高フレームレートカメラと FPGA による空中タイピングシステムの構築

○寺嶋一浩†，小室孝†，石川正俊†
† 東京大学情報理工学系研究科

本研究では，単眼の高フレームレートカメラによる動画像から，掌と指の動きを三次元追跡し，空中でのタイピング動作を認識するシステムを構築した．処理をリアルタイムで行うために，FPGA を用いて演算を高速化するハードウェアを開発した結果，最大 138fps，スループット 29ms でタイピング認識を実現した．

1. はじめに

近年，携帯情報機器が小型化したことから，文字入力やポインティング等の操作のための十分なスペースを，機器の表面上に確保することが困難になっている．ユーザは小型のタッチパネルや，間隔の狭いキーボードなどの限られた平面上で細かい入力操作を要求される事が問題となっている．これを解消する方法として，手と指の三次元動作を認識して，機器への入力とするインタフェースが考えられる．

非接触型の手指認識では，視覚に基づく方式がほとんどを占める．多視点画像から三次元形状を復元する手法[1]では，詳細な手形状を取得できる利点があるが，機器サイズの点で携帯には不向きである．

単眼画像を用いて空間内での手指姿勢の推定を行う研究では，次のようなものが行われている．Hoshino ら[2]は，手の領域画像から特徴量を抽出し，主成分分析で次元を減らしたのち，データベースと比較して推定を行っている．岩井ら[3]は，抽出した特徴点と手形状の三次元モデルとのマッチングにより指位置を推定する方法をとっている．Stenger ら[4]は，手指姿勢の状態を木構造でデータベースと照合して推定を行っている．このように単眼画像による手指認識では，膨大なデータベースを事前に用意する手法が多い．これらの方法は携帯機器のように計算機資源が限られる状況では，適用が難しい．

本研究では，図 1 に示すように，単眼カメラにより撮影した空中での手指動作を入力とし，マーカーを用いずに掌の姿勢と指先のタイピング動作を認識することを目的としている．これにより，非拘束，非接触かつ広い操作空間でタイピングやポインティング操作を行う入力インタフェースが実現できる．また，高フレームレートカメラを用いることで，従来のビデオレートでは捉えることが困難だった素速い手の動作の認識の実現が期待できる．

2. 手指の追跡アルゴリズム

本アルゴリズムの目的は，単眼カメラによって空中でのタイピング動作を推定することにある．日常生活の環境中でカメラを使用することを想定すると，背景や照明条件などのノイズに対するロバスト性や，使用者の手と指に特殊なマーカーを装着させることなく認識できるユーザビリティが要求される．

掌と指の動きの追跡にあたり，掌全体と指先がそれぞれ平面であるとみなし，平面透視変換とテンプレートマッチングにより，カメラと手指の間の相対的な運動を推定する．手の構造・運動モデルを導入し，掌と指の姿勢認識を段階的に適用する事により探索範囲を絞り込み，安定した認識を可能にする．

2.1 画像の前処理

生活環境内でカメラを手持ちで使用することを想定すると，照明や背景の条件が一定でなく，画像にノイズが多く混入する可能性が高い．このため手指の``しわ ” を手掛かりとして探索し，ハードウェア

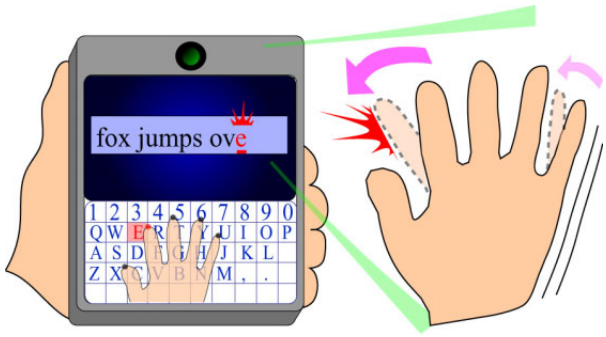


図 1 空中タイピングのコンセプト

での計算に適した形に画像を変換する．まず，前フレームで推定した掌・指の姿勢から現在の姿勢を予測し，最大4096画素に収まるように原画像から切り出して縮小する．平滑化フィルタでノイズを低減したのち，RGB 各成分について勾配を計算し，手指のしわを強調したグレースケール画像に変換する．最後に各画像のヒストグラムを平均化し，照明による影響を低減する．

2.2 掌と指のモデル

掌と指の姿勢を独立して推定する場合，指先の推定において，隣接した指との認識ミスや，極端な外れが発生しやすくなる．このため，掌と指の間に拘束条件を設定し，段階的に推定を行えば，全体の姿勢仮説数を節約し，誤認識を抑えることができる．手指のリンク構造の導入は，人間の骨格構造からみても自然であるといえる．

第一段階として，手首を基準点とする掌全体の姿勢パラメータ $\mathbf{X}_{\mathrm{hand}}$ を推定する．第二段階では指先を基準点とし，指の開きと曲げ角度を表す $\mathbf{X}_{\mathrm{finger}}$ を各指 5 本について推定する．各姿勢パラメータの成分は(1)式の通りであり，計 16 自由度を有する．

$$\mathbf{X}_{\mathrm{hand}} = [\phi_{hand} \quad \theta_{hand} \quad \psi_{hand} \quad x_{hand} \quad y_{hand} \quad z_{hand}]$$
$$\mathbf{X}_{\mathrm{finger}} = [\phi_{finger} \quad \psi_{finger}] \qquad (1)$$

2.3 姿勢推定

画像から対象物体の姿勢を推定するには，特徴点を拘束条件として逆問題を解く方法と，多数の姿勢仮説を用意して検証する順問題の方法が存在する．本研究では画像にノイズが混入する事が予想されるため，安定性を重視し後者を採用した．

まず，姿勢仮説群を用意する．時刻 t-1 で推定した姿勢パラメータ $\mathbf{X}$(t-1)と速度 $\dot{\mathbf{X}}$(t-1)を利用し，(2)式により n 個の仮説群 $\tilde{\mathbf{X}}$i(t) (i=1,..,n)を生成する．ただし，m は平均 0，共分散行列 $\sum$ のガウス雑音である．

$$\tilde{\mathbf{X}}_i(t) = \mathbf{X}(t-1) + \dot{\mathbf{X}}(t-1)\ \Delta t + \mathbf{m} \qquad (2)$$

次に，各仮説に基づいて画像を変形する．時刻 t で取得した画像を I とし，テンプレート画像を T とする．姿勢パラメータ $\tilde{\mathbf{X}}_i(t)$ を用いて $T(p,q)$から $T(p',q')$へ平面透視変換を行なう．これは(3)，(4)式で表現される．f はレンズ焦点距離，$\mathbf{t}$ は並進ベクトル，$\mathbf{n}$ は画像の法線ベクトル，d は原点と画像間の距離を表す．

$$s\begin{pmatrix} p' \\ q' \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{H} \begin{pmatrix} p \\ q \\ 1 \end{pmatrix} \qquad (3)$$

$$\mathbf{H} = \mathbf{R} + \frac{\mathbf{t} \cdot \mathbf{n}^T}{d} \qquad (4)$$

最後に仮説に基づき変形した T' と I を比較する．(5)式に基づき SAD(Sum of Absolute Difference)を計算する．

$$\mathrm{SAD}(\mathbf{T'},\mathbf{I}) = \sum_p \sum_q \left| I(p,q) - T'(p,q) \right| \qquad (5)$$

以上の操作を n 個の仮説群全てについて繰り返す．最後に，仮説群の中から SAD が最小となる仮説を選び出し,時刻 t における姿勢 $\mathbf{X}$(t)として採用する．

3. 性能評価実験

これまでに述べたアルゴリズムの効果を検証するために，事前に高フレームレートカメラの映像を記録しておき，オフラインで本アルゴリズムを適用して性能評価実験を行った．以下の実験では，320×240 画素, 120fps で約 6 秒間撮影した動画像を入力として用いた．姿勢仮説数は，掌 200 個，各指につき 200 個，合計 1200 個とした．

図 2 は，掌で空中に円を描いた場合のトラッキング結果を示している．奥行き方向の移動を抑え，カメラの視野全体を覆うように大きく円を描いた．XY 平面において，楕円形の軌跡が確認できる．また，約 1 秒(120 フレーム)周期の運動とわかる．

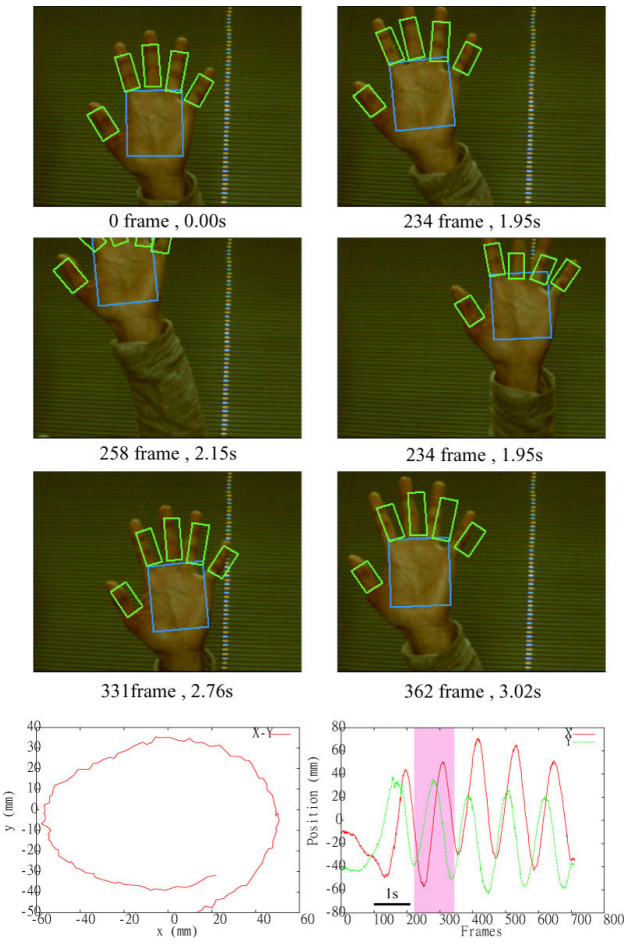


図 2 手で円を描いた場合の追跡結果

図 3 は，掌をカメラの奥行き方向に移動させた認識結果である．Z 方向のグラフから，滑らかに奥行きが推定できている．接近時と遠距離では画像上のサイズ比が 2 倍以上あるにも関わらず，1 枚のテンプレート画像でマッチングが可能となっている．

掌全体の認識に続き，指のタイピング動作の認識について確認した．ここでは指のタイピングを，図 4 に示すように指を内側に軽く曲げる動作と定義する．入力動作は，1.2 回/秒の間欠タイピングと，約 5 回/秒の連続タイピングの 2 種類とした．それぞれ 120fps，60fps，30fps で認識を行った結果が図 5，図 6 である．グラフ中の点線は未加工の認識データ，実線は 20 フレームの平均値を示している．

まず約 1.2 回/秒の認識結果を見ると，120fps と 60fps では約 1.2 秒ごとに明確なパルスが現れており，これは指のタイピング動作を反映しているものと考えられる．これに対し 30fps では，上記の 2 つに比べ不規則で，タイピング動作の見落としも発生している．次に約 5 回/秒の認識結果では，120fps で約 0.2s の周期的なパルスが見られる．60fps では見落としが増え，30fps ではランダムに近い値が観測された．

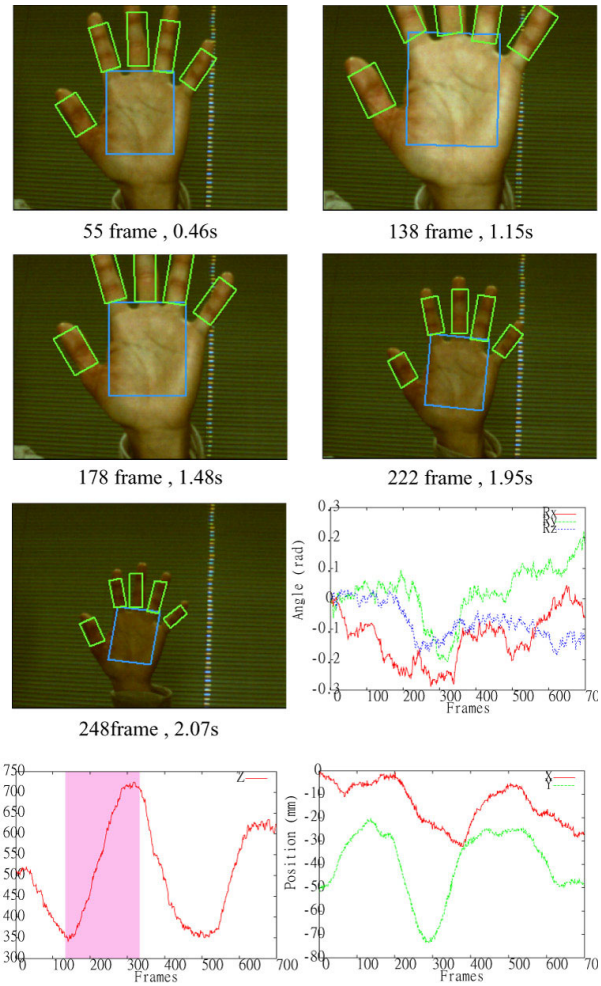


図 3 奥行き方向の運動の追跡結果

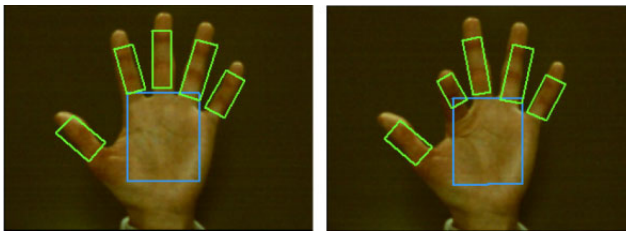


図 4 タイピング動作

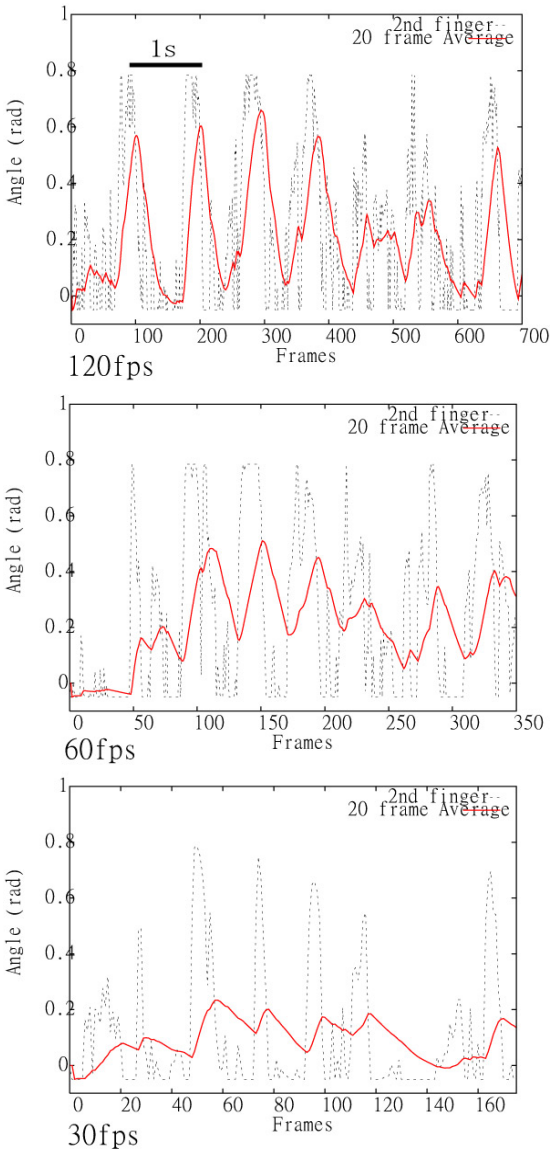


図 5 タイピング動作の認識(1.2 回/秒)

以上の結果から，毎秒 5 回程度のタイピング動作を漏れなく認識するためには，120fps 程度の高フレームレートが要求されることが分かった．

4. ハードウェアの設計と実装

手指の動作認識アルゴリズムを全てソフトウェア上で実行すると，各画素について(3),(4),(5)式が繰り

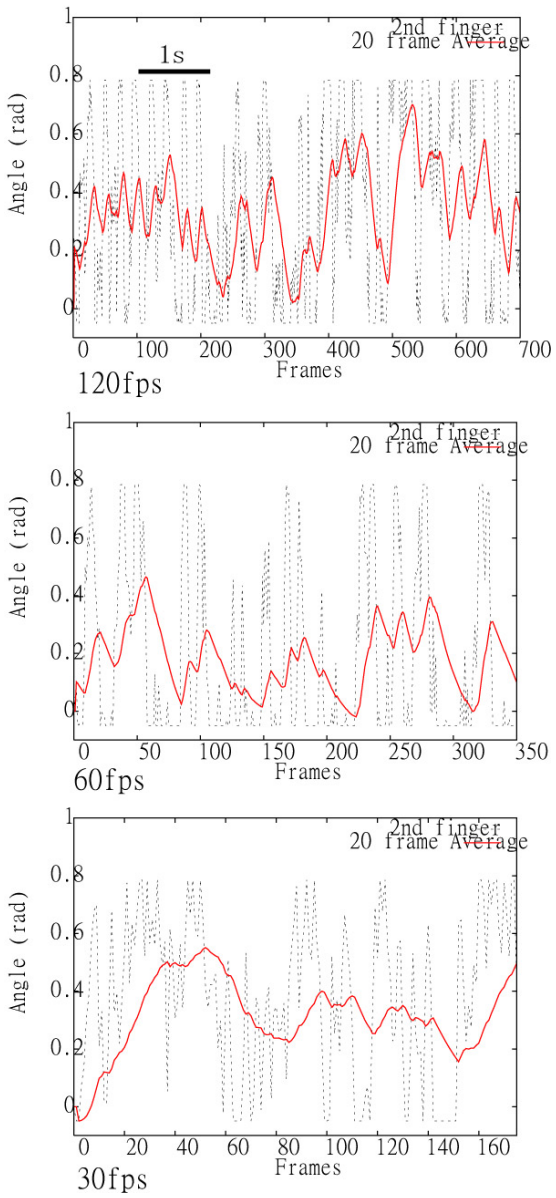


図 6 タイピング動作の認識(5 回/秒)

返し計算されるため，ボトルネックとなる．そこで，平面透視変換とテンプレートマッチングの計算をハードウェア化することで高速化を図った．

4.1 コプロセッサの設計

システム全体のブロック構成を図 7 に示す．コプロセッサの回路のうち，インタフェースや制御に関わる部分を周辺回路と呼ぶ．また，主要な数値演算を行うモジュールを PE(Processing Element)と呼ぶ．各 PE は，演算パイプラインと,ローカルメモリが併設されている．

PE の内部設計では，(3)式の平面透視変換と(5)式のテンプレートマッチング演算をハードウェア化することで処理の高速化を図る．毎回の演算内容は定型的で分岐が存在せず，流れるデータの内容だけが

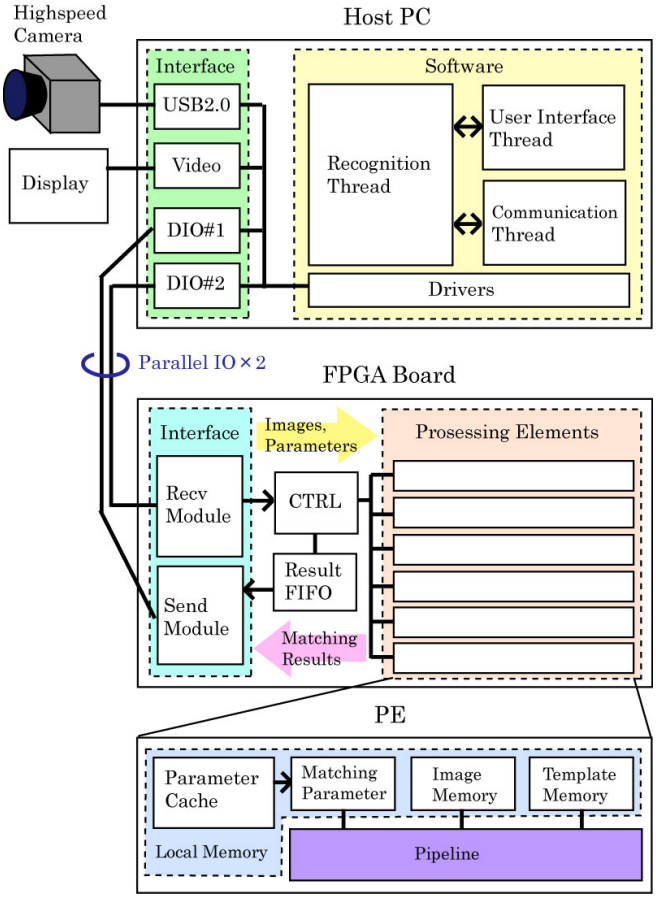


図 7 システムのブロック構成図

異なっている．そのため専用のパイプライン構造であれば，小規模なハードウェアで高いスループットを達成できる．また，(5)式では画素値への局所的な参照が多く発生することから，画像 T と I を保持するローカルメモリを併設し，メモリアクセスに要するコストを削減した．

パイプラインの内部で逐次的に画像を変形する利点として，データベースに比べ必要なメモリが非常に少ないため，携帯機器のように計算機資源が限られる状況で有利である．

また，本コプロセッサでは，SIMD(Single Instruction Multiple Data)による 6 並列化を施した．PE を複数個搭載することで，最小限の設計変更で演算能力を容易に増強することができる．

4.2 システムの実装

PC には Windows XP, Core2Duo 2.66GHz のものを用いた．カメラには 320×240 画素で最大 140fps で撮影できる Artray 社 ARTCAM-036MI を使用した．手の動作範囲を広く取れるように，焦点距離 f=4.2mm の広角レンズを装着した．

コプロセッサの実装には，Humandata 社の XCM-009-LX25 ブレッドボードを使用した．本ボードには Xilinx 社の Virtex4 XC4VLX25 が搭載されている．Verilog HDL で回路を記述し，論理合成と配置配線を行った結果，180MHz で動作する回路が合成された．これは姿勢仮説の検証を，最大 26 万回/秒実行できる性能に相当する．今回の実装では PE を 6 個とし，リソースの最大消費率が 58%であったため，本 FPGA には最大 10 個の PE を搭載できるといえる．

ホストとコプロセッサ間には，画像，パラメータ，結果などのデータ通信が必要である．本実装では，インタフェース社の DIO カード LPC-292144 を 2 枚使用し，転送効率を向上させるために独自のプロトコルを定義した．

5. リアルタイム認識実験

これまでに開発したシステムの動作を確認するため，高フレームレートカメラを用いて手と指の姿勢をリアルタイムに認識する実験を行った．図 8 に示すシステムでは，平均フレームレート 138fps，レイテンシ 29ms で実時間認識が実現された．

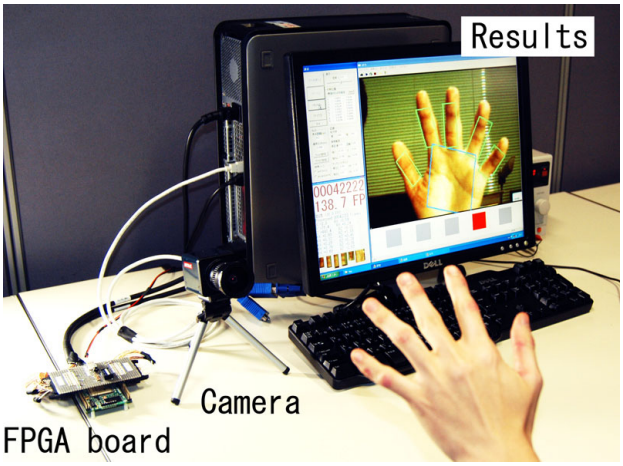
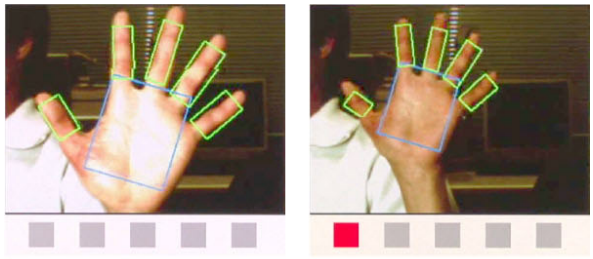
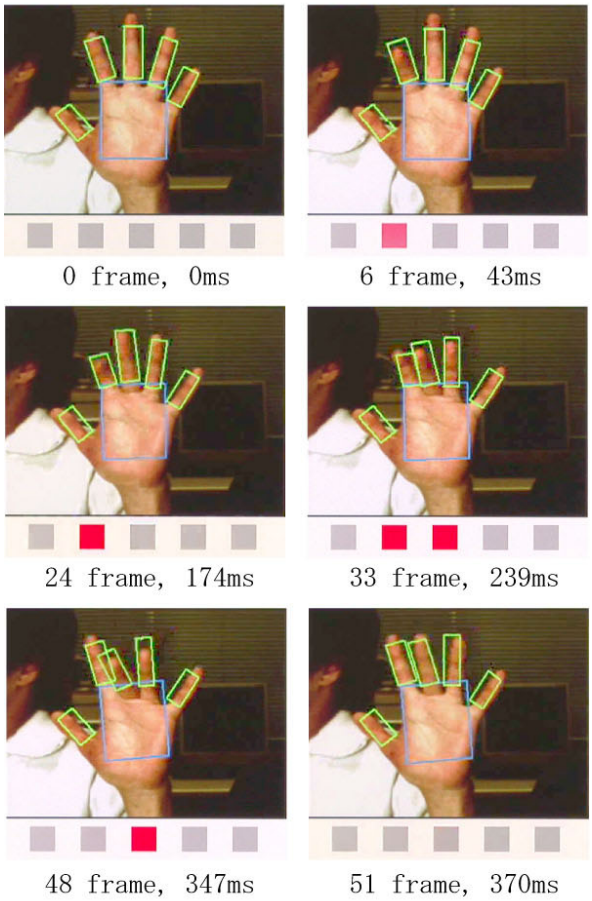


図 8 空中タイピング動作認識システム

実験のようすを図 9 に示す．(a)は奥行き方向の移動を 500ms で行った例である．カメラの接近時と，遠距離の時では掌の局所照明の強度が大きく異なっているが，照明へのロバスト性が高いためトラッキングが継続している．(b)では複数の指を用いた空中タイピング動作を認識した．人さし指は約 250ms，中指は約 180ms の期間，クリック状態だと判定された．



(a) 奥行き方向の実時間認識



(b) 空中タイピング動作の実時間認識

図 9 リアルタイム認識実験

以上の実験から，認識アルゴリズムとハードウェアの統合によって，高フレームレートかつリアルタイムな空中タイピング動作認識システムが実現されたといえる．

6. まとめ

本研究では，単眼の高フレームレートカメラによる動画像から掌と指の動きを三次元追跡し，空中でのタイピング動作をリアルタイムで認識するシステムの構築を行った．手指のしわを利用した変形テンプ

レートマッチングにより，特殊なマーカーを身体に装着する必要をなくし，照明や背景の影響を低減した．また，高フレームレートカメラの導入により，

通常のビデオ信号では判別が困難な素早い打鍵動作の認識に対応した．さらに，認識処理をリアルタイムで行うために，変形テンプレートマッチングを並列演算によって高速化するハードウェアを開発した．これらの手法により，空中タイピング動作に対して，最大 138fps，スループット 29ms で実時間認識を行うシステムを実現した．

今後は，広い入力帯域と三次元の操作空間を活かした応用アプリケーションの開発や，システムの小型化を行う予定である．

参考文献

[1] Etsuko Ueda, Yoshio Matsumoto, Masakazu Imai and Tsukasa Ogasawara. ``Hand Pose Estimation for Vision-based Human Interface'', IEEE Transactions on Industrial Electronics. Vol.50, No.4, pp.676-684, 2003.

[2] K.Hoshino and T.Tanimoto, ``Realtime estimation of human hand posture for robot hand control'', Proc. 6th IEEE International Symposium of Computational Intelligence in Robotics and Automation, Vol.6, No.8172, pp.1-6, 2005.

[3] 岩井儀雄，八木康史，谷内田正彦，``単眼動画像からの手の 3 次元運動と位置の推定''，電子情報通信学会論文誌, J80-D-II, 8, pp.1920-1926, 1997.

[4] Bjone Stenger, Arasanathan Thayanathan, Philip H. S. Torr and Roberto Cipolla, ``Model Based Hand Tracking Using a Hierarchical Bayesian Filter'', IEEE Transactions on Pattern Analysis ans Machine Intelligence, Vol.28, No.9, pp.1372-1384, 2006.