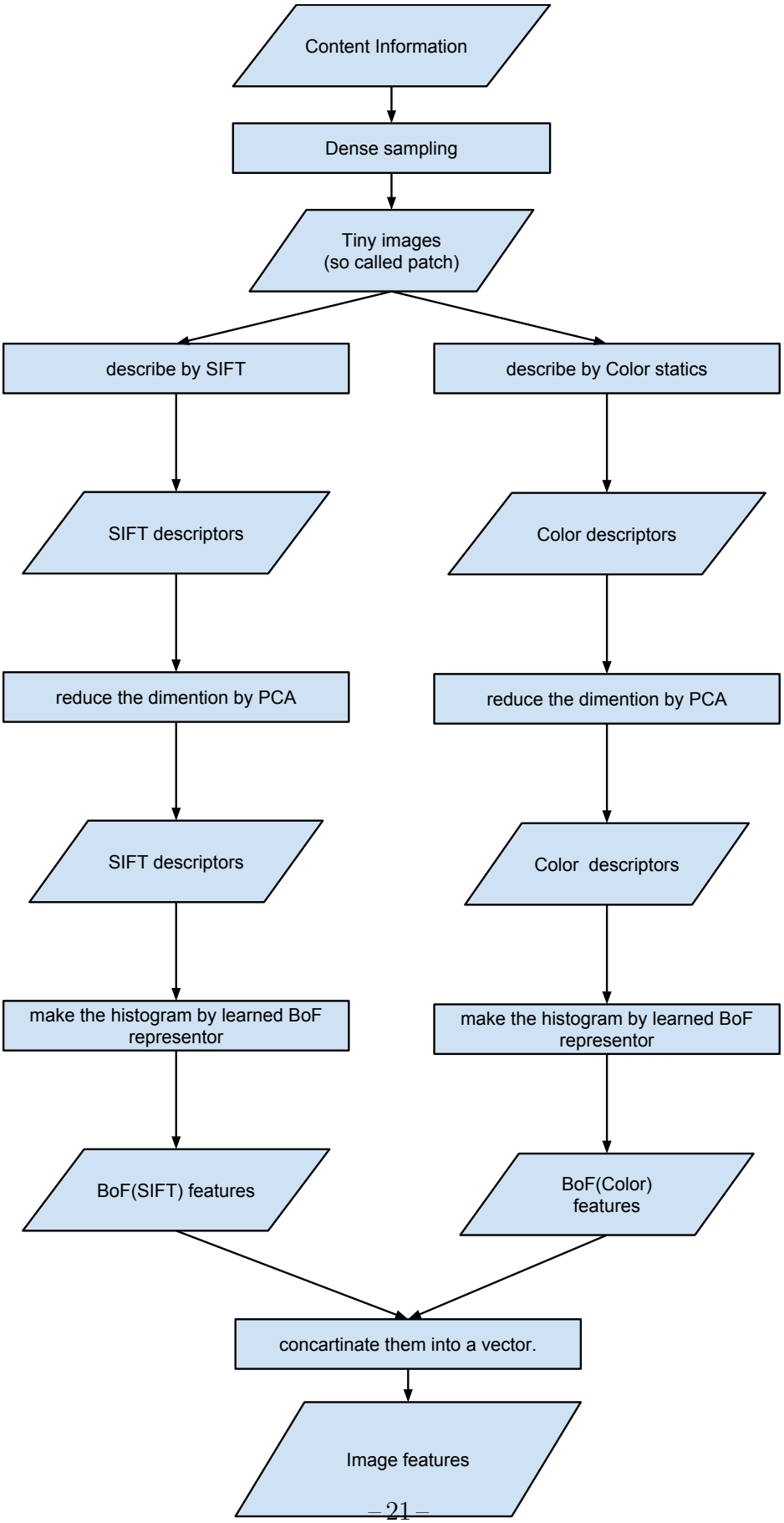

図 3.8: 画像特徴量抽出の流れ

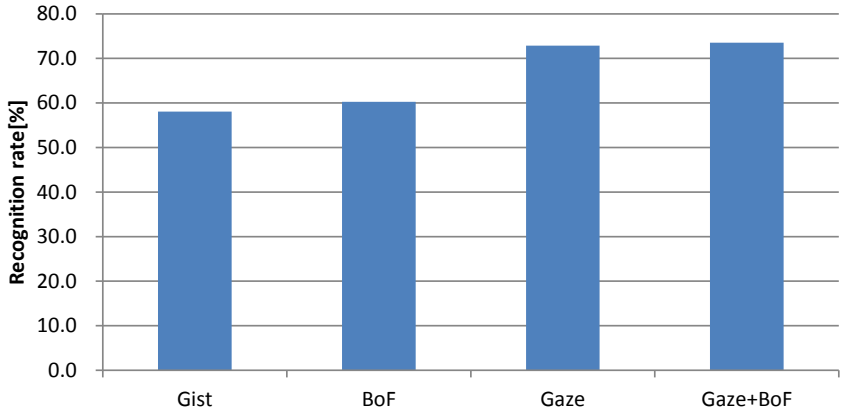

図 3.9: 選好タスクにおける被験者内の各特徴量別推定精度

表 3.3: 選好タスクにおける被験者内の各被験者・特徴量別推定精度 [%]

Subject	Gist	BoF	Gaze	Gaze + BoF
1	57.3	57.8	78.5	77.3
2	54.3	60.9	80.2	80.5
3	55.7	59.2	70.6	70.4
4	62.8	65.4	60.6	65.1
5	56.1	59.3	64.4	63.6
6	60.7	62.5	81.7	82.4
7	56.5	61.1	80.6	80.3
8	61.0	60.3	82.1	81.8
9	56.3	50.3	63.5	63.3
10	61.0	67.7	68.0	71.6
11	56.9	58.4	71.3	72.6
Mean	58.1	60.3	72.9	73.5

表 3.4: 選好タスクにおける被験者内の各特徴量間 p 値

	Gist	BoF	Gaze	Gaze+BoF
Gist	-	0.0537	0.00195	0.000977
BoF	-	-	0.00293	0.00195
Gaze	-	-	-	0.563
Gaze+BoF	-	-	-	-

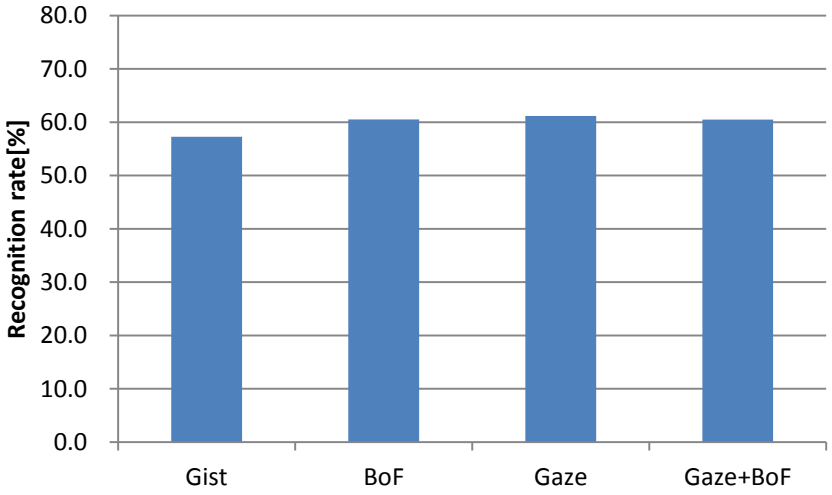

図 3.10: 自由閲覧と選好タスクにおける被験者内の各特徴量別推定精度

3.3.2　自由閲覧と選好タスクにおける被験者内の推定精度

被験者実験のうち，自由閲覧と選好タスクによって得られた選好を被験者内のデータを使用して推定する実験を行った．評価方法は選好タスクを学習セット，自由閲覧をテストセットとしたときの認識率である．画像特徴量や視線特徴量，それらを連結した特徴量にて推定精度を比較した結果を図 3.10 に示す．詳細な結果を表 3.5 に示す．各手法の有効性を調べるため，被験者ごとで算出した精度を用いて Wilcoxon の順位和検定を各手法のすべての組み合わせで行った．その結果を表 3.6 に示す．表 3.6 より Gaze での精度と BoF での精度で，正確な値は出なかったが，有意差はなかった ($p = 0.906$)．Gaze での精度と Gaze+BoF での精度で，正確な値は出なかったが，有意差がないことがわかった ($p = 1.0$)．

表 3.5: 自由閲覧と選好タスクにおける被験者内の各被験者・特徴量別推定精度 [%]

Subject	Gist	BoF	Gaze	Gaze+BoF
1	65.8	58.2	57.0	58.2
2	45.0	63.8	58.8	58.8
3	64.1	60.3	60.3	61.5
4	45.9	47.3	59.5	44.6
5	59.2	59.2	59.2	60.5
6	60.0	66.7	72.0	66.7
7	60.8	72.2	67.1	70.9
8	58.6	57.1	54.3	52.9
9	51.9	58.4	67.5	70.1
10	64.1	69.2	65.4	70.5
11	54.5	53.2	51.9	50.6
Mean	57.3	60.5	61.2	60.5

表 3.6: 自由閲覧と選好タスクにおける被験者内の各特徴量間 p 値

	Gist	BoF	Gaze	Gaze+BoF
Gist	-	0.262	0.221	0.328
BoF	-	-	0.901	0.476
Gaze	-	-	-	1.00
Gaze+BoF	-	-	-	-

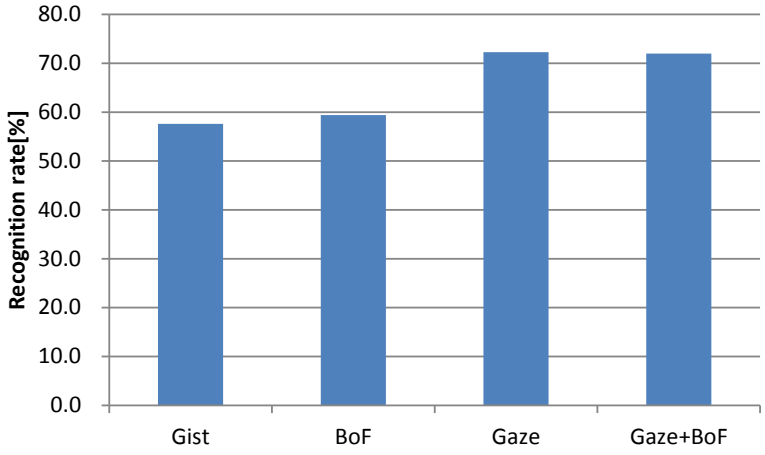

図 3.11: 選好タスクにおける被験者間の各特徴量別推定精度

3.3.3　選好タスクにおける被験者間の推定精度

被験者実験のうち，選好タスクによって得られた選好を被験者間のデータを使用して推定する実験を行った．評価方法は Leave one out で評価指標は認識率である．画像特徴量 (Gist，BoF) や視線特徴量 (Gaze)，それらを連結した特徴量 (Gaze+BoF) にて比較した結果を図 3.11 に示す．また図 3.11 の詳細な結果を表 3.7 に示す．各手法の有効性を調べるため，被験者ごとで算出した精度を用いて Wilcoxon の順位和検定を各手法のすべての組み合わせで行った．その結果を表 3.8 に示す．表 3.8 より Gaze での精度と BoF での精度で，危険率 α =0.01 において有意差が認められた ($p = 0.001 < 0.01$)．Gaze での精度と Gaze+BoF での精度で，正確な値は出なかったが，有意差はなかった ($p = 0.123$)．

3.4　考察

今回，視線特徴量と画像特徴量を比較したが，場合によっては視線が有効であることがわかった．しかし，これらの特徴を結合しても，推定精度に対する有効性は見られなかった．この原因として画像に基づく選好推定の有効性と特徴統合の方法の 2 点に問題があったとみられる．画像に基づく選好が画像特徴量により推定できない理由として好みの傾向が画像に個人間でまったく依存していないことが考えられる．今回の特徴統合は，ただ特徴を連結しただけなので，次元数の偏りや相関する情報を取り除けていないことに問題があると考えられる．たとえば，画像が全体的に明るいことと瞳孔の大きさが大きくなることはほとんど同じである．したがって，画像の情報と視線の情報は互いに相関関係がある

表 3.7: 選好タスクにおける被験者間の各被験者・特徴量別推定精度 [%]

Subject	Gist	BoF	Gaze	Gaze+BoF
1	61.6	59.8	79.0	79.3
2	53.3	57.1	80.5	80.7
3	64.6	66.6	71.1	70.4
4	44.8	51.4	52.4	52.4
5	57.7	56.1	64.4	63.3
6	61.0	63.0	82.7	81.4
7	62.7	63.7	80.1	79.5
8	59.0	59.7	81.3	80.8
9	51.6	52.1	62.8	63.3
10	58.1	65.6	68.2	68.2
11	59.4	58.1	72.3	72.3
Mean	57.6	59.4	72.3	72.0

表 3.8: 選好タスクにおける被験者間の各特徴量間 p 値

	Gist	BoF	Gaze	Gaze+BoF
Gist	-	0.120	0.000977	0.000977
BoF	-	-	0.000977	0.000977
Gaze	-	-	-	0.123
Gaze+BoF	-	-	-	-

にしても，独立な情報は得られると思われるので，その点を効率よく抽出すれば精度の向上が見られるのではないかと考えられる．

第4章

順次提示した場合における画像の選好推定

3章では並列提示された画像の選好を推定したが，このような実験設定の場合，広告のような一枚の画像だけを提示された場合，選好を推定することはできない．そこで，ディスプレイに一枚の画像が表示された時の選好を推定することに着目した．本実験では図4.1のようにディスプレイ上に表示された画像を被験者が比較的好きと思っているか，あるいは好きででないと思っているかを推定することを目的とする．以下の節では本実験とその結果，および考察について詳細に記述する．

4.1 データ収集

まず，本実験を行うために必要なものとして，提示するための画像とそれに対する選好の情報が必要となる．本実験では以下のとおりの実験を行い，これらのデータを収集した．

画像収集では実験をするために画像投稿サイトから画像を収集，またソフトにて作成した．集めた画像はその画像が持つ意味合いで大きく3種類に分けた．画像を3種類に分けた理由は視線情報は見ている画像の意味によって変わると仮定したためである．この3種類の画像を図4.2に示す．1つは特に意味を持たないフラクタル画像である．フラクタルは自己相似図形，すなわち一部を拡大すると拡大前と拡大後が掃除になるという図形のことであるが，ここでは数式で定義される幾何学模様の1つとして考えて取り扱う．もう1つは，一言で説明できる程度の意味を持つシーン画像である．シーン画像は自然や都市などの風景を写した画像である．たとえば，4.2では「レンガ建物」の一言で画像を記述できる画像だけで，物体間の相互作用や人物が写っていない．最後にハイコンテクスト画像は人間などの様々な物体が写っており，一言で記述が難しい画像である．たとえば，図4.2の画像を説明するためには「男性が何らかの電子機器を操作している.」や「男性が手に電子機器にかけ，こちらを見ている.」などの説明するためには複数の記述が必要である．人の間で相互関係が時には見られる．フラクタル画像はFractal Explorer[1]を用いて手動で作成した画像を使用し，シーン画像とハイコンテクスト画像は画像投稿サイトpixabay.com[2]か

[1]http://www.eclectasy.com/Fractal-Explorer/
[2]http://pixabay.com/ja/

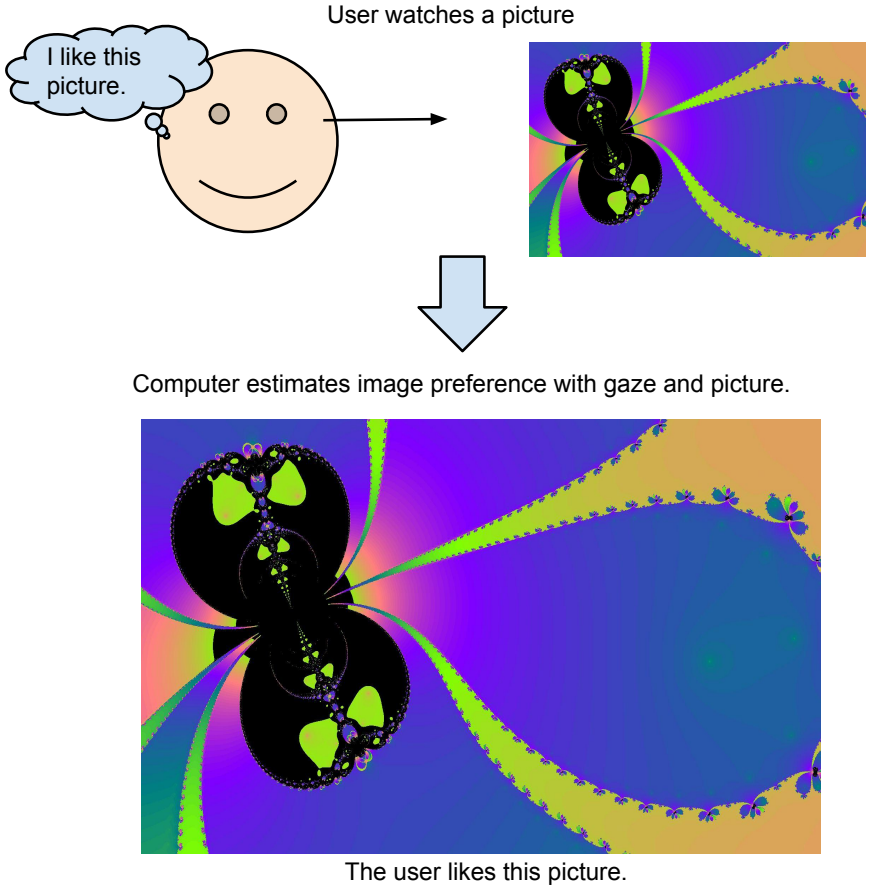

図 4.1: 順次提示した場合における画像の選好推定

ら収集した．またハイコンテクスト画像の一部は SUN Database[33][3] から引用した．以上の 3 種類の区分を総称してカテゴリと以下では呼ぶことにする．それぞれ意味が持つ順に Low context，Mid context，High context とする．これらの画像とディスプレイの解像度と一致させるさせるため，これらの画像を加工した．具体的には画像の上下中央が加工前と加工後が一致するように，リサイズとクロッピングを行った．

表 4.1: 順次画像提示条件下での実験で使用した視線計測器とディスプレイの仕様

Eye tracker	type	Tobii TX300
Display	sampling rate	60[Hz]
	size	23”
	resolution	1920[px]×1080[px]

次にこれらの画像を用いて，選好の情報を集めるための実験を行った．まず，この実験準備としてディスプレイと視線計測器，顎台を用意した．ディスプレイと視線計測器の仕様を表 4.1 に示す．ディスプレイと顎台の距離を約 60cm に固定し，被験者の顔を顎台で固定して実験を行った．このとき，表 4.1 から計算すると，中心視野 1 度はディスプレイ上で約 40[px/deg] に相当する．実験の流れは図 4.3 のとおりである．

実験は大きく自由閲覧と選好タスクに分かれる．この 2 つは図 4.4 ののような流れになる。自由閲覧では実験の意図を伝えず，450 枚の画像をただ視聴してもらった．視聴時には各組の画像を 4 秒間画像を視聴してもらい，それらの間には黒のブランク画像を 1 秒間，視線を固定させるための画像を 1 秒間提示する．ここでは選好情報を入力しない．その後，選好タスクでは，画像に選好をつけることを伝えた上で，自由閲覧と異なる画像を 450 枚視聴してもらった．各組の画像を 4 秒間画像を視聴してもらうが，視聴したあと 7 段階で好きな度合いをキーボードで入力してもらう．7 が画像が好きことを，1 がまったく好きではないことを意味する．この点において，自由閲覧と異なる．自由閲覧と同様に画像の間には黒のブランク画像を 1 秒間，視線を固定させるための画像を 1 秒間提示する．

以上の過程で得られた選好は 7 段階であるが，7 段階を適切に推定することは困難であり，また段階にも曖昧さが伴うため，ここでは 7 段階の情報を元に個人ごとで好きかそうでないかの 2 段階に量子化する．この量子化の方法を説明するため，例えば，図 4.5 のように画像と選好が割り当てられていたとする．この場合，選好の量子化は図 4.5 の計算手順にしたがって行う．量子化の意味をより理解するために図 4.5 について説明する．まず，ある個人のカテゴリ内にあるすべての選好を中央値より大きいか，中央値と等しいか，中央値より小さいかの 3 段階に量子化する．これらにはそれぞれ 1,0,-1 の値を割り当てるとする．数式で記述すると，$\forall p \in P.sign\left(p - median\left(P\right)\right)$ になる．ここで P は図 4.5 に示されるような，ある個人のカテゴリ内にあるすべての選好をのことを意味する．これらの段階のうち，中央値と等しいものは推定が困難であるとかんがえられるため，推定精度の評価からは取り除く．この結果，サンプル数が少なくなるものの，中央値より大きいか，中央値より小さいかの 2 段階になる．これらをそれぞれ好きである，嫌いであるというラベ

[3]http://groups.csail.mit.edu/vision/SUN/

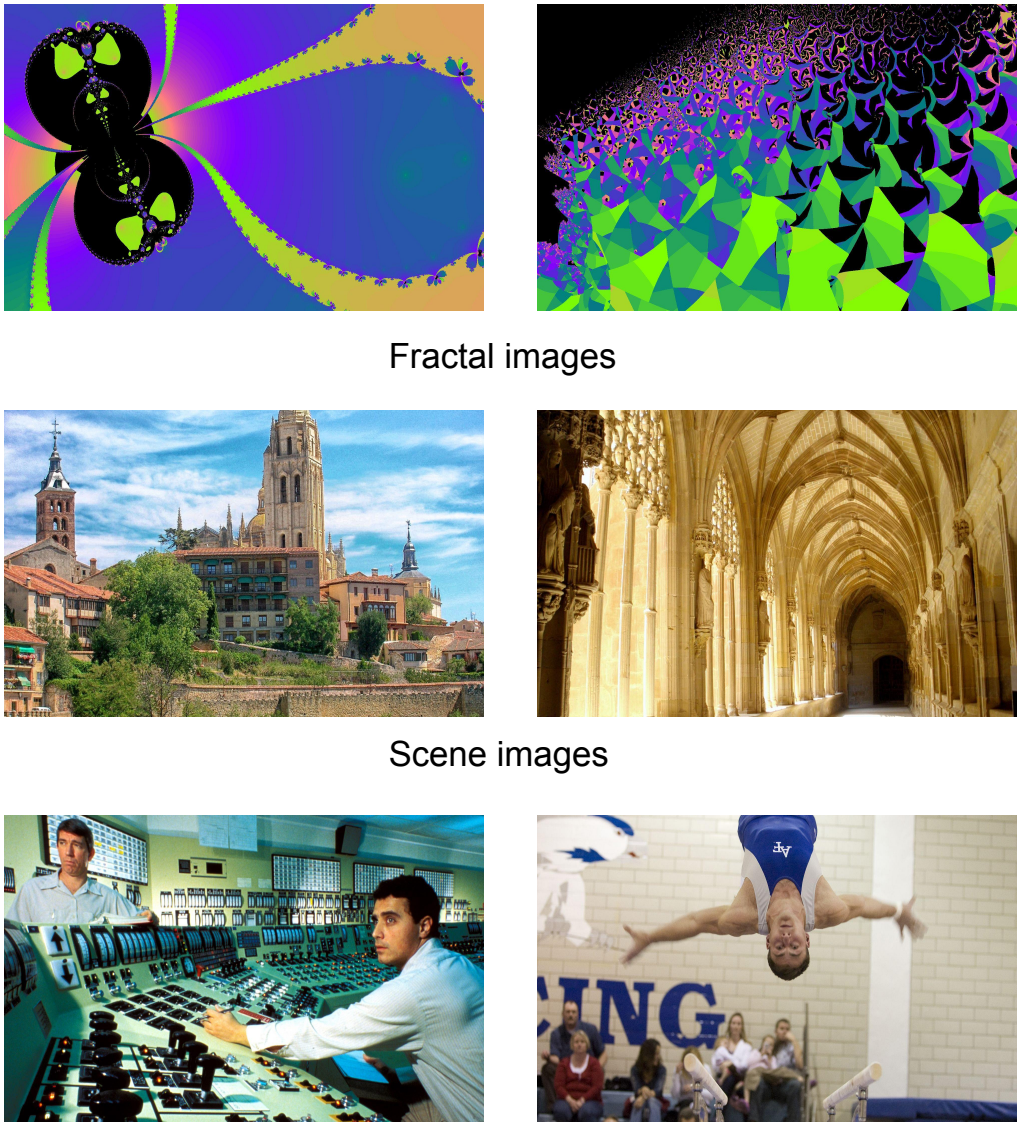

図 4.2: 順次提示条件下の被験者実験に用いる画像の例

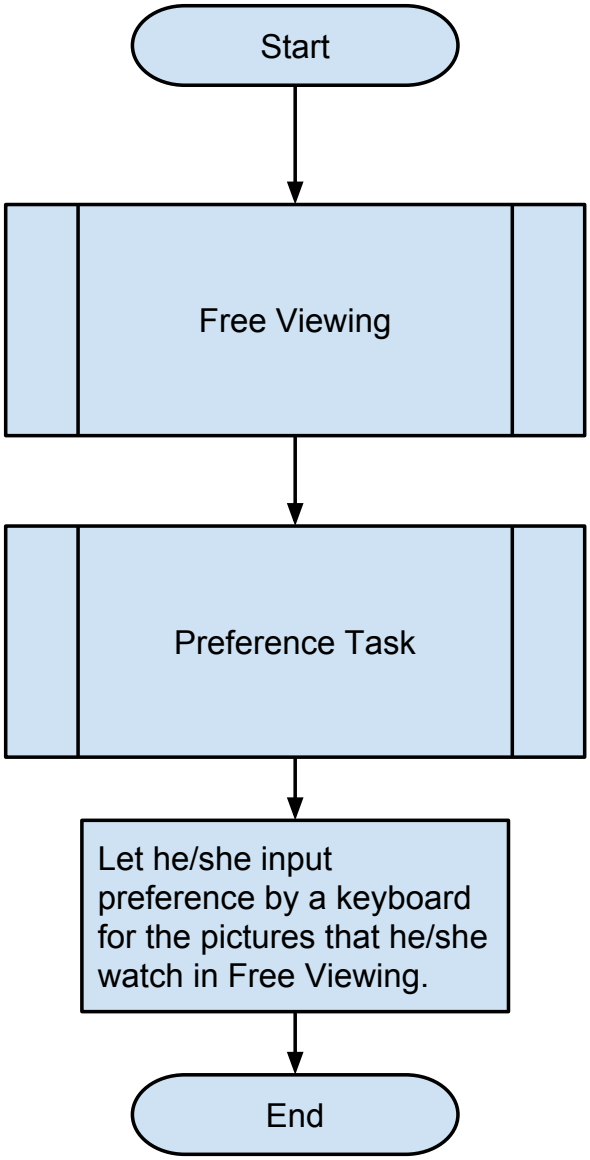

図 4.3: 順次提示条件下における選好情報の収集方法

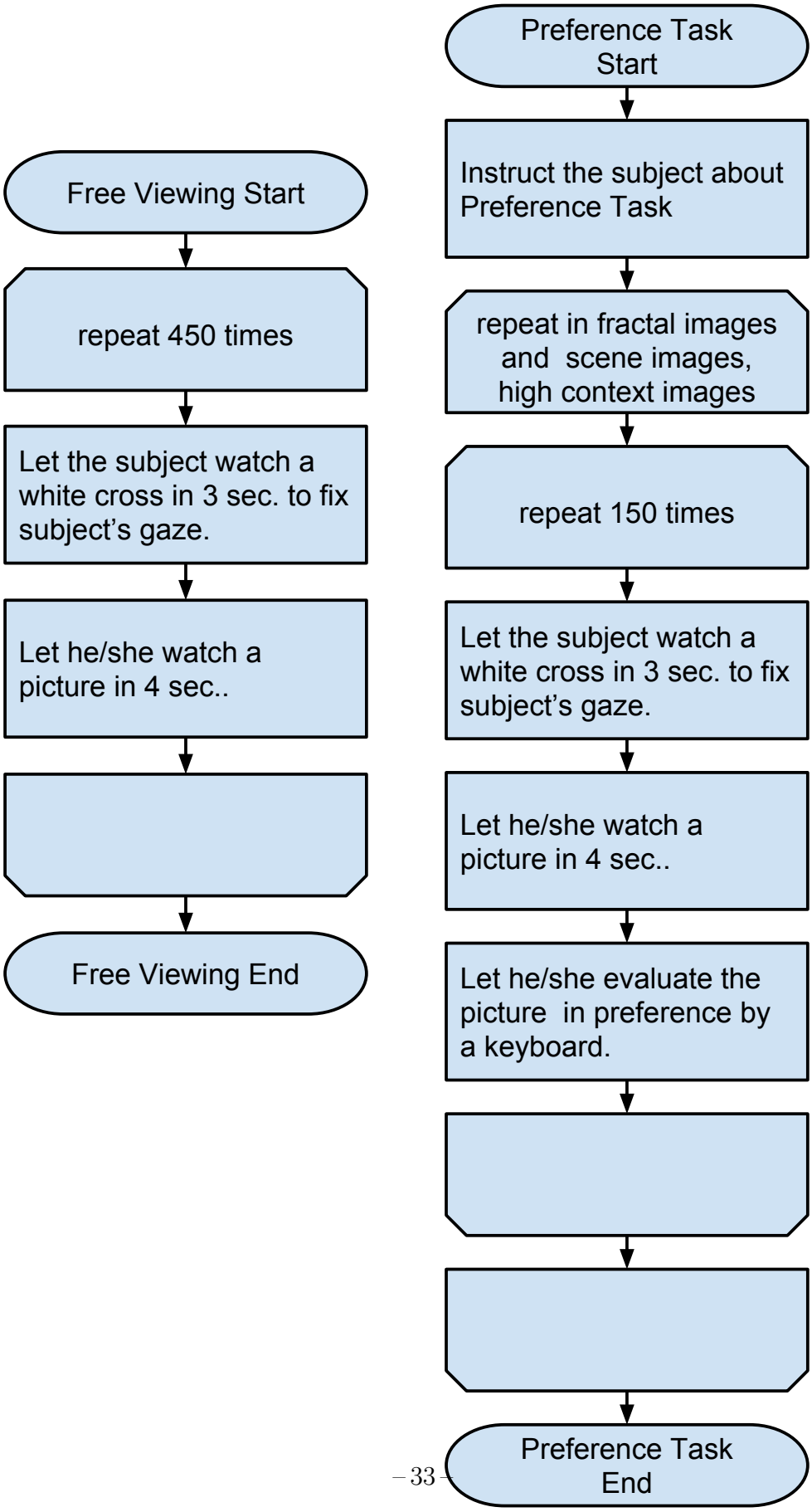

図 4.4: 順次提示条件下における選好情報の収集方法の詳細

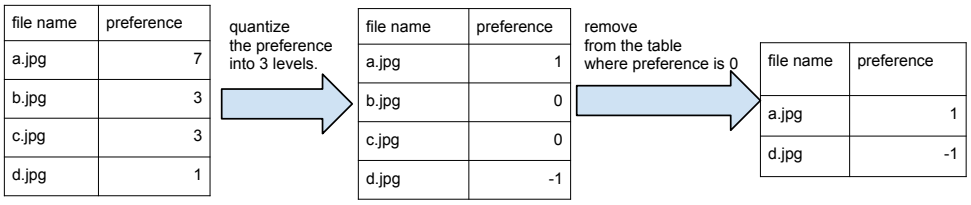

図 4.5: 選好情報の量子化方法とその例

ルを付けることで 2 段階の量子化を行うことができる．この量子化は各カテゴリ，また各個人ごとで行った．

量子化を行った結果，各被験者における得られたサンプル数の内訳を表 4.2 に示す．表 4.2 を見ると，好きのサンプルと好きでないのサンプルとで大きな偏りが見られる．これを改善するため，少ない側のサンプルを多重にサンプリング (オーバーサンプリング) することで偏りをなくした．

表 4.2: 量子化により得られた各被験者における得られたサンプル数の内訳

Subject	Low		High		Mid	
	Like	Dislike	Like	Dislike	Like	Dislike
1	37	7	34	5	30	3
2	66	25	62	40	18	67
3	46	0	63	26	60	39
4	64	52	45	45	56	41
5	41	65	28	71	14	59
6	30	37	23	14	14	58
7	69	25	47	65	72	38
8	47	51	70	56	0	69
9	21	12	17	65	53	9
10	48	48	60	29	65	24
11	48	44	39	52	69	25

以上の実験を被験者 11 人に対し行った．得られたデータを以下に纏める．なお，すべての被験者は 20 歳代から 30 歳代の男性および女性である．

視線情報 $\{(g,t)\}$　視線情報は画像提示開始からの時間 t とディスプレイ上における視線先 g とで構成される時系列データである．g はディスプレイ上の座標 (x,y) の成分を持つ．

選好情報 y　選好情報は y はある画像が与えられた時，好き (1) か (-1) が好きでないかで二値を取る値である．

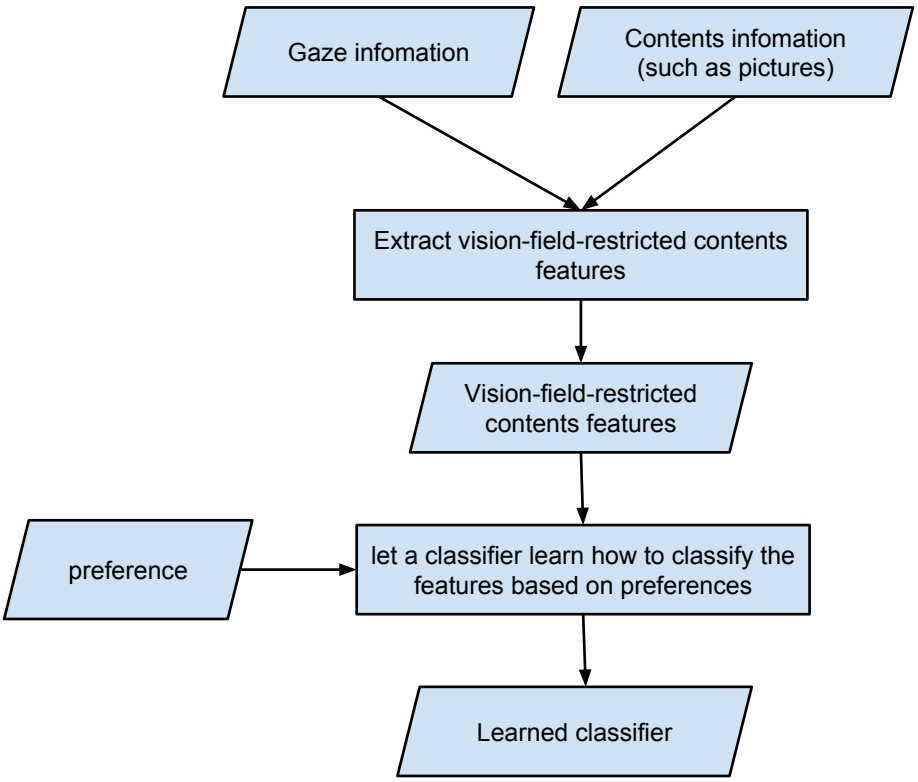

図 4.6: 視野制限付き画像特徴量による選好推定器の学習方法

4.2　提案手法

本提案手法では，基本的には 3 章と同様に Random Forest を用いて各種情報から推定する手法を用いる．本提案手法を図 4.6 に示す．3 章の手法と異なる点は視線情報と画像を組み合わせて特徴量を抽出する点にある．2 章の関連研究の通り，視線が選好に対し意味があると仮定したためである。もし，2.2 節の通り，視野とシーン認識率に関係があるのならば，推定に使う画像情報を視野を基準に変えることで推定精度に変化が起こる可能性がある．そこで，本提案手法では推定に使う画像情報を視野で制限することを考える．これを実現するため，視野制限付き画像特徴量を提案する．

4.2.1　視野制限付き画像特徴量の作成方法

視野制限付き画像特徴量は，画像全体から特徴を抽出するアルゴリズムを中心視野のみ，もしくは周辺視野のみに適用して得られた特徴量である．視野制限付き画像特徴量の例として，4.3.1 節の画像特徴量に視野制限をかけた特徴量を説明する．4.3.1 節の画像特徴量は画像に対し一様に dense sampling を行い，パッチを生成する．ここで視野制限をかけるためにサンプリングの範囲を視野の角度で制限する．制限の仕方は 2 種類ある．1 つは視線先周辺，すなわち中心視野にあるサンプル点のみをサンプリングする方法である．もう

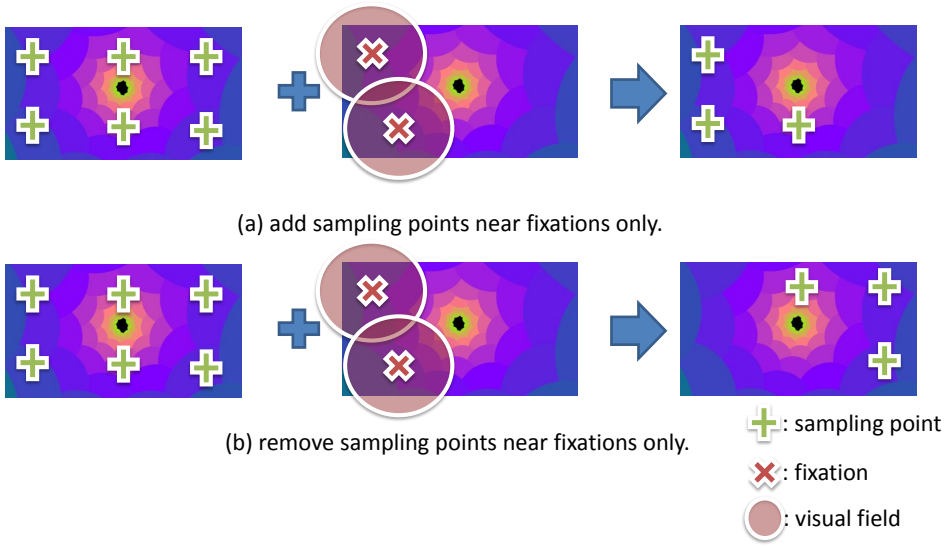

図 4.7: 視野による特徴量抽出範囲の変更

1 つは視線外，すなわち周辺視野のみにあるサンプル点のみをサンプリングする方法である．視野制限をかけたサンプリングの例を図 4.7 に示す．図 4.7 (a) では中心視野にあるサンプル点を利用している例であり，図 4.7 (b) では周辺視野にあるサンプル点を利用している例である．このように視野によりサンプリング範囲を制限することで視野制限付き画像特徴量を計算することができる．

視野制限付き画像特徴量によりユーザの視線情報と画像特徴量を組み合わせることができるため，選好推定に何らかの影響をおよぼすと考えられる．

4.3　実験方法

本実験では提案手法の影響を調べるため，実験を行った．以降の節では，検証のために実験の手続きについて述べる．

4.3.1　特徴量の抽出方法

今回，視野制限付き画像特徴量を計算するため，まず，実験によって得られた視線情報 $\{(g,t)\}$ を視線の移動速度により Fixation と Saccade に区分した．区分するアルゴリズムは視線の移動速度を基にしたアルゴリズム [34] を利用した．本実験ではこの区分に用いた視線の移動速度を 6[deg/s] とした．なお，固視の継続時間が 50[ms] を満たない固視は取り除くこととした．視野制限付き画像特徴量に使うアルゴリズムは節の説明と同様に BoF を用

いる．記述子は節と同様に SIFT と色を用いる．視野制限付き画像特徴量の計算に用いられる中心視野の範囲として，5 度，10 度，15 度の 3 つのパターンを用いた．

4.3.2　実装手法

実装には，画像特徴量抽出は OpenCV で行い，認識は R 言語の randomForest パッケージを用いた．識別器には Random Forest を用いた．決定木は 4 万本使用した．決定木を深くする基準として，元のサンプル数の 2/3 以上のサンプルが葉ノードに存在することとする．

4.3.3　評価方法

ディスプレイ上に表示された画像を被験者が比較的好きと思っているか，あるいは好きででないと思っているかを推定するタスクに関して提案手法の有効性を確かめる評価実験を行った．評価実験は被験者と画像カテゴリの各組み合わせによって行われた．なお，選好情報の量子化によってポジティブサンプルとネガティブサンプルとの比率が 1:3 または 3:1 よりも偏った場合 (被験者 1 すべてのカテゴリ，被験者 3Low context カテゴリ，被験者 8Mid context カテゴリ，被験者 9Mid context カテゴリ) は評価しないものとする．この結果，組み合わせの数は 27 パターン $(= 11(\text{subject}) \times 3(\text{category}) - 6(\text{exception}))$ となる．評価方法は Leave one out で，評価指標は正しく推定できたかの確率，すなわち認識率とする．

4.4　実験結果

提案手法の有効性を確かめるために，ディスプレイ上に表示された画像を被験者が比較的好きと思っているか，あるいは好きででないと思っているかを推定するタスクに関して評価実験を行った．各カテゴリ，視野の定義をした場合のそれぞれ認識率を図 4.8 に示す．また，図 4.8 の具体的な数値を表 4.3 に示す．図 4.8 中の「Ideal」は，各被験者の中にある視野の広さを選んだ中で最も良い認識率を選んだ場合である．視野の広さで選んだ場合の認識率とすべてのサンプル点した場合の認識率を比べると，本手法はすべてのサンプル点を利用するよりも認識率は高かった．また，カテゴリが持つ意味が深くなるほど，視野の広さで選んだ認識率とすべてのサンプル点した場合の認識率の差が大きくなった．中心視野を 5 度とした周辺視野からサンプリングした場合の認識率とすべてのサンプル点を使った場合の認識率はほぼ等しかった．

4.5　考察

カテゴリが持つ意味が深くなるほど，視野の広さで選んだ認識率とすべてのサンプル点した場合の認識率の差が大きくなったのは，画像がもつ意味をユーザが解釈した結果が視線情報に現れたためと考えられる．次に，周辺視野のみのサンプル点を使用した場合とす

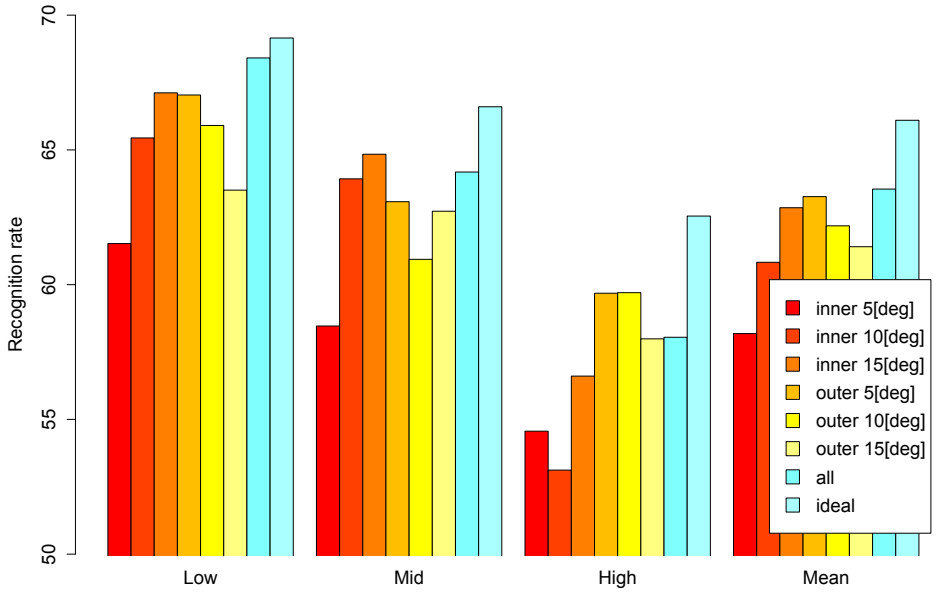

図 4.8: 順次提示条件下における実験結果

表 4.3: 並列条件下における視野制限付き特徴量とすべての特徴量を使用した際の推定精度 [%]

Visual Field	Low	Mid	High	Mean
Central 5[deg]	61.5	58.5	54.6	58.2
Central 10[deg]	65.4	63.9	53.1	60.8
Central 15[deg]	67.1	64.8	56.6	62.9
Peripheral 5[deg]	67.0	63.1	59.7	63.3
Peripheral 10[deg]	65.9	60.9	59.7	62.2
Peripheral 15[deg]	63.5	62.7	58.0	61.4
Original	68.4	64.2	58.1	63.6
Ideal	69.2	66.6	62.6	66.1

べてのサンプル点を使った場合の認識率はほぼ等しく，また中心視野のみから抽出した場合の認識率がひくかったことから，選好に関しては必ずしも中心視野のみが関わってくるわけではないことが判明した．これは画像の選好には中心視野だけの情報でなく，中心視野外の周辺視野を積極的に利用しているためだと考えられる．このことは人間のシーン理解において中心視野よりも周辺視野のほうが役立つという Larson らの研究 [17] も指摘されているように，画像を基にした認識問題は必ずしも中心視野が役立つとは限らない．しかし，各被験者でサンプリング方法ごとで認識率が異なることから，必ずしも周辺視野のみが推定に役立つというわけではない．ここから，被験者ごとでどの視野が推定に役立つかを推定する必要があることがわかった．この結果は Yanulevskaya らの研究 [3] に対して異なる結果，すなわち中心視野がポジティブな印象やネガティブの印象が主に影響しているという結論と異なっており，この違いについて述べるためにはより調査が必要であると考えられる．

第5章

順次提示した場合における動画の選好推定

第3章と第4章では画像の選好を推定したが，このような実験の場合，テレビ広告やテレビ番組のような動画提示された場合，選好を推定することはできない．そこで，ディスプレイに動画が表示された時の選好を推定することに着目する．本実験では，画像の選好で得られた知見を基に動画の選好推定を適用する前に視線情報と画像を組み合わせて動画の選好はそもそも推定可能なのかを目的とする．以下の節では本実験とその結果，および考察について詳細に記述する．

5.1 データ収集

提案手法を評価するために，映像とそれに対応する視線情報を集めた．映像はNHKクリエイティブ・ライブラリー[1]から84本取得した．これらの映像の再生時間は15秒から20秒である．視線情報の収集では映像の明るさと瞳孔の大きさを対応付けるための補正用映像を部屋を暗くした状態で視聴させた．映像の視聴時には，ディスプレイと被験者の目との距離を約60cmとした．また，映像視聴時の視線情報を取得するため，表5.1に示される視線計測器とディスプレイを用いた．次に，映像の視聴と休憩を行った．映像の視聴では21本の映像を再生し，視聴させた．映像の再生順はランダムとした．以上を4回繰り返し，最後に正解ラベルをつけるために全映像に対して3段階による選好の相対評価をさせた．この相対評価では，映像を好きという基準で並べたとき，上位1/3以上に「好き」というラベルを，上位2/3以下を「好きでない」というラベルを付けさせた．なお，この間に相当する「上位1/3よりも下から上位2/3よりも上である映像」は推定には使用しなかった．以上の実験は被験者10人に対して行われた．被験者はすべて20代の男性である．これらのうち，映像視聴中において視線が計測できた割合が85%以下のサンプルは除外した．この結果，得られたサンプル数は525個となった．このうち,「好き」であるサンプルは271個,「好きでない」であるサンプルは254個であった．

[1] http://www.nhk.or.jp/creative/

表 5.1: 順次動画提示条件下での実験で使用した視線計測器とディスプレイの仕様

Eye tracker	type	Tobii TX300
Display	sampling rate	60[Hz]
	size	23”
	resolution	1920[px]×1080[px]

5.2　提案手法

本手法では映像への選好を関連付けた視線情報や映像から抽出した特徴量をコンピュータに学習させることによって，映像への選好が関連付けられていない特徴量から視聴者が持つ映像への選好を推定できるようにする．推定の概要を図 5.1 に示す．学習段階では，視線情報から抽出された特徴量と映像から得られた特徴量と，それらに関連付けられた選好を用いて，特徴選択と識別器を学習させる．今回，使用する選好は「好き」と「好きでない」の 2 種類である．推定段階では，特徴量からラベルを推定する．本手法では特徴選択には Ramdom Forest を，識別器は Random Forest を用いた．

5.2.1　視線特徴量

視線情報は内部状態に関わっていることをこれまでに述べた．そこで，視線情報を選好推定に利用するため，視線情報から表 5.2 のように平均や標準偏差などの統計量を計算した．視線情報を区分する条件として，今回，以下のものを用いた．

サッカード　今回，20[ms] の間に 30[deg/sec] 以上の眼球運動が発生したとき，サッカードが起きたとする．このサッカードの間で視線が止まる状態を固視と呼ぶことにする．

瞬き　今回，100ms から 300ms の間，視線情報を測定できないとき，瞬きを行なっているとした．

瞳孔の直径は内部状態によるものだけでなく，目から入る光量の影響を受ける．そこで光量の影響を考慮するために，表 5.2 中にある補正した瞳孔の直径は光量から予測される瞳孔の直径と観測された瞳孔の直径の比によって計算する．光量から予測される瞳孔の直径 $p\left(Y_t, Y_{t-1}\right)$ は，式 (5.1) で求める．

$$p\left(Y_t, Y_{t-1}\right) = a \times \tan(\log(Y_t)) + b \times \tan(\log(Y_{t-1})) + c \tag{5.1}$$

ここで，Y_t は t フレームにおける映像の輝度である．Y_t は t フレームの画素値の平均 (RGB 値) から ITU-R BT.709 における YCbCr への変換から計算する．係数 a, b, c は映像の輝度と瞳孔の直径を対応付けたデータから回帰分析により推定した．また，瞳孔の直径の周期的変化を表現するために，パワースペクトルを求めた．

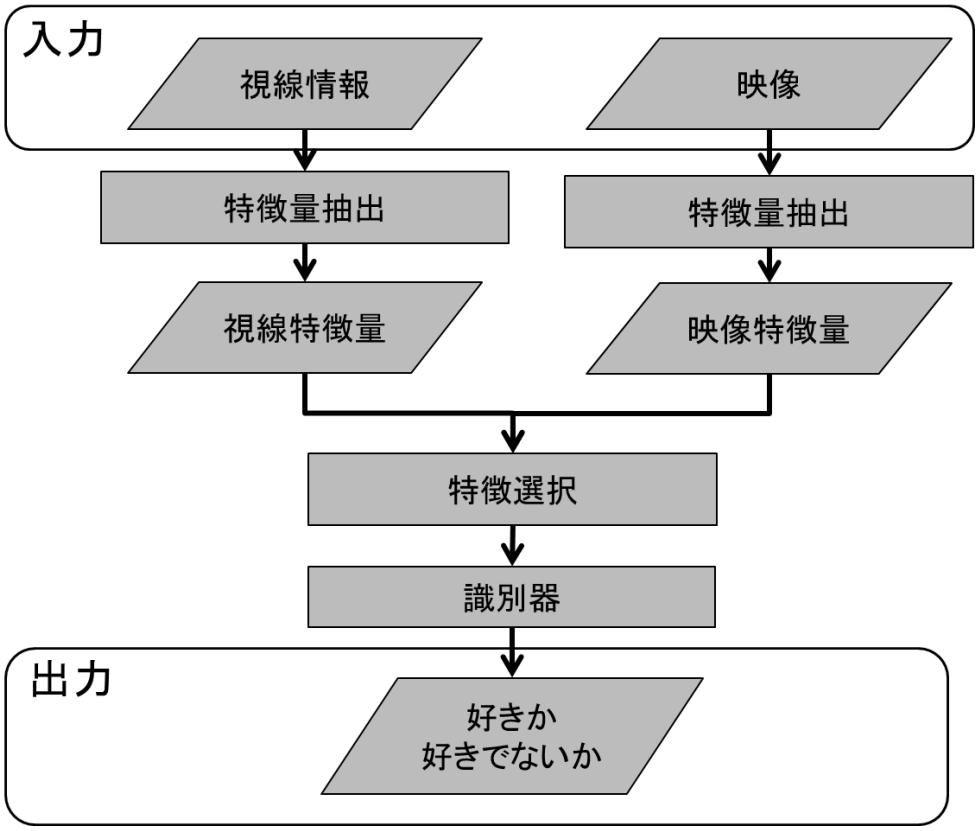

図 5.1: 推定の概要

表 5.2: 視線特徴量の抽出方法

視線情報	抽出方法
固視	1 秒あたりの発生回数 固視継続時間の平均 固視継続時間の標準偏差
サッカード	1 秒あたりの発生回数 視線の移動量 [px] の平均 視線の移動量 [px] の標準偏差
瞬き	1 秒あたりの発生回数 瞬きにかかる時間の平均 瞬きにかかる時間の標準偏差 瞬きにかかる時間の総和 /映像の再生時間
瞳孔の直径	標準偏差 パワースペクトルの面積
補正した瞳孔の直径	標準偏差 パワースペクトルの面積

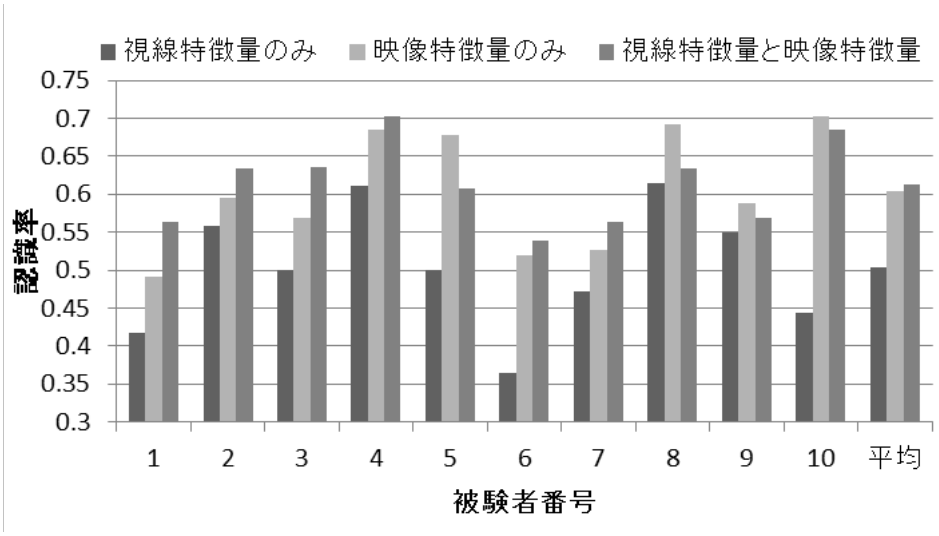

図 5.2: 評価実験結果

5.2.2　画像特徴量

映像から色特徴や形状特徴を抽出し，それらから画像特徴量を抽出する．今回は，映像から画像を 1 コマ抜き出し，その画像を縦横 4 分割し，その分割した領域ごとで色特徴や形状特徴を抽出する．使用する色特徴の抽出方法は HSV ヒストグラムである．なお，今回は HSV のうち，色相と彩度を用いた．形状特徴の抽出方法は Gist[30] を用いた．色特徴と形状特徴は 1 フレームごとに計算し，全フレームで計算した結果の平均と標準偏差を求める．

5.3　実験

評価実験では，収集したデータから視聴者が好きか好きでないと思っているかを提案手法により正しく推定できる割合について調査した．この割合を今回は認識率と呼ぶ．評価実験では被験者ごとで認識率を評価する．評価方法はサンプルごとの Leave one out である．視線情報の有効性を確認するため，特徴量を視線特徴量のみ，画像特徴量のみ，視線特徴量と画像特徴量を連結した場合の 3 通りで行う．実験の結果，認識率は図 5.2 のとおりとなった．選ばれた特徴量の中で，視線特徴量と画像特徴量を連結した場合において，平均 61.4%と最も高い認識率となった．

5.4　考察

画像特徴量のみと 2 つの特徴量を連結させた場合でも認識率に大きな違いが見られなかった原因は特徴選択により選ばれた特徴が画像特徴量が多く推定に使われた特徴がほぼ同じであったためと考えられる．画像特徴量のみで推定できる理由としては，映像の色やスロー

モーション映像などの動きから推定しているためと考えられる．一方で図 5.2 の中でも被験者 3 は特徴選択後も視線特徴量と画像特徴量が推定に使用されており，これが認識率の向上につながったと考えられる．以上の実験結果から，視線情報と画像を組み合わせた動画の選好は推定可能だと考えられる．

第6章

結論

本論文ではマルチメディアコンテンツの選好を推定する方法を実現した．このアプローチとして，視線の情報と画像の情報，それらに関連付けられた選好の情報から機械学習を用いて機械的に推定する手法を用いた．

1章では本論文の目的とアプローチについてまとめた．2章では本論文に関連する研究について調査した結果を報告し，本研究の意義について議論した．3章においては並列提示した場合における画像の選好を推定した結果を分析し議論した．ここで選好とは対象物が好きか好きでないかを表す情報とする．実験の結果，画像の情報よりも視線の情報が推定に有効である場合が有意にあると認められた．4章においては順次提示した場合における画像の選好を推定した結果を分析し議論した．実験の結果，画像のみで推定した時と比べ画像と視線情報を組み合わせて推定したときのほうが精度が平均的に向上した．5章においては順次提示した場合における動画の選好を推定した結果を分析し議論した．実験の結果，画像のみで推定した時と比べ画像と視線情報を組み合わせて推定したときのほうが精度が平均的に向上した．

本論文を通じて判明したのは，マルチメディアの選好を推定する際には画像だけではなく画像と視線情報を組み合わせることで推定精度を向上させられるということである．これは眼球運動がユーザの内部状態を反映することからだと考えられる．

本研究が持つ今後の課題として，画像の選好推定において得られた知見を動画の選好推定に用いることがある．画像の選好推定の手法をそのままに動画の選好推定に用いることは難しい．たとえば，画像では写っている対象が静止しているのに対し，動画では写っている対象が動いていることために，視線情報の取り扱い方が大きく変わる点である．この他にも画像の審美的品質の推定をそのままに動画に適用できない問題もある．

このほかの課題にも，視線や表情などのユーザの行動から推定するマルチメディアコンテンツの選好推定システムを応用した，マルチメディアコンテンツ推薦システムの実現がある．これまでにマルチメディアコンテンツの推薦には選好情報のみで行う研究がある [35]．しかし，選好情報のみではコンテンツの持つ意味までを考慮した推薦はできない．そこでユーザの行動からコンテンツの意味を推察することで，よりよい視聴体験を提供することができると考えられる．

謝辞

本研究を進めるにあたり，充実した研究環境を与えて頂くとともに，多大なる御指導と御鞭撻を賜りました東京大学生産技術研究所 佐藤洋一教授に深く感謝いたします．お忙しい中時間を割いて頂き，研究の進路やテーマについて相談に乗って頂きました．

御多忙の中，本論文の執筆に際し細部に到るまで懇切丁寧な御助言を頂きました東京大学生産技術研究所 菅野裕介特任助教に深く感謝致します．国内外での活動がお忙しい中にもかかわらず，丁寧かつ的確な御指導を頂きました．本当にありがとうございました．

共同研究者である NHK 放送技術研究所の奥田誠様，苗村昌秀様，藤井真人様には研究のアドバイスや実験用のデータを頂きました．深謝致します．

松川徹特任助教には研究室での日常的な活動や，研究の相談まで，幅広く支援していただきました．厚くお礼申し上げます．

種々の事務手続きなどで大変お世話になった秘書の鈴木咲恵さんや今川洋子さん，薄井千緒さんに深くお礼申し上げます．日々の研究室生活やディスカッション，ミーティング等を通して数多くの御助言を頂いた佐藤研究室の皆様に深く感謝いたします．

本研究で使用した実験データの収集にあたり，佐藤研究室のメンバーの方々，私の友人の方々にご協力頂きました．本当にありがとうございました．

最後に研究や将来に挫けそうになりながらも支えてくださった父と母に感謝します．

尾崎 安範

参考文献

[1] K. Helga, “Gross anatomy of the eye”. http://webvision.med.utah.edu/book/part-i-foundations/gross-anatomy-of-the-ey/

[2] S. Bhattacharya, R. Sukthankar, and M. Shah, “A framework for photo-quality assessment and enhancement based on visual aesthetics,” Proceedings of the International Conference on Multimedia, pp.271–280, MM ’10, ACM, New York, NY, USA, 2010. http://doi.acm.org/10.1145/1873951.1873990

[3] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe, “In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings,” Proceedings of the 20th ACM international conference on MultimediaACM, pp.349–358 2012.

[4] 正志西山，孝弘岡部，いまり佐藤，洋一佐藤，“審美的品質識別のための局所領域の組合せによる色彩調和の評価 (光学的解析, 画質改善,¡特集¿画像の認識・理解論文),” 電子情報通信学会論文誌. D, 情報・システム，vol.94，no.8，pp.1324–1334，aug 2011. http://ci.nii.ac.jp/naid/110008686483/

[5] P. Ekman and WallaceV. Friesen, “Measuring facial movement,” Environmental psychology and nonverbal behavior, vol.1, no.1, pp.56–75, 1976. http://dx.doi.org/10.1007/BF01115465

[6] “居眠り、脇見防止技術実用化へ　デンソーが開発本格化,” 2012. http://sankei.jp.msn.com/economy/news/121021/biz12102120280008-n1.htm

[7] 大野健彦，“視線から何がわかるか,” 認知科学，vol.9，no.4，pp.565–579，2002.

[8] R.B. Goldstein, E. Peli, S. Lerner, and G. Luo, “Eye movements while watching video: comparisons across viewer groups,” Journal of Vision, vol.4, no.8, p.643, 2004. http://www.journalofvision.org/content/4/8/643.abstract

[9] 映像情報メディア学会，視覚心理入門: 基礎から応用視覚まで，オーム社，2009. http://books.google.co.jp/books?id=DqWNPgAACAAJ

[10] M. Argyle, Bodily Communication, University paperbacks, Methuen, 1988. http://books.google.co.jp/books?id=crYOAAAAQAAJ

[11] Z.M. Hafed and J.J. Clark, “Microsaccades as an overt measure of covert attention shifts,” Vision Research, vol.42, no.22, pp.2533–2545, 2002. http://www.sciencedirect.com/science/article/pii/S0042698902002638

[12] E.H. Hess and J.M. Polt, “Pupil size as related to interest value of visual stimuli,” Science, vol.132, pp.349–350, Aug. 1960. http://www.ncbi.nlm.nih.gov/pubmed/14401489

[13] R. Subramanian, V. Yanulevskaya, and N. Sebe, “Can computers learn from humans to see better?: inferring scene semantics from viewers’ eye movements,” Proceedings of the 19th ACM international conference on Multimedia, pp.33–42, MM ’11, ACM, New York, NY, USA, 2011.

[14] S. Eivazi, R. Bednarik, M. Tukiainen, M. von und zuFraunberg, V. Leinonen, and J.E. Jääskeläinen, “Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings,” Proceedings of the Symposium on Eye Tracking Research and Applications, pp.377–380, ETRA ’12, ACM, New York, NY, USA, 2012. http://doi.acm.org/10.1145/2168556.2168641

[15] C.A. Curcio, K.R. Sloan, R.E. Kalina, and A.E. Hendrickson, “Human photoreceptor topography.,” The Journal of comparative neurology, vol.292, no.4, pp.497–523, 1990vvvv.

[16] G. Westheimer, “Directional sensitivity of the retina: 75 years of stilescrawford effect,” Proceedings of the Royal Society B: Biological Sciences, vol.275, no.1653, pp.2777–2786, 2008. http://rspb.royalsocietypublishing.org/content/275/1653/2777.abstract

[17] A.M. Larson and L.C. Loschky, “The contributions of central versus peripheral vision to scene gist recognition,” Journal of Vision, vol.9, no.10, pp.●●–●●, 2009. http://www.journalofvision.org/content/9/10/6.abstract

[18] R. Yonetani, H. Kawashima, and T. Matsuyama, “Multi-mode saliency dynamics model for analyzing gaze and attention,” Proceedings of the Symposium on Eye Tracking Research and Applications, pp.115–122, ETRA ’12, ACM, New York, NY, USA, 2012. http://doi.acm.org/10.1145/2168556.2168574

[19] W.-T. Peng, W.-T. Chu, C.-H. Chang, C.-N. Chou, W.-J. Huang, W.-Y. Chang, and Y.-P. Hung, “Editing by viewing: Automatic home video summarization by viewing behavior analysis,” Multimedia, IEEE Transactions on, vol.13, no.3, pp.539–550, june 2011.

[20] Y. Sawahata, R. Khosla, K. Komine, N. Hiruma, T. Itou, S. Watanabe, Y. Suzuki, Y. Hara, and N. Issiki, “Determining comprehension and quality of tv programs

using eye-gaze tracking," Pattern Recognition, vol.41, no.5, pp.1610–1626, 2008. http://www.sciencedirect.com/science/article/pii/S003132030700444X

[21] K. Yukiyasu and T. Frank, "Decoding the visual and subjective contents of the human brain," Nat Neurosci, vol.8, no.5, pp.679–685, may 2005.

[22] H. Ueno, M. Kaneda, and M. Tsukino, "Development of drowsiness detection system," Vehicle Navigation and Information Systems Conference, 1994. Proceedings., 1994, pp.15–20, aug-2 sep 1994.

[23] Y. Negishi, Z. Dou, and Y. Mitsukura, "Estimation system for human-interest degree while watching tv commercials using eeg," Proceedings of the 18th international conference on Neural Information Processing - Volume Part I, pp.46–53, ICONIP'11, Springer-Verlag, Berlin, Heidelberg, 2011.

[24] 高橋正樹，"映像解析による人物動作理解に関する研究," PhD thesis，総合研究大学院大学，2012.

[25] S. Shimojo, C. Simion, E. Shimojo, and C. Scheier, "Gaze bias both reflects and influences preference," Nat Neurosci, vol.6, no.12, pp.1317–1322, 2003.

[26] Y. Sugano, H. Kasai, K. Ogaki, and Y. Sato, "Image preference estimation from eye movements with a data-driven approach," 3rd International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction, pp.●●–●●, PETMEI'13, 2013.

[27] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," Computer Vision (ICCV), 2011 IEEE International Conference on, pp.1784–1791, 2011.

[28] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp.33–40, June 2011.

[29] L. Breiman, "Random forests," Machine Learning, vol.45, no.1, pp.5–32, 2001. http://dx.doi.org/10.1023/A

[30] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," Int. J. Comput. Vision, vol.42, no.3, pp.145–175, May 2001. http://dx.doi.org/10.1023/A:1011139631724

[31] 貴之岡谷，健　増田，浩一黄瀬，啓司柳井，俊和和田，宗樹安田，駿　片岡，和之田中，康史八木，英雄斎藤，コンピュータビジョン最先端ガイド 3，CVIM チュートリアルシリーズ，アドコム・メディア，2010.

[32] D.G. Lowe, “Object recognition from local scale-invariant features,” Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol.2, pp.1150–1157vol.2, 1999.

[33] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” Computer vision and pattern recognition (CVPR), 2010 IEEE conference onIEEE, pp.3485–3492 2010.

[34] T. Judd, F. Durand, and A. Torralba, “Fixations on low-resolution images,” Journal of Vision, vol.11, no.4, pp.●●–●●, 2011. http://www.journalofvision.org/content/11/4/14.abstract

[35] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath, “The youtube video recommendation system,” Proceedings of the Fourth ACM Conference on Recommender Systems, pp.293–296, RecSys ’10, ACM, New York, NY, USA, 2010. http://doi.acm.org/10.1145/1864708.1864770

発表文献

[1] Yusuke Sugano, Yasunori Ozaki, Hiroshi Kasai, Keisuke Ogaki, Yoichi Sato, "Image Preference Estimation with a Data-driven Approach: A Comparative Study between Gaze and Image Features," Journal of Eye Movement Research, 2013 (submitted)

[2] 尾崎 安範, 菅野 裕介, 佐藤 洋一, "視線情報と画像特徴に基づく画像の選好推定," 電子情報通信学会技術研究報告. パターン認識・理解, 一般社団法人電子情報通信学会, 2014 (申込済)