Mouth Motion Recognition for Intelligent Room Using DP Matching

Tatsuya Nakanishi Non-member (Chuo Univ. / CREST, JST, tnakani@sensor.mech.chuo-u.ac.jp)
Kenji Terabayashi Non-member (Chuo Univ. / CREST, JST, terabayashi@mech.chuo-u.ac.jp)
Kazunori Umeda Non-member (Chuo Univ. / CREST, JST, umeda@mech.chuo-u.ac.jp)

Keywords : lip reading, DP matching, intelligent room, gesture recognition, image processing

These days, home appliances in our living environment are becoming full of functions. On the other hand, the increase of their functions makes their operation complicated. For such appliances frequently used in everyday life, intuitive operation is desirable for users. There are studies to make a room or some space itself intelligent. This approach is effective for realizing natural human-machine interface. We are also constructing an intelligent room, in which we can operate home appliances such as a television set by gestures. The operator (i.e., a person with an intention to operate an appliance) does not need a special attachment and can make operations in a natural state. Fig.1 illustrates the room. Color cameras with pan, tilt and zoom functions are installed and gestures of the operator are observed with them.

In the system, a target appliance is chosen by pointing. However, measurement of direction of pointing is not robust according to the relative direction to cameras, etc. Therefore, we consider adding a new, intuitive modality to indicate an appliance. An effective candidate is speech recognition. In this paper, we consider another modality: to recognize mouth motion, i.e., lip reading, using a camera. We propose a method to recognize mouth motion for choosing a home appliance in the intelligent room.

The existing studies dealing with lip reading are divided into two kinds; one is model-based, and the other is image-based. A problem of image-based methods is that they are affected by the size of lip image. We cope with the problem and recognize mouth motion based on Dynamic Programming (DP) matching.

Fig.2 shows the flow of proposing mouth motion recognition. We first detect a face region of the operator as shown in Fig.3. Then mouth region is extracted from the face region. The size of the extracting mouth region can be defined by the size of face region. The mouth region is converted to a low-resolution image. A time series of the low-resolution mouth image is matched to the model data with DP matching. In this process, we introduce a mechanism to cope with a change of mouth position while speaking a word by using nine low-resolution images with small shifts and choosing the best-fit image as shown in Fig.4.

We verified the mouth recognition method by experiments to recognize four words: “fan”, “TV”, “air con” (air conditioner), and “light”. Four kinds of mouth motions of a subject were registered at the distance of 2[m] from a USB camera. Then his mouth motions were recognized at 0.5, 1, 2[m]. It is shown that practical recognition rates are realized for each distance and the proposed method overcomes the disadvantage of image-based method.

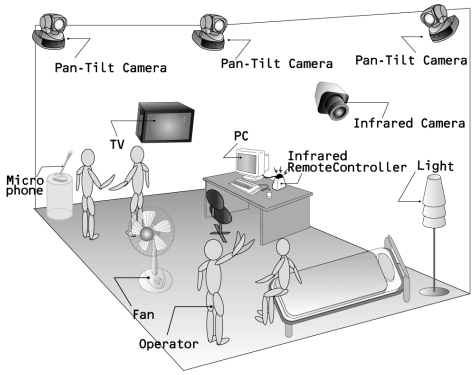


Fig. 1. Conceptual figure of our intelligent room

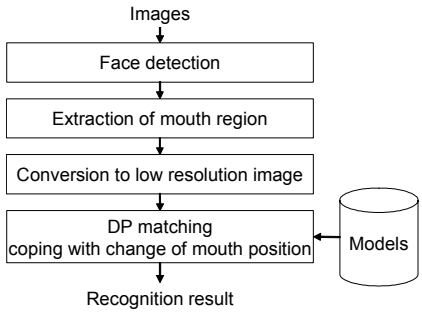


Fig. 2. Flow of mouth motion recognition.

Fig. 3. Detection of face region

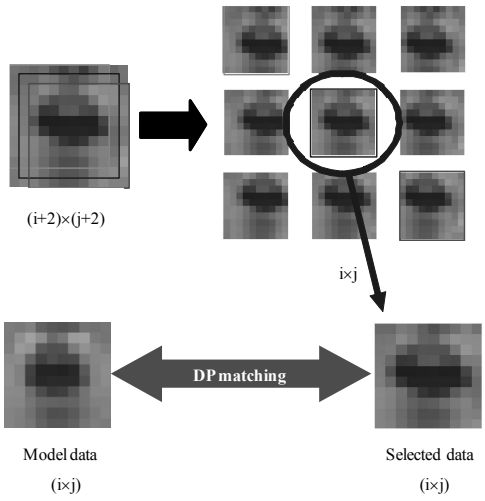


Fig. 4. Nine sub-images for change of mouth position

論　文

インテリジェントルームのための DP マッチングを用いた口唇動作認識

非会員　中西　達也*　　非会員　寺林　賢司**
非会員　梅田　和昇**

Mouth Motion Recognition for Intelligent Room Using DP Matching

Tatsuya Nakanishi*, Non-member, Kenji Terabayashi**, Non-member, Kazunori Umeda**, Non-member

An intelligent room which recognizes gestures and supports people is expected in various situations in recent years. This paper proposes a method to recognize mouth motion, i.e., a method of lip reading, to indicate an object like a home appliance in an intelligent room. The method first detects the operator's face. Then mouth region is extracted from the face region using the fact that inside of mouth is dark. Dynamic Programming (DP) matching is applied to a sequence of low-resolution images of the mouth region and the mouth motion of speaking a word is recognized. The proposed method overcomes the disadvantage of image-based methods that they are not robust to the change of distances between an operator and a camera. Additionally, the proposed method can cope with small displacement of mouth position while speaking by considering one-pixel offsets for low-resolution images and using nine shifted images to obtain the smallest distance. The effectiveness of the proposed method is evaluated by experiments to recognize four words that are typical names of home appliances.

キーワード：読唇，DP マッチング，インテリジェントルーム，ジェスチャ認識，画像処理
Keywords：lip reading, DP matching, intelligent room, gesture recognition, image processing

1. 序　　論

我々の生活環境における家電製品などの機器は，より高機能になってきている反面，操作が複雑になっている。日常生活で用いる機器は，誰にでも直感的に操作できることが望ましい。近年，部屋などの空間そのものを知能ロボット化するコンセプトが提案され，数多く研究されてきている[(1)~(4)]。具体的には，部屋や病室に様々なセンサやアクチュエータを取り付け，ジェスチャ認識などにより人間の意図を理解し，要求に応じるといった支援を行うことが検討されている。このような空間のロボット化は，家電製品などの直感的な操作を行う上でも有効であり，我々も図 1 に示すように，部屋に複数のカメラを設置し，操作者のジェスチャを認識して家電製品の操作を行うインテリジェントルームを構築している[(5)]。このシステムでは，操作者が，何かを装着することなく，自然な状態で，テレビなどの家電製品をジェスチャのみで操作することが可能である。現状のシステムでは，操作する家電製品を選択するのに指差しを用いている。肌色情報を用いて指や腕の領域を複数台のカメラで抽出することで，指差し方向を 3 次元的に認識している。Yoda ら[(6)]も同様に腕による指示に着目し，ステレオカメラを用いた腕の方向認識を行っている。しかしなが

* 中央大学大学院理工学研究科 / JST CREST
〒112-8551　東京都文京区春日 1-13-27
School of Science and Engineering, Chuo Univ. / CREST, JST,
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551
** 中央大学理工学部 / JST CREST
〒112-8551　東京都文京区春日 1-13-27
Faculty of Science and Engineering, Chuo Univ. / CREST, JST,
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551

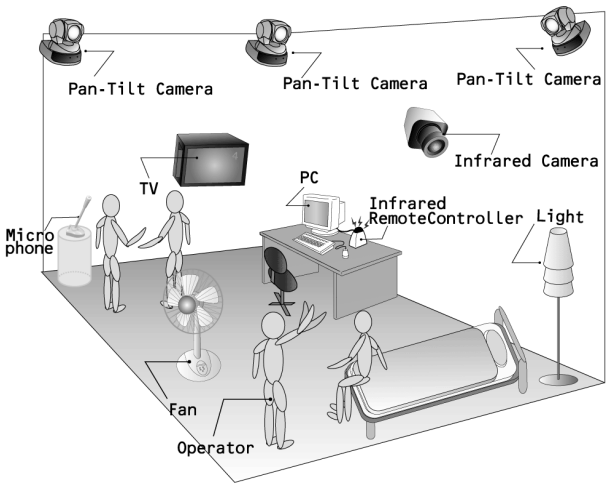


Fig. 1.　Conceptual figure of our intelligent room.

ら，指差し方向の認識は，カメラからの角度や領域の抽出され方によっては方向推定が困難となり，安定性に欠ける。そこで，物体指示のために，指差しに加え新たなモダリティーを導入することを考える。まず考えられるのは音声認識であり，これは人にとって直感的に理解しやすく有用であると考えられる。しかしながら，一般に音声認識は雑音に弱いという欠点があり，操作者がマイクを装着しない状態で安定に音声認識を行うことは困難である。また，部屋内で音を出したくない状況も考えられる。さらに，インテリジェントルームでは，手振りによる操作者の検出・3 次元位置計測およびその結果に基づいたパン・チルト・ズームを行うため，顔を十分な大きさで視野内にとらえることが可能である。そこで，本研究では，口唇動作の認識，すなわち読唇により物体指示を行うことを考え，カメラを用いた口唇動作認識手法の構築を行う。

2. 口唇動作認識処理の流れ

本研究では口唇動作を以下のように行うこととする。

・開始を示すため，発話前に口を開けしばらく静止させる

・終了を示すため，発話後にそのまましばらく動作を停止させる

・普通に静止した状態で行う（部屋の中を歩きながら，などは考えない）

口唇動作認識手法は既にいくつか提案されている。これらはモデルベースのもの[7][8]と画像ベースのもの[9]-[11]に分けられる。モデルベースのものはデータ量が少なく環境の変化に強いが，特徴的なモデルを作成する方法が問題点となる。一方，画像ベースのものはデータ取得が容易であるが，カメラとの距離の変化による唇の大きさの変化の影響を受ける。本研究では，この問題に対応した，画像ベースの手法を用いた手法を構築する。

図 2 に提案手法の流れを示す。まず，入力画像から操作者の顔を検出する。次に，得られた顔周辺の画像から口の周辺の画像を抽出し，さらに得られた口周辺の画像を低解像度化する。低解像度化された口周辺画像の時系列データに対し，登録されているモデルデータと DP マッチング[12][13]を行うことで，認識結果を得る。この時，発話中の口の位置ずれに対応するための処理を適用する。

3. 口唇動作認識手法

本章で，図 2 に示す口唇動作認識手法の各処理の詳細を述べる。

〈3･1〉 顔 検 出 顔検出手法は現在ではデジカメやプリンタにも実装されるなど広く用いられている。本研究では，Intel の OpenCV[14]の顔検出モジュールを用いる。これは Viola と Jones の手法[15]を用いたものである。本モジュールで得られる顔の位置・大きさ（半径）を用いて，入力画像から顔の周辺の画像を一定の比率で抽出する（図 3 参照）。

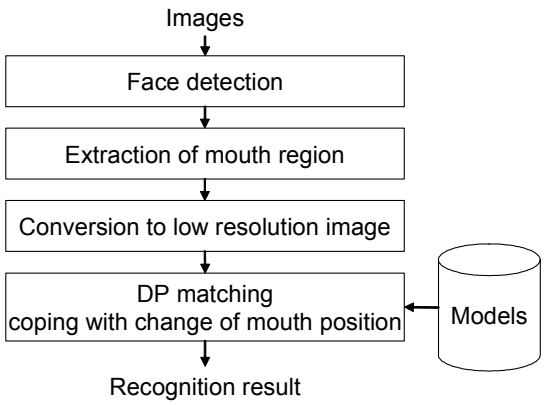


Fig. 2. Flow of mouth motion recognition.

Fig. 3. Detection of face region.

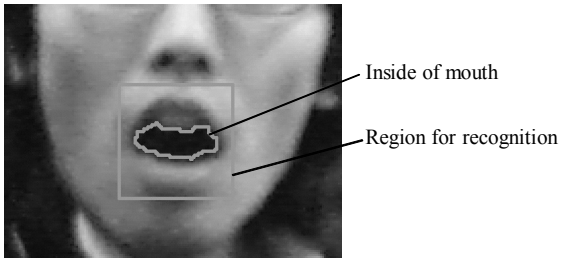


Fig. 4. Mouth region for recognition.

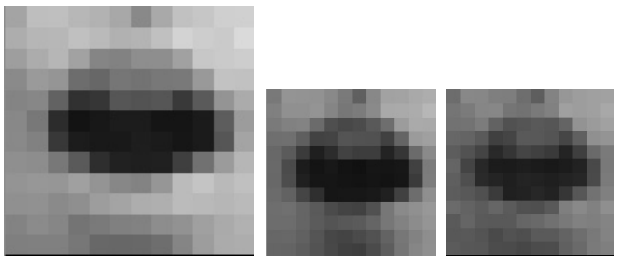


Fig. 5. Low resolution image.

〈3･2〉 口領域抽出 唇は赤色が強いことに着目し，カラー画像の R 画像を認識に用いる。口唇動作を認識するためにはまず口の位置を求める必要がある。本研究では，前章に示したように最初に口を開けることを仮定する。そこで，口を開けたときに口の中が暗く観察されることを利用して口の位置の検出を行う。画素値が閾値以下の領域を口の中とみなす。得られた領域の重心を求め，重心から顔の大きさに対して一定の比率の範囲を認識範囲に決定する。図 4 に抽出された口の中の領域と設定された認識範囲

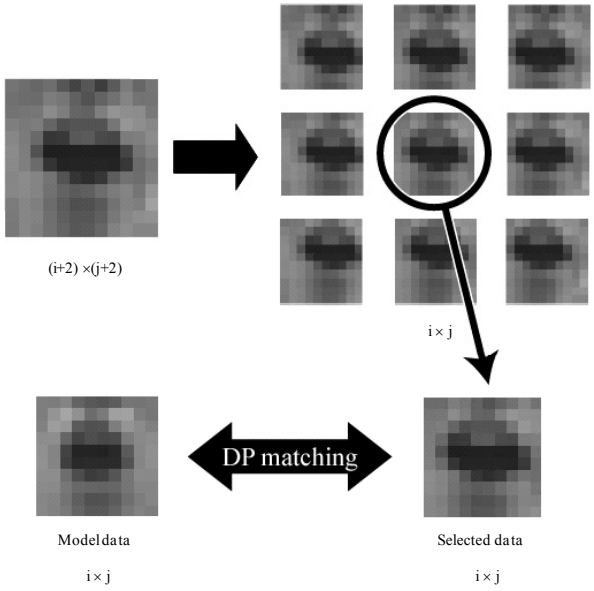


Fig. 6. Nine sub-images for change of mouth position.

の例を示す。なお，認識範囲の画像中での位置は口唇動作の認識中は固定する。

〈3･3〉 低解像度化 決定された認識範囲を低解像度化する。図 5 は図 4 を低解像度化した画像である。R, G, B 画像それぞれを示しているが，認識に用いる R 画像が最もコントラストが大きいことが分かる。最も低解像度化には，データ量を減らすと共に，同じ口唇動作に対しても毎回発生する動作のばらつきや，動作中の口の位置のずれを少量なら吸収できる利点がある。

〈3･4〉 口の位置ずれへの対応 低解像度化では吸収しきれない口の位置のずれが発話中に発生することが考えられる。このずれは認識率の低下の原因となる。これに対応するため，本研究では以下の手法を用いる。

口の低解像度画像を取得する際に，必要とする大きさよりも一回り大きい低解像度画像を取得し，入力データとする。具体的には，必要な低解像度画像の大きさ i×j 画素に対し，(i+2)×(j+2) 画素の低解像度画像を取得する。この画像から 9 パターンの i×j 画素の画像を切り出す。そして DP マッチングを行う時には，各フレームで 9 パターンの画像とモデルデータとの距離のうち最小のものを求め，それをそのフレームでの距離とする。この処理によって，低解像度画像での 1 ピクセル以内の口の位置のずれを吸収することが出来ると考えられる。以上の処理の流れを図 6 に示す。なお，この例では i=j=10 である。

〈3･5〉 DP マッチング 入力されたジェスチャの認識に DP (Dynamic Programming) マッチングを用いる。DP マッチングは入力された未知のパターンデータを事前に登録したモデルデータと比較，照合し類似度を計算する手法である。この手法は時系列データの取扱いが可能で入力データ数と標本データ数が異なる場合にも有効であるため，動画像認識や音声認識などに用いられている[13]。

口唇動作認識を DP マッチングの枠組みで定式化する。モデルデータと入力データのフレーム数をそれぞれ M, N とする。ある一点の画素を考え，m フレーム目のモデルデータの画素値，n フレーム目の入力データの画素値をそれぞれ $mPixelValue[m]$, $PixelValue[n]$ とおく。この時，距離 $PixelValuePD[m][n]$ (m=0,1,2,…,M　n=0,1,2,…,N) を次式で求める。

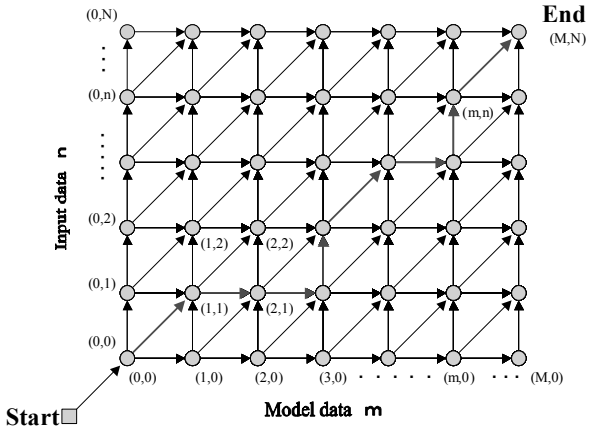


Fig. 7. Optimal path of input data and model data by DP matching.

$$PixelValuePD[m][n] = \frac{\left|mPixelValue[m]-PixelValue[n]\right|}{\sqrt{(mPixelValue[m])^2+(PixelValue[n])^2}} \quad \cdots\cdots(1)$$

(1)式の距離を全ての画素で求め，その平均を平均距離 $PD[m][n]$とおく。(0,0) フレーム目のデータ対から (m,n) フレーム目のデータ対までの距離 $TotalDistance[m][n]$ は次式で与えられる (図 7 参照)。

$$\begin{aligned} TotalDistance[m][n]= & \\ & min\{TotalDistance[m\text{-}1][n\text{-}1]+2PD[m][n], \\ & TotalDistance[m][n\text{-}1]+PD[m][n], \\ & TotalDistance[m\text{-}1][n]+PD[m][n]\} \end{aligned} \quad \cdots\cdots(2)$$

(2)式より，始状態 (0,0) フレーム目のデータ対から終状態 (M,N) フレーム目のデータ対までの距離 $TotalDistance[M][N]$ が求められる。得られた距離を正規化し，モデルデータと入力データとの距離を次式で与える。

$$Value=\frac{TotalDistance[M][N]}{M+N} \quad \cdots\cdots(3)$$

(3)式の値 Value を用いて認識を行う。登録したモデルジェスチャ全てに対して上記の処理を行い，Value が最小となったモデルジェスチャを認識結果とする。

4. プロトタイプシステムの構築

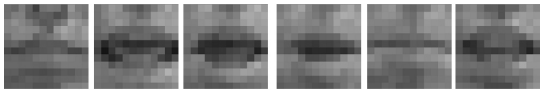
〈4･1〉 ハードウェアとソフトウェア 3 章で述べた手法を実装し，プロトタイプシステムを構築した。図 8 に示すように，PC モニタ上にセットした USB カメラ UCAM-E130HBU (ELECOM) により，対象人物の画像を撮

Fig. 8. Experimental setup.

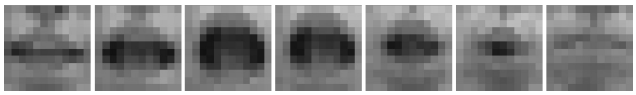
Table 1. Distance between model data.

	“fan”	“terebi”	“eakon”	“raito”
“fan”	0	0.159	0.161	0.156
“terebi”	—	0	0.159	0.140
“eakon”	—	—	0	0.152
“raito”	—	—	—	0

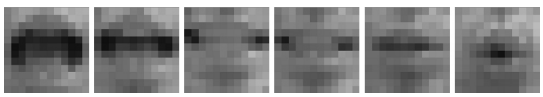
(a) “fan”



(b) “terebi”



(c) “eakon”



(d) “raito”

Fig. 9. Sequence of low-resolution images.

像する。カメラの画素数は 640×480 とした。サンプリングレートは実測 25.6fps（カタログ値 30fps）であった。カメラからの画像は PC (Pentium4 3.2GHz) に入力される。口唇動作認識手法の実装には，画像処理ソフト HALCON ver.7.0 (MVTec)，OpenCV (Intel)，Microsoft Visual C++を用いた。

顔・口領域の抽出に関しては，次のように設定した。顔検出モジュールで得られる顔の半径を r として，顔の位置から横方向に (r,-r)，下方向に (11/8r,-1/8r) の範囲を顔周辺画像とする。さらに，顔周辺画像の中心部の r×r 内で口の位置の検出を行い，得られた位置を中心とした 2/3r×2/3r の領域を口領域とする。

Table 2. Results for different distances

(a) 2m

Output / Input	“fan”	“terebi”	“eakon”	“raito”	Recognition rate
“fan”	20	0	0	0	100%
“terebi”	0	19	0	1	95%
“eakon”	0	4	16	0	80%
“raito”	0	1	0	19	95%

(b) 1m

Output / Input	“fan”	“terebi”	“eakon”	“raito”	Recognition rate
“fan”	20	0	0	0	100%
“terebi”	0	19	1	0	95%
“eakon”	0	1	18	1	90%
“raito”	0	2	0	18	90%

(c) 0.5m

Output / Input	“fan”	“terebi”	“eakon”	“raito”	Recognition rate
“fan”	15	1	4	0	75%
“terebi”	0	20	0	0	100%
“eakon”	0	2	18	0	90%
“raito”	0	4	2	14	70%

低解像度画像のサイズは 10×10 とした。この値は，出来るだけ解像度を下げ，そのなかで口唇動作の違いがわかるよう，定めた。

画像列からの口唇動作の切り出しは，次のように実装した。開始に関しては，現状ではキーボードで指示している。終了に関しては，フレーム間の変化量が小さい状態が 7 フレーム続いたことを検出した時に，その 7 フレーム前に動作が終了していたと判定する。開始・終了の間に得られた低解像度画像を用いて認識を行う。

〈4･2〉 モ デ ル ファン，テレビ，エアコン，ライトの 4 つの家電製品を対象とし，モデルを構築した。被験者が 4 つの機器名を発話し，それぞれの単語のモデルを登録した。被験者とカメラとの距離は 2m（4･4 節では 1m），照明には元々部屋に設置されている蛍光灯を用いた。登録された 4 つのモデルのフレーム数は，24（ファン），21（テレビ），25（エアコン），19（ライト）であった。

式(３)によって求めた各モデルデータ間のマッチング距離を表 1 に示す。各モデルデータ間の距離がほぼ等しいことが分かる。

〈4･3〉 実験による評価 カメラと被験者との距離が 0.5m，1m，2m の 3 つの場合で実験を行った。被験者はモデルデータと同一人物である。4 種の口唇動作を 20 回ずつ行い正しく認識する割合を調べた。実験の際に入力された口

Table 3. Distances between input and output data (upper: average, lower: standard deviation).

Input \ Output	"fan"	"terebi"	"eakon"	"raito"
"fan"	0.181 0.009	0.193 0.008	0.194 0.008	0.194 0.007
"terebi"	0.200 0.004	0.162 0.007	0.178 0.003	0.173 0.004
"eakon"	0.192 0.007	0.188 0.006	0.185 0.005	0.191 0.006
"raito"	0.200 0.005	0.194 0.004	0.197 0.004	0.183 0.004

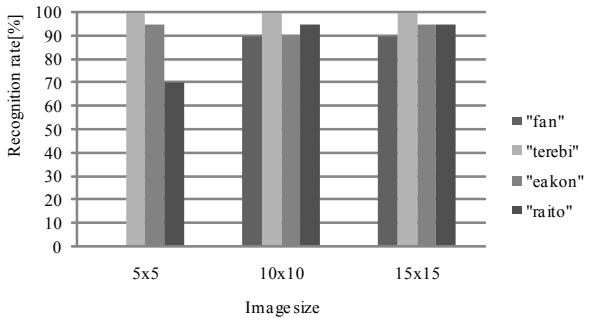


Fig. 10. Recognition rates for various image sizes.

唇動作の数フレームおきの低解像度画像の例を図 9 に示す。この図に示すように，発話時には多少意識して口の動作を明瞭にしている。なお，次節の実験では「ファン」以外は特段意識的な動作はしていない。認識結果を表 2 に示す。また，距離 1m での各口唇動作入力時の入力データとモデルデータの平均距離と標準偏差を表 3 に示す。

表 2 より，口唇動作の平均認識率は，距離 2m，1m，0.5m でそれぞれ約 92%，約 94%，約 84%である。提案手法が実用的な認識率を実現していること，また距離の変化に対しロバストであることが分かる。すなわち，モデル登録時の 2m に対し，距離が半分の 1m で認識率は高いままであり，距離が 1/4 の 0.5m でもまだ比較的高い認識率を維持している。提案手法により，画像ベースの口唇動作認識手法の弱点であるカメラとの距離の変化による唇の大きさの変化に対応できていると言える。

各単語の結果を評価する。全体的として，「ファン」と「テレビ」の認識率が高い。「ファン」は距離 1m，2m で認識率が 100%である。また，「テレビ」はすべての距離で高い認識率となっている。一方「エアコン」と「ライト」の認識率が低い。これは表 3 より説明できる。「ファン」と「テレビ」では，同じ単語との距離が他の単語との距離に対し十分小さく（すなわち高い類似度に）なっている。一方，「エアコン」では，他の単語，特に「テレビ」との距離が小さな値となっており，表 2 の結果で「テレビ」と誤認識される回数が多くなっている。同様に，「ライト」でも，同じく「テレビ」との距離が比較的小さく，表 2 の結果で「テレビ」と誤認識される回数が多くなっている。以上より，当然ではあるが，単語の組み合わせによって，誤認識しやすいもの，しにくいものがあること，また，データ間の距離を求

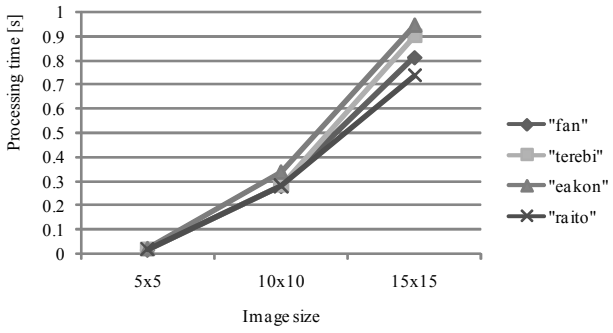


Fig. 11. Processing time for various image sizes.

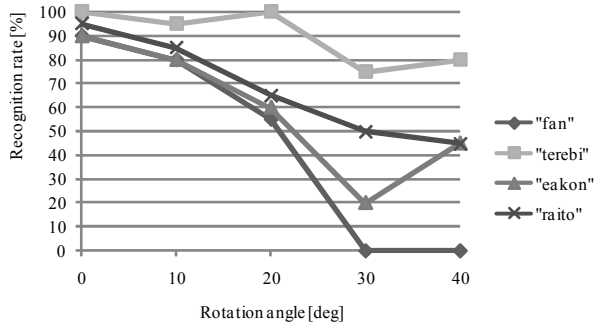


Fig. 12. Recognition rates for rotation.

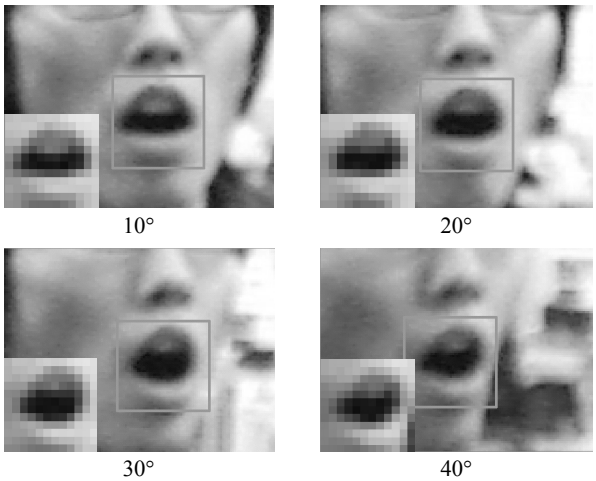


Fig. 13. Extracted face and mouth images with rotation.

めることによって，ある程度誤認識の生じやすさを見積もることができることが示されていると言える。

〈4･4〉 各種要因の影響の評価 前節では距離の影響を評価したが，さらにいくつかの要因の影響を実験的に評価した。本節の実験では，モデルの登録および実験における距離はすべて 1m とした。また，それぞれの実験での各単語に対する実験回数は前節と同様，20 回とした。

(1) 低解像度化 低解像度画像のサイズを 5×5，10×10，15×15 と変えた時の認識率を比較した。結果を図 10 に示す。(a)の「ファン」の認識率は 0%である。また，図 11 に各単語に対する発話終了判定後の処理時間を示す。平均処理時間は，順に 0.019s, 0.296s, 0.849s である。これらの結

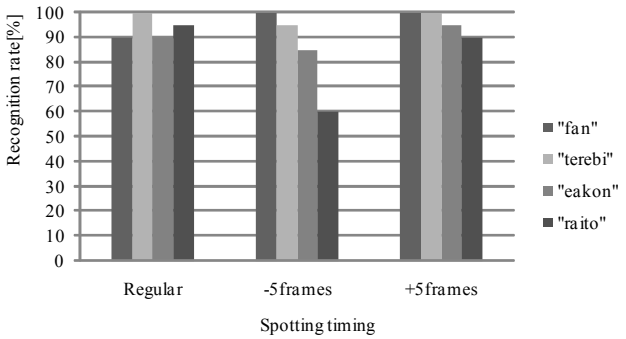


Fig. 14. Effect of errors of spotting timing.

Table 4. Recognition rates for various numbers of words [%].

	4 words	8 words	12 words
"fan"	90	90	50
"terebi"	100	55	95
"eakon"	90	50	10
"raito"	95	50	45
"video"	-	100	35
"telephone"	-	100	60
"camera"	-	55	20
"heater"	-	80	75
"pc"	-	-	75
"kompo"	-	-	20
"oobun"	-	-	80
"renji"	-	-	100
Average	93.8	72.5	55.4

Table 5. Results for different subjects

(a) Subject 1

Output / Input	"fan"	"terebi"	"eakon"	"raito"	Recognition rate
"fan"	18	0	2	0	90%
"terebi"	0	20	0	0	100%
"eakon"	0	2	18	0	90%
"raito"	0	0	1	19	95%

(b) Subject 2

Output / Input	"fan"	"terebi"	"eakon"	"raito"	Recognition rate
"fan"	20	0	0	0	100%
"terebi"	0	19	1	0	95%
"eakon"	2	1	17	0	85%
"raito"	2	0	2	16	80%

(c) Subject 3

Output / Input	"fan"	"terebi"	"eakon"	"raito"	Recognition rate
"fan"	20	0	0	0	100%
"terebi"	0	12	0	8	60%
"eakon"	2	0	17	1	85%
"raito"	4	1	0	15	75%

果より，サイズが 5×5 では認識率が低いのに対し，10×10, 15×15 ではほぼ同様の認識率であり，また，サイズの増加に伴って処理時間が大きく増加することが分かる。認識率と処理時間のバランスから，本論文で用いている 10×10 は適切なサイズであると言える。

(2) 位置ずれへの対応　〈3･4〉節で示した位置ずれへの対応手法の有効性の検証を行った。その結果，対応を行った場合の認識率が，ファン，テレビ，エアコン，ライトの順に 90, 100, 90, 95%であったのに対し，対応を行わない場合は順に 75, 75, 90, 65%であった。対応を行った場合の認識率が高いことから，対応手法の有効性が言える。

(3) 姿勢の影響　インテリジェントルームでは，顔がカメラ正面を向くとは限らない。そこで，斜めを向いた時の認識率の変化を調べた。カメラ正面を向いている状態から 10 度刻みで 40 度まで変化させた時の単語ごとの認識率の変化を図 12 に示す。また，その時の切り出した顔画像領域ならびに口の低解像度画像を図 13 に示す。角度の増加に伴い認識率が低下すること，また単語によって角度の増加の影響が異なることが示されている。なお，角度 50 度以上では，顔画像領域の切り出しサイズが不安定となった。

(4) 切り出しが前後した場合の影響　画像の切り出しの終了時のタイミングを故意に前後させた時の認識率の変化を図 14 に示す。切り出しが 5 フレーム増減した場合の結果である。減少した場合は認識率が低下しているのに対し，増加した場合はほとんど変化がない。これは，実験で用いた 4 単語の終了時の口の形が異なっていること，また DP マッチングのデータの伸縮への対応が働いていることによるものと思われる。

(5) カテゴリ数の変化　単語数を 4, 8, 12 と変化させた場合の認識率の変化を調べた。表 4 に用いた単語ならびに単語ごとおよび平均の認識率を示す。単語数の増加に伴い認識率は低下し，特に 12 単語では平均認識率が 50%近くまで落ち込んでいる。現状のシステムでは 10 単語以上への適用は困難であると考えられる。

(6) 個人差　3 人の被験者での認識率の比較を行った。モデルは，被験者ごとに登録したものを用いた。結果を表 5 に示す。Subject 1 が他の実験と同じ被験者，また他の 2 人は本システムに習熟していない被験者である。3 人とも 20 代の男性である。平均の認識率が順に 93.8, 90, 80%とシステムへの習熟度などにより認識率に差があるものの，各被験者で提案手法が利用可能であることが分かる。また，誤認識の生じ方に個人差があることも示されている。

以上，各種の要因に影響に関する実験結果を示し，提案手法の能力を示した。大まかにまとめると，構築したプロトタイプは一定の認識能力を示しており，位置ずれへの対応も妥当である，ただし姿勢変化に関しては必ずしもロバ

ストではなく，また適用可能な単語数もそれ程多くない。各単語に複数のモデルを持たせる，姿勢毎にモデルを持たせるなどの工夫によりこれらの点を改善していくことが今後必要であると考えられる。

5. 結　論

本論文では，インテリジェントルームで家電製品を指示するために用いることを想定した，口唇動作認識手法を提案した。提案手法は，まず操作者の顔を検出し，次いで顔画像中から口の中が暗いことを利用して口領域を抽出する。低解像度化された口領域画像の時系列データを入力データとして，DP マッチングにより口唇動作を認識し，発話された単語を認識する。低解像度化の際に 1 画素のオフセット分を考慮して 9 枚の画像を用いることにより，発話中の口の位置のずれを吸収する。代表的な家電製品の 4 単語の認識実験により，提案手法の有効性を評価した。提案システムが画像ベースの口唇動作認識手法の弱点であるカメラとの距離の変化による唇の大きさの変化に対応していることも実験により示されている。さらに各種の要因の影響を実験的に評価し，提案手法・構築したプロトタイプの能力を示した。なお，提案手法は，インテリジェントルームでの利用を想定したものであるが，勿論それ以外でも利用可能な独立した手法である。

認識率の向上，ならびにインテリジェントルームへの実装が今後の課題として挙げられる。

（平成 20 年 10 月 2 日受付，平成 21 年 1 月 12 日再受付）

文　献

(1) T. Mori and T. Sato : “Robotic Room: Its Concept and Realization, Robotics and Autonomous Systems”, Robotics and Autonomous Systems, Vol.28, No.2, pp.141-144 (1999)

(2) J. H. Lee and H. Hashimoto, H. : “Intelligent Space -Concept and Contents –”, Advanced Robotics, Vol.16, No.4, pp.265-280 (2002)

(3) T. Mori, H. Noguchi, and T. Sato : “Sensing room -Room-type behavior measurement and accumulation environment-”, Journal of the Robotics Society of Japan, Vol.23, No.6, pp.665-669 (2005) (in Japanese)
森　武俊・野口博史・佐藤知正：「センシングルームー部屋型日常行動計測蓄積環境第 2 世代ロボティックルームー」，日本ロボット学会誌, Vol.23, No.6, pp.665-669 (2005)

(4) B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer : “EasyLiving: Technologies for Intelligent Environments”, Proc. Int. Symp. on Handheld and Ubiquitous Computing, pp.12-27 (2000)

(5) K. Irie, N. Wakamura, and K. Umeda : “Construction of an Intelligent Room Based on Gesture Recognition”, Trans. of the Japan Society of Mechanical Engineers, Series C, Vol.73, No.725, pp.258-265 (2007) (in Japanese)
入江耕太・若村直弘・梅田和昇：「ジェスチャ認識に基づくインテリジェントルームの構築」, 日本機械学会論文集 C 編, **73**, 725, pp.258-265 (2007)

(6) I. Yoda, K. Sakaue, and Y. Yamamoto : “Arm-Pointing Gesture Interface Using Surrounded Stereo Cameras System”, Proc. Int. Conf. on Pattern Recognition, Vol.4, pp.965-970 (2004)

(7) T. Saitoh and R. Konishi : “Lip Reading Based on Trajectory Feature”, *Trans. IEICE*, Vol.J90-D, No.4, pp.1105-1114 (2007-4) (in Japanese)
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」，信学論, **J90-D**, 4, pp.1105-1114 (2007-4)

(8) Y. Ogoshi, H. Ide, C. Araki, and H. Kimura : “Active Lip Contour Using Hue Characteristics Energy Model for A Lip Reading System”, *Trans. IEEJ*, Vol. 128, No. 5, pp.811-812 (2008-5) (in Japanese)
小越康宏・井出寿登・荒木睦大・木村春彦：「読唇のための色特徴量に基づくエネルギーを用いた動的口唇輪郭抽出法」，電学論Ｃ，**128**, 5, pp.811-812 (2008-5)

(9) K. Kiyota and K. Uchimura : “An Utered Word Recognition Using Lip Image Information”, *Trans. IEICE*, Vol.J76-D-II, No.3, pp.812-814 (1993) (in Japanese)
清田公保・内村圭一：「口唇周辺画像情報を用いた発話単語認識」，信学論, **J76-D-II**, 3, pp.812-814 (1993)

(10) Y. Nankaku, K. Tokuda, T. Kitamura, and T. Kobayashi : “Normalized Training for HMM-Based Visual Speech Recognition”, *Trans. IEICE*, Vol.J86-D-II, No.2, pp.163-172 (2003) (in Japanese)
南角吉彦・徳田恵一・北村　正・小林隆夫：「隠れマルコフモデルを用いた視覚音声認識のための正規化学習」，信学論, **J86-D-II**, 2, pp.163-172 (2003)

(11) C. Bregler and Y. Konig : “‘Eigenlips' for robust speech recognition”, Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp.669-672 (1994)

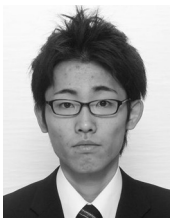
(12) T. Nishimura, T. Mukai, S. Nozaki, and R. Oka : “Spotting Recognition of Gestures Performed by People from a Single Time-Varying Image Using Low-Resolution Features”, *Trans. IEICE*, Vol.J80-D-II, No.6 pp.1563-1570 (1997) (in Japanese)
西村拓一・向井理朗・野崎俊輔・岡　隆一：「低解像度特徴を用いた複数人物によるジェスチャの単一動画像からのスポッティング認識」，信学論, **J80-D-II**, 6, pp.1563-1570 (1997)

(13) S. Uchida and H. Sakoe : “Analytical DP Matching”, *Trans. IEICE*, Vol.J90-D, No.8, pp.2137-2146 (2007) (in Japanese)
内田誠一・迫江博昭：「解析的 DP マッチング」，信学論, **J90-D**, 8, pp.2137-2146 (2007)

(14) http://www.intel.com/technology/computing/opencv/index.htm

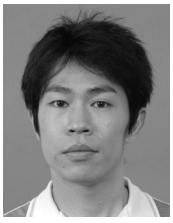
(15) P. Viola and M. Jones : “Rapid object detection using a boosted cascade of simple features”, Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Vol.1, pp.511-5280 (2001)

中　西　達　也



（非会員） 1984 年 7 月生。2007 年中央大学理工学部精密機械工学科卒業。2007 年同大学院理工学研究科精密工学専攻入学，現在に至る。画像を用いたヒューマンインタフェース，画像処理の研究に従事。

寺　林　賢　司



（非会員） 1979 年 10 月生。2002 年，北海道大学工学部システム工学科卒業。2004 年，同大学院システム情報工学専攻修士課程修了。2008 年，東京大学大学院工学系研究科精密機械工学専攻博士課程修了，博士（工学）。同年，中央大学理工学部精密機械工学科助教，現在に至る。ヒューマンインタフェース，バーチャルリアリティ等の研究に従事。日本ロボット学会，日本バーチャルリアリティ学会，日本機械学会等の会員。

梅　田　和　昇

（非会員） 1967 年 2 月生。1989 年東京大学工学部精密機械工学科卒業, 1994 年同博士課程修了。同年中央大学理工学部精密機械工学科専任講師，2007 年より同教授，現在に至る。2003-2004 年カナダ NRC Visiting Worker。ロボットビジョンの研究に従事。画像の認識・理解シンポジウム MIRU2004 長尾賞受賞。博士（工学）。電子情報通信学会，日本ロボット学会，日本機械学会，情報処理学会，IEEE 等の会員。