I_040

ビデオ画像と熱画像を用いた単語認識

Word Recognition based on Video and Thermal Images

齊藤 剛史 †　　小西 亮介 †
Takeshi Saitoh　Ryosuke Konishi

1 はじめに

熱画像計測装置は，物体から放射される熱エネルギーを温度に換算し，温度分布として画像表示するものである．非接触の測定が可能であり，設備監視などにおける異常検出に利用されている．また人物の姿勢判断 [1, 2] や状態判断 [3, 4]，さらに個人認識 [5, 6, 7] への応用も報告されている．一方，熱画像を用いた読唇に関する報告はなく，本論文ではビデオ画像と熱画像による単語認識の有効性について検証する．

熱画像は対象となる物体や人物の表面温度と背景の温度差の違いを利用できるため，容易に対象領域の抽出を実現できる．このことを利用して山崎らは屋内における人物状態のモニタリング [1]，Han らは昼夜に屋外で撮影した熱画像より人物の歩きや走りの動きを識別している [2]．人物の顔を対象とした研究では，善住ら [3] や永峰ら [4] は特に鼻部皮膚温度変化を利用して快－不快状態を判断している．さらに顔面熱画像よる顔認識 [5] や個人認証 [7] がある．熱画像は，一般的なカメラ画像での顔認識などで生じる光源の変化により認識率が変化する問題を解決するために用いられている．

熱画像に関する従来の報告では，人物の全身画像あるいは顔面画像に対する処理が多い．そこで本研究では，局所的な領域である唇に着目し，発話時に生じる唇形状の動きの変化に着目する単語認識を試みる．読唇では認識に適した特徴量を抽出することが重要であり，モデルベース法 [8] と画像ベース法 [9] に大別される．前者は動的輪郭モデルなどを用いて唇の輪郭形状を取得する．後者は，画素の濃度値に主成分分析などを施し特徴量を求める．顔面熱画像は顔の皮膚温度を濃度値で表現しており，カラー画像のように肌と唇の境界が明瞭でない．一方，熱画像は発話時の呼気などによる口内の温度変化が観測できる．そこで本論文では画像ベース法を用いることにする．

本論文では，最初に顔面熱画像から唇領域の抽出をする必要があるが，前述の通り肌と唇の境界が不明瞭であり，唇位置の抽出が困難である．そこで，熱画像と同時にビデオ画像を撮影し，ビデオ画像の情報を利用して熱画像の唇位置を抽出する．そのため，ビデオ画像と熱画像の対応関係を定める方法を提案する．次にビデオ画像から唇領域を抽出し，これに対応する熱画像の唇領域を抽出する．特徴量として，中田ら提案した固有画像波形 [9] を求め，DP マッチングにより認識を行い，ビデオ画像と熱画像の結果より，熱画像の有効性について調べる．

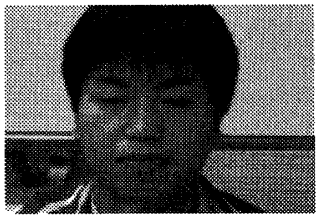
† 鳥取大学工学部，Tottori University

2 アプローチ

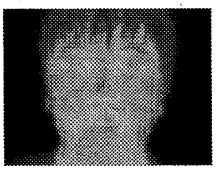
2.1 ビデオ画像と熱画像

ビデオ画像と熱画像は同じ位置から発話者の唇領域を撮影することが望ましい．しかし，両装置を同じ光軸にセットすることは物理的に不可能なため，本研究では，熱画像計測装置の斜め後方にビデオカメラを設置し，熱画像計測装置で撮影される熱画像の顔と同じ大きさで写るようにビデオカメラをズームして撮影を行う．また両装置は共に三脚に固定する．ここで，本研究ではビデオ画像は，Panasonic 社製ビデオカメラ NV-GS200，熱画像はチノー社製熱画像計測装置 CPA-8000 を用いて撮影する．これらの装置を利用することにより，ビデオ画像は 720×480 画素，熱画像は 480×360 画素，共に 30fps で撮影できる．

熱画像の唇領域を抽出するために，ビデオ画像と熱画像の対応点を設けることにより唇領域を抽出する．このとき，対応点は両画像で安定に取得できる位置が望ましい．そこで本研究ではビデオ画像と熱画像をそれぞれ顔全体が写るように撮影する．撮影した画像例を図 1 に示す．図 1(a) はビデオ画像，(b) は熱画像を示し，熱画像の濃度値は温度を意味している．本研究では温度領域は 25 ℃～28 ℃に設定して撮影する．すなわち温度分解能は 0.01 ℃であり，25 ℃以下は濃度値 0，28 ℃以上は濃度値 255 である．



(a) ビデオ画像



(b) 熱画像

図 1　ビデオ画像と熱画像

図 1 を観測すると，ビデオ画像は唇と肌，口内の境界が明瞭である．一方，熱画像は表面温度を示しているため，唇と肌の境界は不明瞭である．しかし，動画像で観測すると，熱画像では発話時の呼気により口内の温度差が時間的に変化していることが確認できる．また撮影時における装置の位置関係より，図 1(a) は少し上方から，図 1(b) は少し下方からの撮影画像となっている．

2.2 認識のアプローチ

前述の通り，ビデオ画像と熱画像は画像サイズが異なる．また，本研究では熱画像を用いた単語認識の有効性の検証が目的であるため，ビデオ画像と熱画像はそれぞれ動画ファイルとして取り込み，オフラインで処理する．そのため同時刻に撮影するものの両画像の同期を取ることが困難であり，両画像の時間の対応をとる必要が

ある．以上のことより，本研究ではビデオ画像と熱画像を用いた単語認識を実現するために，下記のステップで処理を行う．

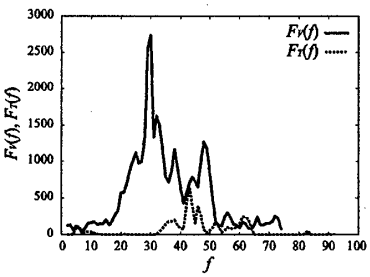
1. 時間マッチング
2. 空間マッチング
3. ビデオ画像による唇領域の抽出
4. 固有画像波形による単語認識

3 マッチング処理

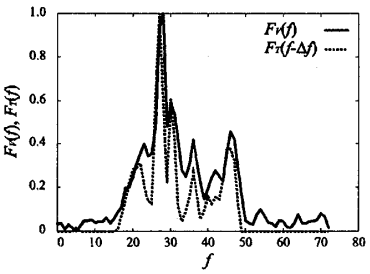
3.1 時間マッチング

前節の通り，ビデオ画像と熱画像は同期がとれていない．そこで，まず時間軸の位置合わせを行う．同時刻に撮影しているため，両画像とも同じ唇の動きである．この動きに着目して，本論文ではオプティカルフローで得られるフレーム毎の動き量を求め，下記の手順により時間マッチングを行う．

ビデオ画像と熱画像のフレームレートは同じであり，撮影範囲もほぼ同じ時間である．そこで最初に，ビデオ画像と熱画像のそれぞれの動画像において，二つのフレーム間におけるオプティカルフロー（ここでは高い精度で求まるブロックマッチング法を用いる）を計算する．ビデオ画像と熱画像におけるフレーム全体でのフローの総和 $F_V(f)$ と $F_T(f)$ をそれぞれ求める．ここで f はフレーム番号を意味す．図 2(a) は被験者に"島根"と発話させたときの $F_V(f)$ と $F_T(f)$ を示している．両波形は分布範囲が異なっているものの，大まかな形状は同じである．また両画像のフレームレートが同じであることを利用して，$F_V(f)$ と $F_T(f)$ に対して，それぞれの最大値をもとに正規化を施し，相関値を計算することにより時間ずれ量 Δf を求める．図 2(a) に対して時間マッチングを適用した結果を図 2(b) に示す．ただし，本研究では Δf を求めるためにオプティカルフローの他にフレーム間差分法による差分量総和を試みている．両手法の比較実験に関しては **5.1** で示すが，前者の方が精度が高かったため本論文ではオプティカルフローを採る．



(a) マッチング前の動き量

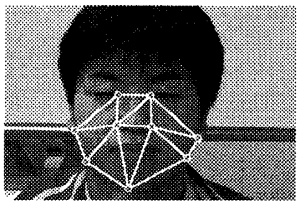


(b) マッチング後の動き量

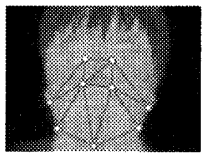
図 2 時間マッチング

3.2 空間マッチング

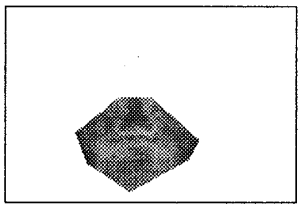
次にビデオ画像と熱画像の空間的な対応ととるための空間マッチングを行う．ここで両画像のサイズは異なる（ビデオ画像の方が画像サイズが大きい）．対応点を自動的に算出できることが望ましいが，本論文では手動で対応点を与える手段を採る．前節で時間マッチングを施したビデオ画像と熱画像の最初のフレーム画像において，唇を周辺とした 9 点（両耳の付け根，顔輪郭と首の交点，顎の下端，両目の内側，鼻の外側）を手動で与える．次に対応点を頂点とする三角メッシュを生成する．そして対応するメッシュ間の画素濃度値を利用して，画像サイズの小さい熱画像をビデオ画像と同じサイズに補間する [10]．この様子を図 3 に示す．図 3(a) と (b) はそれぞれビデオ画像と熱画像の対応点 9 点より生成される三角メッシュである．図 3(c) は補間後の熱画像である．本研究では唇周辺の領域を認識に用いるため，画像全体に補間を行わず，メッシュ内のみとする．



(a) ビデオ画像の対応点



(b) 熱画像の対応点



(c) 補間後の熱画像

図 3 空間マッチング

4 領域抽出および単語認識

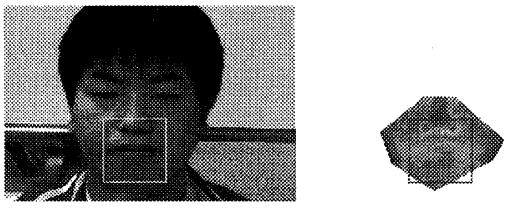
4.1 ビデオ画像による唇領域の抽出

次にビデオ画像と熱画像より唇矩形領域を抽出する．前節までの処理により，両画像は時間的，空間的に対応関係が求まっている．すなわち，ビデオ画像の f フレーム目の座標濃度値 $I_T(x,y,f)$ は熱画像の補間後の座標濃度値 $I_T(x,y,f-\Delta f)$ は同じ位置である．熱画像では唇と肌の濃度値差，すなわち温度差が明瞭でないため，本論文ではビデオ画像より唇領域を抽出し，それに対応する熱画像の唇領域を抽出する．本論文では以下の手順により唇 ROI（Region Of Interest）領域を抽出する．

1. 画像内より皮膚の吹き出物などを誤検出せずに唇の位置を検出するために，最初に原画像をブロック画像に変換する．次に全てのブロックで最も赤いブロックを検出し，このブロックの中心座標を唇基準点とする．
2. 唇の左右端点を求めるために，分離度フィルタ [11] を用いる．分離度フィルタは福井らが提案した手法であり，二つの同心円領域を用いて，線形判別法により両領域の分離程度を表す分離度を求めるものである．分離度フィルタの適用範囲を削減するため

に，まず処理対象領域内に対して p-タイル法（比率を 10% に設定）を適用して暗い領域となる唇の接合部を取り出す．次に p-タイル法により得られた暗い濃度値の画素に対して分離度フィルタを適用し，ガウシアンによる平滑化を施した後に局所最大点を求める．唇基準座標の高さ付近に位置する最大点を左右両方向でそれぞれ 1 点ずつ検出し，これを唇の左右端点とする．

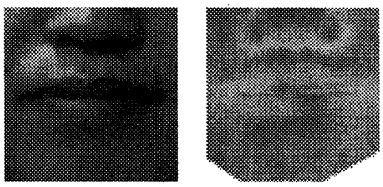
3. 検出された左右唇端点と唇基準点よりビデオ画像の唇 ROI 領域を抽出する．唇 ROI 領域は正方領域とし，一辺の長さを左右唇端点幅，領域中心点を唇基準点とする．次にビデオ画像に対応する熱画像に対する ROI 領域は，ビデオ画像と同じ位置，大きさとする．本論文では，ROI 領域を 64 × 64 pixel とする．ビデオ画像と熱画像の ROI 領域を図 4(a)(b) に示す．



(a) ビデオ画像　　(b) 熱画像

図 4　ROI の抽出

4. 撮影条件（被験者の顔の大きさや装置との距離）によって ROI 領域の大きさは異なる．また本論文では，ROI 領域に対して主成分分析を適用して固有ベクトルを計算する．このとき，処理時間の軽減と領域サイズの違いを考慮するために一定の大きさに正規化する．正規化後の画像を図 5 に示す．



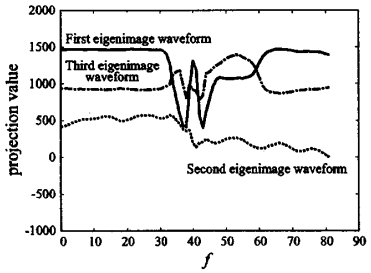
(a) ビデオ画像　　(b) 熱画像

図 5　画像サイズの正規化

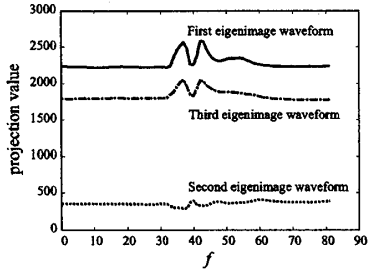
4.2　固有画像波形による認識

中田と安藤は発話中の唇の形状変化を表現する手段として固有画像波形を提案している [9]．これは，唇動画像を固有空間へ射影した射影ベクトルによって数値化するものであり，射影ベクトルの要素は，原画像と固有画像ベクトルとの内積である．射影ベクトルをフレーム f ごとの等間隔に並べた時間応答波形，すなわち固有画像波形 $B(f)$ とすることで，唇の形状変化を表す．本論文では，正規化した ROI 唇画像に対して，ビデオ画像と熱画像の固有画像波形をそれぞれ求める．図 6 に"島根"と発話して求まる第 1 主成分から第 3 主成分の固有画像波形を示す．

同一人物が同じ単語を発話した場合でも，発話時間は異なる可能性がある．時間的伸縮の影響を低減するために，本論文では動的計画法を用いた時間軸伸縮マッチング（以下，DP マッチングと略記する）を用いる．すなわち，データベース用の固有画像波形 $B_{D,w}(f)$ と実験用



(a) ビデオ画像



(b) 熱画像

図 6　固有波形画像（島根）

の固有画像波形 $B_t(f)$ に対して DP マッチングを施し距離 D_{wt} を求める．そして距離の最も小さいデータベース用の固有画像波形 $\arg\min_w D_{wt}$ の単語 W を認識結果とする．

5　実験

男性被験者 3 人に対し中国地方 5 県名（鳥取，島根，岡山，広島，山口）を 5 回ずつ発話させ，その様子を熱画像計測装置とビデオカメラで撮影した．ビデオ画像，熱画像それぞれ 75 シーンずつ撮影した．このとき，75 シーンにおけるフレーム数は 65〜110 の範囲であった．また被験者から熱画像計測装置までは約 60cm，ビデオカメラまでは約 90cm の位置で撮影を行った．

5.1　時間マッチング

75 シーンに対して **3.1** で述べたオプティカルフローによる時間マッチングとフレーム間差分法による時間マッチングをそれぞれ行った．次に，75 シーンにおいて目視によりビデオ画像と熱画像の時間ずれ量を観測して，時間マッチングの結果を評価した．その結果，平均誤差フレーム数はオプティカルフローでは 2.2 フレーム，フレーム間差分法は 3.8 フレームであった．また，誤差フレームが ±3 内を時間マッチング成功と見なす場合，前者は 88.0%，後者は 81.3% であった．処理時間に関して，オプティカルフローでは高い精度を優先してブロックマッチング法を用いたため，フレーム間差分法に比べ，処理時間がかかる．しかし，本研究では熱画像の単語認識への有効性を検証することを目的としており，成功率の高いオプティカルフローを採用した．

5.2　単語認識

前節の時間マッチングで失敗したシーン（誤差フレーム数 ±3 外のシーン）に対しては，目視で求めた時間ずれ量を与えた．また撮影したシーン数が少ないため，本実験では，少量データから正確な認識率を得るため，Leave-one-out 法を適用した．

被験者（A，B，C）ごとに主成分数を 1〜20 に変化させ認識を行った．その結果，主成分数 1 の場合が認

識率が最も高く，3 人の平均認識率はビデオ画像のみでは 76.0%（被験者 A,B,C でそれぞれ 64.0%，88.0%，76.0%），熱画像のみでは 44.0%（被験者 A,B,C でそれぞれ 52.0%，28.0%，52.0%）であった．ここで，主成分分析における寄与率を調べると，主成分数 1，2，3 ではそれぞれ約 40%，約 17%，約 7% であった．熱画像のみの認識は，どの被験者においてもビデオ画像よりも低い．これは熱画像の場合，表面温度の変化が濃度値の変化として表れているが，唇や肌の変化はほとんどなく，主に口内の変化のみが観測できる．一方，ビデオ画像では口内だけでなく，唇の形状の変化も観測ができるため，ビデオ画像の方が高い認識率を得られると推測する．また熱画像で被験者 B の認識率が極端に低いのは，被験者 B の発話時における唇の動きが他の被験者に比べ小さいためであったと推測する．

次にビデオ画像と熱画像で最も認識率が高い主成分数 1 において，両画像の統合認識を行った．統合は DP マッチングにより得られるビデオ画像と熱画像の距離をそれぞれ D_V，D_T とし，両者を統合した距離 D を重み α を用いて $D = \alpha D_V + (1-\alpha) D_T$ と定義する．そしてデータベース内の最小距離の固有画像波形の単語 W を認識結果とする．その結果を図 7 に示す．図中，添字 A，B，C は 3 人の被験者毎の認識率，太線は平均認識率を示している．これよりビデオ画像の方が高い認識率を得られている．また α が高いほど，すなわちビデオ画像に対する重みが大きいほど認識率が高くなる傾向にある．また平均認識率が最も高かったのは $\alpha = 0.9$ のときであり，被験者の平均認識率は 80.0% であった．

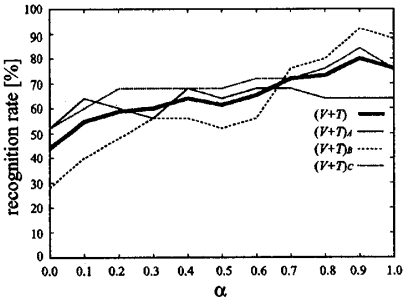


図 7　重みと認識率の関係

また，ここでは ROI 領域を唇領域を含む 64×64 pixel としている．熱画像は呼気による温度変化が観測される．そこで元の ROI 領域を 100% とし，ROI 領域のサイズを 10% ずつ，切り出した領域を用いて認識を行った．すなわち 90% であれば，元の ROI 領域の中央付近 57×57 pixel のみを処理領域とした．その結果，熱画像のみでは 30%（19×19 pixel）において 54.7% が最も高い認識率であった．ビデオ画像のみでは，60%（38×38 pixel）において 90.7% が最も高い認識率であった．ROI のサイズにより認識率の向上は見られるが，熱画像はビデオ画像に比べ低い認識率であることがわかった．

6　おわりに

本論文では，熱画像を用いた単語認識の有効性を検証するために，ビデオ画像と熱画像を用いた読唇を試みた．両画像を時間的，空間的に対応させるため，時間に対してはオプティカルフローより計測できる動き量，空間に対しては手動で与えた対応点をもとに補間による対応する方法を提案した．次にビデオ画像より唇 ROI 領域を抽出し，それに対応する熱画像の唇 ROI 領域を抽出した．最後に固有画像波形による認識を行った．その結果，5 単語における被験者 3 人の平均認識率は，ビデオ画像では 76.0%，熱画像では 44.0%，両者の統合では 80.0% であった．また ROI のサイズを変化させて実験を行い，熱画像では 54.7%，ビデオ画像では 90.7% の認識率を得た．熱画像のみでは十分な認識率を得られないものの，ビデオ画像と統合することにより，精度が向上することを確認した．

今後の課題として，今回手動で与えていた，ビデオ画像と熱画像の対応点の自動化が挙げられる．また被験者数を増やした実験を行い，単語認識で有効な特徴量を検討し熱画像の有用性について検討する．

参考文献

[1] 山崎信寿, 財津篤. 熱画像情報による人物状態の自動判断. 計測自動制御学会論文集, Vol. 33, No. 7, pp. 603–608, 1997.

[2] Ju Han and Bir Bhanu. Human activity recognition in thermal infrared imagery. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 17–, 2005.

[3] 善住秀行, 野澤昭雄, 田中久弥, 井出英人. 鼻部皮膚温度変化による快－不快状態の推定. 電学論 C, Vol. 124, No. 1, pp. 213–214, 2004.

[4] 永峰康司, 野澤昭雄, 井出英人. 聴覚および味覚刺激下での鼻部皮膚温による情動の評価. 電学論 C, Vol. 124, No. 9, pp. 1914–1915, 2004.

[5] Diego A. Socolinsky, Lawrence B. Wolff, Joshua D. Neuheisel, and Christopher K. Eveland. Illumination invariant face recognition using thermal infrared imagery. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Vol. 1, pp. 527–534, 2001.

[6] Jingu Heo, Marios Savvides, and B.V.K. Vijayakumar. Performance evaluation of face recognition using visual and thermal imagery with advanced correlation filters. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 9–, 2005.

[7] 堀込徹郎, 田中久弥, 井出英人. 顔面熱画像の空間周波数特徴による個人識別. 電学論 C, Vol. 122, No. 9, pp. 1645–1650, 2002.

[8] 菅原一孔, 新地俊幹, 岸野誠, 小西亮介. パーソナルコンピュータ上での読唇システムの実時間実現. 計測自動制御学会論文集, Vol. 36, No. 12, pp. 1145–1151, December 2000.

[9] 中田康之, 安藤護俊. 色抽出法と固有空間法を用いた読唇処理. 信学論, Vol. J85-D-II, No. 12, pp. 1813–1822, December 2002.

[10] 高木幹雄, 下田陽久. 新編　画像解析ハンドブック. 東京大学出版会, 2004.

[11] 福井和広, 山口修. 形状抽出とパターン照合の組合せによる顔特徴点抽出. 信学論, Vol. J80-D-II, No. 8, pp. 2170–2177, August 1997.