デジタル・ドキュメント 13－1
（1998．7．17）

# 動的環境における文書映像の収集

中川忠勝　川嶋稔夫　青木由直

北海道大学大学院　工学研究科
〒060 札幌市北区北13条西8丁目

あらまし　動的環境においての文書映像の処理では、文書画像の処理、収集において静的な文書画像の処理と違い、いろいろな性質の違いがあり、文書の動きに対するトラッキングなどの処理などの必要が生まれてくる。また固定カメラによる画像入力ではスキャナなどの静的環境に対して文書に対する解像度などの点で開きがある。

動的環境下で静的環境と比べて特徴のある文書入力システムに的を絞って、文書検出、文書入力画像の解析、文書移動に対するトラッキング、ポインティング動作の認識であるユーザインタフェース部分を紹介し、文書画像の動的映像の解析を検討する。

# An Input Method of Document Image in Dynamic Environment

Tadakatu Nakagawa,Toshio Kawashima, and Yoshinao Aoki

Graduate School of Engineering,Hokkaido University
Kita13jo-Nishi8, Kita-ku, Sapporo 060 JAPAN

**Abstract**　In this report, we propose a dynamic document image input system which automatically understands the document arrangement on a desktop, and recognizes user's pointing action. In the system, a document is viewed with an active camera located above a desk. A doument which is slided in the desktop is automatically detected and tracked by the system. A precise document image is adaptively collected based on the infomation distribution in the document. Pointing action is also analysed to recognize the intention of user. These ideas are tested on an intelligent presentation system.

# 1 はじめに

文書画像の入力は、印刷 (手書きも含む) 文書のデータベース化などを目的として行われることが多く、従来の文書画像の取得や処理形態は、静的入力を前提として行われてきた。つまりスキャナや高解像度のディジタルカメラで、静止した状態の文書を撮影することが一般的であった。

しかしながら、より普通の環境 (例えば机上) の中でインタラクティブに文書を扱おうとすれば、動的な文書の解析が必要となる。つまり、机、あるいは教材呈示用書画カメラのステージの上などに置かれた様々な種類の文書を、文書入力やプレゼンテーションのためにインタラクティブに扱おうとすれば、文書の移動や交換などの文書自体の動きや、ユーザの文書に対する動作の理解が必要である。これらをまとめると以下のような機能を実現する必要がある。

- 文書の文字のサイズにあわせてズームやカメラの位置を制御し、文書に最適な解像度での入力する機能。
- 文書の物理構造や論理構造を解析しの時空間内での一貫性をとる機能。
- ユーザの文書に対する操作を認識し、ユーザの意図にそった操作を文書画像に施す機能。

言い替えると、動的環境の中で文書画像を扱う場合には、文書を高精細に撮影すると同時に、文書に固有のフレームの中でユーザとのインタラクションを行う必要があるということである。

我々はこれまで最初の課題について、文書の構造を解析し、フォントサイズに適応的にズームを制御し、モザイク合成を行うことで、文書を入力する方式を提案してきた。しかし、動的な環境に対応するために不可欠な、第二の課題については着手していなかった。また、指や指示棒などによる文書のポインティング動作の解析についても課題として残されていた。

**カメラ視野内における文書の移動の解析** 文書の視野内への挿入の検出、および移動のトラッキング

**文書フレームでのポインティング動作の理解** 文書に対するポインディング操作をトラッキングの結果に基づいて文書固有のフレームのなかで認識し、ユーザの意図にそった操作を行う

以下では、これらの機能の実現方式について説明を行うとともに、我々の開発した適応的解像度による文書入力方式と統合して構成した、柔軟な文書収集解析環境について報告する。

# 2 書画カメラ下での文書映像のハンドリングに必要な技術

我々が文書映像の収集解析に用いるシステムでは、やや広めの原稿台の上部に固定したズーム、パン、チルトが可能なカメラ (図1) によって、動的に文書画像を撮影することができるようになっているものを前提としている。

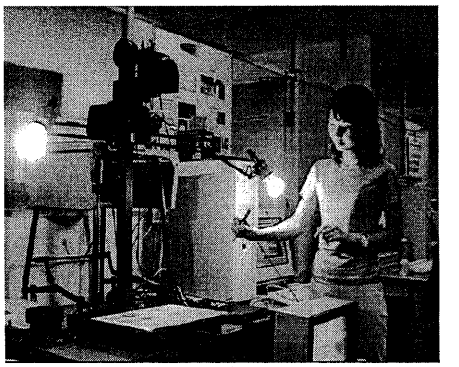

図 1: システム装置

動的環境下での文書画像の性質には、以下にのべるように静的環境ではなかった問題がある。

まず、新たな文書が視野に登場した時には、その文書映像の入力開始の決定をする必要がある。また、ユーザーが文書を移動したり、複数の文書を並べ変えたり、冊子などのページめくりといった操作を行った時には、それらの状況を解析し、視野中の文書と計算機内部にもつ文書イメージとの整合性をとっておく必要がある。

これらの検出認識は、文書の構造を解析することで処理しやすくなる。例えば文書の傾きが分れば、トラッキングや移動後の座標変換の処理において有利である。

次に掲げる項目は、動的環境下で文書画像の収集・処理を行うシステムに必要と思われる機能を、状況解析部分、収集処理部分、出力部分の3部分にわけてまとめたものである。

1 文書映像の状況解析部分

動的な環境における文書の位置の検出、文章の挙動とユーザのインタラクションの様子を動画像解析する。

a 文書検出
文書位置の検出、文書画像の登録・認識、文書移動の検出・トラッキング、文書以外の物体の検出

b ユーザー動作の解析
ポインティングの検出、ポインティング動作の意図解析、書き込みの検出

2 文書画像収集処理部分

文書自体の画像を原稿台や机から抜き出して、以後の処理に必要な解像度で入力する。

a 文書構造の解析
領域分割・領域の内容と構造の識別

b 各領域の適切なズーム値の決定、ズーム画像の取得、統合（モザイキング）

c ユーザー動作の反映（単語、図などの領域再同定など）

3 文書画像の利用部分

入力された文書のイメージを、ユーザの指示にしたがって、加工して表示したり、文書中の特定語の関連データへのリンクなどを実現する。

a 整形出力
傾き補正、指定部分の拡大、文書構造の明示、

b 他情報とのリンク
文字認識によるテキスト化・データーベース化・ハイパーリンク化

本報告では、1の文書映像の状況解析部分の実現が目的であり、以下の点について検討を行った。

1 文書の挿入検出

2 入力文書のトラッキング

3 ポインティング動作の解析

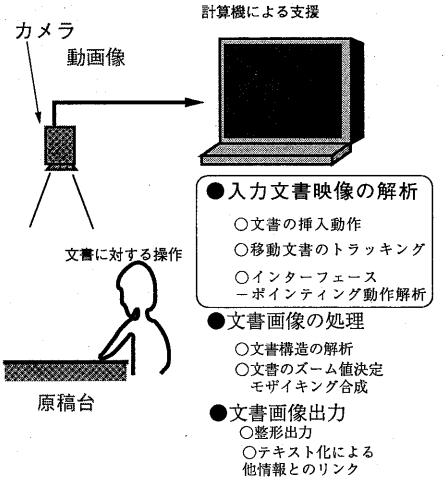


図 2: 全体のシステム

# 3 文書映像の挙動の解析

## 3.1 文書の挿入検出

まず、文章がカメラの視野内に挿入 (登場) された場合の検出方法について考える。我々が利用したシステム構成撮影した映像は、つぎのような特徴を持っている。

1 挿入時点では、文書と同時に文書を動かす手や腕が画像の外枠から文書上にかけて存在する。

2 挿入終了時には文書は静止して、手や腕などは文書から離れていき退出する。

3 文書の位置は挿入開始前の状態の画像との差分から明らかになる。文書の傾きは文書の四辺の傾きあるいは、文書中の文字列 (文) の傾きで検出できる。

以上の特徴を考慮して文書の挿入検出を行い、文書画像の取り込み開始決定を行う。

基本画像の入力まず、視野内の変化が無い時の原稿台の映像を取り込んでおく。これを基本画像と呼ぶ。

この場合、すでに文書や他の物体が置かれていても構わない。基本映像は挿入検出のための差分計算に用いられる。

システム開始時点で何も原稿台に載せないという拘束をユーザーに課したり、システム開始前に予め初期画像取り込んでおく方法も考えられるが、前者は融通性の低下すること、後者は照明光などの環境の変化に適応できないなどの問題点がでてくる。そこで、原稿台上に文書が挿入される前の映像を常に基本として、挙動を解析することにする。

**文書の挿入検出**各時刻の入力画像と基本画像との背景差分により挿入物体の検出を行う。差分画像により手や腕と文書とを分離して、手と文書が離れたところで文書の位置を決定する。差分により検出された変化領域のうち、視野の外枠上に接続するものを手腕とみなし分離を行う。

**文書領域の検出**手腕の分離後、残った差分領域全体を囲む矩形を映像中から抜き取り、その矩形内の文書の傾きと範囲を求めて文書画像とする。文書の紙の色と原稿台の色が異なる場合、この処理により紙全体が抽出できる。しかし、色が似ている場合には文字部分を囲む領域が抽出される。

**文書の傾きの検出**文書の傾きはOCRなどで用いられる文の傾き検出法を利用する(参考文献[5])。一般に文書の動的環境の解析では原稿台の全体の状況を視野に収めるようにズームを下げた状態で撮影していることと、また一般的なカメラの解像度がスキャナなどに比べて格段に低いことから、映像中の文字の細部は潰れて、一行の文字群は直線として画像中に現れる。したがって、画像中の直線群の傾きの平均を計算することで角度が推定できる。この方法は精度は低いものの、簡単なアルゴリズムですみ、後述のマッチングやポインティング位置の座標変換などを行う際の計算には十分な精度である。

求まった基本画像、文書画像と傾きを以降の処理で利用する。

## 3.2 文書画像のトラッキング

動的なシステムでは、静的システムとの一番の違いがこのトラッキングである。ユーザーは必要に応じて文書の位置をずらすことがあるため、その位置あわせが必要となる。静的な場合では文書を固定して画像を取り込むというサイクルの繰り返しであるのに対して、動的な場合はそのサイクルが曖昧でかつ、文書への書き込みや複数の文書の挿入のために、文書をユーザーの都合で動かすからである。このような状況下でユーザーのポインティング動作を正しく解析するためには、文書の移動に追従し、移動する文書のフレームを維持する必要がある。

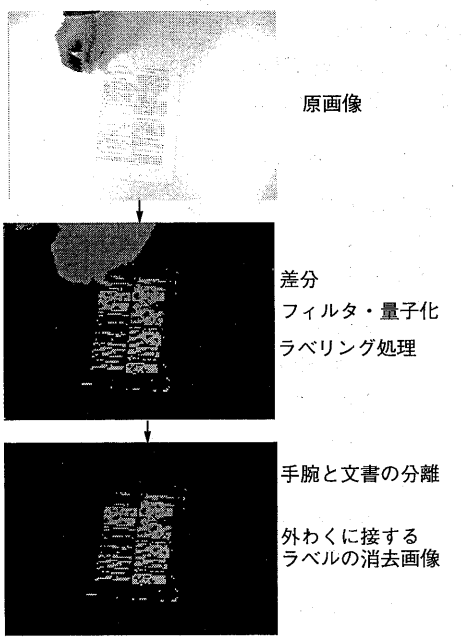


図 3: 挿入検出

ここでも、前節同様に、次のような特徴を利用して解析する。

1 文書の移動は文書と手がかさなっった後に開始される。

2 文書の静止は文書と手が離れる前であり、離れていれば文書は静止した状態である。

3 原稿台上にのっている文書は基本的にアフィン変換で記述可能で、本などの若干のたわみなどがある場合でも2次元に正確にあるいは近似で変換できる。

4 移動中や前後はユーザーは動かすという行為以外の目的、ポインティングや文書取り込みといっ

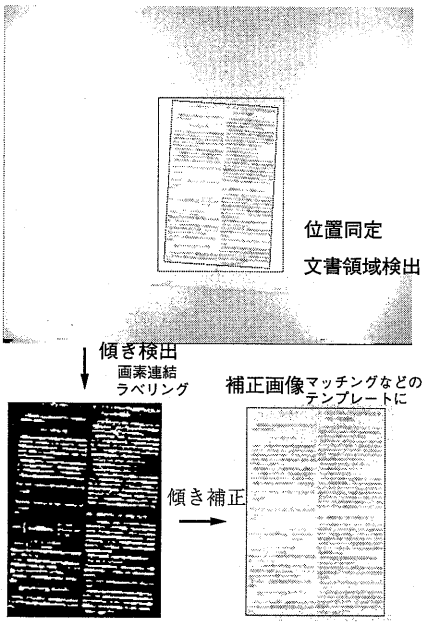


図 4: 傾き検出

た行為を同時に行わないと考えられる。したがってトラッキングを正確に求めるのは移動終了時だけでよく、それまでは粗い推定で構わない。

以上の点を考慮して文書の移動検出を行い、トラッキングを行う。文書が原稿台にある間はトラッキングは常に行っている必要がある。

**参照画像の検出**　まず文書の移動を挿入検出同様に差分を用いて求めるため、文書が静止している段階で文書領域を参照画像として取り込む。また挿入検出時の傾きを補正した画像をマッチングを行うために生成しておく。

**手腕の検出**　始めに挿入検索と同様にラベリング処理で差分画像から手腕の検出を行い、手腕が検出されたら手腕が文書領域に侵入しているかをしらべる。

**文書の静止／移動の判定**　手腕が侵入している間、参照画像との差分により文書が静止しているかどうかを調べる。手腕が文書の一部を覆うので、文書領域を縦横に分割して、各々でマッチングを行い、マッチした数が一定数あれば静止しているとする。静止している場合は後述のポインティング動作などを調べる。静止していなければ、トラッキング処理を行う。

**文書のトラッキング**　文書の領域を挿入検出と同様にして求める。つぎに、文書の傾きを調べ、傾き補正した画像とのマッチング行い位置を補正する。手が文書領域から離れるという静止条件を満たすまでトラッキングを行う。

実際には、文書移動中はインターレース入力の性質や高速物体の残像などから入力画像がぶれるため正確な位置補正をすることは難しいが、移動後の静止時には正確に位置補正をすることができるため問題は少ない。

### 3.3　ポインティング動作の認識

ポインティング動作の認識は文書画像収集後に、特定の文書の部位を指示し、インタラクションを行うために必要である。

ポインティング動作の映像の特徴は以下の通りである。

1 ポインティングの位置はポインタ (指や指示棒) の先端である。

2 ポインティングの位置やその先にはその対象となる文や図などの文書構造がある。

3 ポインティング動作を行っている時は、その動きに特徴がある。つまり、意味をもつポインティング動作は、文書のあるブロック (段落や図、語) を指しており、その物理的な領域を囲んだり、下線を引くような動作をする。

以上の点からポインティング位置を検出しポインティング動作の認識を行う。

今回は我々がすでに報告した適応的ズームによる文書入力システムで解析の途中に得られる、文書のブロック構造のどれかを選択するためのポインティング動作を対象として解析を行う。

**ポインタの検出**　トラッキング処理の項と同様にポインタ (ここでは手腕) が文書領域内に侵入しているかを検出する。そして文書が移動していない場合ポ

インティング位置の検出を行う。

**ポインタ先端位置の検出**　マッチング用に求めた参照画像を利用して現在の画像との差分からポインタを抜きだす。ポインタが文章領域内に侵入している始点を求め、その場所から最遠にあるポインタの領域の点をポインティング位置として求める。始点は文書領域の外枠(境界部分)とポインタがかさなる部分を求めて、その中点の位置とする。

**ポインティング動作の検出**　ポインティング動作の認識は求まったポインタ先端の動きから求める。特定の場所を指示するポインティング動作の特徴は、一般に指す場所まで移動した後、しばらく静止している傾向である(図5)。したがって、ポインティング位置の変位(速度)を求めて、一定の時間変位のない場合をポインティング動作として求めることにする。

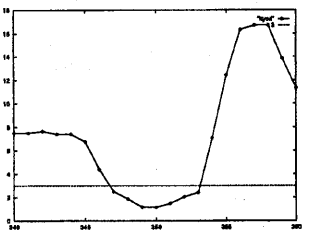

図 5: 指示動作-ポインタの移動距離

この他にも、図などの領域を囲むような動作(図6)も代表的なポインティングの特徴であり、この場合は囲み動作中は位置の変位(速度と方向変移)が一定(図7)であることからそれを検出し、その間のポインティングの位置で囲まれる領域を求める。

# 4　実験結果

文書が原稿台に挿入された後、文書がユーザーにより移動されて、その後ポインティングが行われ、その部分の文書をディスプレイに表示する、というシナリオのもとで、文書の挙動とポインティング動作に対して解析実験を行った。処理の流れはつぎのとおりである。

1. 文書が原稿台に挿入された時点で、挿入検出を行う。

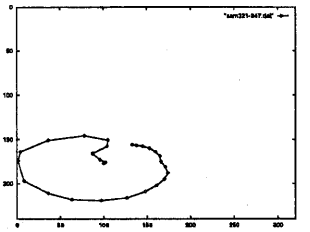

図 6: 囲み動作の動き

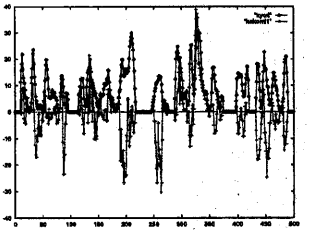

図 7: 囲み動作-加速度と方向変化

2. 文書全体をブロックに分け、個々のブロックに応じた解像度(ズーム値)でモザイク合成をして文書を入力する。
3. 文書の移動をトラッキングする。
4. ユーザのポインティング動作を検出する。
5. ポインティングされたブロックをディスプレイに表示する。

これらのうち、2,5 の処理については別途報告済みであるので、文献[3],[4]を参照していただきたい。

## 4.1　文書の挿入検出結果

原稿台上の物体の検出結果を示す。差分画像をもとの画像の 4 分の 1 に量子化して差分の大きいピクセルをカウントして求めた(図3)。次にいま求めたものをもとにラベリングして、さらに手腕を分離した画像を示す。最後に手腕が画像外にでた時点を手腕部分のピクセルをカウントすることで検出し、その時の文書の領域を求めて、ラベリングされた文からその傾きを求めるめることで、文書の傾き補正を行った(図4)。文書領域の決定には量子化画像の水

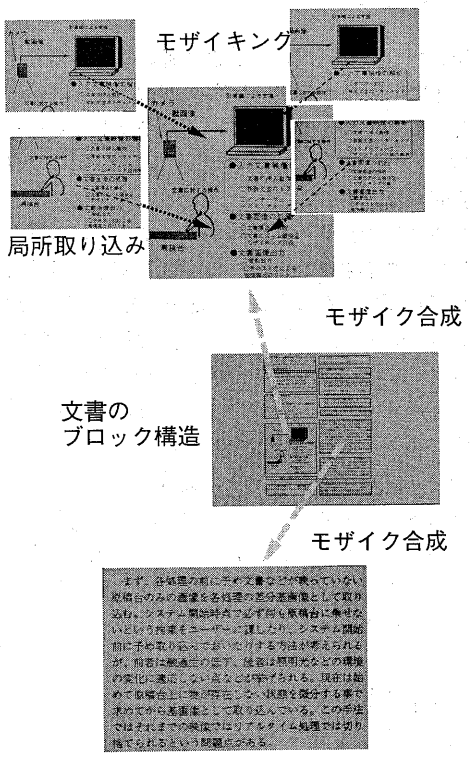


図 8: 文書合成処理

平垂直方向の濃度ヒストグラムを求めることで決定した。

原稿台が文書の色と同じなので紙の枠の検出ができないので文書領域は実際の紙の枠とは異なる。これは文書の文字(図)全体を含む矩形を求めて、適当な余白部分を足した大きさの枠を文書領域としたためである。この様に紙の外枠が検出されない場合は、前もって縦横比を測定しておき当てはめるとよいのだろう。

傾き検出に関しては幾つか処理を行った結果多少のずれは認められるが、トラッキングに用いたマッチング用の画像としては十分な精度だったので、このまま用いた。

## 4.2 文書画像のトラッキングの結果

処理1〜2は文書の挿入処理と重なるので省略する。

文書静止の検出について、実験から推定して文書領域の差分を縦横4分割してそのうち1/4マッチしたら、静止しているとした。実験結果では実験に用いたデータ全てで文書領域の大半を手腕が占めることがなかったので、正しく検出した。誤認で静止条件を満たさない場合も手腕がずっと領域を覆うわけではないので、満たしてから修正すれば問題ない。

トラッキング処理については、次のように文書領域決定時の位置とマッチング処理でてた修正誤差値はほとんどない。よって3章2節でのべたようなことからも多少計算コストのかかるマッチング処理は移動中では行わないで良いのではないかという結果が得られた。

移動フレームのマッチングの誤差

| 誤差 | ~1 | 2 | 3 | 4 | 5~ | 総数 |
|---|---|---|---|---|---|---|
| フレーム数 | 74 | 19 | 7 | 2 | 29 | 131 |

(誤差の単位:ピクセル四方)

(5以上の誤差の大半は文書の移動が速すぎるため)

## 4.3 ポインティング動作の認識

処理1〜2は文書画像のトラッキングと重複するので省略する。

ポインティング位置の同定は現在の文書画像領域内で検出を行い、最初の挿入検出で求めたキーとなる文書画像に座標変換を行っている。

求まったポインティング位置は低解像度の文書画像では精度的に文を指し示すのがやっとであり正確に求めることができない。現在のままでは、図や段落などの文のかたまりを指す程度の精度しかないので、単語レベルの認識のためには動的にズーム処理などを行う必要がある。

ポインティング動作の認識ではポインティング位置から速度方向の変位をもとめて実験を行った結果、時間になおして0.5秒程度以上の静止で指示のポインティング動作と認識できる(図5)。囲みの動作(図6)も経験則から以下の判断式で検定を行った(図7)。

|フレーム間の移動方向の変位| $< \frac{\pi}{4}$ のとき

$$\sum_{\text{開始フレーム}}^{\text{静止フレーム}} \text{フレーム間の移動距離} \times \cos\left(\pi - \frac{(\text{フレーム間の移動方向の変位})}{2}\right) \times \frac{\text{定数}\,a}{\text{定数}\,a - |\text{フレーム間の移動の変位の変位}|}$$

$$\text{定数}\ a = 5.0 \times \frac{2\pi}{360} \ldots (1)$$

上式がある程度大きいとき

1式は速度、方向変位とも一定なほど条件を満たすようになっている。

### 4.4 システムの構築

これらの各処理を章頭のべたシステムとして実際に構築してみた(図9). 文書合成は図(8)のようになる(参考文献[3][4])。

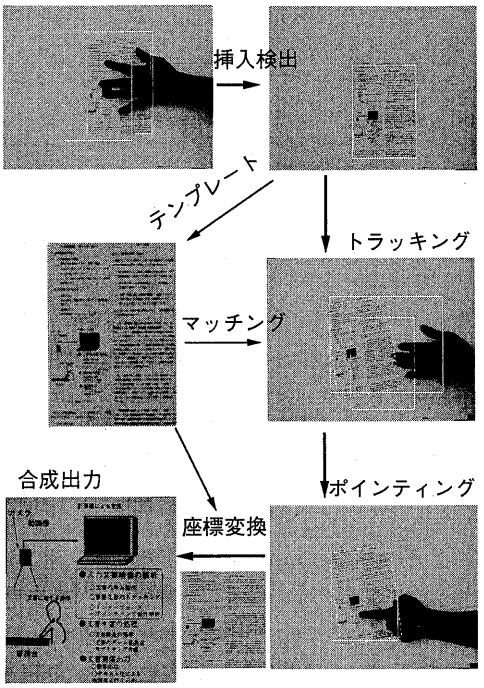


図9: 処理の流れ

全体の問題点として

実際の映像は、被写物体が静止していてもカメラの特性やノイズ照明などのせいで各フレームによって各画素の濃度は結構ばらつきがあり前処理にかなり工夫を強いられた。

またズームやパン、チルトが行えるアクティブカメラをより効率良く利用したり、複数のカメラを用いて例えば1つを原稿台全体を投影し1台を文書追跡専用に用いるといったことも考えられる。

## 5 まとめ

現在のシステムでは複数枚文書が重なったり本などの様に紙にたわみなどの歪みがある場合、またページをめくったり各ドキュメントのつながりといった複数文書の構造をどの様に扱って行けばよいか今後の課題である。

文書画像の動的収集の効率化にはさらに静的収集にはあまり必要なかった文書画像の特徴を利用した処理アルゴリズムの開発、改良が必要となる。また、ズームやチルト、パンといった動作が可能なアクティブカメラなどを効果的に用い、いかにその特徴をみつけ利用しで処理を行うかで、さらに質の高い画像収集の利用ができるのであろう。

現在、プレゼンテーション支援システムとしての文書収集処理システムなどを(参考文献[3] [4])構築している。今後このようなシステムの開発·統合をはかることで、インタラクティブなシステムの実現をめざしている。今後の課題である。

## 参考文献


[1] Pierre Wellner: Interacting with Paper on The Digitaldesk.,Vol.36 No7 in Comunications of the ACM,pp.87-96,1993.

[2] Shanon X.Ju, Michael J.Black, Scott Minneman, Don Kimbe: Analysis of Gesture and Action in Technical Talks for Video Indexing.,IEEE.CVPR97,pp.595-601,1997.

[3] 戸田真志, 山口敬人, 川嶋稔夫, 青木由直:能動視覚を用いたプレゼンテーション支援, 電子情報通信学会技術研究報告, PRMU97-127,pp.114-121,1997-10.

[4] 戸田真志, 川嶋稔夫, 青木由直:画像情報分布に基づく視覚ズームの適応的制御, 電子情報通信学会論文誌,vol.J81-D-2,No.1,pp84-92,1998-1.

[5] Anil K.Jain,Fellow,Bin yu: Document Representation and its Application to Page Docomposition.,IEEE Transactions on Pattern Analysis and Machine Intelligence.,vol.20,no.3,March 1998.