音声言語情報処理 7－3
（1995. 7. 20）

同時複数話者の会話音声およびジェスチャのリアルタイム統合理解による Novel Interface System

伊藤慶明† 木山次郎† 関進† 小島浩† 張建新‡ 岡隆一†

†新情報処理開発機構つくば研究センタ
〒305 つくば市竹園 1-6-1 つくば三井ビル 16 階
E-mail: itoh@trc.rwcp.or.jp
‡メディアドライブ株式会社
〒360 埼玉県熊谷市筑波 3-195 熊谷駅前ビル 7 階
E-mail: chou@mediadrive.co.jp

あらまし— 本稿では、人間と計算機の新しい対話形態，インタフェース・システムの提案を行う。本システムでは、マルチユーザによる音声とジェスチャのマルチモーダルな入力が可能で、これらの認識技術を統合することによって音声とジェスチャの同時かつ相補的な理解を実現する。さらに、システムの理解内容を合成音声と画像を通してリアルタイムにかつ漸次的にユーザにフィードバックすることによって、複数の人間と計算機との知的で、かつ豊かなコミュニケーションを実現する。本方式は、一種の思考の支援と考えることもでき、これを次世代のインタフェースと位置付ける。我々は、このインタフェースを実現するために、frame-wise and realtime spotting 技術を用いて、複数話者による音声とジェスチャの同時認識／理解リアルタイム統合インタフェースシステムを試作した。

Novel Interface System by Real-time Integration of Conversational Speech Understanding and Gesture Understanding by Multiple Users

Yoshiaki ITOH† Jiro KIYAMA† Susumu SEKI† Hiroshi KOJIMA†
JianXin ZHANG‡ and Ryuichi OKA†

†Tsukuba Research Center, Real World Computing Partnership
Tsukuba Mitsui Bldg. 16F,1-6-1,Takezono,Tsukuba-shi,Ibaraki,305
‡Media Drive Corporation
Kumagaya-Ekimae Bldg. 7F,3-195,Tsukuba,Kumagaya-shi,Saitama,360

Abstrct This paper proposes a new type of dialog system, or interface system between men and computers. This system allows multi-modal input of speech and gesture by multiple users, and enables simultaneous and complimentary understanding for speech and gesture by integrating both recognition technologies. It realizes intellectual and affluent communication between multiple users and computers by real-time and gradual feedback of understanding state in the system, using synthesis speech and graphics image. The system can be thought as a novel interface system as it gives users a sense of reality and unity. We realized such a real-time interface system that integrates speech understanding and gesture understanding by multiple users.

1. はじめに

人間が計算機を利用する場合のインタフェースは、マン・マシンインタフェースと呼ばれ、計算機の進歩とともに、その形態も、キーボードとキャラクタ端末から、ビットマップディスプレイ，マウスさらにはポインティングデバイスを用いたグラフィカルユーザインタフェース（GUI）へと進歩してきた。

日常の人間同士のコミュニケーションにおけるインタフェースは、音声のみで行われるわけではなく、身ぶり，手ぶりなどのジェスチャや表情や視線などもそれぞれインタフェースの重要な要素になっている。音声は言語的な表現手段である一方、ジェスチャなどは音声と相補的な役割を担っている。本来、マン・マシンインタフェースにおいても、人間同士のコミュニケーション同様に、コンピュータを意識することなく、自分の意図を音声あるいはジェスチャ等のマルチモーダルなインタフェースによって表現できることが望ましいであろう。

従来の単純なマルチメディア環境では、個々の要素技術が独立に提供されているに過ぎない場合が多い。我々は、複数のメディアがある目的のために相補的な役割を担い、かつ同時に用いることで、片方が他方の表現の確認，補強する手段となるような、より高度なインタフェースが必要であると考える。そこで我々は音声の認識と動画像の認識技術を統合し、人間の自然な発話とジェスチャを同時にかつ相補的に理解可能なシステムの構築を目指した。

音声言語入力を用いる多くのシステムでは、一人の人間が計算機に向かうことを想定している。この場合、計算機に対して伝えたい明確な要求があって、計算機のコマンドのように文法に則った正確な文音声入力を要求することが多い。しかし、計算機が人間の知的作業を支援するためには、人間の思考を妨げるこうした制約や計算機のための pause 要求などをできるだけ排除し、人間の考えていることがそのまま発話として現れるようにする必要があると考える。すなわち、任意のタイミングの、しかも不要語や言い直し、非文等を含む発話を適切に処理できなければならない。さらに、知的作業は複数の人間による討論によって行われることも多い。この討論の知的支援を計算機が行うためには、本来の不特定話者の音声認識技術が必要であり、音声が重なり合う状態なども考慮する必要があるだろう。

さらに、我々は画像によるリアルタイムの漸次的なフィードバックは豊かなコミュニケーションの手助けになると考えている。すなわち、理解内容の漸次的視覚化（漸次的な視覚的フィードバック）は、発話を介してユーザの思考が反映されたものであり、人間の思考の支援あるいは思考の促進とも捉えることができる。ユーザにとってみると自分の考えていることが自発話を通じて時々刻々と情報検索の結果が画像表示によって会話中にもフィードバックされ、それがトリガとなって次の思考へ影響を与えることも想定される。このような自分の思考の延長線上に画像表現を与えることで、臨場感、計算機との一体感をユーザに与えることができる。ここで、人間の会話の途中における適切な視覚的フィードバックや計算機の割り込む応答性が重要な役割を担うことになろう。

上述した、音声と動画像理解の統合，複数ユーザの自然な発話の同時認識／理解，およびリアルタイムの視覚的漸次的なフィードバック、これらの機能を備えたインタフェースは、人間と計算機の新しい対話形態，すなわち、次世代のインタフェースと考えることもできるであろう。我々はこれらの機能を、人間同士の会話進行とともに実現するには frame-wise and realtime spotting 技術が必要であると考える。この考えに基づき、今回、複数話者による音声とジェスチャの同時認識／理解をリアルタイムに行い、それに対する漸時的な視覚的フィードバックを与える新しい統合インタフェースシステムを試作したので報告を行う。

以下、第 2 章で、対象としたタスクと本システムの構成について述べる。第 3 章では本システムの音声認識部とジェスチャ認識部の認識方式について説明し、第 4 章で音声とジェスチャ認識の統合方式について説明する。第 5 章で本システムの動作例を紹介し、最後に今後の課題とまとめを述べる。

2. タスクとシステム構成

2.1 タスク

タスクとして「家の配置設計」を採用した。ここでいう「設計」とは専門家による詳細設計ではなく、一般ユーザが自分 (達) の家の大雑把な配置を決定していくというものである。ユーザは、通常、「こうしたい」という漠然としたイメージを抱いているが、予

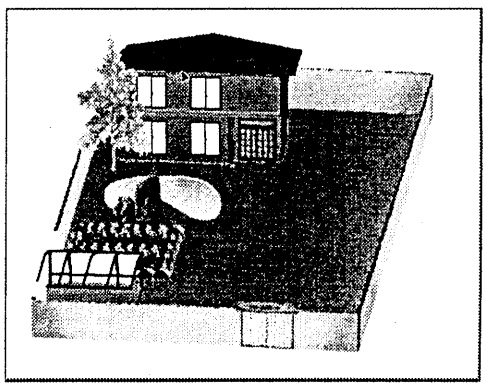

図 1: 提示画像例

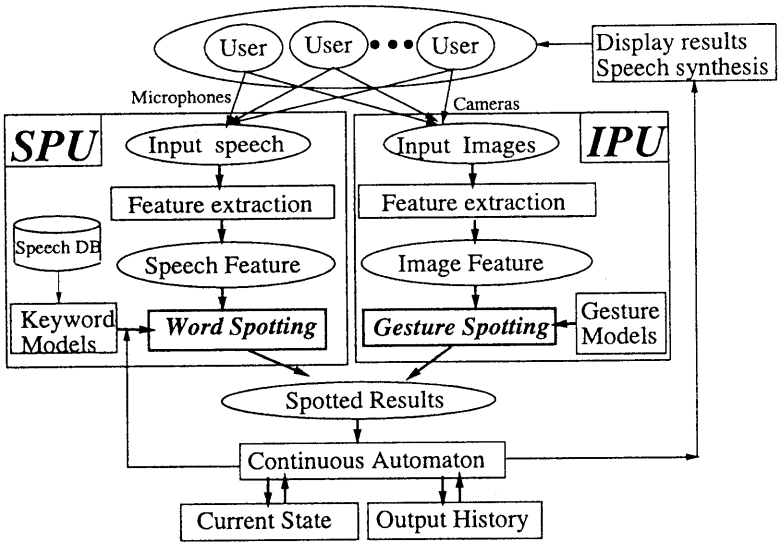

図 2: 音声動画像認識対話システムの構成

め仕様が確定しているわけではない。ユーザ間の会話や一時的な変更要求に対する視覚的フィードバックを通して、試行錯誤的，漸時的に自分の要求を満足する仕様にしていく。このため、視覚的な漸時的フィードバックは対話を進める上で重要な要素となる。また、配置問題では例えば、「こっち」というような指示のようにジェスチャが音声言語と相補的に大きな役割を担うことが多い。このように、会話音声による思考の促進，視覚的な漸時的フィードバックの役割，音声とジェスチャの相補的入力，さらにユーザにとってのアミューズメント性などを考慮すると、「家の配置設計」はそれに適合するものだと考えられる。ユーザには図1に示すような画像が提示され、希望するイメージをユーザが発声，あるいは会話，ジェスチャで指示すると、即座に画像に反映されていくことになる。

2.2 システム構成

システムの構成例を図2に示す。本システムでは、各ユーザにマイクロフォンおよびカメラが向けられ、それぞれ音声認識部 (Speech Processing Unit : SPU)，動画像認識部 (Image Processing Unit : IPU)，に送られる。音声認識部における複数話者への対応は、後述するように、各ユーザ毎のモデル等を用いて認識を行うことも、全ユーザの入力を一系統で認識を行うことも可能である。音声認識部では、フレームと同期した認識処理によってリアルタイムにスポッティングの結果を連続オートマトンへ出力する。動画像認識部では、各ユーザ毎に特徴分析を行い、ジェスチャスポッティングを行う必要がある。動画像認識部においても、フレームと同期した認識処理によってリアルタイムにジェスチャのスポッティング結果を連続オートマトンへ出力する。

音声認識部と動画像認識部から得られるスポッティング結果の統合方式としては、[4] のように連続単語スポッティングの枠組を音声と動画像に拡張して、連続パターンスポッティングなどによって統合することも考えられるが、今回、連続オートマトン [5]-[8] を用いた。連続オートマトンは、意味ネットワークのような相互結合によるオートマトンで、あるまとまった意味，あるいは概念をスポッティングする方式と位置付けられる。現在の状態、およびこれまでの認識履歴との整合性を考慮した上で、連続オートマトンによってある概念のスポッティングが行われる。このスポッティングされた概念は、ユーザ側に即座にシステムの理解状況として、前節のグラフィクスによってフィードバックされる。これにより、ユーザはある概念を発話，あるいはジェスチャによって示した瞬間に、その概念の実現イメージをグラフィクスを通して体感することができる。ここで、誤認識が発生した場合には、人間同士の会話と同様に、即座に修正すればよいのである。

人間へのフィードバックは、グラフィクスを用いて、過去の理解状況とともに現在の理解状況をリアルタイムに表示する。また、コンピュータとのユーザフレンドリな対話を実現するために、画面上に女性を模したエージェントを表示し、ユーザからの入力がしばらくない場合には、エージェントが適度に口と顔を動かしながら、最新の認識結果を音声合成で答え、ユーザの注意を喚起し会話の維持に努めている。現在のシステムでは図5のような画面を提示している。

現時点で、活性している連続オートマトンの状態を各スポッティング・プロセスに反映していないが、例えば、ワードスポッティングにおけるキーワードの対象範囲を指定するというようなフィードバックも可能である。

3. 認識方式

3.1 音声認識部

人間の自然発話では、必ずしも文法に則った文音声が正しく発声されるわけではなく、非文，省略，倒置，言い淀み，言い直しなどを含んでいることが多く、人間同士の会話ではこの現象は人間がコンピュータに対して発話する場合と比べるとさらに増加する傾向にある[11]。このような会話音声に対し、前章で述べたような機能を実現するためには、(1) 複数ユーザの任意時刻の自然な発話音声の認識／理解，を (2) フレーム同期処理，によって (3) リアルタイム，で処理する必要がある。

このための要素技術として、我々は連続DP等の始終端フリーのワードスポッティング技術が適していると考えている。しかし、単純なスポッティングの応用では挿入誤りが多いため、発話者の意味的なまとまりをスポッティング（意図スポッティング，概念スポッティング）する方式も有効と考えられるが、今回はワードスポッティングの出力結果を後述する連続オートマトンによって制御する。以下では、本システムで採用した音声認識方式について述べる。

音声認識方式

音声認識方式としては、始終端フリーのワードスポッティングを適用する。今回は連続DPを用いた。単語モデルは、不特定話者モデルとすべく、予め作成した不特定話者の3連続音素片モデル (Context-dependent Model) の音声データベースを用いて、自動的に単語表記から以下の手順でキーワードモデルを作成した[1]。

1．キーワードを音素片表記に変換

2．キーワード中の各音素片について前から順番に、前後の音素片で3連続音素片とし、同一のラベルを持つ3連続音素片を音声DBから抽出

3．音声DB中の3連続音素片の真中の音素片の特徴量系列を、キーワードの特徴量系列に付加

サンプリング周波数15kHzでA/D変換を行ない、フレーム周期8msec、フレーム長17msecで、36次元のボカシスペクトルベクトル場[3]を特徴量として用い、キーワードスポッティングは連続DP（CDP）によって実現した。

複数話者への対応方式としては、

・各話者に対する認識モデル（単語モデル，対話モデル）を用意

・1つの認識モデルを用意

等、考えられるが、今回は複数の話者に対し、1つの認識モデルで対応した。従って、前章のシステム構成において、各ユーザの音声を1系統にMixingして、音声認識部への入力とした。

3.2 ジェスチャ認識方式

人間同士の会話では、ジェスチャを交えることも多く、このジェスチャが相手の意図を理解する上で重要な要素となることも多い。音声は言語的な表現手段である一方、ジェスチャは音声と相補的な役割を担っており、音声では表現しずらい、大きさや相対的な位置関係を表現するには適している場合が多い。ジェスチャは非言語的な内容を含む表現手段の典型で、画像表現への繋がりと位置付けることもできる。我々は、この人間のジェスチャを動画像として捉え、音声と同様の枠組でスポッティングによる認識が可能であることを既に検証している[14]-[17]。

以下に、本システムで用いたジェスチャ認識方式について説明する。

ジェスチャ認識方式

ジェスチャの認識方式としては、音声認識方式と同じスポッティング技術を適用する。今回は音声と同様、連続DPを用いた。ジェスチャのモデルは、予め、対象となるジェスチャ、すなわち動画像を撮影しておき、それを特徴量系列に変換して、ジェスチャの標準パターンを作成する。これは、キーワードスポッティングにおいて、各キーワードを発声して標準パターンを作成する方法と同じである。

フレーム周期は33msec、以下の処理によって求めた特徴量を用いた。

1．時空間ベクトル場抽出として時間差分抽出

2．上記物理量に対する空間的リダクション

3．上記物理量に対する時間的平滑化

4．上記物理量に対するなまし処理

連続DPに伴う局所距離は市街地距離を用いた。

4. 認識統合方式

音声と動画像の認識統合方式には、連続オートマトンを用いる[5]-[8]。連続オートマトンは、図3に示すような意味ネットワークと類似した相互結合オートマトンによって表現し、概念スポッティングを実行

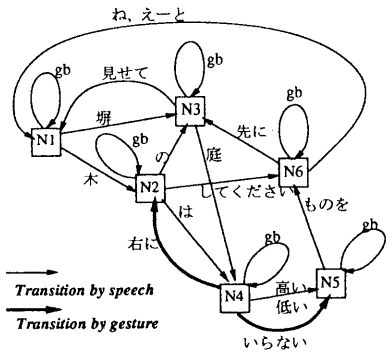

図 3: 連続オートマトンのネットワーク表現例

するアルゴリズムである。連続オートマトンは、認識単位の粒度フリーかつ並列アーキテクチャを実現するアルゴリズム [8] の一つである。ノードの状態が画像を表し、アークが語彙を表すラベルとなっている。また、図 4に示すように、各時刻の入力信号がオートマトンのすべてのアークに付加されたラベルと比べられ、その距離が計算される。同時に、すべてのノードの状態が、その関係するアークの距離を用いて更新される。ノードにはアーク遷移の履歴が記憶され、ノードの更新では、選択されるアークを通じて、遷移してくる他のノードのもつ履歴を引き継ぐ。すべてのノードの状態は画像としての出力の候補となる。処理方式の特徴を以下に示す。

(1)一つの概念は一つのノードに対応し、ノードは充足度と出力となる画像特徴ベクトルをもつ。このとき履歴により状態、すなわち出力画像は同じノードにおいても変化する。

(2)アークは (部分) 画像または画像の操作を示す記号を表し、このアークのラベルが、音声またはジェスチャの語彙となる。(すべての) アークにつくラベルと、各時刻の入力との類似度、すなわち音声／画像認識部の認識結果が入力となる。

(3)各時刻ごとに、その時刻の入力記号によって、すべてのノードの状態とアークの充足度を一斉に (2) の類似度を用いて更新する。

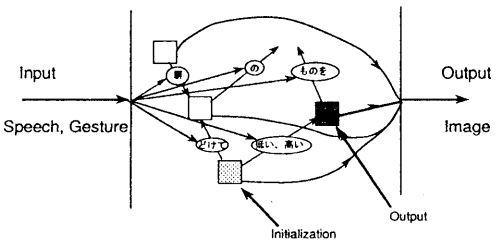

図 4: 連続オートマトンの概念図

(4)出力を行ったノードに履歴を渡したノードの状態はリセットし、すべてのノードで任意の時刻で出力することができる。

このアルゴリズムによる、音声, 動画像の認識統合の動作例を図 3を用いて示す。

(a) 「.... 木は低いものを先に....」

(b) 「.... 木はこっちに (「右に」というジェスチャ) してください...」

という 2 つの発声を考える。上記の (a) の音声だけの場合は→ $N_2 \rightarrow N_4 \rightarrow N_5 \rightarrow N_6 \rightarrow N_3 \rightarrow$となる。一方、(b) の場合には、「右に」という音声をジェスチャで行い、$\rightarrow N_2 \rightarrow N_4 \rightarrow N_2 \rightarrow N_6 \rightarrow$という遷移が実行される。また、入力の不確定要因から競合状態が生じた場合、各アークの活性度 (入力信号と類似性が高い場合に活性する) の連鎖関係より、競合解消が行われ、結果の統合／出力が行われる。この遷移過程で各ノードにおける画像出力は閾値を用いて行う。

5. システムの構成とその動作例

現在音声認識部のキーワードは 50 単語, 動画像認識部は「前へ」, 「後ろへ」, 「不要」を表す 3 種類の動画像を標準パターンとして用いている。ジェスチャの認識に関しては、今回ハードウェア上、1 人のユーザのみの入力が可能とした。本システムでは、これらの認識, 統合系はすべてフレーム同期にリアルタイムで処理している。音声認識と動画像認識は異なるフレーム周期で処理されているが、音声側の短い周期 (8msec) で連続オートマトンのインクリメンタルな更新および出力を行っている。

次に、本システムの画面を図 5に示す。また、2 人のユーザとの対話例を図 6に紹介する。

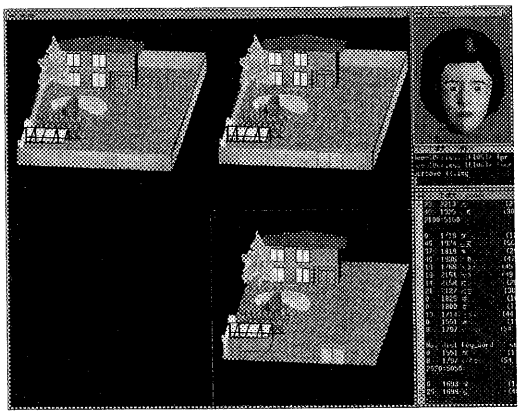

図 5: 本システムの画面例

A：家はどういうタイプがいいですか？

B：私は家は洋風の2階建てがいいです。

＜洋風2階建ての家が表示＞

A：そうすると塀はどうしましょう。

B：塀はコンクリートでどうかな。

＜塀はコンクリートに変更＞

A：やっぱり、塀はレンガのほうがいいな。

＜塀はレンガに変更＞

A：そう、そう、そうすると、門も洋風にしないとおかしいね。

＜門が洋風に変更＞

B：家はもうちょっとこっち（前の方というジェスチャ）にして。

＜家が前に移動＞

A：じゃあ、植木は向こう（後ろの方というジェスチャ）にして。

＜植木が後ろに移動＞

A：ううん、植木はもっと向こう（後ろの方というジェスチャ）の方がいいね。

＜植木が後ろに移動＞

A：それで、石は、

B：いらないよね。

＜石がなくなる＞

図 6: 会話例：下線＝スポッティングされる部分
(2重下線：ジェスチャ)

6. おわりに

本稿では、人間と計算機の新しい対話形態，次世代のインタフェースとして、音声と動画像理解の統合，複数ユーザの自然な発話の認識／理解，およびリアルタイムの視覚的漸次的なフィードバックなる機能を備えた新しいインタフェース方式を提案した。我々は今回、frame-wise and realtime spotting 技術をベースとして、上記の統合インタフェースシステムを試作した。本方式は、今後の新しいインタフェースの方向を示しているのではないかと考える。

謝辞　本研究の機会を与えて下さった新情報処理開発機構 島田潤一所長に深く感謝致します。本研究では電総研の音声DB　"ETL-WD・I-1" を使用した。

参考文献

[1] 岡 隆一："部分整合法の出力へのベクトル連続DP適用による文スポッティング型連続音声認識", 信学論 (D-II), J76-D-II, No.5 (1993).

[2] 伊藤慶明, 木山次郎, 岡隆一:"文スポッティング音声認識における部分文認識と未知語処理方式", 信学論 (D-II), J77-D-II, No.8 (1994).

[3] 岡隆一："音素スポッティングにおけるスペクトルベクトル場のボカシ処理の効果について", 信学技報, SP88-100 (1988-12).

[4] 木山次郎, 伊藤慶明, 岡隆一："リアルタイム発話視覚化システムの試作", 情処研報, SLP2-3 (1994-7).

[5] 岡隆一, 木山次郎, 伊藤慶明："概念スポッティングのための画像オートマトン", 音講論 3-4-12 (1995-3).

[6] 岡隆一, 伊藤慶明, 木山次郎, 張建新："思考過程におけるサブサンプション・アーキテクチャとしての概念スポッティング", RWC 情報統合ワークショップ '95 (1995-4).

[7] R. Oka, J. Kiyama, H. Kojima, Y. Itoh, S. Seki, and S. Nagaya, "Real-time Integration of Speech, Gesture, Graphics and Data-base" '95 RWC Symposium (1995-6).

[8] 岡隆一, 木山次郎, 伊藤慶明："認識単位の粒度自由・並列アーキテクチャとその実現のためのReference Interval-free 連続DP", 情処研報 (1995-7).

[9] K. Nagao and A. Takeuchi, "A New Modality for Natural Human-Computer Interaction: Integration of Speech Dialogue and Facial Animation", Proc. ISSD-93, (1993-11)

[10] 竹林洋一："音声自由対話システム TOSUBURG II ーユーザ中心のマルチモーダルインタフェースの実現に向けて", 信学論 (D-II), J77-D-II, No.8 (1994).

[11] 上條俊一, 秋葉友良, 伊藤克亘, 田中穂積："音声対話データの分析と発話理解への応用", 情処研報,SLP3-6(1994-10).

[12] 吉岡理, 南泰浩, 山田智一, 鹿野清宏, "電話番号案内を対象としたマルチモーダル対話システムの作成", 音講論,1-8-19, (1993-10).

[13] 畑崎香一郎, 野口淳, F. Ehsani, 渡辺隆夫："発話同時理解による音声対話インタフェースの検討", 音講論,3-4-13 (1993-3).

[14] 高橋勝彦, 関進, 小島浩, 岡隆一:" ジェスチャー動画像のスポッティング認識," 信学論 (D-II),J77-D-II, 8, (1994).

[15] K. Takahashi, S. Seki, and R. Oka: "Spotting Recognition of Human Gestures from Motion Images," Time-Varying Image Processing and Moving Object Recognition 3,Ed. V. Cappellini, Elsevier, (1994-6).

[16] K. Takahashi, S. Seki, H. Kojima, and R. Oka: "Recognition of Dexterous Manipulations from Time-Varying Images", MNAO'94, (1994-11).

[17] S. Seki, K. Takahashi, and R. Oka: "Gesture Recognition from Motion Images by Spotting Algorithm," ACCV'93, (1993-11).