2014 年度 卒業論文

相補的口唇トラッキング

2015 年 2 月 16 日

大阪大学 基礎工学部
システム科学科 生物工学コース
加藤 弘樹

主査:_____　日付:_____

相補的口唇トラッキング

加藤 弘樹

概 要

口唇動作は発話と深い関係があり，口唇動作を用いた読唇や発話訓練支援などの研究が行われており，口唇形状を計測することは重要であると考えられる．本研究では，距離画像センサと高解像度カメラを用いる口唇トラッキングシステムを提案する．

本システムは，ユーザを拘束しない条件下において高精度かつ高速な計算速度を有する口唇トラッキングを行うため，高解像度カメラと距離画像センサを用いる．高解像度カメラを用いることで高精度な口唇トラッキングが期待されるが，計算コストが増加し実時間的な計測が難しい．距離画像センサを用いて口唇位置を追跡することで口唇を含む最小限の領域を切り出し，切り出しにより得た領域において口唇トラッキングを行い，計算速度の向上を実現する．まず，距離画像センサで骨格推定を用いた顔の追跡を行い，顔モデルを用いることで口唇位置を取得する．次に距離画像センサを用いて取得した口唇位置は距離画像センサと高解像度カメラの座標変換により高解像度カメラにおける口唇位置へと変換され，その座標を元に切り出しを行う領域を決定する．

本システムによる口唇領域抽出の計算速度がカメラのフレームレートに対して充分であることを検証するため，口唇領域の抽出を行い，計算時間を計測した．結果として，ユーザの顔を追跡し領域を抽出する場合の平均の計算時間が 1.04×10^{-4} 秒であり，カメラのフレームレートである 30 fps に対して充分な計算速度で口唇領域を抽出することが出来た．

キーワード : 口唇，動作習得，センサフュージョン，画像処理

Complemental lip tracking

Hiroki Kato

Abstract

Mouth motion has a close relationship with speech and there are many studies about lip reading, pronunciation training and so on. In this study, the author proposes a lip tracking system that uses multiple cameras, distance image one which tracks the position of the lip and a high resolution one which tracks lip with high precision.

In this proposed system, high resolution camera and distance image sensor are used to track lip allowing natural movement. A high resolution camera enables the system to track lip with high precision, but the calculation costs high. In order to solve the problem, the system trims the picture including lip region and then tracks lip. A division of the lip region is implemented by using distance image sensor. To trim lip region from the picture portrayed by a high resolution camera, the coordinates of deciding region in the distance image sensor are mapped to that in high resolution camera by using transformation.

In the experiment, the author measured the machine time of division of the lip region. As a result, it spent about 1.04×10^{-4} s for tracking the face of the user. The author could confirm that the machine time of this system is short enough to use with cameras which can record videos at 30 fps.

Keywords : lip, motion learning, sensor fusion, image processing

目 次

図 目 次

表 目 次

第1章　序論

発話による会話は人間社会において重要な情報伝達手段の一つである．日常的な挨拶や他愛のない世間話のみならず，宣誓や発表などの重要な場面においても発話による情報発信が行われる．一般に発話において情報の伝達を担うのは音であり，発話と関連して音声に関する研究が広く行われている．一方，会話においては音だけでなく発話者の口唇動作を視覚的に捉えることもあり，幼児の発話習得において口唇動作の観察が必要であると考えられている [1]．発話においては音声と同様に口唇動作が重要な役割を担うと考えられ，発話と口唇動作に関する研究が行われている．視覚情報により計測される口唇動作は環境雑音に影響されないため，雑音環境下などにおける発話内容理解の手段やヒューマンインタフェースとしての機械読唇が研究されており [2]，また発話訓練を想定した場合に音声だけでなく口唇動作を手本として提示することで発話習得補助を行う研究 [3] などが存在する．雑音環境下における読唇の例を図 1.1 に示す．口唇動作と発話に関する研究では室内で単一の人物を対象とする場合が多いが，口唇形状の計測に関してユーザの口唇のみを撮影するためにカメラとユーザの位置関係を固定するなどの束縛を伴う．

口唇形状の計測手法として様々な手法が存在し，画像情報を用いる画像ベース法と口唇の形状を利用した計測を行うモデルベース法に大別される．動的輪郭モデルや Active Appearance Models に代表されるモデルベース法は口形の形状変化に基づいて計測を行うため，画像ベース法に対して認識精度が高く広く利用されるが，計算コストが高く，実時間計測ではユーザの動きを制限するなどして口唇のみを計測するなどの工夫が求められる．ユーザの動きを制限せず顔や体全体を撮影する場合，高精度に口唇形状を計測するためには高解像度のカメラを用いることが望ましいが，モデルベース法を用いた口唇形状計測においては計算時間が必要となるため，発音習得補助やインタフェースなど実時間的に計測する必要がある場合に計算量の多い高精度な形状計測が困難となる．従って，実時間で高精度な口唇形状計測を行う場合，画像全体から口唇を含む最小限の領域を切り出し形状計測に用いるのが望ましいと考えられる．

口唇を含む領域を切り出す方法として，口唇色領域を抽出する方法，テンプレートマッチングによる方法，顔検出を用いて口唇を含む領域を広く抽出する方法 [4] などが挙げられるが，ユーザの動きに対応することが難しく，不十分であると考えられる．これらの方法に対し，深度画像を利用してユーザの骨格を推定し，顔を追跡することでユーザの動きに対応した実時間的な計測が可能であると考えられる．従って，本研究では高精度かつ高速な計算速

度が得られる口唇形状の計測を目的とし，深度画像を撮影するための距離画像センサと高解像度カメラを相補的に用いるセンシングシステムを提案する．本手法では，距離画像センサを利用した顔の追跡に基づき，高解像度カメラから口唇を含む最小限の領域を抽出し，抽出した画像において口唇形状の計測を行う．距離画像センサにより取得した口唇の三次元座標を高解像度カメラにおける点と対応付けるため，距離画像センサと高解像度カメラは予め較正を行う．相補的なトラッキングでは高解像度カメラを用いた高解像度での計測と距離画像センサを用いた計算速度の高速な口唇位置推定を行うことにより，相互に欠点を補うトラッキングが可能であると考えられる．

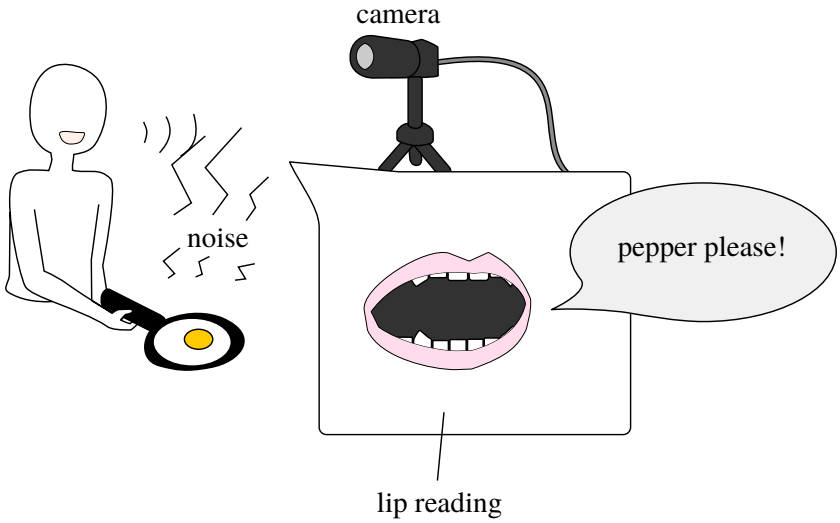


図 1.1: 雑音環境下における読唇

第2章　口唇形状とその計測手法

本章では，口唇形状の計測について述べる．

2.1　口唇形状の計測

人は発音において口を用いて調音を行う．子音では/p/，/b/，/m/，/f/，/v/は唇を用いて調音する音であり，また母音は舌と唇の形により作られている [5]．従って，口唇形状は発音において重要であると考えられ，口唇形状により発話内容を理解する読唇などの研究が行われている [2]．母音発音時の舌と唇の形の例を図 2.1 に示す．口唇形状の変形を考えるにあたり，計測が必要なパラメータとして口唇の三次元形状や輪郭形状，口内領域の形状や面積などが考えられる．発音時の口唇形状を計測するにあたり，人が口唇動作を視認する際口唇の輪郭形状に着目すると考えられるため，口唇の輪郭形状を扱う．本研究では屋内においてユーザの口唇形状計測行う場合を想定するため，ユーザの前面に設置されたカメラを用いて唇の形状を取得する．以下では，画像全体から唇のみを抽出するための技術について述べる．

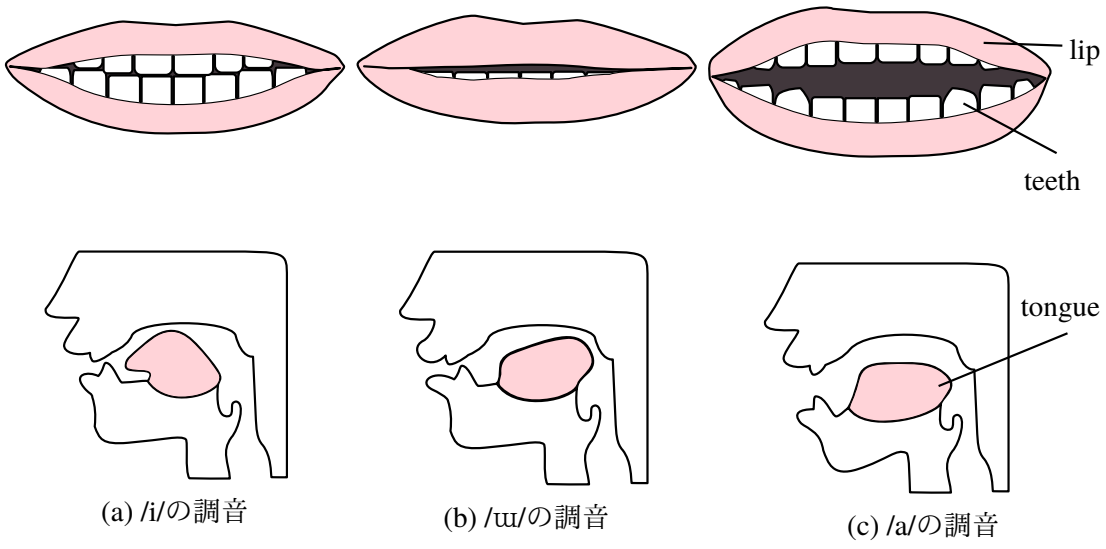
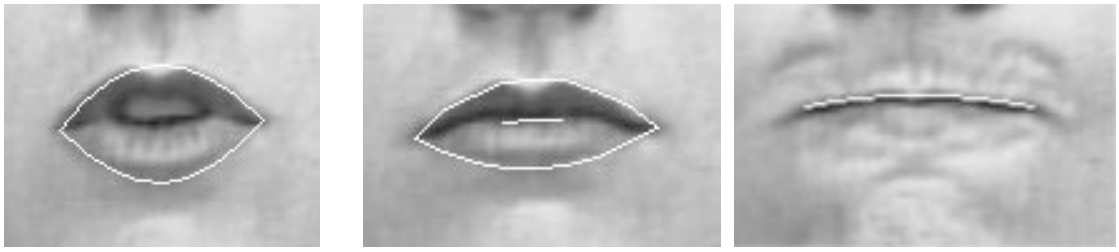


図 2.1: 母音発音時の舌と唇の形の例

2.2 口唇形状計測に関する研究

口唇形状の計測手法は，画像ベース法とモデルベース法に大別出来る．画像ベース法では，画像中の特徴点や輪郭などを利用して唇を計測するが，モデルベース法では唇の幾何的な情報を用いて画像中より唇を計測する．また，計測においてマーカを用いる場合 [3] とマーカを用いない場合 [6] がある．本研究では，機械読唇や動作習得支援を目的としたシステムに用いる計測を対象とするため，唇の計測は長期間において断続的に行われると考えられ，マーカの装着はユーザの負担となる．従って，マーカを用いない計測が適切であると考えられ，以下ではマーカを用いない場合の計測手法について記述する．

Tian らは唇の形状，色，動きの情報を用いることで口唇形状を計測する画像ベースの唇抽出手法を提案した [7]．Tian らの手法では口唇は四つの特徴点により追跡され，特徴点の座標をもとに口唇画像テンプレートを用いて口唇領域を決定している．また，口唇の形状と色情報により口の状態を図 2.2 に示す open，relatively closed，tightly closed の三つに分類し，tightly closed の場合はテンプレートではなく色情報を用いることで輪郭を決定する．訓練が不要であり異なる個人において計測出来るが，口唇形状が非対称である場合に輪郭の抽出精度が落ちるなどの問題が存在する．



(a) open　(b) relatively closed　(c) tightly closed

図 2.2: 口の三状態 [7]

Kass らは，対象物に対して適切なエネルギー関数を定義し，エネルギーが最小となるように曲線を導くことで曲線を画像中の線や輪郭へと収束させることにより輪郭を抽出する動的輪郭モデル (Snakes) を提案し，唇形状の取得への応用例を示している [8]．唇形状が変化した場合に自動で唇形状を追跡している様子を図 2.3 に示す．一般に，曲線は曲線自身の形状に依存するエネルギーと画像から受けるエネルギーに影響を受けるが，エネルギーを適切に定義することにより計測対象の形状の情報を考慮することが可能である．エネルギーの定義により目的の形状となる輪郭を探索出来るが，計算コストが高く，パラメータ調整が難しい．

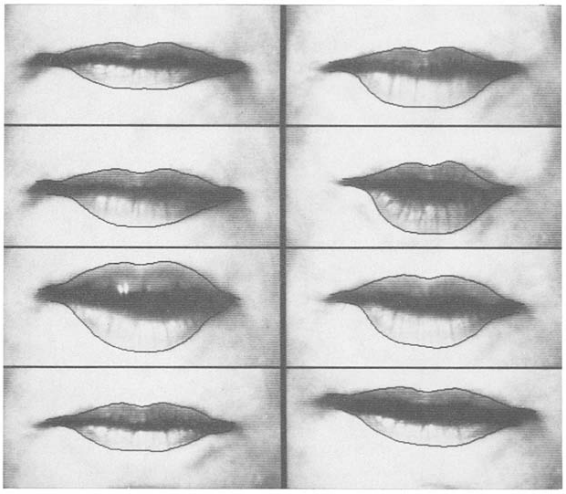


図 2.3: 動的輪郭モデルによる唇形状の追跡 [8]

Active Appearance Models(AAMs) は物体の外観と形状をモデリングすることにより物体の抽出や追跡を行うモデルであり，主に顔認識の手法として用いられる．AAMs では，入力画像の対象物体の外観及び形状を各々モデルの線形和を用いて表す．線形和に用いたパラメータにより対象物体を表現することで対象物体を点の集合として表す．AAMs におけるモデルの例を図 2.4 に示す．Matthews らは AAMs における計算手法を改良することにより計算コストを軽減した [9] ため，AAMs を用いた実時間での顔認識が可能である．顔の方向によらず高精度な顔認識が可能であり，口唇形状の計測も可能であるが，学習データが必要であり使用準備に時間が必要であると考えられる．

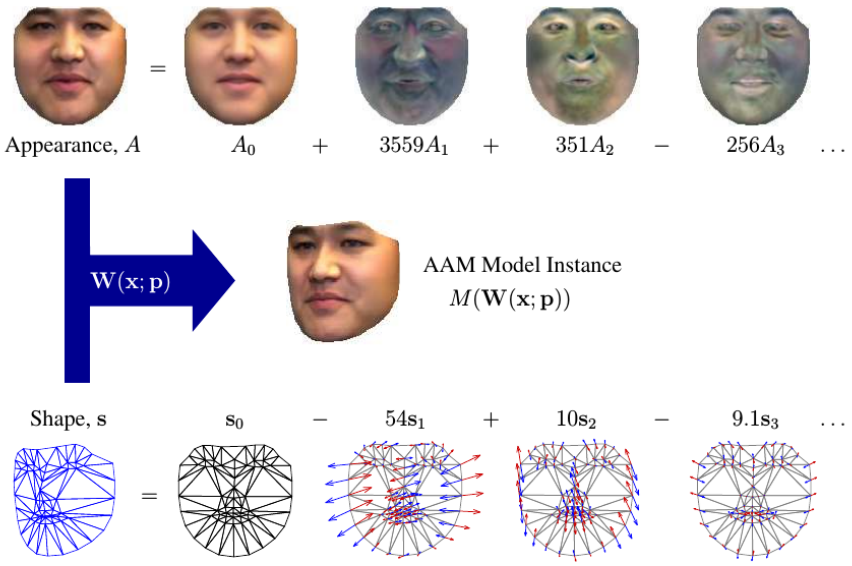
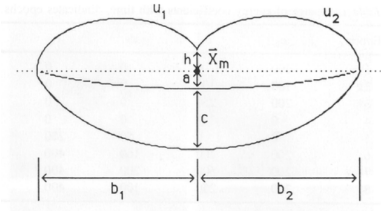
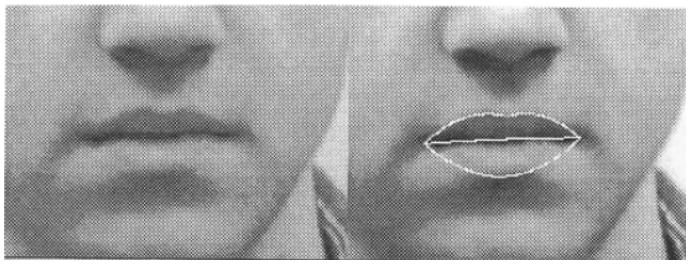


図 2.4: AAMs の例 [9]

Yuille らはパラメータの変化により変形する形状テンプレートを用いて顔の造作を抽出する手法を開発した [10]．閉口時の唇のテンプレートと唇の検出結果を図 2.5 に示す．テンプレートは，画像の明度値に起因するエネルギーが最小となるように変形し，顔の造作と形状が等しくなる．Yuille らの手法では造作を検出出来るだけでなく，形状を説明する特徴量であるパラメータを取得することが出来るが，詳細な形状を抽出する場合にパラメータが多くなりモデルが複雑になると考えられる．



(a) 閉口のテンプレート



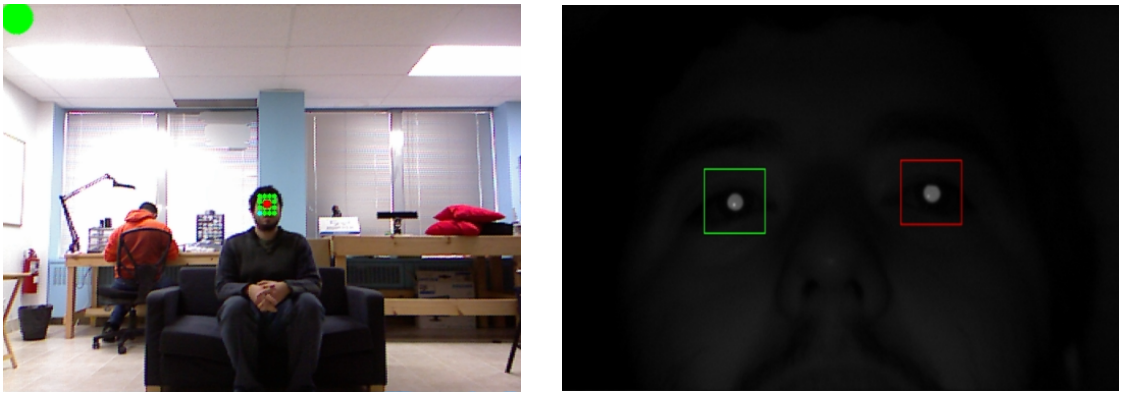
(b) 検出結果

図 2.5: テンプレート及び検出結果 [10]

2.3 距離画像センサと他のセンサを用いた動作計測に関する研究

人の動作計測では画像情報が広く用いられてきたが，近年では安価な距離画像センサの登場により深度画像の計測が容易になり，他のセンサにより取得した情報と深度画像を用いた高精度な計測や広範囲での動作計測が研究されている．以下では距離画像センサと他のセンサを用いた動作計測に関する研究について述べる．

注視点検出は心理学の研究やヒューマンインタフェースとして用いられる動作計測であるが，注視点を計測するためには注視点の存在する平面と計測に用いるカメラ及びユーザの位置関係を知る必要があり，ユーザと計測システムの位置を固定した計測が広く用いられる．Hennessey らは深度画像によりユーザの位置を計測することで，ユーザが居間で自由に動く条件下での注視点検出を実現した [11]．Hennessey らの計測システムは広範囲の撮影を行う距離画像センサとユーザの目を撮影するための狭い視野を持つ赤外線カメラにより構成されている．広範囲を撮影する距離画像センサを用いてユーザの顔を追跡し，ユーザと計測システムの間の距離を計測する．目を撮影するカメラはパンチルト機構を備え，距離画像センサにより取得したユーザの顔の位置を用いてユーザの目のみを撮影する．Hennessey らのシステムによる目の追跡結果を図 2.6 に示す．ユーザの位置は距離画像センサにより取得されるため，ユーザの位置が変化した場合でも注視点検出が可能である．



(a) 広範囲撮影カメラによる顔追跡　　(b) 赤外線カメラにより撮影した目

図 2.6: 広範囲撮影カメラと赤外線カメラによる目の追跡 [11]

Bo らは加速度センサ，ジャイロセンサと共に距離画像センサを用いたリハビリテーションのための関節角計測システムを提案した [12]．Bo らのシステムでは加速度センサにより取得する重力加速度とジャイロセンサにより取得する角度変位を用いて関節角を推定する．ジャイロセンサを用いた計測ではオフセットが蓄積し連続的に計測可能な時間が限られるが，距離画像を用いた骨格推定により関節角を計算し，ジャイロセンサを較正し蓄積するオフセットの影響を抑えることで長時間の関節角計測が可能である．また，距離画像センサを用いて色画像を取得することで計測した関節角をユーザの画像に重畳することが出来る．そのため，計測した角度を視覚情報として提示することが可能であり，リハビリテーションに適した情報提示が可能である．計測により取得した角度による骨格の重畳表示を図 2.7 に示す．

図 2.7: 骨格の重畳表示 [12]

Caputo らは距離画像センサと高い解像度を有するカメラを用いた三次元手振り認識システムを提案した [13]．Caputo らのシステムでは深度画像により手の位置計測を行い，取得した位置情報を用いて高解像度のカメラにおいて手の探索を行う．深度画像における手の位置の計測結果を図 2.8 に示す．高解像度を有するカメラを用いることでユーザが数 m 離れた場合での手振り認識が可能である．高解像度カメラと距離画像センサは凡そ同じ視点を持ち，同じ空間を撮影する．カメラによる画像と距離画像センサによる深度画像の対応は較正により取得する．色画像により手形状を計測し，深度情報による手の位置変化の計測により三次

元的な手の動作を計測する．手形状と手の動作を統合し，三次元的な手振り計測を行う．

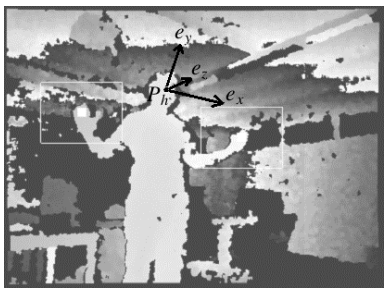


図 2.8: 深度画像における手の位置 [13]

2.4 本研究の位置づけ

従来の手法では詳細な口唇形状を取得するためには比較的高い計算コストが必要であり，ユーザの口唇動作を実時間的に計測すること難しいと考えられる．そのため本研究では，計算コストの低い抽出手法と高精度な抽出手法を組み合わせることで相互の利点を利用し，高精度な抽出手法における計算コストを低減させた比較的計算コストの低い口唇追跡システムを提案する．具体的には，距離画像センサと高解像度のカメラを組み合わせて用い，口唇の位置推定には距離画像センサを用い，詳細な口唇形状の計測には高解像度カメラによる画像を用いる相補的トラッキングを行うことで，実時間での詳細な口唇形状の計測を実現する．相補的トラッキングでは，距離画像センサを用いた実時間での口唇位置推定と高解像度カメラを用いた詳細な口唇形状計測を組み合わせることで相互に欠点を補うトラッキングが可能であると考えられる．

第3章　システムの構成

本章では，提案する相補的口唇トラッキングシステムの構成について述べる．

3.1　システムの概要

相補的口唇トラッキングシステムの構成を図 3.1 に示す．ユーザは同じ視点を持ちユーザの正面に存在する二つのカメラにより計測される．低解像度のカメラ及び距離画像センサによりユーザの口唇の位置を推定することで高解像度画像における探索範囲を限定し，口唇形状の決定に必要なコストを軽減することが可能である．

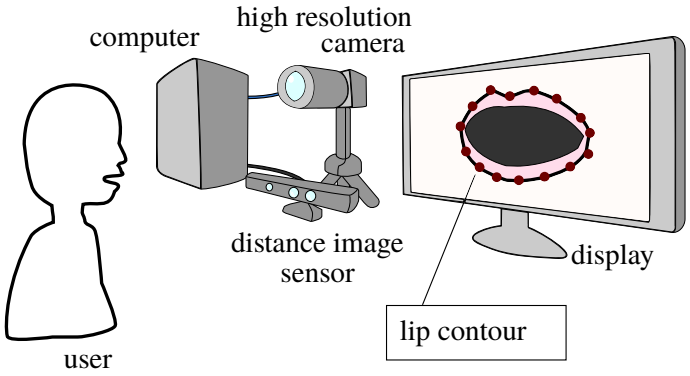


図 3.1: 相補的口唇トラッキングシステムの概要

システムの処理の流れを図 3.2 に示す．

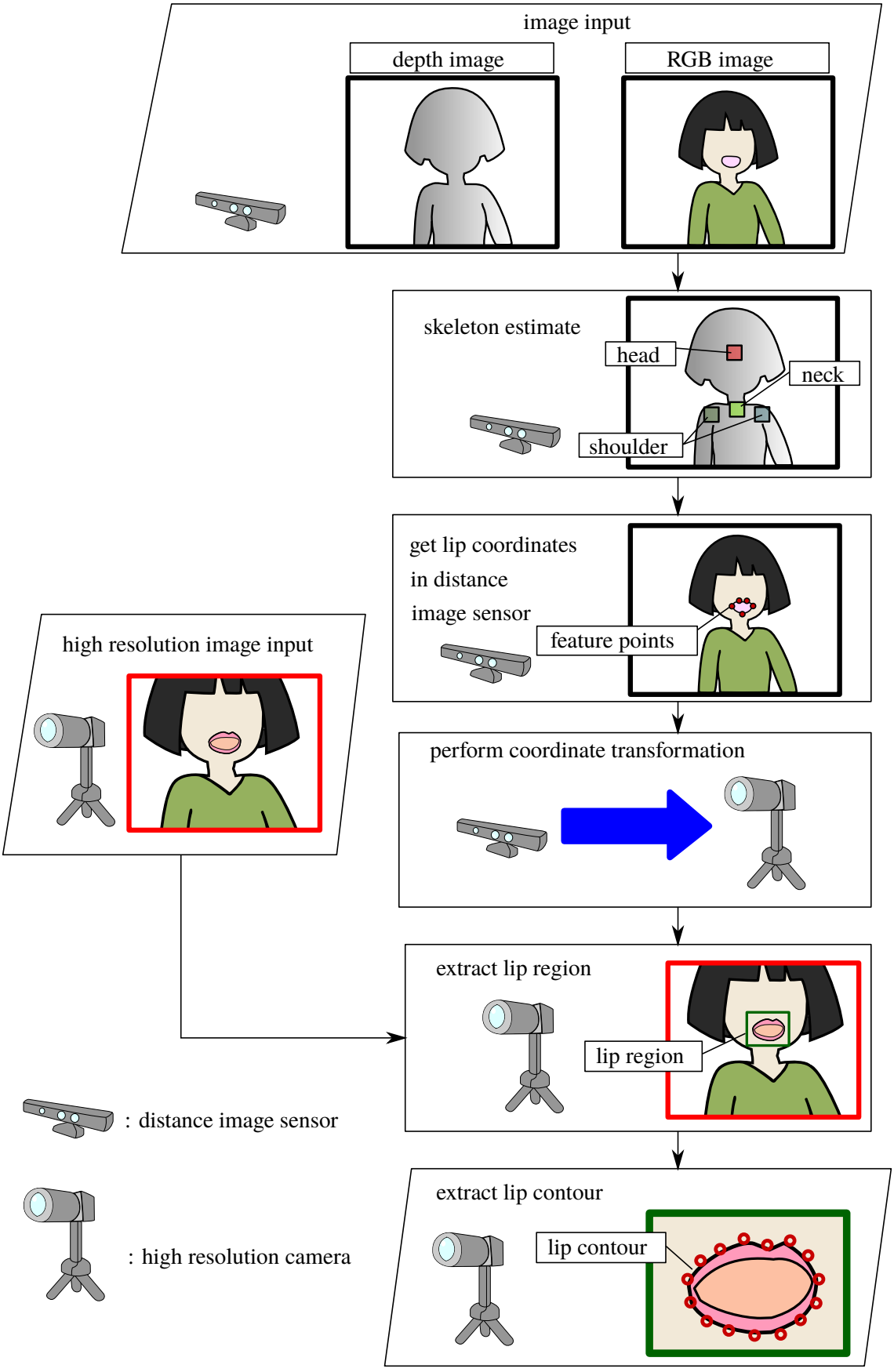


図 3.2: システムの処理の流れ

高精度の口唇形状計測を行う場合，高解像度画像を用いた計測が行われるが，高い解像度の画像においては口唇の探索コストが高く，実時間での口唇計測が難しいと考えられる．本システムでは，ユーザを撮影する複数のカメラを用いる．全てのカメラはユーザを含む視野を持ち，ユーザを含む領域に存在する点の座標の共有が可能である．そのため，距離画像センサを用いてユーザの口唇位置を推定し，距離画像センサにおける座標系から高解像度カメラにおける座標系への座標変換を行うことで，高解像度カメラ座標におけるユーザの口唇位置の推定を行い，口唇形状の計算に必要な探索時間を短縮することが可能である．以下では具体的に手法を説明する．

3.2 口唇位置の推定

まず，ユーザの口唇位置の推定について説明する．ユーザの口唇位置は距離画像センサを用いて距離画像センサ座標系における三次元座標として推定される．本システムではユーザの姿勢は拘束されず顔姿勢の変化を考慮する必要があるため，口唇位置の推定には顔姿勢に依存せず顔認識が可能である AAMs を用いる．AAMs を用いることで画像中における顔の特徴点の座標と顔姿勢を取得することが可能であるが，AAMs は比較的計算コストが高いため画像において予め顔の領域を取得する．顔の位置については，深度画像を用いユーザの骨格を推定することで推定を行う．骨格の推定には Shotton らの手法 [14] を用い，座位を想定した学習データを用いる．AAMs により顔モデルである Candide-3 を顔にフィッテイングし，モデルの各頂点座標を取得する．Candide-3 を図 3.3 に示す．

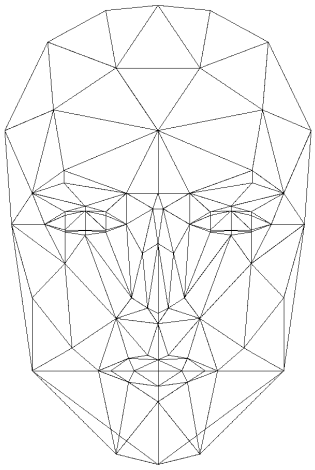


図 3.3: 顔モデル Candide-3 のワイヤフレーム [15]

3.3 距離画像センサと高解像度カメラの座標変換

距離画像センサ座標系から高解像度カメラの座標系への座標変換は，カメラ較正により設定される世界座標系を基準とした座標変換を用いて行われる．具体的には，まず距離画像センサ座標系における口唇座標を世界座標系における座標へと変換し，次に世界座標系における口唇座標を高解像度カメラの座標系における座標へと変換することで高解像度カメラにおける口唇の座標を取得する．但し，距離画像センサ座標系及び世界座標系は三次元座標系であり，高解像度カメラ座標系は二次元座標系である．

以下では，本システムにおいて用いる各座標系とその較正について説明する．世界座標の設定のために格子点が N_{grid} 個存在する一松模様を用いる．一松模様の i 番目の格子点の座標を $\mathbf{p}_i$，格子点の集合を $P = [\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{N_{\text{grid}}}]$，各格子の幅を W として，一松模様における右下の格子点 $\mathbf{p}_0$ を原点，一松模様を含む平面 ϕ と垂直な z 軸を持つ，図 3.4 に示す世界座標系を考える．一松模様の格子の幅が既知であるため，一松模様より世界座標における N_{grid} 個の点の座標を得ることが出来る．

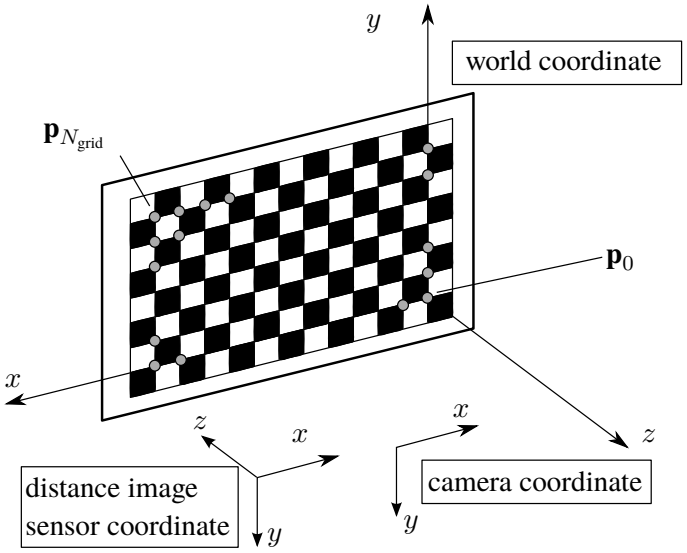


図 3.4: 世界座標系と高解像度カメラ座標系及び距離画像センサ座標系

世界座標系から高解像度カメラ座標系への変換として，透視投影変換を考える．世界座標における任意の点 Q の座標を $\mathbf{p}_{\text{world}} = [x_{\text{world}}, y_{\text{world}}, z_{\text{world}}]^{\text{T}}$ とすると，高解像度カメラに投影された点 Q の座標 $\mathbf{m}_{\text{cam}} = [u, v]^{\text{T}}$ は式 (3.1) で求めることが出来る．但し，$\widetilde{\mathbf{p}}_{\text{world}}$ 及び $\widetilde{\mathbf{m}}_{\text{cam}}$ はそれぞれ $\mathbf{p}_{\text{world}}, \mathbf{m}_{\text{cam}}$ の同次座標であり，$\mathbf{A}$ はカメラの内部パラメータ行列，$\begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix}$ は並

進回転の同次変換行列である．

$$\widetilde{\mathbf{m}}_{\text{cam}} = \mathbf{A}\begin{bmatrix}\mathbf{R} & \mathbf{t}\\ 0 & 1\end{bmatrix}\widetilde{\mathbf{p}}_{\text{world}} \tag{3.1}$$

$$= \mathrm{B}_{\text{wc}}\widetilde{\mathbf{p}}_{\text{world}} \tag{3.2}$$

$\mathbf{A}$ は，主点の座標 $[c_x, c_y]^{\mathrm{T}}$ 及び高解像度カメラ座標系における u 軸，v 軸方向の倍率である f_x, f_y を用いて式 (3.3) のように表される．

$$\mathbf{A} = \begin{bmatrix} f_x & 0 & c_x & 0\\ 0 & f_y & c_y & 0\\ 0 & 0 & 1 & 0\end{bmatrix} \tag{3.3}$$

世界座標系から高解像度カメラ座標系への変換行列を計算することは，既知の座標 $\mathbf{p}_i$ を用いて $\mathrm{B}_{\text{wc}} = \mathbf{A}\begin{bmatrix}\mathbf{R} & \mathbf{t}\\ 0 & 1\end{bmatrix}$ を求めることであり，その手法として Zhang の提案する手法 [16] を用いて計算する．

回転行列 $\mathbf{R}$ を列ベクトル $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ を用いて $\mathbf{R} = [\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3]$ とする．平面 ϕ と垂直な方向に z 軸を考えるため，世界座標系における原点が ϕ 上に存在するとき，ϕ 上の $\widetilde{\mathbf{p}}$ の z 座標は全て 0 である．$\mathbf{p}_{\text{world}} \in P$ とおくと，式 (3.1) より式 (3.4) を得る．

$$\begin{aligned}\begin{bmatrix}u\\ v\\ 1\end{bmatrix} &= \mathbf{A}\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1\\ r_{21} & r_{22} & r_{23} & t_2\\ r_{31} & r_{32} & r_{33} & t_3\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix} x_{\text{world}}\\ y_{\text{world}}\\ 0\\ 1\end{bmatrix}\\ &= \mathbf{A}\begin{bmatrix} r_{11} & r_{12} & t_1\\ r_{21} & r_{22} & t_2\\ r_{31} & r_{32} & t_3\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix} x_{\text{world}}\\ y_{\text{world}}\\ 1\end{bmatrix}\end{aligned} \tag{3.4}$$

$\mathbf{p}'_{\text{world}} = [x_{\text{world}}, y_{\text{world}}]^{\mathrm{T}}$ として，$\widetilde{\mathbf{p}}'_{\text{world}}$ を $\mathbf{p}'_{\text{world}}$ の同次座標とすると，ホモグラフィ変換行列を $\mathbf{H}_{\text{cam}}$ として，式 (3.4) より式 (3.5)，式 (3.6) を得る．

$$\begin{cases}\widetilde{\mathbf{m}}_{\text{cam}} = \mathbf{H}_{\text{cam}}\widetilde{\mathbf{p}}'_{\text{world}} & (3.5)\\ \mathbf{H}_{\text{cam}} = \mathbf{A}\begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix} & (3.6)\end{cases}$$

$\mathbf{H}_{\text{cam}} = [\mathbf{h}_1\ \mathbf{h}_2\ \mathbf{h}_3]$ とおくと，式 (3.5) より式 (3.7) を得る．但し，λ は任意定数である．

$$\begin{bmatrix}\mathbf{h}_1 & \mathbf{h}_2 & \mathbf{h}_3\end{bmatrix} = \lambda\mathbf{A}\begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix} \tag{3.7}$$

$\mathbf{r}_1$，$\mathbf{r}_2$ は正規直交しているため，式 (3.7) より式 (3.8)，式 (3.9) を得る．但し，$\mathbf{A}^{-1}$ は $\mathbf{A}$ の逆行列，$\left(\mathbf{A}^{-1}\right)^{\mathrm{T}}$ は $\mathbf{A}^{-1}$ の転置行列である．

$$\mathbf{h}_1^{\mathrm{T}}\left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{h}_2 = 0 \tag{3.8}$$

$$\mathbf{h}_1^{\mathrm{T}}\left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{h}_1 = \mathbf{h}_2^{\mathrm{T}}\left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{h}_2 \tag{3.9}$$

式 (3.8) 及び式 (3.9) はカメラの内部パラメータに対する束縛を与えるため，異なる四つ以上の点に関して $\mathbf{H}_{\text{cam}}$ を求めることでカメラの内部パラメータ行列 $\mathbf{A}$ を決定することが可能である．式 (3.5) より式 (3.10)，式 (3.11)，式 (3.12)，式 (3.13) を求めることが出来るため，$\mathbf{A}$ より $[\mathbf{R}\ \mathbf{t}]$ を決定することが可能である．

$$\mathbf{r}_1 = \lambda \mathbf{A}^{-1}\mathbf{h}_1 \tag{3.10}$$

$$\mathbf{r}_2 = \lambda \mathbf{A}^{-1}\mathbf{h}_2 \tag{3.11}$$

$$\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2 \tag{3.12}$$

$$\mathbf{t} = \lambda \mathbf{A}^{-1}\mathbf{h}_3 \tag{3.13}$$

距離画像センサ座標及び世界座標は共に三次元座標である．世界座標における i 番目の格子点の座標 $\mathbf{p}_i$ を $\mathbf{p}_i = [x_{\text{world}_i}, y_{\text{world}_i}, z_{\text{world}_i}]^{\text{T}}$ として，距離画像センサ座標における i 番目の格子点の座標を $\mathbf{q}_i = [x_{\text{k}_i}, y_{\text{k}_i}, z_{\text{k}_i}]^{\text{T}}$ とする．距離画像センサ座標系から世界座標系への変換行列を D_{wk} とおくと，$\mathbf{p}_i$ と $\mathbf{q}_i$ の関係は式 (3.14) で表すことが出来る．

$$\text{D}_{\text{wk}}\, \widetilde{\mathbf{p}}_i = \widetilde{\mathbf{q}}_i \tag{3.14}$$

世界座標の原点は平面 ϕ 上に存在するため，$\mathbf{p}$ について式 (3.14) より式 (3.15) を得る．

$$\text{D}_{\text{wk}} \begin{bmatrix} x_{\text{world}_1} & \cdots & x_{\text{world}_n} \\ y_{\text{world}_1} & \cdots & y_{\text{world}_n} \\ 0 & \cdots & 0 \\ 1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} x_{\text{k}_1} & \cdots & x_{\text{k}_n} \\ y_{\text{k}_1} & \cdots & y_{\text{k}_n} \\ z_{\text{k}_1} & \cdots & z_{\text{k}_n} \\ 1 & \cdots & 1 \end{bmatrix} \tag{3.15}$$

一松模様を z 軸に沿って δz だけ平行移動すると，$z = \delta z$ で表される平面 ϕ' に含まれる点の座標 $\mathbf{p}'$ を取得することが出来る．$\mathbf{p}'$ に関して式 (3.14) より式 (3.16) を得る．

$$\text{D}_{\text{wk}} \begin{bmatrix} x_{\text{world}_1} & \cdots & x_{\text{world}_n} \\ y_{\text{world}_1} & \cdots & y_{\text{world}_n} \\ \delta z & \cdots & \delta z \\ 1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} x'_{\text{k}_1} & \cdots & x'_{\text{k}_n} \\ y'_{\text{k}_1} & \cdots & y'_{\text{k}_n} \\ z'_{\text{k}_1} & \cdots & z'_{\text{k}_n} \\ 1 & \cdots & 1 \end{bmatrix} \tag{3.16}$$

行列 $\text{K}_{\text{world}}, \text{K}_{\text{k}}$ を式 (3.17) で定義すると，式 (3.15)，式 (3.16) より，式 (3.17)，式 (3.18)，式 (3.19) を得る．

$$\text{K}_{\text{world}} = \begin{bmatrix} x_{\text{world}_1} & \cdots & x_{\text{world}_n} & x_{\text{world}_1} & \cdots & x_{\text{world}_n} \\ y_{\text{world}_1} & \cdots & y_{\text{world}_n} & y_{\text{world}_1} & \cdots & y_{\text{world}_n} \\ 0 & \cdots & 0 & \delta z & \cdots & \delta z \\ 1 & \cdots & 1 & 1 & \cdots & 1 \end{bmatrix} \tag{3.17}$$

$$\text{K}_{\text{k}} = \begin{bmatrix} x_{\text{k}_1} & \cdots & x_{\text{k}_n} & x'_{\text{k}_1} & \cdots & x'_{\text{k}_n} \\ y_{\text{k}_1} & \cdots & y_{\text{k}_n} & y'_{\text{k}_1} & \cdots & y'_{\text{k}_n} \\ z_{\text{k}_1} & \cdots & z_{\text{k}_n} & z'_{\text{k}_1} & \cdots & z'_{\text{k}_n} \\ 1 & \cdots & 1 & 1 & \cdots & 1 \end{bmatrix} \tag{3.18}$$

$$\mathrm{D_{wk}}\ \mathrm{K_{world}} = \mathrm{K_k} \tag{3.19}$$

式 (3.19) において $\mathrm{K_{world}}$ の疑似逆行列を用いることで式 (3.20) を得る．

$$\mathrm{D_{wk}} = \mathrm{K_k}\mathrm{K_{world}^T}\left(\mathrm{K_{world}}\mathrm{K_{world}^T}\right)^{-1} \tag{3.20}$$

距離画像センサ座標系における口唇座標を $\mathbf{q}_{\mathrm{face}}$，世界座標系における口唇座標を $\mathbf{p}_{\mathrm{face}}$，高解像度カメラ座標系における口唇座標を $\mathbf{q}_{\mathrm{face}}$ とすると，式 (3.2)，式 (3.14) より $\mathbf{q}_{\mathrm{face}}$ と $\mathbf{m}_{\mathrm{face}}$ の関係は式 (3.21) となる．

$$\begin{aligned}\widetilde{\mathbf{m}}_{\mathrm{face}} &= \mathrm{B_{wc}}\ \widetilde{\mathbf{p}}_{\mathrm{face}} \\ &= \mathrm{B_{wc}}\ \mathrm{D_{wk}^{-1}}\ \widetilde{\mathbf{q}}_{\mathrm{face}}\end{aligned} \tag{3.21}$$

式 (3.21) を用いることで距離画像センサ座標系と高解像度カメラ座標系の変換が可能である．

3.4 高解像度カメラ画像における口唇領域の抽出

距離画像センサ画像において取得した口唇の特徴点の座標を式 (3.21) を用いて高解像度カメラ座標系における座標へ変換し，口唇領域の抽出を行う．口唇領域は口唇に外接する長方形領域として与えるが，長方形の中心及び高さ，幅は座標変換により求めた特徴点座標により取得する．口唇の上端，下端，左端，右端に存在する特徴点の座標をそれぞれ $\mathbf{m}_{\mathrm{lip}}^{u} = [u_{\mathrm{lip}}^{u}, v_{\mathrm{lip}}^{u}]$，$\mathbf{m}_{\mathrm{lip}}^{d} = [u_{\mathrm{lip}}^{d}, v_{\mathrm{lip}}^{d}]$，$\mathbf{m}_{\mathrm{lip}}^{l} = [u_{\mathrm{lip}}^{l}, v_{\mathrm{lip}}^{l}]$，$\mathbf{m}_{\mathrm{lip}}^{r} = [u_{\mathrm{lip}}^{r}, v_{\mathrm{lip}}^{r}]$ とする．このとき，口唇領域の中心座標 $\mathbf{m}_{\mathrm{lip}}^{c}$ 及び高さ l_{height}，幅 l_{width} は式 (3.22) となる．口唇領域における中心座標及び長辺，短辺を図 3.5 に示す．

$$\begin{cases}\mathbf{m}_{\mathrm{lip}}^{c} = \left[\dfrac{u_{\mathrm{lip}}^{l} + u_{\mathrm{lip}}^{r}}{2}, \dfrac{v_{\mathrm{lip}}^{u} + v_{\mathrm{lip}}^{d}}{2}\right] \\ l_{\mathrm{width}} = |u_{\mathrm{lip}}^{r} - u_{\mathrm{lip}}^{l}| \\ l_{\mathrm{height}} = |v_{\mathrm{lip}}^{u} - v_{\mathrm{lip}}^{d}|\end{cases} \tag{3.22}$$

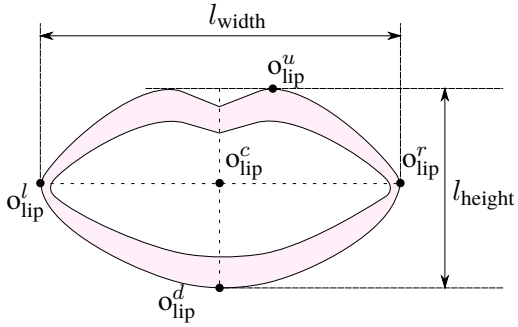


図 3.5: 口唇領域

3.5 口唇形状の計測

抽出した口唇領域において口唇の輪郭線を抽出する．口唇輪郭線の抽出には Kass らの提案した動的輪郭モデルである Snakes を用いる [8]．Snakes では輪郭はエネルギーを極小にする点列として求めることが出来るが，エネルギーの定義を様々に変えることで対象物形状の先験知識を利用することが出来る．

Snakes において用いるエネルギーとは画像座標における輪郭点の座標 $[u, v]$，輪郭点座標の媒介変数 s を用いて表される輪郭 $\mathbf{o}(s) = [u(s), v(s)]$ により定義され，一般には輪郭の形状により定義される内部エネルギー E_{int}，外部エネルギー E_{con} 及び画像の輝度値により決定される画像エネルギー E_{image} の和で表される．Snakes の持つエネルギーの総和 E^*_{snake} は式 (3.23) で表される．

$$\begin{aligned} E^*_{\text{snake}} &= \int_0^1 E_{\text{snake}}(\mathbf{o}(s))ds \\ &= \int_0^1 \{E_{\text{int}}(\mathbf{o}(s)) + E_{\text{image}}(\mathbf{o}(s)) + E_{\text{con}}(\mathbf{o}(s))\}ds \end{aligned} \tag{3.23}$$

E_{int} は輪郭の形状の滑らかさを決定するものであり，$\mathbf{o}(s)$ の s による一次微分及び二次微分を用いて式 (3.24) で定義される．α 及び β は任意定数である．

$$E_{\text{int}} = \frac{(\alpha(s)|\mathbf{o}'(s)|^2 + \beta(s)|\mathbf{o}''(s)|^2)}{2} \tag{3.24}$$

E_{image} は輪郭 $\mathbf{o}(s)$ を画像上の輝度変化の大きい座標へと動かす働きがあり，画像の輝度値 $I(u(s), v(s))$ を用いて一般的に式 (3.25) で定義される．

$$E_{\text{image}}(\mathbf{o}(s)) = -|\nabla I(\mathbf{o}(s))|^2 \tag{3.25}$$

本システムでは口唇形状に着目するが，口唇と肌では輝度変化が小さい場合が多く，口唇形状計測においては十分な輝度変化を得ることが難しい．そのため，本システムでは画像エネルギーとして HSV 表色系における色相に着目し，色相 $I_{\text{Hue}}(u(s), v(s))$ を用いて画像エネルギーを定義する．色相において赤色は 0 deg 付近で変化するが，0 deg 付近では色の変化が微小な場合においても画素値が循環し値が大きく変化する場合があるため，三角関数を用いて 0 deg 付近での色相の変化により画素値の変化量が一定となるように変換を行う．変換した色相を I'_{Hue} とし，式 (3.26) で定義する．但し，彩度 I_{Sat} の低い画素では色相に対する雑音の影響が大きくなるため，彩度が閾値 I_{th} より低い画素では色相を 0 とする．閾値は判別分析法を用いて決定する．

$$I'_{\text{Hue}} = \begin{cases} 128 + 128 \sin I_{\text{Hue}} \cdot \dfrac{2\pi}{360} & (I_{\text{Sat}} > I_{\text{th}}) \\ 0 & (I_{\text{Sat}} \leq I_{\text{th}}) \end{cases} \tag{3.26}$$

本システムでは画像エネルギーとして輪郭画像を用いるが，輪郭抽出の方法として雑音の影響を受けにくく，計算速度が得られるため DoG を用い，画像エネルギーを式 (3.27) で定義する．但し，k は定数である．G を式 (3.28) で表す．

$$E_{\text{image}} = G(u(s), v(s), k\sigma)(I'_{\text{Hue}}(u(s), v(s)) - G(u(s), v(s), \sigma)I'_{\text{Hue}}(u(s), v(s)) \tag{3.27}$$

$$G(u, v, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{u^2 + v^2}{2\sigma^2}\right) \tag{3.28}$$

E_{con} は $\mathbf{o}(s)$ を期待される局所最小値に近づける働きがあり，E_{con} の定義は様々存在するが，本システムにおいては距離変換画像の輝度値を元に決定する．本システムにおいて E_{con} を式 (3.29) で定義する．

$$E_{\text{con}} = E_{\text{dist}} - E_{\text{oral}} + E_{\text{pot}} \tag{3.29}$$

E_{dist} は距離変換画像であり，式 (3.27) で得た輪郭画像に対して二値化を行い，二値化画像において非 0 の画素値を持つ画素から画素値が 0 である画素までの最小の距離を画素値として持つ．本システムでは口唇の輪郭を考えるが，口唇と口内の境界では色相の変化が大きく，Snakes が口内の輪郭へ収束する場合が存在する．そのため，Snakes を口内から離す働きを持つエネルギーとして E_{oral} を用いる．E_{oral} は口内輪郭の輪郭画像を元に作成した距離変換画像であるが，口内輪郭の輪郭画像は HSV 表色系における明度 (V) を二値化することにより口内領域を求め，その領域の輪郭を抽出することで求める．二値化に用いる閾値は判別分析法により決定する．E_{pot} は画像中の座標によるエネルギーであり，画像の中央より離れるに従い増加するエネルギーである．画像中央の座標を $\mathbf{o}^c$ として，E_{pot} を式 (3.30) で定義する．

$$E_{\text{pot}} = |\mathbf{o}(s) - \mathbf{o}^c|^2 \tag{3.30}$$

Snakes におけるエネルギー最小化の手法は様々存在し，一般的には動的計画法 (DP) [17] を用いる．しかし本システムでは，DP と比べ計算速度を得ることが出来るため計算アルゴリズムとして貪欲法を用いる [18]．計算において輪郭は離散的に扱い，$\mathbf{o}(s)$ を N_{cont} 個の点の集合として式 (3.31) で表す．$[\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_{N_{\text{cont}}}]$ を制御点とよぶ．

$$[\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_{N_{\text{cont}}}] = \left[\begin{bmatrix} u_1 \\ v_1 \end{bmatrix}, \begin{bmatrix} u_2 \\ v_2 \end{bmatrix}, \cdots, \begin{bmatrix} u_{N_{\text{cont}}} \\ v_{N_{\text{cont}}} \end{bmatrix}\right] \tag{3.31}$$

制御点は空間的に離散的であるため，制御点の座標の空間微分は差分法を用いて計算し，一次微分及び二次微分はそれぞれ式 (3.32)，式 (3.33) とする．

$$\mathbf{o}'_i = \mathbf{o}_{i+1} - \mathbf{o}_i \tag{3.32}$$

$$\mathbf{o}''_i = \mathbf{o}_{i+2} - 2\mathbf{o}_{i+1} + \mathbf{o}_i \tag{3.33}$$

制御点 $\mathbf{o}_i$ により決定されるエネルギーを e_i と表すと，e_i は $\mathbf{o}_i$ の二次微分に依存するため，$e_i(\mathbf{o}_i, \mathbf{o}_{i+1}, \mathbf{o}_{i+2})$ と表すことが出来，全体のエネルギーは式 (3.34) のようになる．

$$E^*_{\text{snake}}(\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_{N_{\text{cont}}}) = e_1(\mathbf{o}_1, \mathbf{o}_2) + e_2(\mathbf{o}_2, \mathbf{o}_3) + \cdots + e_{N_{\text{cont}}}(\mathbf{o}_{N_{\text{cont}}-1}, \mathbf{o}_{N_{\text{cont}}}) \tag{3.34}$$

貪欲法では各制御点 $\mathbf{o}_i$ は $e_{i-1} + e_i$ を最小とするよう移動し他の制御点の移動を考慮しないため，制御点の最適解への収束が保証されないが，制御点の初期位置を適切に設定することで実用上の問題はないと考えられる．

第4章　発音習得支援システムの実装と評価

本章では提案した発音習得支援システムの実装と評価について述べる．

4.1　システムの構成

本システムは，コンピュータ，高解像度カメラ，距離画像センサ，ディスプレイで構成されている．実装したシステムの概要を図 4.1 に示す．

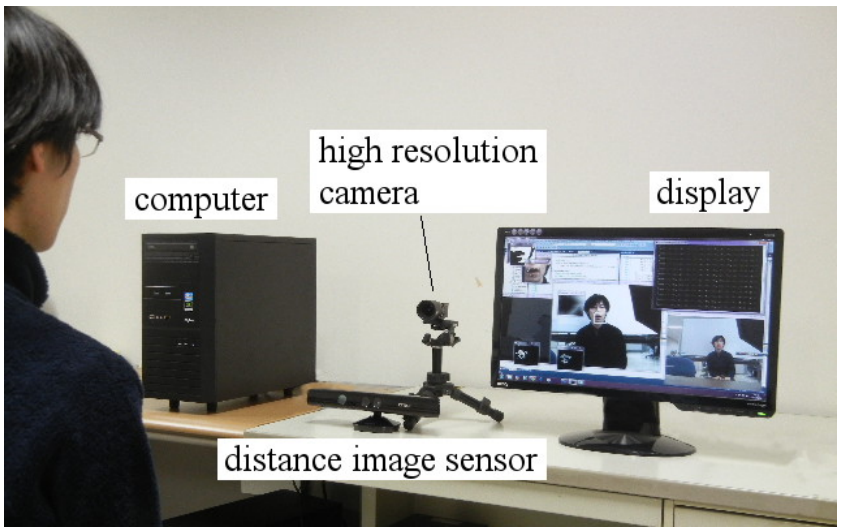


図 4.1: システムの概要

また，システムに用いたコンピュータ，高解像度カメラ，距離画像センサの仕様をそれぞれ表 4.1，表 4.2，表 4.3 に示す．

表 4.1: コンピュータの仕様

項目	
OS	Microsoft Windows 7
CPU	Intel Core i7-3770 3.40 GHz
RAM (GB)	8.0
GPU	NVIDIA GeForce 9600GT

表 4.2: 高解像度カメラの仕様

項目	
製品名	Flea3 FL3-U3-32S2C-CS (POINT GREY 社)
フレームレート (fps)	30
解像度	2080 × 1552

表 4.3: 距離画像センサの仕様

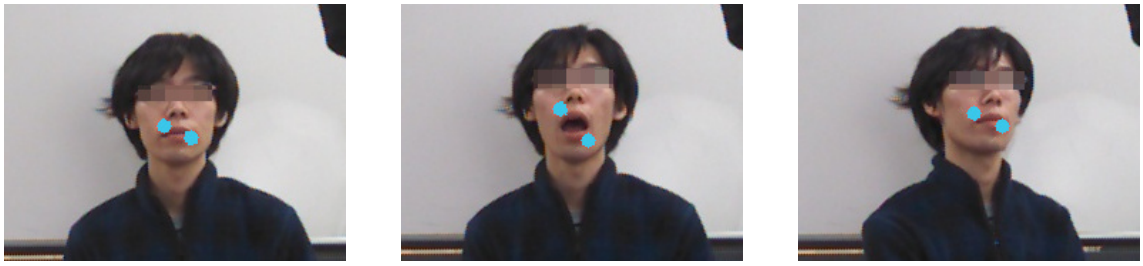
項目	
製品名	Kinect for Windows (Microsoft 社)
最大フレームレート (fps)	30
解像度	640 × 480

4.2 システムの実装

本節では，構築したシステムの実装について述べる．

距離画像センサによる口唇位置追跡

提案システムでは，距離画像センサを用いて口唇位置を追跡し，追跡した口唇位置を高解像度カメラ座標へ投影することで高解像度カメラ座標での口唇位置を取得する．距離画像センサによる口唇位置推定の結果を図 4.2 に示す．図 4.2 における青い点は距離画像センサの情報を元に推定した口唇矩形領域の左上点及び右下点を表す．距離画像センサにおいて取得した口唇位置を元に高解像度カメラにおける高解像度カメラにおいて口唇位置を取得した結果を図 4.3 に示す．図 4.3 における矩形は抽出領域を表す．抽出領域に Snakes を適用した結果を図 4.4 に示す．図 4.4 中の白丸は Snakes における制御点を表す．

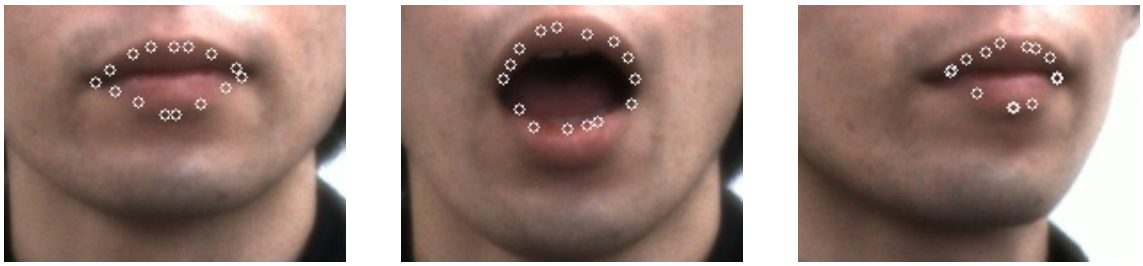


(a) 閉口時正面　(b) 開口時正面　(b) 閉口時斜め

図 4.2: 距離画像センサによる口唇位置追跡の結果

(a) 閉口時正面　(b) 開口時正面　(b) 閉口時斜め

図 4.3: 変換により求めた高解像度カメラ座標における口唇位置



(a) 閉口時正面　(b) 開口時正面　(b) 閉口時斜め

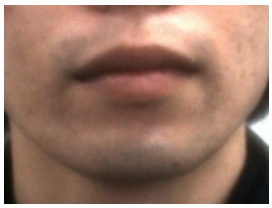
図 4.4: トリミング画像における Snakes の結果

また，較正により求めた世界座標から高解像度カメラ座標への変換行列及び世界座標から距離画像センサへの変換行列をそれぞれ式 (4.1)，式 (4.2) に示す．

$$\mathrm{B_{wc}} = \begin{bmatrix} -9.36 & 1.26 \times 10^{3} & 6.90 \times 10^{2} & 3.27 \times 10^{5} \\ -1.22 \times 10^{3} & -1.14 \times 10 & 6.45 \times 10^{2} & 3.89 \times 10^{5} \\ 3.57 \times 10^{-2} & 3.78 \times 10^{-2} & 0.998 & 5.02 \times 10^{2} \end{bmatrix} \tag{4.1}$$

$$\mathrm{D_{wk}} = \begin{bmatrix} 0.124 & 7.43 & 0.907 & -2.13 \times 10^{2} \\ -7.12 & 0.120 & 0.792 & 48.2 \\ 1.59 & 0.399 & 6.19 & 7.15 \times 10^{3} \\ -1.86 \times 10^{-9} & 3.72 \times 10^{-9} & 0.00 & 1.00 \end{bmatrix} \tag{4.2}$$

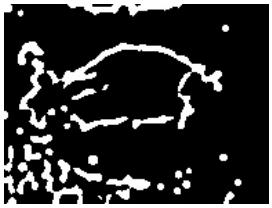
以上の結果を用いて距離画像センサを用いて切り出した画像において Snakes を用いて口唇形状の計測を行った．正面姿勢において取得した口唇画像において Snakes を適用するために用いた画像を図 4.5 に示す．図 4.5(a) に対して Snakes を適用し，図 4.5(h) が Snakes を適用した結果である．図 4.5(h) における白丸は Snakes の制御点を表す．



(a) 元画像

(b) HSV 表色系へ変換したときの色相画像

(c) 色相画像を元に作成した DoG 画像

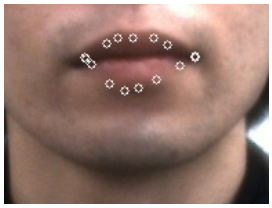


(d) DoG 画像に二値化処理を施した画像

(e) 距離変換画像

(f) HSV 表色系へ変換したときの二値化明度画像

(g) 画像 (f) を元に作成した距離変換画像



(h) Snakes を重畳表示した画像

図 4.5: Snakes を適用するために用いた画像

図 4.6: 実装したシステムの動作の様子

実装したシステムを用いて口唇輪郭の抽出を行った様子の動画を図 4.6 に示す.

4.3 実験

実装したシステムによる領域抽出が計算速度において他の手法と比べ高速であるかを確認するために領域抽出に要する計算時間を計測する実験を行った．実装したシステムを用いて口唇領域の抽出を行い，領域の抽出に要した時間の測定を行った．高解像度カメラの解像度を 1600×1200 pixels とし，フレームレートは 30 fps とした．距離画像センサの解像度を 640×480 pixels とした．ユーザはカメラに対して正面を向き，口は閉じている状態であった．領域の抽出を行う際に各処理において要した平均時間を表 4.4 に示す．同じ条件下で顔検出を用いるシステムを用いて口唇領域の抽出を行った．顔検出には Haar-Like 特徴量を用いた顔検出器を用い，検出した顔領域の下半分の領域を口唇領域として抽出した．顔検出を行う際，画像を縮小して処理を施すことで検出に要する計算時間を短縮する手法が広く用いられるため，画像の縮小率を様々に変えて時間の計測を行った．但し，縮小率は元画像の大きさに対する縮小画像の大きさの比率とする．抽出に要した平均時間を表 4.5 に示す．

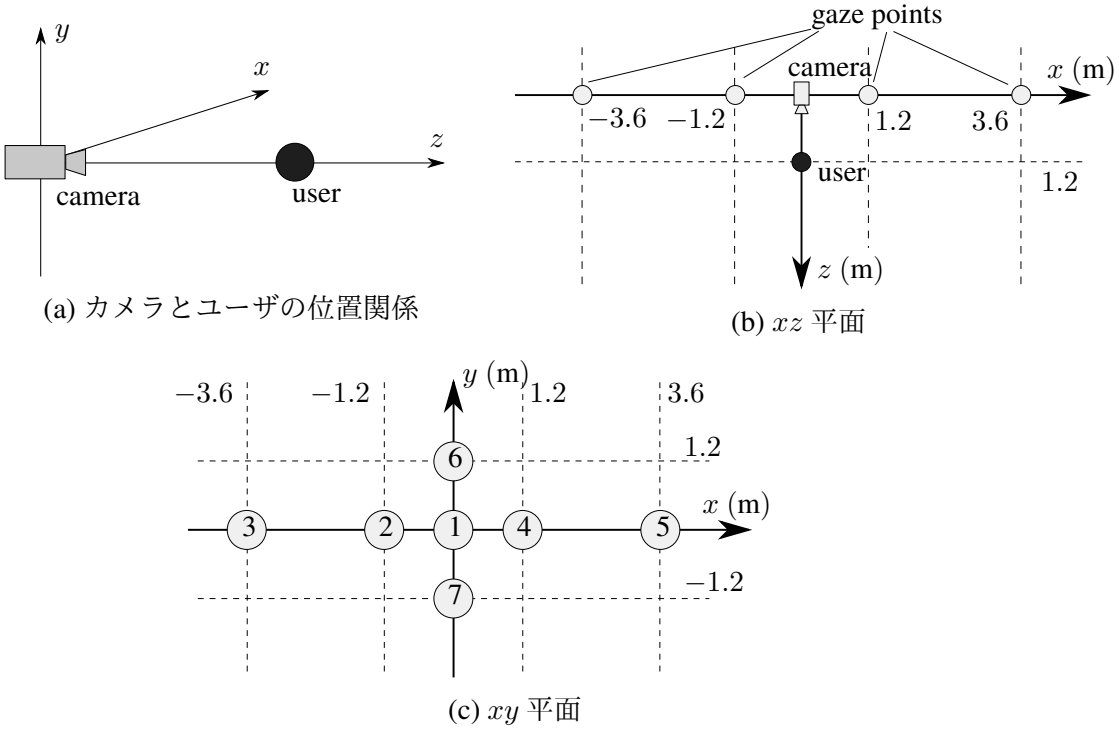
表 4.4: 提案手法を用いて口唇領域の抽出に要した時間

処理内容	時間 (s)
顔の位置推定	4.64×10^{-2}
口唇位置の取得	3.60×10^{-7}
座標変換	1.04×10^{-4}
計	4.65×10^{-2}

表 4.5: 顔検出を用いて口唇領域の抽出に要した時間

縮小率 (%)	時間 (s)
100	3.43×10^{-1}
80	2.29×10^{-1}
40	6.94×10^{-2}
20	2.51×10^{-2}
10	1.17×10^{-2}

構築したシステムが顔姿勢に依存せずに口唇領域を抽出可能であることを確認するために様々な姿勢における口唇領域の抽出を行った．ユーザはカメラの正面に位置し，注視する点を変えることで姿勢を変化させた．カメラの光軸方向を z 軸とし，z 軸に垂直で地面と水平である軸を x 軸，地面と垂直である軸を y 軸とする直交座標系を設定した．カメラとユーザ及び注視点の位置関係を図 4.7 に示す．



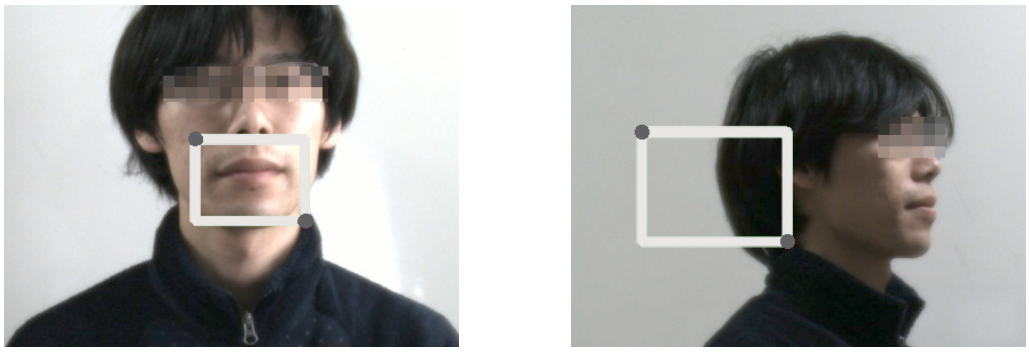
(a) カメラとユーザの位置関係

(b) xz 平面

(c) xy 平面

図 4.7: 姿勢変化のために注視した点とユーザの位置関係

実験を行った姿勢として，閉口状態において図 4.7 (c) における点 1 から点 7 のそれぞれを注視し，計七姿勢における計測を行った．また，点 1 を注視した際には開口と閉口の二つの姿勢において計測を行った．抽出した結果，八姿勢のうち点 5，点 6 を注視した姿勢を除く六姿勢で正しい抽出結果を得た．正しく抽出した結果と誤った抽出を行った結果を図 4.8 に示す．



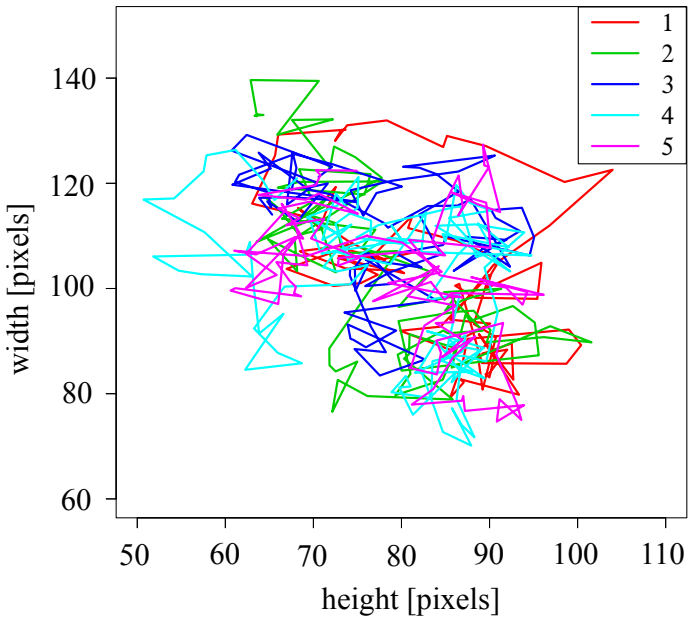
(a) 正しい抽出結果

(b) 誤った抽出結果

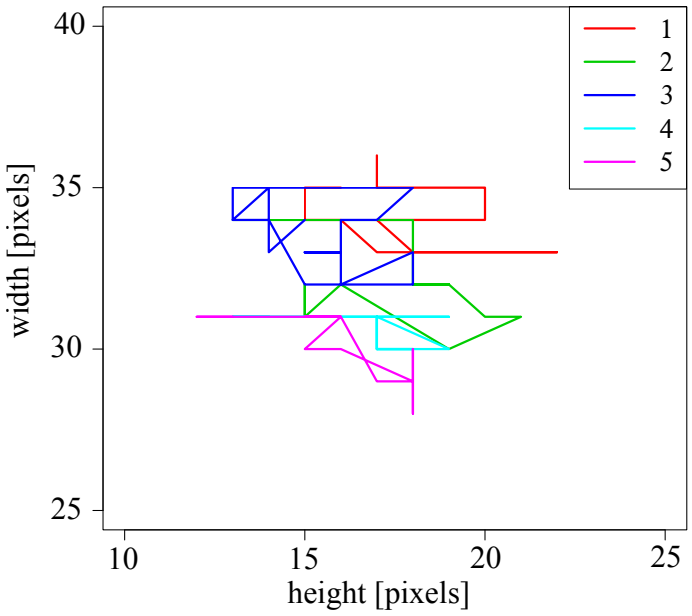
図 4.8: 提案システムを用いた口唇領域抽出結果

構成したシステムを用いて精度の高い口唇動作計測を行うことが可能であることを確認するために，構成したシステムを用いて単語発話時の口唇動作計測を行った．発話した単語は英単語の “hello” であり，計測時のシステムとユーザの距離は約 1 m とした．発話は約 3 秒間行い，5 回計測を行った．同じ条件下で距離画像センサのみを用い，AAMs を用いて口唇

動作計測を行った．システムでの計測結果と距離画像センサのみを用いた計測結果より口唇の高さと幅を取得した．取得した口唇の幅と高さを図 4.9 に示す．但し，図 4.9 における各色は各試行を表す．



(a) システムの計測結果



(b) 距離画像センサの結果

図 4.9: 単語発話時の口唇計測結果

4.4 考察

計測に要した時間に関して考察する．顔検出を用いた手法において抽出に要した時間と縮小率の関係を図 4.10 に示す．但し，図 4.10 において点線は提案手法を用いた抽出に要した時間を表す．領域抽出に要した時間を比較すると，顔検出を用いた手法において画像の縮小率を 40 % より小さくした場合において計算時間が提案手法より短くなるが，提案手法に比べ精度の低い抽出であると考えられる．また，構築したシステムにおける計算速度では高解像度カメラのフレームレートである 30 fps を満たさないが，実際のシステムにおいては一度顔の位置推定を推定すると顔の追跡を行うため，顔の追跡を行っている際には口唇領域の抽出に要する時間は 1.04×10^{-4} 秒であり高解像度カメラのフレームレートを満たす計算速度で計算可能である．従って，構築したシステムは高速な計算速度を有すると考えられる．

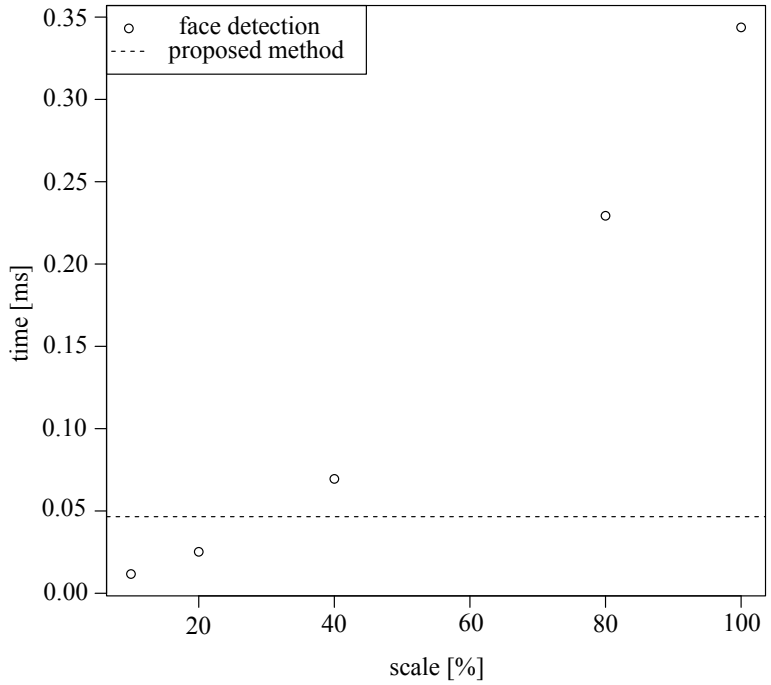


図 4.10: 口唇領域の抽出に要した時間

構築したシステムを用いて様々な姿勢における領域抽出の精度に関して考察を行う．構築したシステムにおいて，ユーザが右を向いた場合に正しく口唇領域を抽出することが出来なかった．システムに用いた距離画像センサは距離情報を得るために赤外線パターンを対象に投影し，パターンのゆがみから距離情報を取得する Light Coding と呼ばれる方式を用いて距離情報を取得しているが，Light Coding では物体の輪郭にあたる部分の距離情報を取得することが難しい．従ってユーザが右を向いた場合，距離画像センサによる口唇の特徴点の距離情報を取得することが困難となり，距離情報が正しく取得出来なかったため誤った座標における領域を抽出したと考えられる．そのため，距離画像センサとして物体の輪郭部分の距離

情報の取得が可能なデバイスを用いることで誤った抽出を防ぐことが出来ると考えられる．

単語発話時の口唇動作計測に関して考察を行う．構成したシステムを用いた計測結果と距離画像センサによる計測結果を比較するために，各結果の正規化を行い平均を計算した．口唇の高さや幅を計測したデータは一様分布すると考えられるため，各試行において最大値と最小値を用いて空間的に正規化を行った．時刻 j における口唇の高さを l^j_{height}，幅を l^j_{width} とし，高さの最小値を $l_{\text{height,min}}$，高さの最大値を $l_{\text{height,max}}$，幅の最小値を $l_{\text{width,min}}$，幅の最大値を $l_{\text{width,max}}$ とすると正規化した高さ l'^j_{height}，幅 l'^j_{width} は式 (4.3)，式 (4.4) で表される．

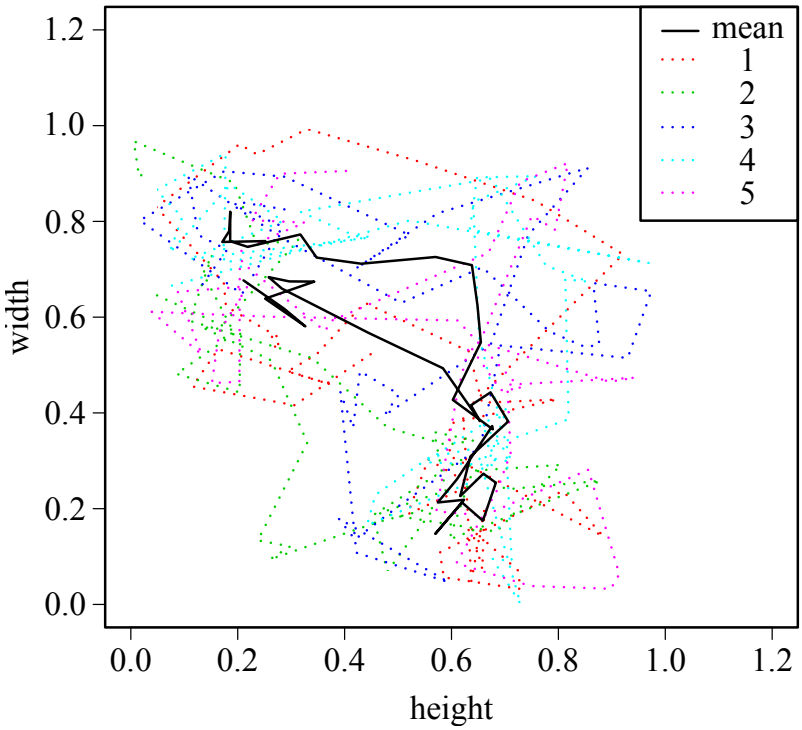
$$l'^j_{\text{height}} = \frac{l^j_{\text{height}} - l_{\text{height,min}}}{l_{\text{height,max}} - l_{\text{height,min}}} \tag{4.3}$$

$$l'^j_{\text{width}} = \frac{l^j_{\text{width}} - l_{\text{width,min}}}{l_{\text{width,max}} - l_{\text{width,min}}} \tag{4.4}$$

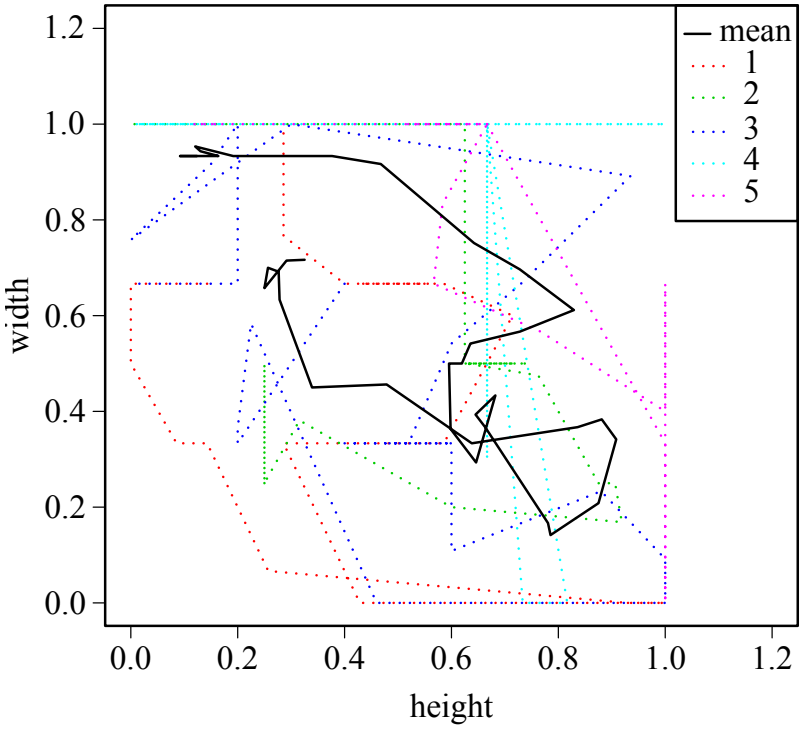
各試行において計測時間が異なるため，時間的な正規化を行った．計測開始時の時刻 τ を $\tau = 0$，計測終了時の時刻を $\tau = 1$ とする正規化を行い，標本抽出を行った．標本抽出においてデータの存在しない時刻のデータを参照する場合，前後のデータを用いて線形補完を行い標本抽出を行った．距離画像センサにより取得したデータの点数が約 40 点であったため，全てのデータにおいて 40 点のデータ点を抽出した．抽出したデータの平均を計算し，試行間の分散を計算した．抽出したデータを図 4.11 に示す．分散の平均を表 4.6 に示す．表 4.6 より，構築したシステムによる計測は距離画像センサのみを用いた計測に比べ分散が小さく，精度の高い計測を行うことが可能であると考えられる．

表 4.6: 試行間における分散の平均

	高さ	幅
構築したシステム	2.72×10^{-2}	3.00×10^{-2}
距離画像センサ	6.49×10^{-2}	4.15×10^{-2}



(a) システムの計測結果



(b) 距離画像センサの結果

図 4.11: 正規化した計測結果

第5章　結論

本研究では，カメラと距離画像センサを用いた，相補的な口唇トラッキングシステムを提案した．距離画像センサを用いて口唇位置を推定し，高解像度カメラにおける口唇領域を抽出した．抽出した口唇領域において口唇トラッキングを行った．

構築したシステムを用いた口唇領域抽出に必要な時間を計測した．顔検出を用いた口唇領域検出手法による領域抽出に要した時間と構築したシステムによる時間を比較したところ，同程度の計算時間でより高精度な口唇領域抽出が可能であることが確認出来た．本システムによりユーザを拘束しない，高い精度と高速な計算速度を有する口唇トラッキングが可能となった．ユーザの動きを拘束しない口唇トラッキングにより，口唇動作を用いた読唇や発話訓練支援の実用性が広がると考えられる．

謝 辞

本研究は，大阪大学基礎工学部で行ったものである．

研究を行うにあたり，研究環境を提供して頂き，本論文の添削指導や研究室での日々のゼミにおいて多大なる御指導を頂きました大阪大学 大学院基礎工学研究科 大城理教授に深く感謝するとともに，篤く御礼申し上げます．本研究のみならず，研究生活において様々な観点からの助言を頂き，とても多くのことを学ぶことが出来ました．普段のゼミや，研究のテーマに関して常に的確な指導をしてくださいました大阪大学 大学院基礎工学研究科 井村誠孝准教授に深く感謝致します．担当教員として様々な相談に応えて頂き，幾度となく助けて頂きました大阪大学 大学院基礎工学研究科 吉元俊輔助教に心より御礼申し上げます．研究の右も左も分からない私に，学問的なことのみならず研究生活におけるノウハウを教えてくださいました．

また，大城研究室の先輩方である井手口裕太氏，加藤雄樹氏，園原佑壮氏，長坂信吾氏，中藤寛己氏，豆野裕信氏，和田章宏氏，加藤高浩氏，上西健太氏，川口純輝氏，古澤大樹氏にお礼申し上げます．先輩方は普段から進捗を心配して頂き，ゼミでは分からないところを教えて頂き，研究生活の様々なところで支えてくださいました．さらに，研究活動において一番の相談相手として共に努力した同期の桑谷達之氏，武村浩志氏，日夏俊氏に感謝します．

最後に，いつも変わらず支えてくれた家族に感謝します．

本研究における被験者実験は，基礎工学研究科における人を対象とした研究に関する倫理委員会の承認 (26-13) を得て行ったものである．

参考文献

[1] 村田孝次. 幼児のことばと発音 -その発達と発達障害-. 培風館, 東京, 1970.

[2] 齊藤剛史, 小西亮介. トラジェクトリ特徴量に基づく単語読唇. 電子情報通信学会論文誌 D, Vol. 90, No. 4, pp. 1105–1114, 2007.

[3] Olov Engwall. Introducing visual cues in acoustic-to-articulatory inversion. In *INTERSPEECH*, pp. 3205–3208, 2005.

[4] 宮崎剛, 中島豊四郎. 読唇技能保持者をモデル化した機械読唇のための特徴的口形検出方法に関する研究. 立石科学技術振興財団 助成研究成果集, Vol. 21, pp. 2–7, 2012.

[5] 川越いつえ. 英語の音声を科学する. 株式会社大修館書店, 東京, 1999.

[6] Rainer Stiefelhagen, Uwe Meier, and Jie Yang. Real-time lip-tracking for lipreading. In *Eurospeech*, 1997.

[7] Ying-li Tian, Takeo Kanade, and Jeffrey Cohn. Robust lip tracking by combining shape, color and motion. In *Proceedings of the 4th Asian Conference on Computer Vision*, pp. 1040–1045, 2000.

[8] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, Vol. 1, No. 4, pp. 321–331, 1988.

[9] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 135–164, 2004.

[10] Alan L Yuille, Peter W Hallinan, and David S Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision*, Vol. 8, No. 2, pp. 99–111, 1992.

[11] Craig Hennessey and Jacob Fiset. Long range eye tracking: Bringing eye tracking into the living room. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 249–252, 2012.

[12] Antonio Bo, Mitsuhiro Hayashibe, and Philippe Poignet. Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect. In *EMBC'11: 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3479–3483, 2011.

[13] Manuel Caputo, Klaus Denker, Benjamin Dums, and Georg Umlauf. 3D Hand Gesture Recognition Based on Sensor Fusion of Commodity Hardware. In *mensch & Computer*, Vol. 2012, pp. 293–302, 2012.

[14] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, Vol. 56, No. 1, pp. 116–124, 2013.

[15] Jörgen Ahlberg. "Candide". `http://www.icg.isy.liu.se/candide/`. 2015 年 1 月 20 日参照.

[16] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 11, pp. 1330–1334, 2000.

[17] 上田修功, 間瀬健二, 末永康仁. 弾性輪郭モデルとエネルギー最小化原理による輪郭追跡手法. 電子情報通信学会論文誌 D, Vol. 75, No. 1, pp. 111–120, 1992.

[18] Donna J Williams and Mubarak Shah. A fast algorithm for active contours and curvature estimation. *CVGIP: Image understanding*, Vol. 55, No. 1, pp. 14–26, 1992.