I-60

音声と顔画像を用いた個人認識

Person Recognition Using Speech and Face

山本真理† 柴田沙矢香† 南角吉彦† 宮島千代美† 徳田恵一† 北村正†
M. Yamamoto S. Shibata Y. Nankaku C. Miyajima K. Tokuda T. Kitamura

1. はじめに

バイモーダル情報を用いることにより，個人認識システムの高性能化を目指す研究が行われている [1]–[5]．音声と顔画像を用いた個人認識もその一つである [6],[7]．従来，顔画像のモデル化には，固有顔 [8] 等が用いられ，音声のモデル化には隠れマルコフモデル (HMM) や混合ガウスモデル (GMM) [9] が多く用いられている．

本研究では，HMM に基づいた顔画像モデルと GMM に基づいた音声モデルを組み合わせたバイモーダル個人認識について検討する．文献 [10] では，Y 方向にフレームをシフトしながら切り出した顔画像の 2 次元 DCT 係数を HMM でモデル化し，顔画像認識を行っている．本研究では，顔画像を Y 方向のみでなく，X 方向についても同様にシフトして HMM でモデル化し，これらを組み合わせることにより，顔画像認識部の性能向上を図る．本システムを個人識別及び個人照合実験において評価し，個人照合実験では，バックグラウンドモデルによる尤度正規化と，話者別の閾値設定についても検討する．

2. 音声と顔画像のモデル化と統合方法

図 1 に音声と顔画像を用いた個人認識システムのブロック図を示す．Y 方向にシフトした場合に得られる横長の画像のフレーム系列と，X 方向にシフトした場合に得られる縦長の画像のフレーム系列をそれぞれ独立の HMM でモデル化し，重み付けして統合する．次に，顔画像の統合スコアと，音声の GMM の対数尤度についても重み付けにより統合する．

2.1 音声のモデル化

音声は，サンプリング周波数が 12kHz で，長さ 32ms のフレームを周期 8ms で切り出し，各フレームにブラックマン窓をかけ，14 次のメルケプストラム分析を行う．得られたメルケプストラム係数のうち，0 次を除く 14 個の係数を特徴量として用い，32 混合の GMM によりモデル化する．

2.2 顔画像のモデル化

顔画像は図 2 に示す手順でモデル化する．画像サイズは 160 × 140 ピクセルである．まず，Y 方向にシフトする場合は，160 × 14 のフレームサイズで下方向に 1 ラインずつ移動してフレームを切り出し，各フレームに対して 2 次元 DCT を行う．得られた 2 次元 DCT 係数のうち，60 × 5 の領域を特徴量として用い，11 状態 1 混合 HMM によりモデル化する．同様に X 方向についても，14 × 160 のフレームサイズで 1 ラインずつ右方向にシフトしながら切り出し，得られた 2 次元 DCT 係数のうち，6 × 53 の領域を特徴量とし，同じく 11 状態 1 混合 HMM によりモデル化する．

2.3 統合方法

X, Y 方向の顔画像の HMM の対数尤度を，重み係数 α を用いて統合する．これを GMM の対数尤度と重み係数 β を用いて統合する．

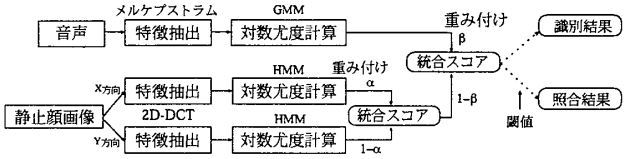


図 1: 音声と顔画像を用いた個人認識システム

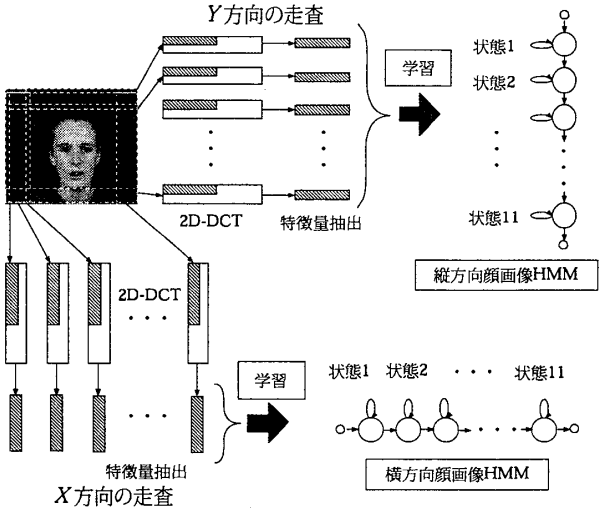


図 2: HMM による顔画像のモデルの作成法

3. 個人認識実験

本実験では，XM2VTS データベース [4],[5] を用い，個人識別及び照合実験を行った．このデータベースには音声と顔画像が 4 時期に亘って収録されている．3 時期分のデータで学習し，残りの 1 時期をテストデータとする leave-one-out 法により評価した．実験に用いたデータ数は，顔画像は 1 時期につき 1 人 10 枚，音声は 1 時期につき 1 人 6 文章 (3 種類の文章× 2 回分) である．

識別実験は 289 名で行い，結果は図 3 のようになった．重み係数 α と β は，識別率が最も高くなる値を事後的に選んだ．顔画像の識別においては，X 方向のシフトに比較して，Y 方向にシフトしたモデルの方が識別率が高くなっている．これは，両眼画像の情報など，X 方向にシフトした場合よりも Y 方向にシフトした場合の方が有効な特徴を取り出すことができるためであると考えられる．また，X, Y 方向を統合させることにより，X もしくは Y 方向単独の場合に比べ，高い識別率が得られた．更に，これを音声と組み合わせることによって識別率が改善され，83.8%の識別率となった．

照合実験では，289 名のうち，登録話者に 100 名，登録外話者に 99 名，バックグラウンドモデルの作成に 90 名を使用した．照合誤り率を図 4 に示す．重み係数 α と β 及び照合判定の閾値は，照合誤り率が最も低く

†名古屋工業大学 知能情報システム学科, Department of Computer Science, Nagoya Institute of Technology

なる値を事後的に選んだ．図4の結果は，本人棄却率 (FRR) と詐称者受理率 (FAR) が等しくなる等誤り率 (EER) で示している．識別の場合と同様の傾向が見られ，X, Y 方向の顔画像の統合，更に音声の統合によって，照合誤り率が減少した．顔画像の統合では，尤度の正規化を行わなかった場合には照合結果に変化は見られなかったが (A, B)，尤度の正規化を行った場合には誤り率が減少した (C, D)．更に，音声と統合することによって，誤り率が減少し，最も良い結果として4.2%の照合誤り率が得られた．

図5, 6は，それぞれ閾値を全話者共通とした場合と話者別とした場合の音声と顔画像の統合結果を示したものである．これらの図に示すように，尤度の正規化を行うことによって，誤り率が大きく減少した．また，閾値を話者別に設定した場合の方が，共通の場合に比べ，誤り率が全体的に低くなっている．音声と顔画像を統合した場合の結果においては，閾値を全話者共通にした場合に比べ話者別とした場合の方が誤り率が若干低くなっているが，これは，話者別に閾値を設定した場合の方が EER を求めるためのテストデータ数が少なくなるためであると考えられる．

4. むすび

本研究では，音声と顔画像を用いた個人認識システムについて検討した．顔画像を X, Y 方向でシフトし，HMM を用いてモデル化を行い，さらに音声の GMM と統合させた結果，識別率，照合率ともに大きく改善した．

今後の課題としては，固有顔などの他の手法との比較，DCT 係数使用領域の検討，顔画像の位置の正規化等が挙げられる．

謝辞 本研究の一部は，中部電力基礎技術研究所研究助成，科学研究費補助金若手研究 (B) No.14780274，堀情報科学振興財団研究助成，及び人工知能研究振興財団研究助成による．

参考文献

[1] Proc. AVBPA'97, '99, 2001.

[2] 坂野 鋭, "多重バイオメトリクスの研究動向," 電子情報通信学会総合大会講演論文集, PD-2-6, pp.282–283, Mar. 2002.

[3] M2VTS Project, http://www.sic.rma.ac/be/Projects/M2VTS/content.html

[4] The XM2VTS Database, http://xm2vtsdb.ee.surrey.ac.uk/

[5] K. Messer, J. Matas, J. Kitter, J. Luttein, and G. Maitre, "XM2VTSDB : the extend M2VTS database," Proc. AVBPA '99, 1999.

[6] B. Duc, G. Maître, S. Fischer, and J. Bigun, "Person authentication by fusing face and speech information," Proc. AVBPA'97, pp.311-318, Mar. 1997.

[7] S. Ben-Yacoub, J. Luettin, K. Jonsson, J. Matas, and J. Kittler, "Audio-visual person verification," Proc. CVPR'99, pp.580–585, June 1999.

[8] M. Turk and A. Pentland, "Eigenfaces for recognition," Vision and Modeling Group, The Media Laboratory, Massachusetts Institute of Technology.

[9] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixure speaker models," IEEE Trans. on Speech and Audio Processing, vol.3,no.1,pp.72-83, Jan, 1995.

[10] A.V. Nefian and M.H. Hayes III, "Markov models for face recognition," Proc. IEEE, 1998.

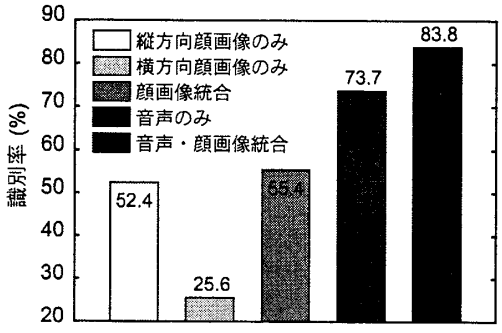
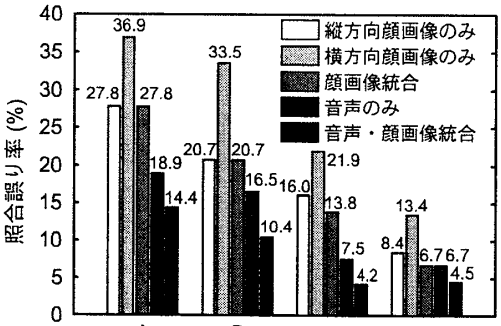


図 3: 統合前後の識別率の変化



尤度正規化	なし	なし	あり	あり
閾値設定	共通	話者別	共通	話者別

図 4: 統合前後の照合誤り率の変化

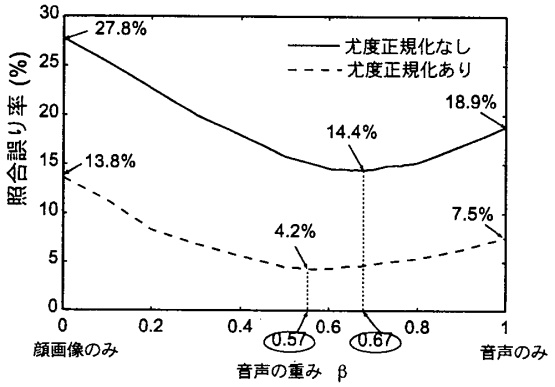


図 5: 重み β と照合誤り率の関係（閾値：全話者共通）

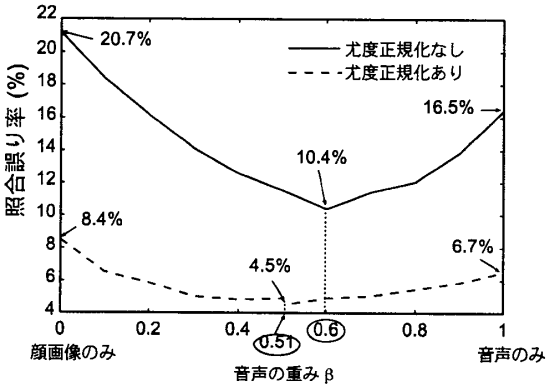


図 6: 重み β と照合誤り率の関係（閾値：話者別）