百認一取：音声認識と画像認識を統合した対戦システムの構築

吉川唯杜[1]　木村翔真[1]　伊藤壮真[1]　岡本学[1]　筒口拳[1]

概要：本研究は，人間とコンピュータがリアルな場で対戦可能な百人一首システムの構築をめざす百認一取プロジェクトの一環である．これまでに我々は，読み上げ音声から歌番号を識別する音声処理と，取札配置画像から取札領域を抽出して文字認識を行い，歌番号を判定する画像処理についてそれぞれ検討を進めてきた．本研究ではこれらのサブシステムを統合し，取札位置の認識，歌番号の識別，および該当取札への投光を一連の処理として実行する対戦システムを構築した．システムの動作実験から，初学者から中級者の人間と対戦することが十分に可能であるという結果が得られた．

キーワード：百人一首，画像認識，音声認識，対戦システム

Hyakunin Isshu Project: Carta System Integrating Speech Recognition and Image Recognition

YUITO YOSHIKAWA[†1]　SHOMA KIMURA[†1]　SOMA ITO[†1]
MANABU OKAMOTO[†1]　KEN TSUTSUGUCHI[†1]

Abstract: This study is part of the Carta Project, which aims to develop a carta system that can play against humans in real-world settings. We previously investigated an audio-processing subsystem that identifies the poem ID from recited speech and an image-processing subsystem that extracts card regions from layout images, performs character recognition, and determines the poem ID. Here, we integrate these subsystems and build a competitive system that recognizes card positions, identifies the poem ID, and projects light onto the target card. Experiments show it can compete with beginner-to-intermediate players.

Keywords: Hyakunin-Isshu, Image recognition, Speech recognition, Game system.

1. はじめに

本研究は，リアルな場で人間とコンピュータが百人一首の対戦を可能とするシステムの構築をめざす「百認一取プロジェクト」の一環である．「リアルな場」とは，実物の取札を畳の上などの現実世界に配置し，読み上げられた歌に該当する取札を物理的に取る，または指定することを意味する．一連の流れとして，人間は視覚で取札の位置を確認し，聴覚で歌を識別し，手で取札を取得する．一方コンピュータは「カメラから取得した画像の認識による取札位置の識別」，「音声認識による歌の識別」，「プロジェクション等による取札の指定」を行うことを想定している．これまでに我々は，取札配置画像から取札領域を抽出する研究[1]や，取札領域を文字認識（OCR; Optical Character Recognition）にかけて歌番号を付与する研究[2]，読み上げ音声を音声認識し読札番号を識別する研究[3][4]をそれぞれ行ってきた．これらの認識精度が十分に活用できる段階であるという仮定のもと，各機能を統合して，システムが試合での一連の動きを可能かの検証を行った．

本稿では，まず，画像処理と音声処理に関するこれまでの研究成果について説明する．次に，取札位置の認識，歌番号の識別，ならびに該当取札への投光を一連の処理として実行する統合システムの提案手法を述べる．さらに，統合システムの精度評価実験を行い，得られた結果に基づいて性能および誤り要因について考察する．加えて，実際のイベントにおける使用事例を示し，実環境下での動作状況と運用上の知見を報告する．最後に，本研究のまとめと今後の課題について述べる．

2. 先行研究

2.1 画像処理の先行研究精度結果

吉川らの研究[5]によると，札領域抽出及び OCR による歌番号付与の精度は，自然光や部屋の照明のみの光を使用する場合は 95.8%であるが，提案手法である照光をした状態で取札配置画像を取得すると，98.2%に向上することを報告している．この OCR 精度は札領域抽出精度と同じく，母数を 50 枚とする成功率の平均値である．OCR は札領域を正しく抽出された後に行わなければ成功しないため，その精度は札領域抽出に依存する．例えばもし札領域抽出成功

1 崇城大学　情報学部
Faculty of Computer & Information Sciences, Sojo University

率が 98%である場合，OCR の精度は 98%以下となる．

同先行研究では，場にある 50 枚の札の内，OCR が原因である 1 枚程度の誤認識があった．理由として，照光によって文字が白飛びした可能性が十分にある．文字を認識しやすくなる先鋭化の処理などを加えていないこともあり，100%に至らなかったことが複数回起こったと考えられている．

本研究においてはこの画像処理システムで実用上問題ないと判断し，現状のまま組み込んで使用している．

2.2 音声処理の先行研究精度結果

木村らの研究[4]によると，読み上げられた音声を認識する際のシステムの流れは，スピーカから再生された音声をマイクで受け取り，オープンソースの音声認識システム Julius[6]を使用して，事前作成した百人一首用の辞書に従って判定させている．音声認識システムのハードウェア構成は図 2 で示している．

また，音声処理システムの評価実験において、百人一首の各歌の第 1 句から第 5 句（5・7・5・7・7）を含む辞書（Dictiornary1）を作成し用いた場合で，アナウンサーによる読み上げ音声 100 首の認識精度は 90%であった（表 1）．それに対して，第 1 句と第 2 句のみを含む辞書（Dictionary2）を作成し用いた場合の認識率は，100 首中 94%であった．百人一首の歌には，第 1 句と第 3 句の内容が重複する歌が存在するため，Dictionary2 を用いることで，第 1 句を認識できた際に第 3 句と誤認識するケースに対処している．また，Dictionary1 と Dictionary2 で共通する結果としては，音素的に似た他の句と誤認識したケースが見られた．

先行研究における課題として，入力された音声の音量が閾値以下になった時点で認識を行うため，百人一首特有の歌のような読み上げ方では句と句の間の区切りを見出すことができず，使用する音声の種類（読み手や読み上げ方）に応じて誤認識が増えるケースも見られた．

本研究の統合システムにおいて，リアルタイムの百人一首システムとしての性能を検証するため，実装には Dictionary1 を用いており，その他のシステムはそのまま使用している．

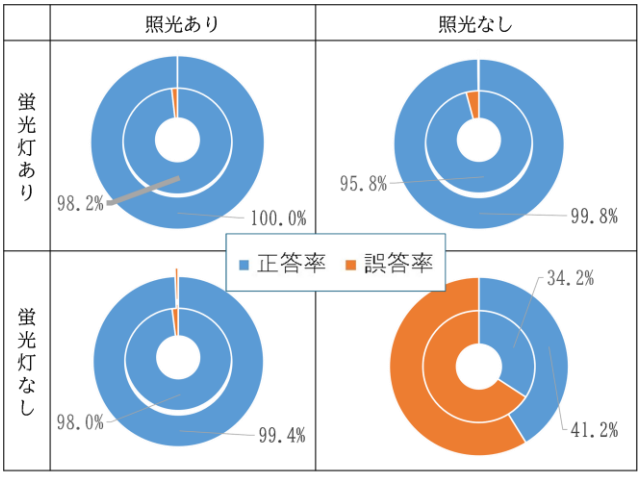

図 1　取札認識率（外円）と OCR 精度（内円）の先行研究結果（文献[5]より引用）

3. 提案手法

3.1 処理の流れ

先行研究で述べた画像処理システムと音声処理システムを組み合わせて，コンピュータがフィールド内でどこに何番の札があるかを知り，読まれた読札と同じ歌番号の札を取るという一連の動作ができるように，以下の処理順にて実装した．①は起動時にのみ実施し，以降は②～⑥を繰り返す．歌の読み上げ前に②，③を行い，歌の読み上げ時には④～⑥を行う．⑥から②に戻るタイミングは，読み上げ音声ファイル[7]の再生が終わった後である．

① プロジェクタを使って長方形の光を畳の上に当て，光が当たっている範囲をフィールドの範囲として座標を取得し，保持する．
② フィールドと定義した範囲内をグレースケール化した後にノイズ処理を行い，二値化よって札領域を抽出し，札の面積がある座標を保存する．
③ 札領域それぞれに対して OCR を行い，OCR 結果と歌の類似度が最も高かった歌番号を，札領域に付与する．
④ 音声認識モジュールは音声処理の歌番号出力待機状態にして，並行処理の形で読み上げ音声ファイルをランダムに再生する．
⑤ 再生された音声をマイクから Julius へ入力し，

図 2　音声認識実験のハードウェア構成（文献[4]より引用）

表 1　先行研究での比較実験の結果（100 首中）（文献[4]より引用）

音声	認識率（Dictionary1）	認識率（Dictionary2）
NHK 音声	90%	94%

Julius の中で事前作成した辞書に従って最も近い歌番号を出力する．

⑥ ⑤にて出力された歌番号と合致する札領域に赤色の光を照射する（札の取得を意味する）．

3.2 ハードウェア構成

本統合システムは図 3 のようにデバイスを構成，配置している．デバイスはパソコン，プロジェクタ，ビデオカメラ，マイク，スピーカを使用している．

4. 統合システムの実験

4.1 実験内容

本システムの認識精度評価を行う．畳の上に歌番号 1～50 の計 50 枚の取札を並べ，配置を変更した 10 パターンを実行する．評価指標は正解札に赤の光を照射できた札数を取札取得数とし，「精度 ＝ 取札取得数÷50 枚」という計算によって求める． 10 パターンの精度から算出する平均値（平均精度とする）で評価を行う．

4.2 実験結果

平均精度の結果は 92.0%であった．音声認識が原因の誤認識では歌番号 40 番，44 番，50 番を間違えることが多発した（表 2）．原因は基本的に決まり字の文字数である．決まり字とは，歌を読まれ始めて何文字目で取札を特定できるかを示す，先頭からの数文字のことである．例えば歌番号 15 番と 50 番は第 1 句が同じ「きみがため」であり，第 1 句が読み上げられた時点ではどちらの札か分からない．しかし本システムはこの時点で歌番号を出力したため，誤認識が起こった．画像認識が原因の誤認識は，フィールド内の札数が 3 枚以下である場合に，札領域をうまく抽出できずに取得できないことが多発した．

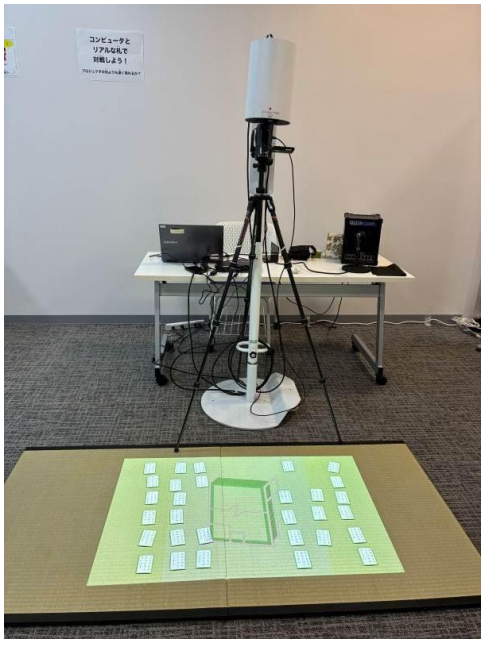

図 3 システム動作時のハードウェア構成

5. イベントでの使用事例

崇城大学ではテクノファンタジーという一般の方々が展示技術などを体験できるイベントが毎年開催されている．そこで本研究の統合システムを設置して，不特定多数からの感触を得ることができた．具体的な数値はないものの，イベント特有のざわざわした会場内でも基本的に 8 割以上は正解札に赤の光を照射することができていた．認識結果が誤っていた場合は，基本的に音声処理が間違った歌番号を出力していることが原因であった．しかし，その誤認識も 6 時間稼働して 20 回に収まっている．フィールド内の札数が 3 枚以下に減少することはなかったため，画像処理が原因の誤認識は 1 回に留まった．

来場される方々は歌を覚えていない方がほとんどであったため，歌番号が出力されてから 7 秒後に赤の光を照射するという難易度設定を設けたところ，未経験者や初学者でも楽しめる難易度になった．この 7 秒という数字は，読み上げる場所で言うと下の句 14 文字 7 文字目前後を読み終える（第 4 句を読み終える）タイミングである．

フィールド内には 20 枚の取札を置き，人間が読みやすいよう，取札は全て人間がいる方向に向けた状態で行った．来場者はお子様連れが多く，親と子の合計 4 人～6 人と同時に対戦した．

課題としては，システム側の赤い光の照射とユーザの手が札を撮るタイミングが同時だった場合の判定が挙げられる．特に上級者が札を払う動作は非常に速く，判定が困難であるため対策として高速カメラの導入などが考えられる．

表 2 実験結果の成功枚数と誤認識した歌番号

試行回数	成功枚数	誤認識札番号	
		音声	画像
1	47	4, 40, 44	-
2	47	-	13, 26, 35
3	45	40, 50	11, 30, 39
4	46	50	5, 12, 34
5	46	44, 50	16, 40
6	45	44, 26	46, 29, 16
7	45	44, 40, 50	2, 9
8	47	40, 50	3
9	47	44	40, 50
10	45	44, 50	6, 21, 37

6. おわりに

本研究では，先行研究で実装した画像処理および音声処理を統合し，取札位置の認識，歌番号の識別，ならびに該当取札への投光を一連の処理として実行するリアルな場での百人一首対戦システムを構築した．

本システムの平均精度率を実験した結果，1 試合につき 4 枚ほど正解札を取得できない結果であった．原因は音声処理と画像処理のどちらも存在している．音声処理は決まり字が多い札に対する課題があり，特定の歌番号が検出された際に歌番号の出力を保留し，認識する区間を延長する等の処理が有効であると考えられる．一方，画像処理はフィールド内の札数が 3 枚以下に減少した際に課題があり，明確な原因は不明であるが，札領域抽出の失敗が判明している．そこで 3 枚以下の条件時，HSV により緑の領域を抽出[1]する処理の導入を検討している．今後はこれらの課題に対処し，基本的に 50 枚すべてを取得できるシステムをめざす．

その他の改善点として，先行研究で指摘した照光条件を原因とする OCR の白飛びや，百人一首特有の音声認識の難しさは，引き続き誤認識の原因となり得る．

更に，実環境での動作確認として，イベント会場に本システムを設置して運用した結果，会場特有の雑音化においても，概ね 8 割以上の割合で正解札に照射できることを確認した．動作中の誤認識は，平均精度率の実験と同種のものに加えて，OCR を起因とする誤りが 1 回のみ生じた．また，現状の性能でも初学者から中級者程度の人間との対戦が十分に可能であることが分かった．

さらに，イベントでは競技経験者や想定外の挙動を行う参加者が多くは見られなかったため，競技かるた特有の技や細かなルールに対して十分に適応できているかは未検証である．今後は，競技者を含む多様な参加者を対象とした評価や，本番の試合環境に合わせて運用条件を変化させるなど，実戦環境におけるシステムの有用性を確認する必要がある．

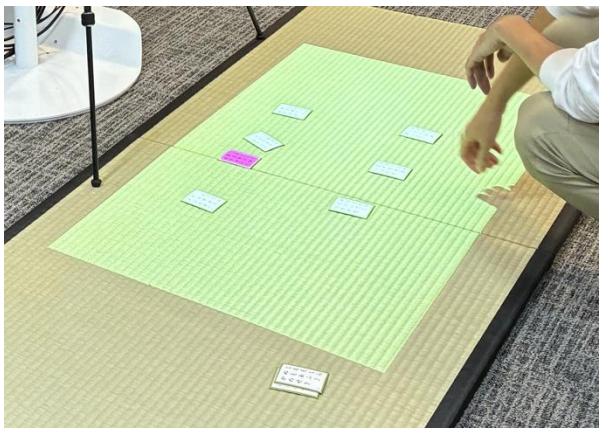

図 4 イベント稼働時の様子

参考文献

[1] 伊藤壮真，長嶺和紀，吉川唯杜，角田唯隼，佐藤礼一朗，岡本学，筒口拳：百認一取：色情報を用いた実画像からの取札領域抽出，火の国情報シンポジウム 2024．
[2] 吉川唯杜，佐藤礼一郎，安慶直哉，長嶺和紀，岡本学，筒口拳：百認一取(2):取札画像のサイズに対する文字読み取り精度 の評価，電子情報通信学会九州支部学生会講演会，D-27，2023．
[3] 濱武右京，木村翔真，黄思韵，柴田美桜，筒口拳，岡本学：百認一取(3):歌読み上げのための音声認識システムの検討，電子情報通信学会九州支部学生会講演会，D-28，2023．
[4] 木村翔真，濱武右京，黄思韵，柴田美桜，筒口拳，岡本学：百認一取：歌読み上げのための音声認識システムの評価，火の国情報シンポジウム 2024．
[5] 吉川唯杜，伊藤壮真，木村翔真，岡本学，筒口拳：百認一取：カメラ・プロジェクタを用いたリアル取札認識システムの構築，２０２５年度電子情報通信学会九州支部学生会講演会・講演論文集，D-30，2025．
[6] 河原達也，李晃伸：連続音声認識ソフトウエア Julius，人工知能学会誌，Vol.20，No.1，pp.41-49，2005．
[7] NHK ：NHK アーカイブス，〈https://www.nhk.or.jp/archives/creative/〉（閲覧 2026/02/13）．