ヒューマンインタフェース 80－3
音声言語情報処理 23－3
（1998．10．16）

入力モダリティの多様化とその統合・利用について

河野恭之，鈴木薫，知野哲朗，田中克己，屋野武秀，金澤博史

{kono,suzuki,chino,tanaka,yano,kanazawa}@krl.toshiba.co.jp

(株) 東芝 関西研究所

〒658-0015 神戸市東灘区本山南町 8-6-26

人間の知覚や情動により合致したヒューマンインタフェース（HI）の実現のためには，複数の入出力モダリティを組み合せ使い分けることのできるようなHIのマルチモーダル化が重要である．筆者らは，いくつかのマルチモーダルインタフェース，及び擬人化インタフェースの試作を行ってきた．その過程を通じ，複数モダリティの利用時における各モダリティやその組み合わせに起因する性質，及びメディア処理技術に起因する性質が浮かび上がってきた．本稿では，これらのシステムを紹介すると共にマルチモーダル・擬人化インタフェース設計時の視点を挙げる．

On Dealing with Multiple Modalities

Yasuyuki KONO, Kaoru SUZUKI, Tetsuro CHINO,
Katsumi TANAKA, Takehide YANO, and Hiroshi KANAZAWA

Kansai Research Laboratories, Toshiba Corporation

8-6-26 Motoyama-Minami, Higashi-Nada, Kobe 658-0015, Japan

Human-human communication is essentially multimodal. Thus a multimodal human-computer interface (MMIF) which enables a user to communicate with a computer without paying special attention to input modalities, *i.e.*, an MMIF which accepts such input that anybody does to the others is required. We have been developing some MMIFs and anthropomorphic interfaces. Some common characteristics on multimodal/anthropomorphic interfaces appeared through the development process. They are originated from the characteristics of each modality, arrangement of modalities, and media processing technologies. This paper briefly describes the developed interfaces and the characteristics.

1 はじめに

計算機能力の進歩に伴い，オーディオやビデオ等のマルチメディアデータをリアルタイムに処理できる環境が急速に整いつつある．進歩してきたメディア処理技術を利用し，より人間の知覚や情動に合致するヒューマンインタフェースの総称であるPUI(Perceptual User Interfeces)[竹林98]を指向した研究が盛んに行われている．マルチモーダルインタフェースの研究もその流れで捉えられ，複数の入出力モダリティを組み合わせ，また使い分けることで，機械とのより自然なインタラクションを目指していると考えられる[Maybury94, 速水98]．

一般に，人間が各モダリティに求める知覚的・感覚的な性質は異なっており，またそのモダリティに対応するメディア処理技術の到達レベルはまちまちである．マルチモーダルシステム構築には，これらの性質を鑑みた設計が求められる．筆者らが所属する東芝関西研究所では，音声処理や画像処理等のメディア処理基盤技術の深耕と，それらの技術を用いたより人間の知覚に合致したヒューマンインタフェースを目指して研究開発活動を行っている．そしていくつかのマルチモーダルインタフェースの試作を通じ，複数モダリティの利用時における各モダリティやその組み合わせに起因する性質，及びメディア処理技術に起因する性質がいくつか浮かび上がってきた．本稿では，筆者らが構築したマルチモーダルインタフェースを概観すると共に，これらの性質についての整理を行う．

2 マルチモーダルインタフェース

ユーザにとっての自然さとは，効率的，すなわち楽に効果的に自分の意思を相手に伝達できることと考えられる．実際，人間は相手に対し，言葉や身振り，手振り，表情といった様々なモダリティを利用して意図を表現し，効率的に伝達している．

その意味で，マルチモーダルインタフェースを構築する際には，ユーザの負担にならないようなモダリティを選択し，またユーザが自然に操作できるモダリティの組合せを選択する必要がある．例えば，データグローブやアイトラッカーといった装着型のデバイスよりは画像処理を用いることで特別なデバイスを装着しないで済む方が望ましいと考えられる．また例えば，タッチパネルとキーボードといった複数のデバイスからの入力を組み合せる等，操作が競合するモダリティの組み合せは避ける必要がある．逆に，システムの出力としては，注目してほしい操作対象が明確にユーザに伝わり，また操作やシステム状態のフィードバックが明確なことがわかりやすく自然なインタフェースに必要と考えられる．

以降本節では，これまでに試作したマルチモーダルシステムを紹介し，上記の視点から分析する．

2.1 ナレーションエージェント

筆者らは以前，非エキスパートが訓練無しに使いこなせるような将来のマン・マシンインタフェースの形態として，擬人化が一つのキーポイントとなると考え，マルチモーダルな擬人化ユーザインタフェースのパイロットシステムとしてナレーションエージェントシステム Rachel を試作した [鈴木 96]．このシステムは人物を検出すると，所定の挨拶を音声出力した後で本題のナレーションを始める．また，人物がシステムの前を離れるとこれを検出し，ナレーションを中断してお別れの挨拶をする．発話中には音声の有無に合わせて口を自動開閉し，また人物を常に追跡して視線と顔向きを変える．

図 1に Rachel の構成を示す．このシステムではビデオカメラから得られる動画像を入力とし，スピーカーから挨拶やナレーション音声を，ディスプレイには 3 次元 CG によってリアルタイム動作するエージェント映像が表示される．ナレーション音声出力は，実在人物の声の録音を状況に応じてナレーション制御モジュールで選択し再生することで行う．

カメラ入力は画像処理され，(1) システムの前に

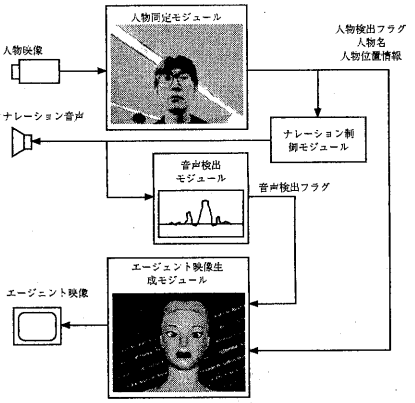


図 1: ナレーションエージェント Rachel の構成

人物が存在するか否かを判定し，(2) 人物が存在する場合，その人物が誰であるかを顔認識により判定する．人物検出と顔の個人照合は，形状抽出とパターン認識を組み合せることにより行っている [福井 97]．顔検出 / 照合処理は 5 ～ 6 枚 / 秒の速さで実行される．人物を検出するとエージェントはその方向に顔を向けて挨拶と自己紹介を開始する．個人照合できたユーザには名前を呼んで話し始める．ユーザが離れると話を止めて待ち受け状態になる．

オーディオ出力ラインから音声区間が検出されると，エージェント画像の口が動くよう制御される．発話中のエージェントの頭部は小刻みに上下に揺れるよう制御される．CG エージェント映像は約 7000 ポリゴンのデータから生成され，約 2 ～ 3frame/ 秒 (ONYX RE2) の更新速度で描画される．

Rachel のトップレベルでは，人物の存在判定と人物照合という共に 1 台のカメラの情報に基づき制御している．既に述べたようにこれらの認識情報は 0.2 秒程度毎に連続的に得られるが，認識失敗や誤認識等によりそのデータにはノイズが不可避である．このため例えば顔領域検出が所定回数以上連続して成功した場合に人物が存在すると判断する等によりノイズの影響を避けると共に，登場・退出という離散データに変換してシステムの制御に用いている．

Rachel を試作した目的の一つは，リアルな表象と動作を持つ擬人化インタフェースが有効か否かを検証することであった．社内公開実験の結果，顔を持ちしゃべる計算機への期待と関心が高いことは確認された．しかし音声と口の同期が不十分なことに起因して自然性の評価は該して低かった．また，「こ

図 2: マルチモーダル秘書エージェントの画面例

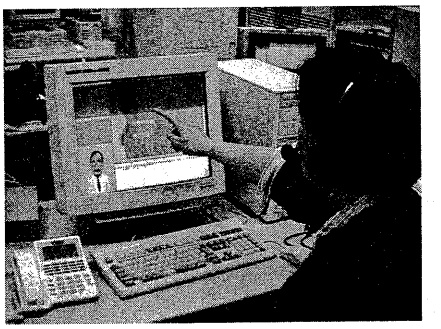


図 3: マルチモーダル秘書エージェント使用の様子

んなにリアルなのに決まったことしか言えないのか」といった画面上の表象から期待される知的有用性とのギャップが多く指摘された．このことから，リアルな表象にはその他の要素のリアルさも期待され，インタフェースの見かけとその他の要素とのバランスが設計上重要であることがわかった．

2.2 マルチモーダル秘書エージェント

人間が自然に生成するマルチモーダル表現を受理できるインタフェース構築のためには，複数のモダリティから非同期に与えられる入力の統合・解釈が重要である．音声やジェスチャ等の各モダリティの入力要素の獲得には一般に認識処理を伴う．このためマルチモーダル入力の統合・解釈において，

認識結果の曖昧性 入力となる各モダリティの認識結果は，曖昧性を含む認識結果候補集合となる．モダリティが増えると解析候補数はその組み合わせとなり，しばしば膨大な数になる．

遅着データ マルチモーダル入力は各モダリティの認識部に非同期に与えられ，個々に認識した後に統合・解釈するのが一般的である．認識コストはモダリティ毎に異なり，統合時の順序や時間遅れが保証できない．かなりの割合のマルチモーダル入力が複数モダリティの同時入力でなく，順次入力という報告もある [Oviatt97].

という問題があることがわかってきた [Maybury94]. マルチモーダル秘書エージェントシステムは，これらの課題にフォーカスして試作された [河野 98].

図 2に本システムの画面例を示す．本システムは，オフィス業務におけるノウハウを持つ人間本人に代わり擬人化エージェントが問い合わせに答えるというタスクをアプリケーションとして想定している [中山 98]．図 3に示すように本システムの利用者は，タッチパネルを用いたポインティング及びサークリングジェスチャ入力と，マイクからの音声入力を併用したマルチモーダル入力が行える[1]．ユーザの文発声を受理する連続単語音声認識モジュールを備え，フットスイッチをトリガとして処理を開始する．システムは文字画像情報に加え，合成音声による音声出力，画面左下部に表示される男性あるいは女性の 3 次元 CG エージェントの身振り，表情といったノンバーバルメッセージを用い，オフィス業務ノウハウとシステムの状態をユーザに伝達する．エージェントは合成音声に同期して口を動作させる．

ユーザは文書名のリスト上で指示したい文書付近を手でサークリング / ポインティングしながら「この議事録を見せて」などとシステムに話しかけることができる．図 2の画面中にユーザのサークリングの軌跡が薄い円形で示されているが，ここでユーザの発話が「この議事録を見せて」でありその発話が正しく認識されると，エージェントは「はい，技術仕様検討会議議事録を示します．」と合成音声で答えると共に，該当するドキュメントを提示する．

本システムの核となるマルチモーダル統合・解釈モジュールは，マルチモーダル参照同定問題 [Kobsa86] の問題解決器である．筆者らはその枠組として，ATMS ベースのマルチモーダル統合・解釈メカニズムを開発した [Kono98]．この方式では，過去の推論の依存関係を記録し再利用することで，上記認識結果の曖昧性の問題と遅着入力を効率的に扱える．

[1] ジェスチャ入力は画像認識等の際の認識ノイズの影響を避けるため，接触型のタッチパネルを採用した．

同様に曖昧性を効率良く扱う枠組には，型付き組成構造の単一化に基づくマルチモーダルパーザ[Johnston98]がある．このパーザは軍事演習シナリオ設計システムQuickSet上で実動している．彼らの方式では候補単位の枝刈りが可能だが，過去の解析過程の再利用やより細かい枝刈りは困難である．

本システムはRachelと異なり，音声を用いてCGキャラクタとインタラクションできる．このため，社内外での展示において擬人化インタフェースのまた違う効果・性質が見られた．まず，エージェントという見る対象があることで対話の相手を認知し，相手を見て話せるという効果が認められた．

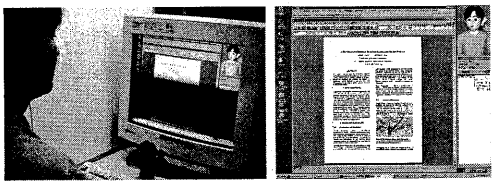
Rachelの経験を基に過度な期待を避けるため，本システムにはマンガ的な容貌を与えた．更にエージェントが生きているように見えるために，視線の揺れや首の振れ，瞬きなど生理的な動きを常に与えた[2]．これがエージェントの生気感を向上させた．

本システムは音声認識部やジェスチャ認識部から認識結果が得られるまで上記の生理的な動作しか示さないことに起因して，「エージェントが聞いているかどうかわからない」との感想が多く聞かれた．

2.3 GazeToTalk

現在の音声インタフェースでは，ボタンを押す等の形で音声入力開始のトリガを与えるPush To Talk方式の採用が一般的である．しかしこの方式では非接触，ハンズフリーという音声入力の利点を生かしたシステムの実現が難しい．GazeToTalkは動画像からの視線認識技術と擬人化インタフェース技術を用い，視線による音声入力スイッチを実現したシステムである[知野98]．すなわちGazeToTalkは，(1)画面上の擬人化エージェントに対するユーザの注視を検知し，(2)エージェントの表情・身振りでフィードバックを返すことで，(3)ユーザとシステムとのアイコンタクトを実現し，(4)それにより音声入力受付可否を制御するシステムである．

図4(a)にあるようにCRTの前下部にビデオカメラが設置され，図4(b)のように画面の右上部にCGエージェントが表示される．ビデオカメラから得られた動画像は画像処理され，(1)画像からの顔領域の抽出及び位置トラッキングによるユーザ検出と，(2)目鼻等の部品領域候補の同定及び目周辺のパターン照合による注視位置の判定と，が行われる[Yamaguchi98]．これらの処理はソフトウェアで実現されており，CRT面9分割の分解能で5回/秒(Indy R5000)の高速な視線検出を実現している．



(a) (b)

図4: GazeToTalk使用の様子(a)と画面例(b)

本システムの前にユーザが現れると，視線による非言語メッセージが利用可能なことを伝えるためにまずCGエージェントが現れる．この状態でユーザがエージェントを注視しそれが検出されると，システムは音声認識を開始すると共に“眉間の閃き”と呼ばれる表情をユーザへのフィードバックとして提示する．そして，耳に手をかざしたエージェントの身振りにより音声受付可能状態がユーザに伝えられると共に，適度なタイミング毎にうなずきが提示され，通信路の維持がフィードバックされる．その後，画面の他領域への視線の移動が検出されると，エージェントが視線をそらすと共に音声認識が停止される．この状態でユーザがシステムの前から離脱すると，エージェント表示は消える．また音声受付可能状態にある時，すなわちアイコンタクトが成立している際にユーザの離脱が検出されると，エージェントは驚きの表情を浮かべた後に消える．

GazeToTalkでは視線という非言語メッセージに注目し，それを元に音声認識開始/停止といった制御を行うと共に，うなずき表情等の非言語メッセージをフィードバックチャネルに用いている．非言語メッセージの特徴の一つに，意図が瞬時に伝達可能なことがある．このため，レスポンスタイミングがシステム全体の重要な要因となる．本システムでは，音声区間の確定と認識処理自体に時間のかかる音声認識の結果を視線認識結果と統合せず，視線を単独で音声入力のスイッチとして利用している．

2.4 マルチモーダルショートカット

既に述べたようにGazeToTalkでは，視線認識結果と音声認識結果を統合しているわけではない．

[2] Rachelにおいてもこのような動作を採り入れていたが，本システムでは動きを滑らかにし，また実時間で意図したタイミングと速度が出るようになったことでより動作の自然性が増している．

これは例えばマルチモーダル秘書エージェントでは，システムで扱っている音声とタッチジェスチャの両モダリティの入力が共に離散的に得られるのに対し，入力が一定期間毎に継続的に与えられる性質を持つ視線認識のモダリティを扱うことに起因する．更に視線認識等の画像処理技術は音声認識以上に成熟しておらず，継続的に得られる認識結果には頻繁にノイズが混在し不安定である．このような性質を持つモダリティからの入力を扱うには，離散的なモダリティだけからなるマルチモーダル入力統合・解釈手法とは異なるアプローチが必要と考えられる．マルチモーダルショートカットは視線検出と音声認識をユーザの入力手段とし，ユーザの注意がどこにあるかを推定してユーザによるオブジェクトの選択を支援するシステムである [田中 98].

マルチモーダルショートカットでは前記のオブジェクト選択といった比較的単純なタスクに対し，継続的なものを含む複数モダリティからの入力を用いる．このような継続的なモダリティにおいてより顕著となる不安定な入力に対しても適切な動作を目指し，本システムではエージェントモデルを導入した．これは個々の選択対象をエージェントとして表現し，エージェントは入力に基づき自己の注目度推定と行動を自発的に行う，というものである．

各エージェントは，(1) ユーザ入力の観測，(2) それに基づく意図の推定，(3) 行動，という 3 つの段階からなるサイクルを独立に行う．このサイクルの特徴は，各段階で取り扱う情報を不確実性を持ったままで残し，学習により各情報間の関係を取り扱う点である．認識システムより得られる認識結果は類似度という形式で取り扱われる．意図推定段階ではその類似度に基づいてユーザの意図が推測されるが，その過程で意図それぞれに対する確率に変換される．行動決定の段階では，意図確率から最適の行動を決定し実行する．各段階での決定はあらかじめ獲得された学習データに基づいて行われ，学習データは環境の変化に従い適宜更新される．このようなアプローチにより，不確実で安定性に欠ける入力においてもその場で最適な行動を取ることができる．意図の推定・学習にはベイズ推定の手法を用いている．このような構成により，タスクは単純であるがキーボード・ショートカットのような素早いインタラクションを実現することが可能になる．

図 5にシステムの画面の例を示す．画面上の矩形領域は選択対象であり，選択対象エージェントと呼ぶ．各エージェントの持つ意図は，自分が選択されているか否かの 2 種類である．各エージェントで意図確率が算出されると，行動決定の段階では意図確率を基に取るべき行動を決定する．本システムにおいて各エージェントは，図 5中のレベルメータを用いて自らの活性度情報をユーザに知らせる．ユーザはそのフィードバックに基づいて次の行動を決定し，それが次のマルチモーダル入力に反映される．

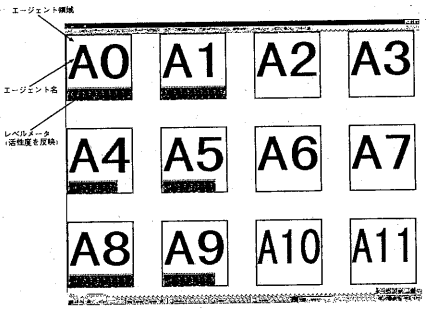


図 5: マルチモーダル・ショートカットシステム画面

3 考察

3.1 マルチモーダル入力統合

ここでは入力モダリティの取り扱いについて論じる．マルチモーダルインタフェース研究が目標とするところの一つに，ユーザにとってより知覚的に自然な操作の実現がある．自然なという言葉を狭義に解釈すると，人間が普段行っているようなコミュニケーション方法を崩さない必要があると考えられる．モダリティの組み合せにおいて操作が競合しないような選択が必要というだけでなく，あるモダリティの操作に際して人間が普段行う行動に加えて何らかの操作を要求するべきでないことになる．

秘書エージェントシステムでは音声入力開始時にフットスイッチの押下が必要だが，人間どうしの対話には通常このような明示的なトリガは必要ない[3]．トリガが必要なのは，ユーザの独り言や周辺ノイズを音声認識しないように制限する必要があるという技術的な制約に起因する．このシステムの展示等の機会に一般の人による操作も試みたが，このような「不自然」な操作を理解してもらい習得させること

[3] 近い行動として，同じ音声モダリティによる「呼びかけ」や「相槌」に加え，「目配せ」等の非言語メッセージがあるが，スイッチを押すという操作には程遠い．

は困難を伴った．GAZETOTALKでは操作を指令する対象を見るという人間の自然な行動を利用し，付加的なデバイスの装着なしに上記の問題の解決を目指した．視線検出はまだ計算コストを要する上に認識率も十分ではないが，画像処理による非言語メッセージの取得と利用に大きな可能性を示した．

本稿で示したシステムの試作と展示を通じて，組み合せるモダリティ間の親和性にいくつかの性質があることがわかってきた．その大きな要素がデータの入力タイミングである．秘書エージェントシステムで用いたATMSベースの統合・解析手法は，音声とタッチパネルへのジェスチャといった離散的な入力のあるモダリティの統合には有効であった．しかし，視線のような連続的に入力が得られるモダリティとの統合はより難しい．筆者らは，マルチモーダルショートカットのようなアプローチにより，単純なタスクならば入力性質の異なるモダリティの統合が可能なことを示したが，タスクが複雑化した場合が依然課題である．モダリティ間の親和性を左右するものとして筆者らは，(1) 適用タスクの複雑さ，(2) モダリティの機能，(3) モダリティからの入力到着形式（離散と連続），(4) 入力認識の性能（速度，認識率），といった要素を考えている．

これら各モダリティの性質と複数モダリティを組み合せた際の親和性について，適用タスクの性質も含めて分類と整理を行うことが必要と考えられる．

3.2 擬人化インタフェース

擬人化インタフェースの利点の一つに，表情や身振り等の非言語メッセージを用いたフィードバックがある．また，ユーザの心理的抵抗感を和らげる効果もある [河野 98]．しかしシステムの知的有能性に対するユーザの予見にCGキャラクタの見かけが大きく影響する．キャラクタに人を使うかどうかを含め，システムの設計時にCGキャラクタの見かけのリアルさ等の要素の検討が必要である．

インタフェースに顔があればユーザが対話の相手を認知でき，そちらを向いて発話できる．逆に顔以外に注目して欲しいところがある場合に，そちらを見てくれないこともある．また，入力のためにエージェントを見ることを要求するGAZETOTALKは，逆にユーザが注目しているところを注視できないという副作用（モダリティの競合）がある．ユーザの視線を制御するノウハウの蓄積が，実用的な擬人化インタフェースの実現に必要である．

4 おわりに

本稿では，筆者らが試作してきたマルチモーダル，及び擬人化インタフェースを紹介し，定量化してはいないものの試作から得られた知見を述べた．本稿で示した視点を更に整理して分析・試作を進め，更に自然なインタフェースの実現を目指す．

参考文献

[知野 98] 知野哲朗，福井和広，山口修，鈴木薫，田中克己: GAZETOTALK: メタコミュニケーション能力を持つ非言語メッセージ利用インタフェース，インタラクション98論文集，情報処理学会，pp.169–176，1998.

[福井 97] 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.2170–2177，1994.

[速水 98] 速水悟，竹澤寿幸: マルチモーダル情報統合システムの研究動向，人工知能学会誌，Vol.**13**，No.2，pp.206–211，1998.

[Johnston98] Johnston, M.: Unification-based multimodal parsing, In *Proc. ACL'98*, 1998.

[Kobsa86] Kobsa, A.: Combining deictic gesture and natural language for referent identification, In *Proc. COLING86*, pp.356–361, 1986.

[河野 98] 河野恭之，屋野武秀，池田朋男，知野哲朗，鈴木薫，金澤博史: ATMSベースのマルチモーダル入力統合方式を用いたインタフェースエージェントシステム，人工知能学会誌，Vol.**13**，No.2，pp.212–220，1998.

[Kono98] Kono, Y., Yano, T., Ikeda, T., Chino, T., Suzuki, K., and Kanazawa, H.: An animated interface agent applying ATMS-based multimodal input interpretation, *Applied Artificial Intelligence Journal*, 1998 (to appear).

[Maybury94] Maybury, M.T.: Research in multimedia and multimodal parsing and generation, *Artificial Intelligence Review*, Vol.**8**, No.3, 1994.

[中山 98] 中山康子，真鍋俊彦，竹林洋一: 知識情報共有システムの開発と実践，情処学論，Vol.**39**，No.5，pp.1186–1194，1998.

[Oviatt97] Oviatt, S., DeAngeli, A., and Kuhn, K.: Integration and synchronization of input modes during multimodal human-computer interaction, In *Proc. CHI'97*, pp.415–422, ACM Press, 1997.

[鈴木 96] 鈴木薫，山口修，福井和広，田中英治，倉立尚明，松田夏子: 人に近いインタフェースを目指して，情処研報，**96-HI-69**，pp.47–53，1996.

[竹林 98] 竹林洋一: Perceptual user interfaces –GUIからPUIへ–，人工知能全国大会，AIレクチャー，AI-L1，1998.

[田中 98] 田中克己: 視線・音声入力に基づくマルチモーダル・ショートカット機能の提案・試作と評価，情処研報，**98-SLP-22**，pp.7–14，1998.

[Yamaguchi98] Yamaguchi, O., Fukui, K., and Maeda, K.: Face recognition using temporal image sequence, In *Proc. Automatic Face and Gesture Recognition (FG98)*, pp.318–323, 1998.