# 時空間情報伝搬に基づく多眼動画像の対話的セグメンテーション

渡部　善雄† 　中島　諒† 　ファンヴェトクォク† 　高橋　桂太†† 　苗村　健†

† 東京大学大学院情報理工学系研究科　〒 113-8656 東京都文京区本郷 7-3-1
†† 東京大学 IRT 研究機構　〒 113-8656 東京都文京区本郷 7-3-1
E-mail: †{watanabe,nakashima,viet,keita,naemura}@nae-lab.org

**あらまし**　本論文では，多眼動画像から特定の物体領域を切り出す対話的セグメンテーションについて述べる．1 枚の画像を対象とする場合には，ユーザが物体領域の手がかりを与えながら精度よくセグメンテーションする手法が開発されているが，これを数千のフレームからなる多眼動画像に直接適用するのは現実的ではない．そこで本研究では，一部の画像のみにユーザが手がかりを与えてセグメンテーションし，残りの画像を自動処理するフレームワークを提案する．自動処理では，セグメンテーション済みの画像から，順次，近接するセグメンテーションされていない画像へ物体の形や色の情報を伝搬させ，それらによって定義されたエネルギー関数をグラフカットで最小化することによりセグメンテーションを行う．実験では，25 眼の多眼動画像 200 フレーム（合計 5000 枚）を用い，5 枚の画像に手がかりを与えるだけで残りの画像を精度よくセグメンテーションすることができた．

**キーワード**　多眼動画像，セグメンテーション，グラフカット

## 1. はじめに

近年，3 次元的な映像技術に関する研究が盛んになっている．その 1 つの技術として，自由視点映像合成 [1] が挙げられる．自由視点映像合成とは，空間内に配置された複数のカメラを用いて撮影された画像群を処理して，任意の視点から見た映像を合成する技術である．ユーザが見たい視点からの映像を得られるため，高い臨場感を得ることができ，次世代の映像技術として注目されている．本研究では自由視点映像合成の入力として用いられる多眼動画像から特定の物体に対応する領域を切り抜く，多眼動画像セグメンテーションについて考える．セグメンテーションされた多眼動画像を用いると，物体領域が切り抜かれた自由視点映像が合成できるようになり，それを他の映像に重畳するなど映像表現の幅が広がる [2]．

一枚の画像についてセグメンテーションする場合には，ユーザが物体領域について手がかりを与えながら半自動処理で精度の高いセグメンテーションを行う手法が開発されており [3], [4]，OpenCV 等でも入手可能である．しかし多眼動画像セグメンテーションの場合には，視点数分の画像が時系列に並ぶため画像数が膨大になる．したがって，すべての画像に対してユーザが入力を与えるのは困難である．そこで，ユーザが手がかりを与える画像の数を最低限にし，残りの画像を自動でセグメンテーションするフレームワークを構築する．このフレームワークでは，セグメンテーション済みの画像から物体の形や色の情報を伝搬することによって，近接する他の画像を自動的に処理する．この伝搬を順次繰り返すことによって，ユーザが手がかりを与えていない画像についても自動ですべての画像がセグメンテーションされる．

提案するフレームワークは手動セグメンテーションと自動セグメンテーションからなる．手動セグメンテーションでは，ユーザが選択した 1 枚の画像に対して GrabCut [4] を用いて物体領域の手掛かりを与えながらセグメンテーションを行う．自動セグメンテーションでは，動画像セグメンテーションの分野で培われてきた技術を利用する．動画像では時系列に連続するフレーム間では物体領域の形や色は大きく変わらないと考えられるため，あるフレームで得られた物体の形や色の情報を対応点の追跡によって次のフレームへ伝搬し，セグメンテーションに利用する手法が提案されている [5], [6]．したがって，あるフレームについてセグメンテーションを与えれば，順次情報を伝搬しながら残りの画像を自動で処理することができる．キョら [7] はこの考え方を多眼静止画像の視点間の伝搬に適用した．動画像の場合にはフレームは時系列の一次元方向に並ぶが，多眼画像の場合には視点は二次元方向に並んでいるため，考えうる伝搬のパターンが多くなる．そこでキョらは，最も信頼できるフレームから情報を伝搬する，選択的情報伝搬を提案した．本研究では多眼動画像セグメンテーションを行うため，キョらの手法を新たに時系列方向にも拡張する．さらに，時空間のボリューム（時系列と視点）の中で複数の画像に対してユーザが手がかりを与えられるようにする．

提案するフレームワークを実装したソフトウェアを用いて実験を行った結果，25 眼，200 フレーム（合計 5000 枚）の多眼動画像に対して，5 枚の画像を手動処理するだけで，残りの画像を自動で精度よくセグメンテーションできることを確認した．

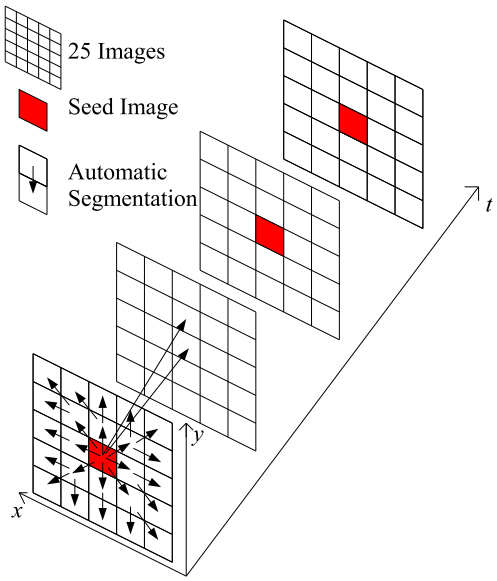



図 1　手動セグメンテーションと自動セグメンテーション

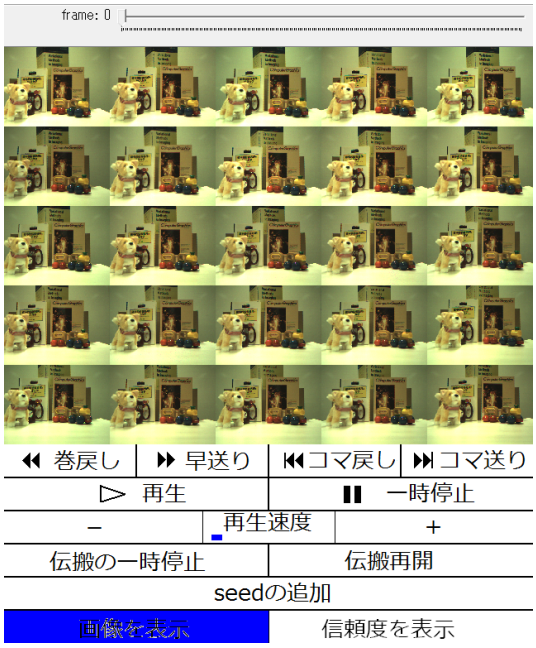



図 2　25 眼映像を閲覧するインタフェース

## 2.　多眼動画像セグメンテーションのフレームワーク

本研究では，多眼動画像セグメンテーションのフレームワークを提案する．提案するフレームワークは，ユーザが入力を与える手動セグメンテーションと，すでにセグメンテーションされた参照画像の情報を伝搬して近接する他の画像をセグメンテーションする自動セグメンテーションの 2 段階からなる．図 1 のように，25 眼の動画像のうち一部の画像についてユーザがセグメンテーションを与えると，残りの画像が自動でセグメンテーションされる．

手動セグメンテーションでは，25 眼動画像を閲覧できるインタフェースを用いて，セグメンテーションを与えたい画像を選び，GrabCut を用いてセグメンテーションを行う．ユーザがセグメンテーションを与えた画像を seed 画像と呼ぶ．多眼動画像においては，時系列方向への伝搬距離が長くなるため，1 枚の画像からでは情報伝搬が機能しないケースも多くなる．そのため，複数の画像について対話的に seed 画像を与えることができるインタフェースを構築し，自動セグメンテーションの途中においても任意に seed 画像を追加できるようにする．

### 2.1　手動セグメンテーション

手動セグメンテーションでは，ユーザのシンプルな入力から正確なマスクを与える．

図 2 は任意の時刻における多眼画像（25 眼）を閲覧できるインタフェースである．このインタフェースを用いてユーザはセグメンテーションを行う画像を 1 つ選択する．セグメンテーションを与えたい画像をクリックすると，図 3(a) のようなウィンドウが現れる．ここで，GrabCut [4] の手法を用いたインタラクティブセグメンテーションを行う．まず初めに，図 3(b) のように，物体領域を囲むバウンディングボックスを与える．このバウンディングボックスの情報を利用して色分布モデルが作成されセグメンテーションが実行された結果が図 3(c) である．修正したい部分がある場合は，図 3(d) のように，ユーザは前景または背景ボタンをクリックして，それぞれ画面上で前景・背景領域をマウスストロークによってマークする．これを繰り返すことにより図 3(e) のような正しいセグメンテーション結果が得られる．

### 2.2　自動セグメンテーション

自動セグメンテーションでは，すでにセグメンテーション済みの画像を参照画像とし，まだセグメンテーションされていない対象画像をセグメンテーションすることを繰り返すことによって，すべての画像がセグメンテーションされる．キョらの提案した選択的情報伝搬 [7] を新たに時系列方向へ拡張する．

#### 2.2.1　伝搬の順序

時系列方向に $t$ 軸，同一時刻の水平方向に $x$ 軸，鉛直方向に $y$ 軸をとり，多眼画像の視点および時刻の座標を $(x, y, t)$ で表す．提案するフレームワークでは複数の seed 画像を与えることができる．seed 画像の座標を $(x_1^s, y_1^s, t_1^s), \cdots, (x_n^s, y_n^s, t_n^s)$ とする．ただし，$t_1^s < t_2^s < \cdots < t_n^s$ である．$(x_i^s, y_i^s, t_i^s)$ の seed 画像から伝搬を始める場合を考える．この時の伝搬は，時刻ごとに $t_i^s, t_i^s - 1, t_i^s - 2, \cdots, t_{i-1}^s + 1$（下り），$t_i^s + 1, t_i^s + 2, \cdots, t_{i+1}^s - 1$（上り）のような順序にする．同一時刻内の画像に関しては，$D = (x - x_i^s)^2 + (y - y_i^s)^2$ の小さい順に選択する．$D$ が同じ画像については，$y$，$x$ の順に大きいものから選択していく．

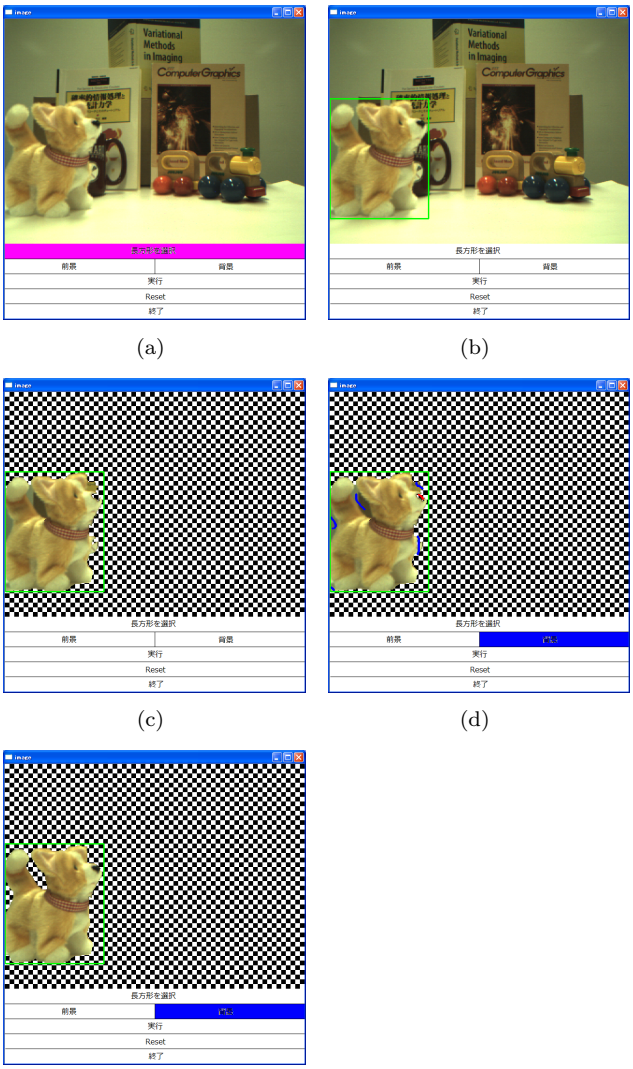



図 3　GrabCut を用いた手動セグメンテーション．(a) インタフェース画面，(b) バウンディングボックス，(c) GrabCut の最初の実行結果，(d) ストロークを用いた前景・背景の指定 (赤い線：前景，青い線：背景)，(e) 最終的な実行結果．

例えば図 4 において，赤色の画像を seed 画像とした場合，伝搬の順序は図に書かれている数字の順になる．

**2.2.2**　情報伝搬によるセグメンテーション

情報伝搬によるセグメンテーションでは，対象画像 $I_o$ についてセグメンテーションを行うために，すでにセグメンテーションされている参照画像 $I_c$ の情報を用いる．

$I_c$ と $I_o$ について特徴点を抽出して対応をとり，参照画像 $I_c$ で物体領域に含まれる特徴点についてのみ位置の差のベクトル（オプティカルフローベクトル [8]）を求める．オプティカルフローベクトルの平均をとったベクトルに従って，物体領域のマスクを参照画像 $I_c$ から対象画像 $I_o$ へ平行移動する．さらに Bai ら [6] の提案したローカル識別器の手法を用いて，マスクの境界線を小さいウィンドウに分けて，再びオプティカルフローベクトルを求めて移動する．このようにして対象画像 $I_o$ のマスクの初期値が得られる．このマスクの初期値の境界線からの距離を利用して，shape prior [5] を作る．

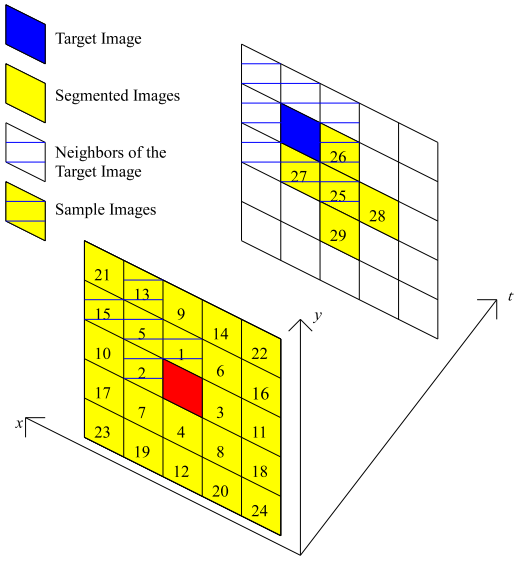



図 4　伝搬の順序と参照画像の候補の選び方

セグメンテーションの問題は，次の式で表されるエネルギーを最小化する問題として解くことができる．

$$E(\boldsymbol{A}) = \lambda \sum_{p \in P} R_p(A_p) + \sum_{(p,q) \in N} h_{pq} \delta(A_p, A_q) \qquad (1)$$

$A_p$ は対象画像 $I_o$ の画素 $p$ のラベル (前景：1, 背景：0)，$\boldsymbol{A}$ は画像全体のラベルの組である．第 1 項は参照画像 $I_c$ の色情報に基づいて，各画素が前景，背景である尤度を表す．第 2 項はラベルの変化を滑らかにする条件と shape prior を含む．$\lambda$ は正の重みであり，第 1 項と第 2 項のバランスを調整する働きをする．

$R_p(A_p)$ は以下の式で表される．

$$R_p(A_p) = -\log \theta(I(p), A_p) \qquad (2)$$

$I(p)$ は画素 $p$ の色の値，$\theta(I(p), A_p)$ は前景（$A_p = 1$）または背景（$A_p = 0$）の色分布における $I(p)$ の相対頻度である．ここで色分布は参照画像 $I_c$ の前景，背景から求められるので，色情報が伝搬されていることになる．

$h_{pq}$ は以下の式で表される平滑化項で，ラベルの変化を滑らかにする条件と shape prior を表す．

$$h_{pq} = (1-\mu)\frac{e^{-\kappa(I(p)-I(q))^2}}{\mathrm{dist}(p,q)} + \mu\left[1 - \exp\left(-d\left(\frac{p+q}{2}\right)^2 / \sigma_s^2\right)\right] \qquad (3)$$

$\mathrm{dist}(p,q)$ は画素 $p, q$ 間の距離，$d(\frac{p+q}{2})$ は推定された初期マスクの境界線と画素 $p, q$ の中央との距離である．$h_{pq}$ は $p$ と $q$ のラベルが異なる場合にのみ加えられる項で，第 1 項がラベルの反転に対して与えられるコスト，第 2

項が初期マスクの境界線からの距離が離れた位置でのラベルの反転に対して与えられるコストである．$\mu$ は各コストの重みである．

このエネルギー関数を，Graph Cut [9] を用いて最小化することによって，対象画像 $I_o$ のセグメンテーションが得られる．

**2.2.3** 選択的情報伝搬

対象画像 $(x^o, y^o, t^o)$ をセグメンテーションする場合，対象画像の近傍にある画像のうち，すでにセグメンテーションされているものを参照画像の候補として選ぶ．ここで対象画像の近傍にある画像とは

$$(x-x^o)^2+(y-y^o)^2+(t-t^o)^2 \leqq 2 \tag{4}$$

を満たす $(x, y, t)$ にある画像と定義する．例えば図 4 において，青色の画像をセグメンテーションする場合，斜線部分が (4) 式を満たす画像，黄色の部分がすでにセグメンテーションされた画像である．

参照画像の候補が複数存在する場合，次節で定義する信頼度を用いて最も良い参照画像を推定する．

複数の seed 画像を用いる場合には，以下のような手順で実行する．各 seed ごとに，seed 画像が追加された順に伝搬を行う．近傍の 2 つの seed 画像（$\text{seed}_i$ と $\text{seed}_j$）がそれぞれ座標 $(x_i^s, y_i^s, t_i^s)$ と $(x_j^s, y_j^s, t_j^s)$ にあるとする．ここで，$t_i^s < t_j^s$ であり，先に $\text{seed}_i$ から伝搬が行われると仮定する．この時，この seed 画像間については，はじめに $\text{seed}_i$ から $t_i^s, t_i^s+1, \cdots, t_j^s-1$, の順に伝搬を行ってセグメンテーションを進める．次に $\text{seed}_j$ から伝搬を行い，信頼度がより大きい場合には結果を更新する．$t_j^s, t_j^s-1, \cdots, t_i^s+1$, の順に伝搬を行うが，ある時刻 $t$ において，すべての $(x, y)$ について信頼度が大きいものに更新されなくなった場合，伝搬を止める．

**2.2.4** 信　頼　度

本手法では，複数の seed 画像からの伝搬においても信頼度を用いるため，正確な信頼度を得ることが重要になる．そこで本研究では信頼度の再検討を行う．キョら [7] が定義したように，参照画像のマスク $\text{mask}_c$ の信頼度が $R_c$ のとき，参照画像 $I_c$ から伝搬された対象画像 $I_o$ のマスク $\text{mask}_{c\to o}$ の信頼度 $R_{c\to o}$ を，

$$R_{c\to o} = u_{c\to o} \cdot R_c \tag{5}$$

とする．このうち信頼度が最大になる参照画像を選び

$$R_o = \max_c R_{c\to o} \tag{6}$$

とする．$u_{c\to o}$ は参照画像 $I_c$ のマスクを基準にした対象画像 $I_o$ のマスクの信頼度であり，本研究では 3 つの定義を用いる．1 つはキョらの提案した定義 [7] であり，2 つは今回新たに検討するものである．

(a) 従来手法（キョら [7]）

従来の信頼度の定義では，対象画像から参照画像を再びセグメンテーションして得られたマスク $\text{mask}_{c\to o\to c}$ と元の参照画像のマスク $\text{mask}_c$ との一致度を求める．

$$u_{c\to o} = 1 - \frac{\text{mask}_c, \text{mask}_{c\to o\to c}\text{で異なる画素数}}{\text{mask}_c\text{の物体領域の画素数}} \tag{7}$$

(b) 参照画像と対象画像のマスクの形の一致度

本研究で用いた情報伝搬の手法は，shape prior を用いている．このため，物体が変形する部分では信頼度を低くしたい．そこで，参照画像 $I_c$ のマスク $\text{mask}_c$ と参照画像 $I_c$ を基にした対象画像 $I_o$ のマスク $\text{mask}_{c\to o}$ の間の形の一致度を用いた信頼度を定義する．

参照画像 $I_c$ のマスクをオプティカルフローベクトルに従って平行移動したマスクを $\text{mask}'_c$ とする．$u_{c\to o}$ はマスク $\text{mask}'_c$ とセグメンテーションされた対象画像のマスク $\text{mask}_{c\to o}$ の一致度とする．

$$u_{c\to o} = 1 - \frac{\text{mask}'_c\text{と } \text{mask}_{c\to o}\text{で異なる画素数}}{\text{mask}'_c\text{の物体領域の画素数}} \tag{8}$$

(c) ヒストグラムの相関

参照画像と対象画像間で物体領域の色分布はほぼ同じと仮定すると，ヒストグラムの一致度を用いることが考えられる．

$$u_{c\to o} = \sum_y \sum_u \sum_v \text{hist}_c[y][u][v] \cdot \text{hist}_{c\to o}[y][u][v] \tag{9}$$

ここで，$\text{hist}_c[y][u][v]$，$\text{hist}_{c\to o}[y][u][v]$ は，それぞれ参照画像，対象画像の物体領域における YUV 画像の正規化したヒストグラムである．

ただし，カメラの位置による照明条件の違いなどにより，同じ領域でも色ヒストグラムが異なることがあるので，2 枚の画像間で RGB についてそれぞれ平均値の比を求め，平均値が一致するように画素値に定数倍を掛けて補正している．例えば，画像 $I_c$ の前景の R の値が $r_1^c, \cdots r_{N_c}^c$，画像 $I_o$ の前景の R の値が $r_1^o, \cdots r_{N_o}^o$ である時，

$$\bar{r}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} r_i^c \tag{10}$$

$$\bar{r}_o = \frac{1}{N_o}\sum_{i=1}^{N_o} r_i^o \tag{11}$$

$$\bar{r}_c \geqq \bar{r}_o\text{のとき}, r_i'^c = \frac{r_i^c \bar{r}_o}{\bar{r}_c}(i=1,\cdots,N_c) \tag{12}$$

$$\bar{r}_o > \bar{r}_c\text{のとき}, r_i'^o = \frac{r_i^o \bar{r}_c}{\bar{r}_o}(i=1,\cdots,N_o) \tag{13}$$

とする．$r_i'^c$ または $r_i'^o$ は補正後の値である．G,B についても同様である．

## 3. 実 験 結 果

ViewPLUS 社製 25 眼 PCI-Express カメラ「ProFUSION25」で撮影した多眼動画像（25 眼× 200 フレームの 5000 枚）を用いて実験を行った．電池駆動で歩く犬のおもちゃを使った撮影の様子を図 5，実験に用いた PC のスペックを表 1 に示す．ソフトウェアは Visual Studio C++ 2008 で OpenCV2.1 を使い作成した．

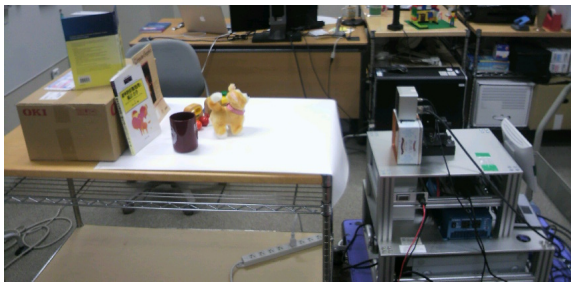


図 5　多眼カメラを用いた動画撮影

表 1　実 験 環 境

| | |
|---|---|
| OS | Microsoft Windows Vista Business Service Pack 2 32bit |
| CPU | Intel Core2 Extreme X9770 3.20GHz |
| Main memory | 4.00 GB RAM |

### 3.1　信頼度と精度の関係

新たに定義した信頼度と精度の関係を見るために実験を行った．信頼度については，セグメンテーションを実行すると自動で計算される．精度については，手動で与えた正確なマスクを ground truth とし，自動でセグメンテーションされたマスクと ground truth との一致度を，そのセグメンテーションの精度とする．一致度は次の式によって定義する．

$$1 - \frac{\text{2 つのマスク間でラベルが異なる画素数}}{\text{ground truth のマスクの画素数}} \qquad (14)$$

視点間と時系列方向は性質が異なるので，分けて実験を行った．

#### 3.1.1　視点間の伝搬

時刻 $t = 181$ の 25 眼画像のうち，中央の画像 1 枚についてについて手動でセグメンテーションを与えた．残りの 24 枚の画像について，選択的情報伝搬による自動セグメンテーションを実行した．

信頼度と精度のグラフを図 6 に示す．(a)–(c) はそれぞれ 2.2.4 の (a)–(c) の信頼度の定義に対応する．この実験で，決定係数 $R^2$ の値は色ヒストグラムの相関を用いた場合に最も高くなった．しかし，精度よりも信頼度がかなり小さな値になった．これは，同じ領域でもカメラの特性や照明条件などによってカメラ間で色ヒストグラムが異なるためであると考えられる．

#### 3.1.2　時系列間の伝搬

中央視点のカメラで撮影した画像から連続した時刻 $t = 181$–$192$ の 12 フレームを選択し，隣り合った画像の組を 11 組作った．この時刻では犬が歩いたり尻尾を振ったりするため，物体の形の変化が大きい．それぞれについて $t$ の小さい方に手動でマスクを与え，もう一方を情報伝搬により自動セグメンテーションした．

自動でセグメンテーションされた 11 枚について，信頼度と精度を図 7 に示す．この実験で，$R^2$ の値は色ヒストグラムの相関を用いた場合に最も高くなった．色ヒストグラムによる信頼度 (c) は，時系列間の物体の形の変化に有効であると考えられる．

### 3.2　対話的セグメンテーション

25 眼× 200 フレームの 5000 枚の画像すべてをセグメンテーションする実験を行った．ここで，信頼度は実験において良好な結果が得られたこと，実行時間がキョら [7] の信頼度 (a) の約 1/2 であることから，ヒストグラムを用いる手法 (c) を利用した．はじめに，手動で $t = 100$ の画像についてセグメンテーションを与え seed 画像とした．次に，$t$=118，134，151，160 の順に seed 画像を追加し，残りを自動でセグメンテーションした．この時，すべてのセグメンテーションを終えるまでにかかった時間は約 7 時間であった．$t$=0，40，80，120，160，199 について，セグメンテーションの結果を図 8 に示す．また，ground truth と比較したときの精度を図 9 に示す．自動でセグメンテーションされた画像についても良好な結果が得られている．$t = 0$ は直近の seed 画像から時刻が遠く離れているのでやや精度が低いが，追加で seed 画像を与えれば改善が可能である．

## 4.　ま　と　め

本論文では，多眼動画像セグメンテーションのためのフレームワークを提案した．膨大な数の画像を含む多眼動画像のうち，数枚の画像だけを手動でセグメンテーションすると，残りの画像は情報伝搬に基づいて自動でセグメンテーションされる仕組みを提案し，そのユーザインタフェースを実装した．そして 25 眼× 200 フレームの 5000 枚の画像について実験を行い，提案したフレームワークの有効性を確認した．

**謝辞**　本研究の一部は，NICT（独立行政法人情報通信研究機構）の高度通信・放送研究開発委託研究「革新的三次元映像技術による超臨場感コミュニケーション技術の研究開発」によるものです．

### 文　　献

[1] 高橋 桂太，苗村 健：“視点依存奥行きマップ実時間推定に基づく多眼画像からの自由視点画像合成”，映像情報メディア学会誌，Vol. 60, No. 10, pp. 1611 – 1622, 2006.

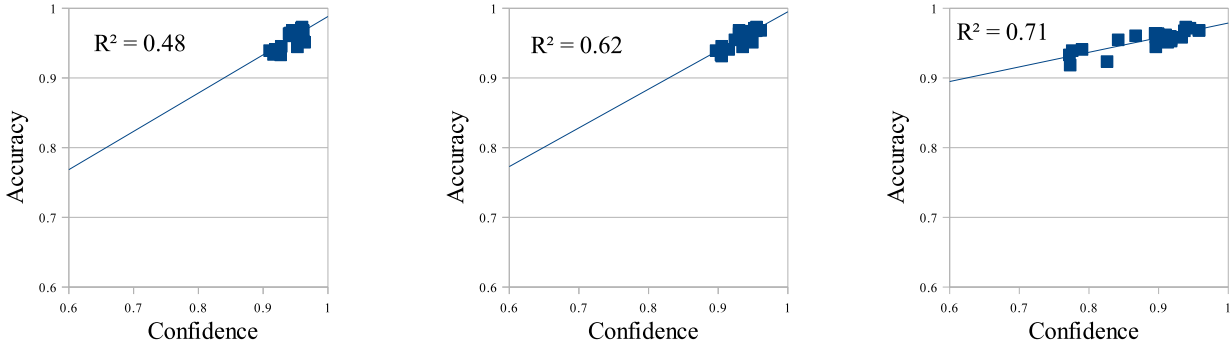
[2] 柏木 陽佑，中島 諒，ファンヴェトクォク，高橋 桂太，苗村健：“半透明マスク付き多眼画像を用いた自由視点映像合成”，3 次元画像コンファレンス，2011.

[3] Yuri Boykov, Marie-Pierre Jolly: “Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images,” IEEE ICCV, vol. I, pp. 105–112, 2001.

[4] Catsten Rother, Vladimir Kolmogorov, Andrew Blake: “ GrabCut-Interactive Foreground Extraction using Iterated GraphCuts,” ACM SIGGRAPH, Vol. 23, pp. 309–314, 2004.

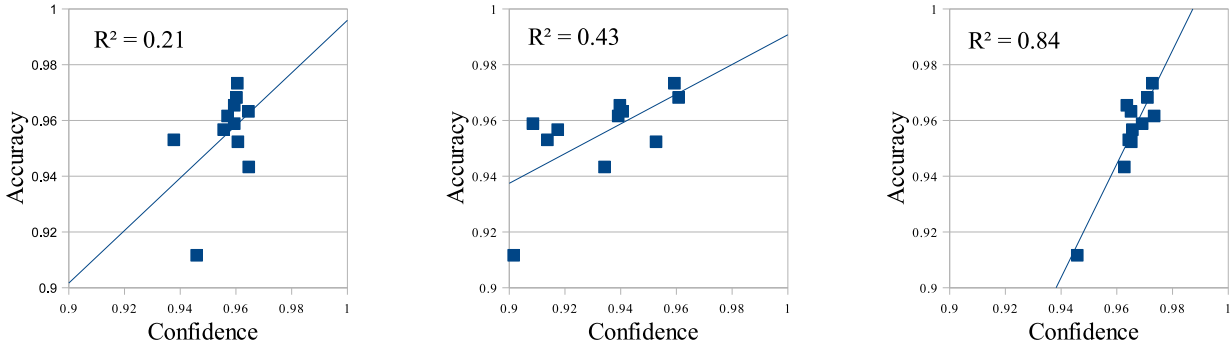
[5] Daniel Freedman, Tao Zhang: “Interactive Graph Cut Based Segmentation With Shape Priors,” IEEE CVPR, Vol. 1, pp. 755–762, 2005.

[6] Xue Bai, Jue Wang, David Simons, Guillermo Sapiro:




(a) (b) (c)

図 6　視点間の信頼度と精度の比較実験




(a) (b) (c)

図 7　時系列間の信頼度と精度の比較実験

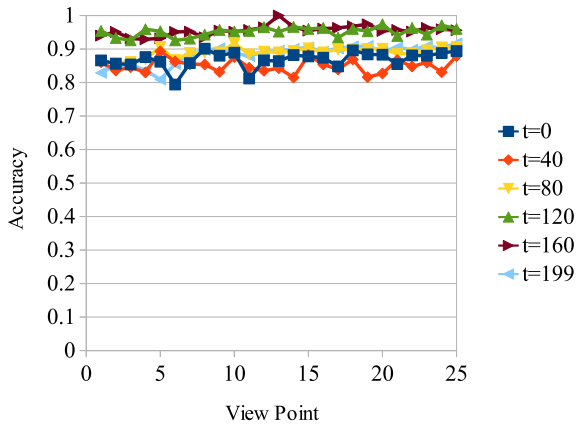



図 9　セグメンテーションの精度

"Video SnapCut: Robust Video Object Cutout Using Localized Classifiers," ACM SIGGRAPH, Vol. 28, Issue 3, 2009.

[7] キョ タオ，中島 諒，ファン ヴェトクォク，高橋 桂太，苗村 健：“画像間の選択的情報伝搬に基づく多眼画像セグメンテーション”，3 次元画像コンファレンス，pp. 67–70, 2010.

[8] Bruce D. Lucas, Takeo Kanade: "An Iterative Image Registration Technique with an Application to Stereo Vision," Proceedings of Imaging Understanding Workshop, pp. 121–130, 1981.

[9] Yuri Boykov, Vladimir Kolmogorov: "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," IEEE TPAMI, Vol. 26, no.9, pp. 1124–1137, 2004.

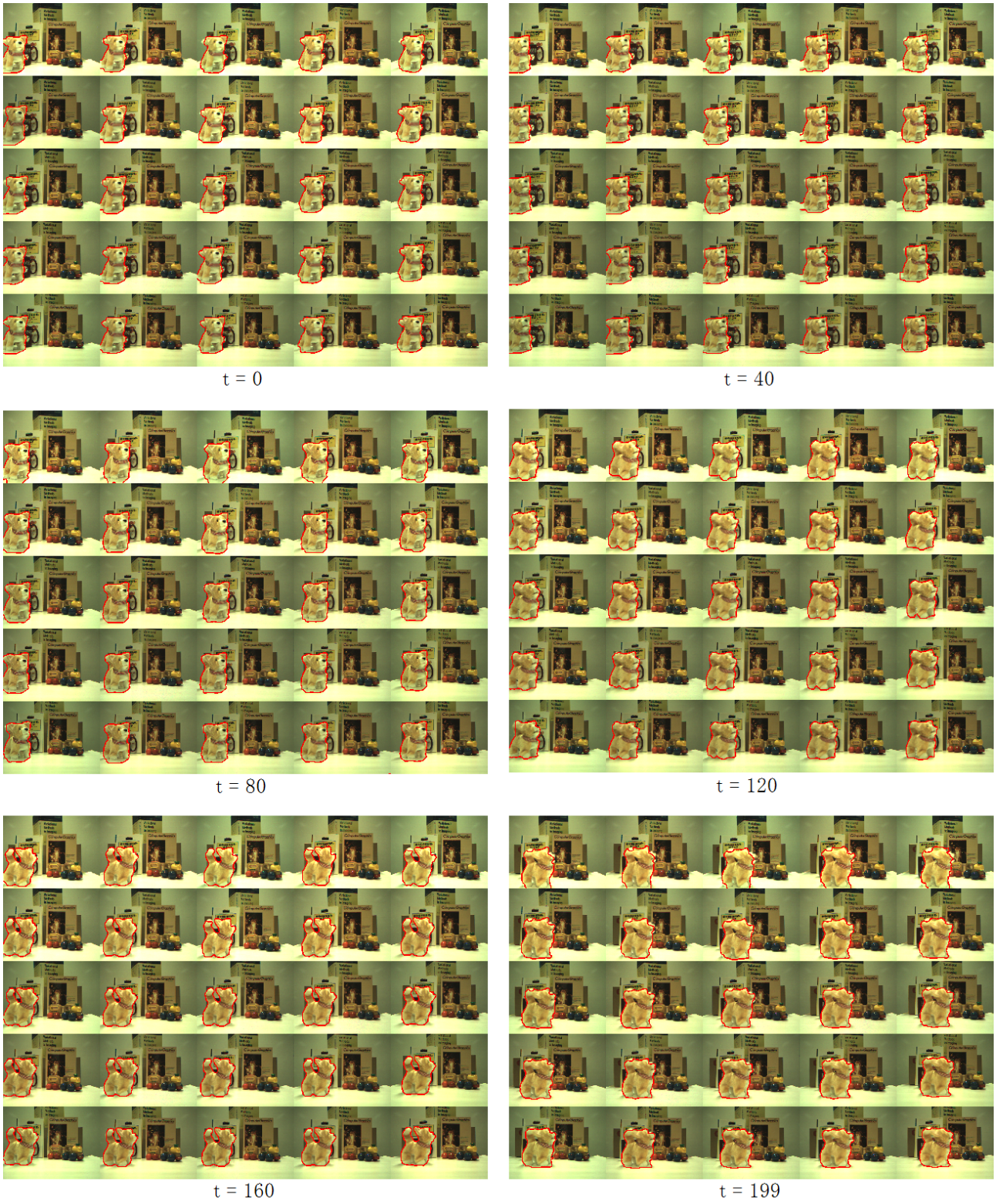



図 8　セグメンテーションの結果