

新型インテリジェントルームの構築

○ 永易武[†], 浅野秀胤[‡], 高橋真人[†], 寺林賢司[†], 梅田和昇[†]

○ Takeshi Nagayasu[†], Hidetsugu Asano[‡], Masahito Takahashi[†]

Kenji Terabayashi[†] and Kazunori Umeda[†]

[†]: 中央大学/JST CREST, {nagayasu,takaha}@senseor.mech.chuo-u.ac.jp
{umeda,terabayashi}@mech.chuo-u.ac.jp

[‡]: Pioneer, hidetsugu_asano@post.pioneer.co.jp

本論文では、入江らによって提案されたインテリジェントルームを基に、各要素技術の性能を向上させた新たなインテリジェントルームを提案している。インテリジェントルームとは、位置や道具に拘束されることなく、ジェスチャを用いた家電機器の操作が可能な知能ロボット化された部屋である。従来のシステムの要素技術に対し、「手振り検出」では指振り検出による位置計測分解能の向上、「肌色登録」では指振り検出との同時処理による処理時間の短縮、「指の本数認識」では Snakes を用いた新手法による環境に対するロバスト化をそれぞれ実現している。

<キーワード> インテリジェントルーム, ジェスチャ認識, 画像処理, ヒューマン・インターフェース

1. 序論

近年、様々な技術の発展に伴い、我々の生活環境の情報化、インテリジェント化、ネットワーク化が進んでいる。一方、多機能化することで、操作の複雑化、階層化といった問題が生じている。我々の周囲にある機器の多くはボタンやリモコンを使って操作される。しかし、ボタンは使用位置が限定される欠点があり、リモコンはそれ自体を所持しなければならず、場合によっては探してから操作するという二度手間になってしまう。我々が周囲の機器を操作する場合、位置の拘束を受けず、道具の拘束を受けずに操作できることが望ましい。そこで、直感的かつ非接触なインターフェースとしてジェスチャが考えられる。これまでに、動画像を用いたジェスチャ認識の研究は数多く報告されており[1][2], それらを用いた商品も発表されている[3][4]. またそれらを利用し部屋全体を知能ロボット化したインテリジェントルームの研究も行われている[5][6]. 入江ら[7]は手振り検出から始まるジェスチャ認識を利用して、位置や道具の拘束を受けない家電操作システムを構築している。

本研究では入江らの手法を基に、各要素技術を改動的画像処理実利用化ワークショップ DIA2011 (2011.3.3-4)

良した新たなインテリジェントルームを構築する。

2. インテリジェントルームの概要

本研究で扱うインテリジェントルームとは、室内のすべての家電機器をジェスチャのみで操作できる知能ロボット化された部屋のことであり、一般的な家庭などやオフィスなどでの使用を想定している。

まず、二台以上のカメラで「指振り検出」を行うことで、操作者を発見すると共にその三次元位置を取得する。取得した操作者の位置へカメラを Pan-Tilt-Zoom し、ジェスチャ認識範囲を限定する。その間、検出された指振り領域内の色情報を用いて「肌色登録」を行う。操作毎に肌色登録を行うことで、照明の影響や個人間の肌色の違いに対してロバスト化が望める。その後機器に対応した番号を指の本数で表し、「指の本数認識」を行うことで操作機器を決定する。具体的な機器の操作は「指の本数認識」および「手の動き認識」を用いて行う。

システムは Pan-Tilt-Zoom カメラ, PC, 赤外線学習リモコンから構成される。カメラと PC はネットワークで接続されており、PC からのカメラの操作、カメラから PC への画像の転送をネットワーク

を介して行う。また、画像から認識したジェスチャに応じて、PC から赤外線学習リモコンを用いて家電操作を行う。

以下の章で、各要素技術について従来研究からの改良点を述べながら、手法について詳しく説明する。

3. 操作者特定のための指振り検出

操作者を特定するために画像内でフーリエ変換を用いた指振り検出を行い、同時に3次元位置を計測する。

従来研究[7]では手全体を大きく振る必要があったが、本手法では5m以上離れていても指を上下するだけで検出が可能である。また検出の分解能があがっているため、3次元位置計測の精度も向上している。従来システムでは、手振り位置にZoomした直後は画面の中心に手が映っておらず、微調整を行う必要があった。しかし本システムではその必要はない。そのため、処理の削減による処理時間の低減を実現している。

3.1 各画素に対するフーリエ変換

指振りを行う場合、その領域の濃淡値は、背景と指が交互に現れることで、一定の周期変化を生じる。これを利用して指振りを検出する。まず、ノイズの低減と演算量の低減のため、入力画像を低解像度化し、さらにグレースケール化する。低解像度画像の各画素に対して時間軸方向にフーリエ変換を行う。その際、ちらつきによる微小な変化などの影響を低減するためハミング窓を使用する。

3.2 指振り動作の検出

フーリエ変換の結果を元に、各画素に対して、以下に示す手法で指振り検出処理を行う。

1. 3~6Hzでの振幅のピーク強度を求める
2. この強度と、2Hzでの振幅強度の差を求める
3. この値を時間的に蓄積していく
4. この蓄積値が閾値を超えた場合手振り候補画素として検出する。
5. 以上の処理を複数のカメラで行い、エピソード拘束を満たす画素を指振り画素として特定する。

Fig.1 に指振り検出の様子を示す。カメラとの距離は約6mである。(a)が原画像であり白丸で囲まれた場所で指振りを行っている。(b)は低解像度画像であり、指振り検出された画素が白く表示されている。



(a) Input image (b) Low-resolution image

Fig.1 Detecting finger waving

3.3 二台のカメラによる三次元計測

二台以上のカメラで指振り検出を行い、三角測量の原理から操作者の位置を特定する。得られた3次元位置は、カメラからの距離を算出しZoom倍率を決定する際に用いられる[9]。

4. 肌色登録

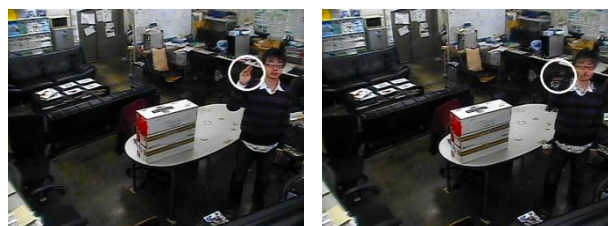
指振り領域内の画素を用いて指部分の色を肌色データとして登録する[10]。

従来研究[7]では肌色登録するためにはカメラに手のひらを向けて一定時間静止する必要があった。しかし本システムでは指振り検出と肌色登録を統合することで処理時間や操作者の負担を軽減し、よりスムーズな操作を実現している。

4.1 背景登録

指振り領域から指部分を抽出するため、数フレーム間変化の無い画素を背景画像として登録し、逐次更新する。短時間での変化の有無を判断するため、家具や壁などの静止物体のほか、手振をする際にほぼ静止する操作者の体も背景として登録される。一方、指振り部分は相対的に高速で動いているため登録されない。

背景登録の様子を Fig.2 に示す。(a)が入力画像であり、(b)が登録された背景画像である。指振りを行っている部分を白丸で囲んでいる。背景や操作者の胴体、腕などは背景として登録されているが、動きの速い指振り部分は登録されていないことが分かる。



(a)Input image (b)Registered Background image

Fig.2 Background image registration

4.2 背景差分

指振り領域内の画素値を平均し、背景画像と過去 n フレームのキャプチャ画像で差分を取得する。キャプチャ画像には指が映っているものと、そうでないものが混在している。背景画像に指は含まれていないので、領域内を指が通過しているフレームは差分が大きくなる。つまり、背景差分による差分を評価することで、肌色登録の候補となるフレームを抽出することができる。

差分計算する評価関数として、正規化した色相(H)、彩度(S)、輝度(V)それぞれの差の絶対値を合計した S_d を式(5)のように定義する。

$$S_d(t) = |I_H(t) - I_H(0)| + |I_S(t) - I_S(0)| + |I_V(t) - I_V(0)| \quad (t = 1, 2, \dots, n) \quad (5)$$

ここで $I_H(0)$, $I_S(0)$, $I_V(0)$ は背景画像の色相成分、彩度成分、輝度成分であり、 $I_H(t)$, $I_S(t)$, $I_V(t)$ は t フレーム目のキャプチャ画像の色相成分、彩度成分、輝度成分である。この評価関数をもとに、

$$S_d(t) \geq \max(S_d) / 2 \quad (t = 1, 2, \dots, n) \quad (6)$$

を満たすフレームを指が含まれるフレームとして抽出する。さらに、抽出されたフレームの領域内には指の部分と背景の部分が存在する。領域内の低解像度化する前の画像を用いて、式(6)の条件によって1画素単位で指部分を選択し、肌色情報として登録する。このとき、体の揺れなどから生じる、背景画像とキャプチャ画像のずれを軽減するため、背景画像をガウシアンフィルタを用いて平滑化する。

4.3 肌色登録

取得した肌色情報には肌色に近い背景の画素値やノイズが含まれている。それらの外れ値を除去するため、H, S, Vそれぞれの中央値を基準にデータを一定範囲に制限し、さらに各値のヒストグラムの両端から一定割合を外れ値として除外し、それ以外のデータを肌色情報として登録する。

登録した肌色情報のうち、輝度(V)は同じ環境でも照明の当たり方によって値が大きく変化する。安定した肌色抽出を行うために、照明に影響されにくいH, Sを特徴量として用いる。以下に示す式(7)、式(8)、式(9)より、マハラノビス距離を計算するための特徴量ベクトル $\mathbf{X}$ 、平均ベクトル $\mathbf{A}$ 及び共分散行列 $\mathbf{V}$ を求め、最終的な肌色データとして登録する。

$$\mathbf{X} = [H \ S]^T \quad (7)$$

$$\mathbf{A} = \frac{1}{N} \sum_{t=1}^N \mathbf{X}_t \quad (8)$$

$$\mathbf{V} = \begin{bmatrix} \sigma_H^2 & \sigma_{HS} \\ \sigma_{HS} & \sigma_S^2 \end{bmatrix} \quad (9)$$

N は登録する画素値集合の要素数である。キャプチャ画像から肌色抽出を行う際は、

$$dm^2 = (\mathbf{X} - \mathbf{A})^T \mathbf{V}^{-1} (\mathbf{X} - \mathbf{A}) \quad (10)$$

の式から肌色データに対するマハラノビス距離を求め

$$dm^2 \leq k \quad (11)$$

を満たす画素を肌色画素として抽出する。 k はしきい値である。

5. ジェスチャ認識

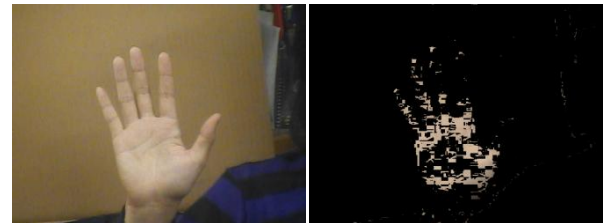
5.1 指本数認識

抽出された肌色データに対し Snakes[11]を用いて手領域を囲み、囲まれた領域の面積を用いて指の本数を認識する。

従来研究[7]ではモルフォロジー処理を利用して指領域の抽出を行っており、正確な認識には手領域が正確に抽出されている必要があった。しかし、実環境では、背景に肌色に近い色が存在し、それらも抽出してしまう場合や、手領域のみ抽出できているが領域に欠損があり正しく抽出できていない場合などが多い。本システムでは Snakes を用いて手領域の輪郭から得られる面積を特徴量とするため、領域の抽出が不安定な場合でも、領域の輪郭さえ抽出できていれば安定した認識が可能である。

5.1.1. 肌色領域の抽出

登録した肌色情報により手領域を抽出する。Fig.2に示すように背景の色が手の色と近い場合、背景のノイズも含め手領域を抽出するか、欠損を伴いながら手領域のみを抽出しなければならない。従来手法ではどちらの場合でも正確な本数認識は困難である。一方、本手法では輪郭のみ得られれば認識は可能である。よって、式(11)の閾値 k を小さく設定し、背景のノイズを排した画像を用いる。



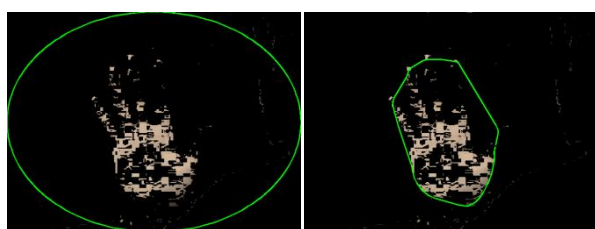
(a)Input image (b)Skin color image

Fig.3 Extracted skin color

5.1.2. Snakes による手領域の包囲・面積の算出

Snakes を用いて手領域の包囲を行う。Snakes とは物体の形状に近い閉曲線をエネルギー最小化問題として求めるものである。初期閉曲線とパラメータを適当に与えることで、与えた初期閉曲線を逐次的に変形していき、任意の形状に近い輪郭線を求めることができる。Fig.4(a)は初期輪郭を与えた状態、(b)は求められた輪郭曲線である。

手領域に対して求められた輪郭線内の画素数を面積として計測する。



(a) Initial image (b) Result image

Fig.4 Recognition of number of fingers by Snakes

5.1.3. 指本数の認識

求めた手領域の面積によって指本数の認識を行う。認識手法に以下の 2 つを用い、従来手法を含め認識結果を比較する。

- ・マハラノビス距離を用いた認識

トレーニングデータを用いて、各指本数時の面積の平均 μ_i と標準偏差 σ_i を求め、入力データ x に対して以下の式を用いて各指本数に対するマハラノビス距離 D_i を求める。

$$D_i = \frac{|x - \mu_i|}{\sigma_i} \quad (i = 1, 2, \dots, 5) \quad (12)$$

その後距離 D_i が最も小さいものを指本数として認識する。

- ・確率密度の累積による認識

トレーニングデータを用いて、各指本数時の面積の平均 μ_i と標準偏差 σ_i を求め、入力データ x に対して以下の式を用いて確率密度 $f_i(x)$ を求める。

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right) \quad (i = 1, 2, \dots, 5) \quad (13)$$

この処理を時間軸方向に複数フレームに対して行い、結果を各本数ごとに足し合わせていく。いずれかの本数の累積値が一定の値に達した場合、その本数を認識結果とする。

5.2 手の動作認識

手の動作認識には DP マッチングを用いる。肌色抽出によって各フレームに対し手領域の抽出を行う。

抽出された手領域の重心を前フレームと比較し、上下、左右、斜めの 8 方向のどの方向への移動か検出し特徴量とする。求めた特徴量を用いて DP マッチングを行い、登録されている全てのモデルとの類似度を比較し、最も近いモデルデータの動作を入力動作として認識する。モデルは任意のものを登録することが可能であるが、本システムで現在使用している動作は、"Up", "Down", "Right", "Left" の 4 種類である。

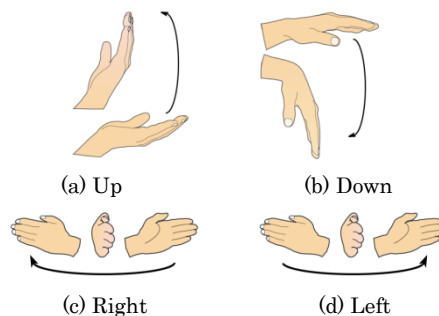


Fig.5 Gesture models

6. 実験

6.1 実験システム

画像処理には OpenCV を用い、各種演算処理は PC(Core2Quad Q9400 2.66GHz, DDR2 6GB)で行った。カメラは Pan-Tilt-Zoom 機能を搭載し、ネットワークに接続可能な AXIS 233D を用いた。赤外線学習リモコンは玄人志向 KURO-RS を使用した。

6.2 指振り認識

3 章で求めた手法に基づき、指振りの検出実験を行った。被験者 2 人を対象に計測距離を変化させて認識率及び認識時間を算出した。二台のカメラの対応した画素で指振りを検出すれば指振りを認識できたとする。実験回数は各距離 20 回ずつである。各距離で認識に要した時間を Table 1 に示す。

処理時間は距離に依らず約 2 秒である。認識率はすべての距離で 100%であった。この手法の頑健さを示している。

Table 1 Recognition time of finger waving (s)

	4m	5m	6m
平均 A	2.18	2.31	2.24
標準偏差 A	0.14	0.24	0.10
平均 B	2.12	2.09	2.14
標準偏差 B	0.20	0.26	0.16

6.3 肌色登録

4章で述べた手法に基づき、肌色登録の実験を100回行い、背景登録、肌色登録、肌色抽出の処理時間を計測した。Table 2に結果を示す。

背景登録の処理時間は十分速く、リアルタイム処理に障害とはならない。肌色登録は指振り検出を行った後のカメラのズーム中に行われる。そのためこの処理時間は十分な速さである。肌色抽出はジェスチャ認識の前処理として用いられる。多くの処理時間がかかっているが、ジェスチャ認識を行うには十分な速さである。

Table 2 Processing time (ms)

	背景登録	肌色登録	肌色抽出
平均	45.6	40.1	310
標準偏差	4.5	6.7	5.5

6.4 指の本数認識

5.1節で述べた手法に基づき、被験者2人に対して指の本数認識実験を行った。まず、被験者ごとに4章で説明した肌色登録を行う。その後、1~5本の指本数認識を行い、認識率及び処理時間を算出した。実験回数は各指本数100回ずつである。5.1節で述べた2手法と従来の認識手法を用いて比較を行った。

事前に各被験者で各指本数20回ずつデータを取得し、平均、標準偏差を算出しモデルデータとした。モデルデータの作成に使用したデータをFig.6に示す。被験者までの距離は、モデルデータ作成時、認識時とも6mとした。認識率をFig.7, Fig.8に、処理時間をTable 3に示す。

提案手法は全ての指の本数で従来手法にくらべて良い結果が得られた。特に確率累積の手法が安定した認識率を得ることができている。

なお、被験者Aの2本の場合には従来手法の認識率も高いが、これは正確な指領域の抽出には失敗しながら結果的に2本と判定されたものであり、従来手法の信頼性は低い。

処理時間は従来手法が一番短く、次いでマハラノビス距離、確率密度累積の順番になっている。確率累積の標準偏差が大きいのは、入力データにより認識まで必要なデータ数が変化するためである。今回の実験では必ず3回以内で認識できていた。実際の認識時間には肌色抽出の処理時間も加算されるが、今回の実験では認識に2秒以上かかることはなかった。

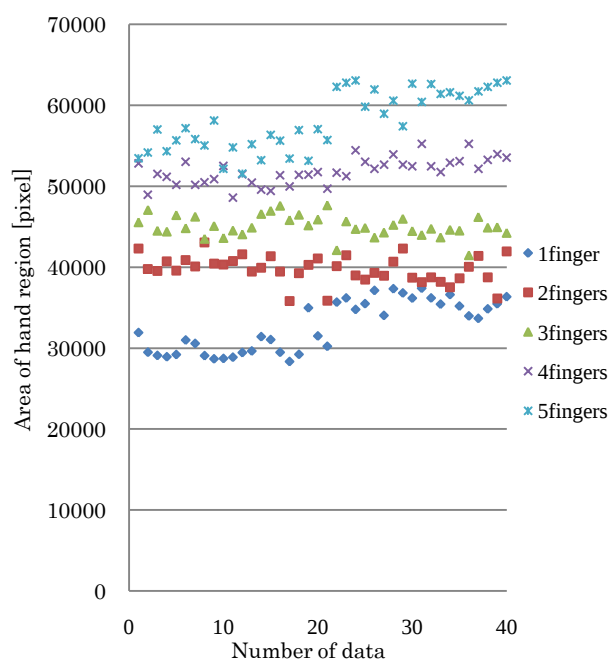


Fig.6 Model data of number of fingers

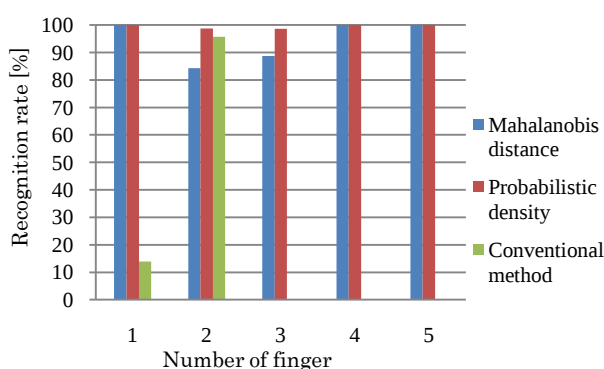


Fig.7 Recognition rate of number of fingers of A

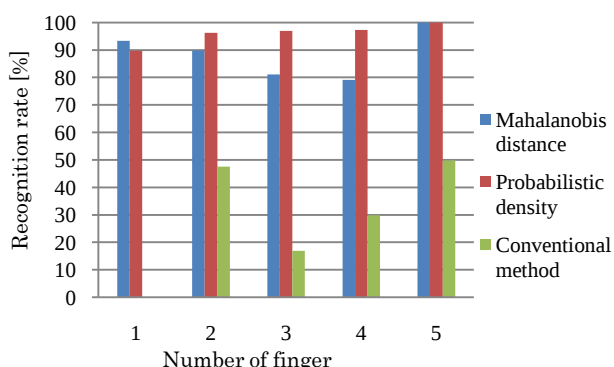


Fig.8 Recognition rate of number of fingers of B

Table 3 Processing time (s)

	マハラノビス 距離	確率密度 累積	従来 手法
平均	0.30	0.43	0.18
標準偏差	0.02	0.20	0.02

6.5 手の動作認識

5.2 節で述べた手法に基づき、被験者 2 人に対して各ジェスチャの認識実験を行った。実験回数は各動作に対して 20 回ずつである。カメラから被験者の距離は、肌色登録時と同じとした。まず、被験者ごとに 4 章で説明した肌色登録を行う。その後 4 種類のジェスチャを行い、認識率及び処理時間を算出した。

実験の結果、平均処理時間は 2.5s、標準偏差は 0.1s であった。Table 4, 5 に認識率の結果を示す。両被験者共に高い認識率を実現しているのが分る。Up, Down 動作に対して Left, Right 動作の認識率が低いのは、上下動作に対して左右動作のほうがぶれやすいためだと考えられる。

Table 4 Recognition rate of hand motion of A

UP	Down	Left	Right
100%	100%	95%	80%

Table 5 Recognition rate of hand motion of B

UP	Down	Left	Right
100%	95%	95%	85%

7. 結論

本研究では、入江らのインテリジェントルームを基に、各要素技術での性能向上を実現し、システム全体の利便性向上、処理時間の低減を実現した新型インテリジェントルームを構築した。今後は、さらに利便性を向上させるために、口唇動作認識や 3D ジェスチャ認識の導入を考えている。

参考文献

- [1] 小荒健吾, 西川敦, 信戸健太郎, 石井薫, 安井陽一, 宮崎文夫: “動きの変化に着目した動作パターンの 1 次元符号化に基づく身振り認識”, 情報処理学会論文誌, Vol.44, No.2, pp.466-477 (2003)
- [2] 益満健, 小林哲則: “部分隠れマルコフモデルとジェスチャの認識への応用”, 情報処理学会論文誌, Vol.41, No.11, pp.3060-3069 (2000)
- [3] 株式会社日立製作所, ジェスチャインターフェース, <http://www.hitachi.co.jp/New/cnews/month/2009/10/1005.html>
- [4] パナソニック株式会社, Easy-Gesture-TV, <http://panasonic.co.jp/exhibihion/ceatec2008/concept3/index.html>
- [5] 森武俊, 佐藤知正: “ロボティックルーム 1・2・3 の開発”, 電子情報通信学会誌, Vol.91, No.5, pp.402-410 (2008)
- [6] 橋本秀紀, 新妻実保子, 佐々木敦: “空間知能化: インテリジェント・スペース”, 日本ロボット学会誌, Vol.23, No.6, pp.674-677 (2005)
- [7] 入江耕太, 若村直弘, 梅田和昇: “ジェスチャ認識に基づくインテリジェントルームの構築”, 日本機械学会論文集 C 編, Vol.73, No.725, pp.258-265 (2007)
- [8] 浅野秀胤, 織茂達也, 高橋真人, 寺林賢司, 太田睦, 梅田和昇: “フーリエ変換を用いた小さな手振の検出”, ビジョン技術の実利用ワークショップ (ViEW2010), pp.264-267 (2010)
- [9] 徐剛, 辻三郎: “三次元ビジョン”, 共立出版 (1998)
- [10] K. Irie, M. Takahashi, K. Terabayashi, H. Ogishima and K. Umeda: " Skin Color Registration Using Recognition of Waving Hands", J. Robotics and Mechatronics, Vol.22, No.3, pp.262-272 (2010)
- [11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, Vol.1, No.4, pp.321-331 (1988)