テンプレートマッチングを用いた口唇動作認識

Mouth Motion Recognition Using Template Matching

○学　滝田清（中央大）　　永易武（中央大）
浅野秀胤（Pioneer）
正　寺林賢司（中央大）　正　梅田和昇（中央大）

Kiyoshi TAKITA, Chuo University, takita@mech.chuo-u.ac.jp
Takeshi NAGAYASU, Chuo University
Hidetsugu ASANO, Pioneer
Kenji TERABAYASHI, Chuo University
Kazunori UMEDA, Chuo University

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. Therefore, an intelligent room which recognizes gestures and support operators is required in various places in recent years. In this paper, we propose a method to recognize mouth motion from images. The proposed method uses template matching and recognizes mouth motion for indicating a target object in an intelligent room. DP matching is applied to similarity measure obtained by template matching. Several experiments are performed to demonstrate the effectiveness of the proposed method.

***Key Words*:** Intelligent Room, Mouth Motion Recognition, Template Matching, Image Processing

1. 序論

日常生活において欠かすことのできない家電製品が，多機能化，高機能化している．しかしその一方で，操作の複雑化という問題も生じている．近年，家電製品のように人に身近な製品を，人のジェスチャを用いて直感的に操作することを目的とした研究[1]が多くなされている．我々は，図 1 に示すように，部屋の四隅にカメラを設置し，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[2]．先行研究[3]で家電製品の操作のための手法の一つとして，口唇動作認識を用いている．このシステムでは口唇動作を認識するため，特徴量として口元の画像を低解像度化した画像[3]，唇の形状，口を開いた時に現れる開口部分の形状[4]を用いている．しかし，実際にシステムをインテリジェントルームに実装すると，画像を圧縮することで発生するブロックノイズなどの影響から，安定して特徴量を取得することが出来ない[5]．そこで本研究では，インテリジェントルームにおいても安定した特徴量の取得が可能な手法として，テンプレートマッチングを用いた口唇動作認識を提案する．

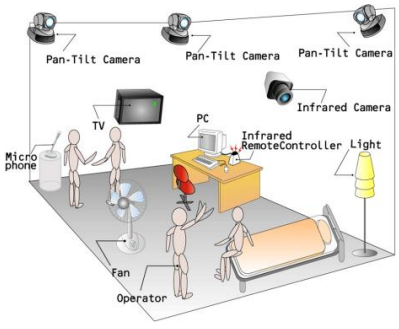


Fig.1　Conceptual figure of our intelligent room

2. 認識範囲決定手法

口唇動作を認識するため，取得した画像に対し顔検出，唇検出を順に行い，認識範囲を決定する．この時，図 2 に示すように操作者とカメラとの距離が異なる場合でも認識範囲内の唇の割合が一定になるように認識範囲を決定する．

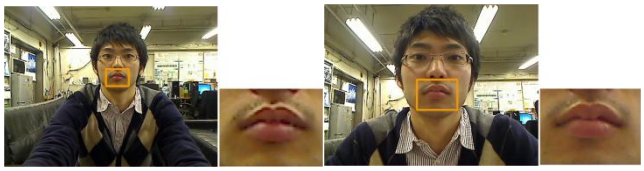


Fig.2　Region for recognition of mouth motion

2.1 顔検出

取り込んだ画像に OpenCV の顔検出モジュール[6]を用いることで顔の位置と大きさを特定する．本モジュールでは顔とされる箇所を円と仮定するので，円の中心を顔の中心，円の直径を顔の大きさとする．

2.2 唇検出

検出した顔の中心，大きさから経験的に口元の位置を特定し，口元領域を抽出する．抽出した口元領域に対し，図 3 の唇の画像のテンプレートを用いてテンプレートマッチングを行うことで唇を検出する．唇のテンプレートは 7 枚の唇の画像を合成して作成する．この時，テンプレートのサイズを変えて 4 枚用意しておき，類似度が一番高いテンプレートの幅，高さを唇の幅，高さとする．こうすることでカメラとの距離が異なっても認識範囲内の唇の割合を一定にする．類似度は次式の正規化相関で得る． この際，テンプレートの大きさを M×N，テンプレートの位置(i , j)における画素値を T(i , j)，テンプレートと重ね合わせた対象画像の画素値を I(i , j)とする．

$$\bar{I} = \frac{1}{MN}\sum_{j=0}^{N-1}\sum_{i=0}^{M-1} I(i,j) \quad , \quad \bar{T} = \frac{1}{MN}\sum_{j=0}^{N-1}\sum_{i=0}^{M-1} T(i,j)$$

$$R_{ZNCC} = \frac{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}((I(i,j)-\bar{I})(T(i,j)-\bar{T}))}{\sqrt{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}(I(i,j)-\bar{I})^2 \times \sum_{j=0}^{N-1}\sum_{i=0}^{M-1}(T(i,j)-\bar{T})^2}} \qquad (1)$$

検出した唇のテンプレートの中心と幅，高さから認識範囲を決定し，認識範囲の画像を 200×150 にリサイズする．認識範

囲の位置，大きさは口唇動作の認識中は固定する．

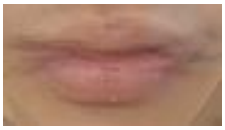


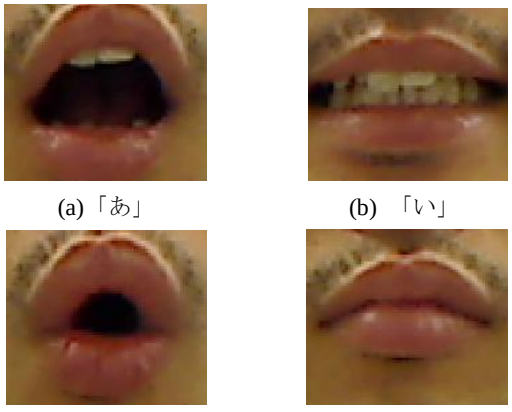
Fig.3　Template image of lip

3.　口唇動作認識手法

操作者はシステムが認識範囲を決定した後に口唇動作を始め，特徴量の時系列データと事前に登録しておいたモデルデータを DP マッチングで比較し，入力単語を認識する．

3.1 特徴量の取得

口唇動作を認識するための特徴量として，テンプレートマッチングで得ることのできる類似度を用いる．テンプレートは口の形状が異なる 4 枚の画像を用意する．具体的には「あ」，「い」,「う」, と発声している際の 3 枚の画像と口を閉じ,「ん」の状態の画像の計 4 枚である．「え」,「お」に関しては「い」,「う」の時の形状と似ているため本研究では用いない．それぞれの画像の例を図 4 に示す．テンプレートのサイズは認識範囲よりも小さい 150×130 にし，認識範囲内で口の位置がずれても影響がでないようにする．各フレームで 4 枚の画像それぞれをテンプレートとし，各テンプレート 1 回のテンプレートマッチングを行う．その結果，得られた 4 つの類似度を特徴量として取得する．類似度の計算には式(1)を用いる．



(a)「あ」　(b)「い」

(c)「う」　(d)「ん」

Fig.4　Template images for recognition

3.2 モデルデータとの比較

モデルデータとの比較には DP (Dynamic Programming)マッチングを用いる．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．m フレーム目のモデルデータで，4 枚あるテンプレートの内の t 枚目をテンプレートとしている際のテンプレートマッチングの計算結果を MTemplate[m][t]，n フレーム目の入力データで，4 枚あるテンプレートの内の t 枚目をテンプレートとしている際のテンプレートマッチングの計算結果を Template[n][t]と置く．この時，距離 TPD[m][n][t] (m=0,⋯,M　n=0,⋯,N　t=1,⋯,4)を次式で求める．

$$\mathrm{TPD}[m][n][t] = \frac{|\mathrm{MTemplate}[m][t] - \mathrm{Template}[n][t]|}{\sqrt{(\mathrm{MTemplate}[m][t])^2 + (\mathrm{Template}[n][t])^2}} \quad (2)$$

始状態 (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の距離 TTD[m][n][t]を次式で求める．

$$\begin{aligned}\mathrm{TTD}[m][n][t] = \min \{&\mathrm{TTD}[m-1][n-1][t] + 2\mathrm{TPD}[m][n][t],\\ &\mathrm{TTD}[m][n-1][t] + \mathrm{TPD}[m][n][t],\\ &\mathrm{TTD}[m-1][n][t] + \mathrm{TPD}[m][n][t]\}\end{aligned} \quad (3)$$

式(3)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 TTD[M][N][t]を求める．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 TValue[t]を次式で求める．

$$\mathrm{TValue}[t] = \frac{\mathrm{TTD}[M][N][t]}{M+N} \quad (4)$$

式 (4) より得られた TValue[1] ， TValue[2] ， TValue[3] , TValue[4]を次式により足し合わせ，AllTValue を求める．

$$\mathrm{AllTValue} = \sqrt{\mathrm{TValue}[1]^2 + \mathrm{TValue}[2]^2 + \mathrm{TValue}[3]^2 + \mathrm{TValue}[4]^2} \quad (5)$$

上記の処理を登録したモデルデータ全てに対して行い，モデルデータと入力データとの距離が最小となったモデルデータを認識結果とする．

4.　インテリジェントルームへの実装

インテリジェントルーム[2]に口唇動作認識を実装する．インテリジェントルームでは手振りを検出することで操作者の特定を行い，検出した手振り位置にズームし，カメラの視野内に操作者の手を十分な大きさで捉えることで手のジェスチャを識別している．このシステムへの実装方法として，まず，手振り検出により操作者の特定を行う．図 5 に示すのがカメラから取得された画像で，図 6 が手振り検出した際の様子である．次に，検出した手振り位置に一度目のズームを行い，一度目のズームが終了した時点で 2.1 節で説明した方法で顔検出を行う．図 7 にその様子を示す．最後に顔の中心に二度目のズームを行うことでカメラの視野内に操作者の顔を十分な大きさで捉える．その後は上記で示した方法で認識範囲を決定し，口唇動作の認識を行う．図 8 に顔へのズームを行った後の顔検出の様子を示す．

Fig.5　Input image using pan-tilt-zoom camera

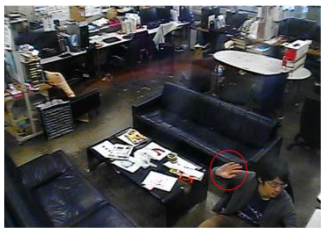


Fig.6　Waving hand detection

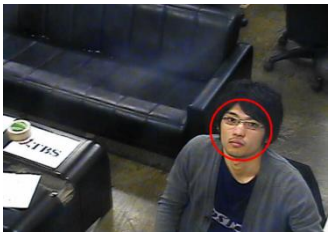


Fig.7　Face detection after the first zoom

Fig.8　Face detection after the second zoom

5.　実験

構築した口唇動作認識システムの有用性を検証するため実験を行った．5.1 節で固定カメラを用いた際のシステムの有用性を検討し，5.2 節でインテリジェントルームに実装した際の有用性を検討する．

5.1 固定カメラを用いた実験

5.1.1 実験システムおよびパラメータ設定

実験システムは Web カメラ C905m (Logicool，画素数 640×480)，PC (Core i7 CPU 930 2.80GHz，DDR3 6.00GB)，画像処理ソフト (Intel OpenCV)により構成される．

認識範囲の決定の際に用いるパラメータは次のように設定した．顔検出で得られる顔の半径を r として，顔の中心から横方向に(-r/2,r/2)，縦方向に(0,r)の範囲を抽出し，抽出した範囲に対し唇検出を行う．唇検出で用いる 4 枚のテンプレートのサイズはそれぞれ 80×44，100×55，115×63，130×77 にした．唇検出で得られる唇の幅，高さを w，h として，唇の中心から横方向に(-3/4w,3/4w)，縦方向に(-h/2-5,9/8w-h/2-5)の範囲を抽出し，認識範囲とする．単位は pixel である．

5.1.2 認識実験

カメラを被験者の顔とほぼ同じ高さ，前方 0.4[m]に設置し，蛍光灯下で 1 名の被験者 A に対して実験を行った．認識する単語として「Video」，「TV」，「Light」，「Up」，「Down」を対象とした．モデルデータとして被験者が各単語を一回発話した時の特徴量の時系列データを用いた．各口唇動作を 50 回ずつ行い，認識率を調べた．

まず，自身のテンプレート，モデルデータを用いて実験を行った．結果を表 1 に示す．

この場合，認識率の平均が 98%と高い認識結果が得られた．これは特徴量が安定して取得できたからだと考えられる．自身のテンプレート，モデルデータを用いた場合、提案した口唇動作認識システムが有用であるといえる．

次に被験者とモデルデータを登録した人物が異なる場合の認識率を調べるため，被験者 A に対して他人のテンプレート，モデルデータを用いて実験を行った．結果を表 2 に示す．

この場合，認識率の平均は 82%だった．単語ごとに認識結果をみると，「Video」，「TV」，「Light」，「Down」と発話した時は，比較的高い認識結果が得られた．一方で「Up」と発話した時は低い認識結果だった．このことから発話する単語によって認識率に大きな差が出ると考えられる．

Table 1　Results of subject A using his own template and model data [%]

input \ output	Video	TV	Light	Up	Down
Video	98	0	0	0	2
TV	2	98	0	0	0
Light	0	0	98	2	0
Up	0	0	4	96	0
Down	0	0	0	0	100

Table 2　Results of subject A using another person's template and model data [%]

input \ output	Video	TV	Light	Up	Down
Video	98	0	0	0	2
TV	0	98	0	0	2
Light	2	0	96	0	2
Up	8	0	2	38	52
Down	16	2	2	0	80

5.2 インテリジェントルームのカメラを用いた実験

5.2.1 実験システムおよびパラメータ設定

実験システムはパン・チルト・ズームカメラ AXIS 233D(AXIS, 画素数 640×480)，PC(Core i7 CPU 930 2.80GHz, DDR3 6.00GB)，画像処理ソフト(Intel OpenCV)により構成される．

認識範囲の決定の際に用いるパラメータは 5.1.1 項と同じ値を用いた．

5.2.2 認識実験

設置したパン・チルト・ズームカメラが被験者の顔から高さ 1.4[m]，前方 2[m]，3[m]，4[m]になる位置で，照明環境としては蛍光灯下で実験を行った．図 9 にそれぞれの位置でのカメラへの映り方を示す．モデルデータとして被験者が各単語を一回発話した時の特徴量の時系列データを用いた．1 名の被験者 A に対し各口唇動作を 50 回ずつ行い，認識率を調べた．認識する単語として「Video」，「TV」，「Light」，「Up」，「Down」を対象とした．

モデルデータ，テンプレートは自身のものを用いた．モデルデータとテンプレートはカメラが前方 2[m]になる位置のものを用いた．各距離で認識範囲の RGB 画像とそれをグレースケール化した画像の 2 種類，テンプレートのサイズを 10×10 を 1 ピクセルに低解像度化したものと低解像度化しない画像を用いた 2 種類を組合わせ，計 4 種類の実験を行った．認識

結果を表 3 に示す．以下の表で「color」は色情報を表し，RGB 画像ならば「RGB」，グレースケールならば「Gray」と表記し，「low」は低解像度化して 1 ピクセルにしたサイズを表す．

Table 3　Results of subject A using his own template and model data with the PTZ camera in the intelligent room

(a) Distance is 2m [%]

feature \ Word		Video	TV	Light	Up	Down	Average
color	low						
RGB		98	100	92	84	96	93.2
RGB	10×10	90	98	96	94	94	94.4
Gray		82	100	96	82	96	91.2
Gray	10×10	90	100	80	98	82	90

(b) Distance is 3m [%]

feature \ Word		Video	TV	Light	Up	Down	Average
color	low						
RGB		86	94	70	94	56	80
RGB	10×10	88	98	86	24	90	77.2
Gray		92	88	100	40	96	83.2
Gray	10×10	100	100	78	2	94	74.8

(c) Distance is 4m [%]

feature \ Word		Video	TV	Light	Up	Down	Average
color	low						
RGB		92	62	42	96	16	61.4
RGB	10×10	82	74	6	0	90	50.4
Gray		52	98	14	96	4	52.8
Gray	10×10	84	100	38	66	84	74.4

2[m]の位置ではテンプレートによらず高い認識結果が得られた．これはテンプレート，モデルデータを取得した際の距離と認識実験を行った距離が同じだったため，モデルデータと類似した特徴量の時系列データが取得できたからだと考えられる．

3[m]の位置では，低解像度化を行わないテンプレートを用いたほうが高い認識結果が得られた．これは，実験を行った位置とテンプレート，モデルデータを取得した際の位置は異なるが，テンプレートの画像との違いが少ないため，低解像度化を行わないほうがモデルデータと類似した特徴量の時系列データが取得できたからだと考えられる．

4[m]の位置ではグレースケールで低解像度化したテンプレートを用いた場合，一番高い認識結果が得られた．これは，グレースケールで低解像度化したテンプレートを用いることで，距離の違いにより現れるテンプレートとの光の当たり方，口の傾きの違いを一番吸収できたからだと考えられる．

(a) 2m　(b) 3m

(c) 4m

Fig.9　Images of intelligent room camera

6.　結論

本研究では，インテリジェントルームに口唇動作認識システムを実装し，実験によりシステムの有用性を検証した．操作者とカメラの距離に応じて高い認識結果が得られるテンプレートの種類が異なるので，今後は距離に応じたテンプレートの選択を行うことで，安定して高い認識率を実現できるシステムの構築を目指す．

7.　謝辞

本研究は科研費(19100004)の助成を受けたものである．

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]　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, Vol.73, No.725, pp.258-265, 2007.

[3]　中西達也，寺林賢司，梅田和昇，“DP マッチングを用いた口唇動作認識”，電気学会論文誌 C，Vol.129, No.5, pp.940-946, 2009.

[4]　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，Vol.J90-D, No.4, pp.1105-1114, 2007.

[5]　滝田清，永易武，浅野秀胤，寺林賢司，梅田和昇，“DP マッチングを用いた口唇動作認識における特徴量の検討”，動的画像処理実利用化ワークショップ 2011 講演論文集，pp.302-307, 2011

[6]　OpenCV-1.1pre リファレンス マニュアル
http://opencv.jp/opencv-1.1.0/document/index.html