情報処理学会論文誌　コンピュータビジョンとイメージメディア　Vol. 1　No. 2　32–40 (July 2008)

音と映像の相関を用いた画像分割による話者領域の切り出し

劉　　玉宇[†1,†2]　佐藤洋一[†1]

本論文では音と映像の相関を用いて映像中の話者領域を自動で切り出す手法を提案する．これまでにも音と映像の相関を手がかりとした音源位置推定手法がいくつか提案されていたが，各画素ごとに独立な処理に基づいていたため，断片化された領域しか得られないという共通の問題が存在した．これに対し，本研究ではグラフカット最適化による画像分割処理に音と映像の相関分析を組み入れるという新たな枠組みを用いることにより，領域の断片化を抑制しつつ複雑背景から話者領域を切り出すことを実現する．複雑かつ動きをともなう背景中で話している人物の映像を用いた実験により提案手法の有効性を示した．

Speaker Segmentation Using Audiovisual Correlation

Yuyu Liu[†1,†2] and Yoichi Sato[†1]

Audiovisual correlation has been used successfully for audio source localization. However, the previously proposed techniques were mainly based on local processing and, as a result, suffered from the common problem of estimated sound sources being highly fragmented. In this work, we propose a novel technique based on audiovisual correlation analysis for segmenting moving speakers appearing in complex backgrounds. The main idea of our approach is to use audiovisual correlation analysis in the context of image segmentation, so that moving speakers in complex backgrounds can be segmented out with very little or no fragmentation. First, we introduced a spatiotemporally local measure for audiovisual correlation, whose locality is the key to realize our idea. Then, we forced soft constraints in both temporal and spatial domains to incorporate visual information like boundary, region, and intra-frame motion. Finally, we used graph cut-based optimization to obtain a final segmentation. Experiments using video sequences of moving speakers in cluttered non-stationary backgrounds demonstrate the effectiveness of our technique.

1. はじめに

ビデオ映像に含まれる音源位置を推定する技術はテレビ会議や監視システムなどさまざまなアプリケーションへの応用へが期待され，これまでにもさまざまなアプローチに基づく音源位置推定手法が提案されてきた．その中でも，音と映像の相関分析を用いたアプローチは，マイクロホンアレイなどの特殊な装置を利用することなく一般的なビデオ映像の中から音源の位置を推定することができるという点で近年特に注目されている．

このアプローチに基づく音源位置推定に関する研究としては，心理学分野における研究[7]から始まり，Hershey ら[8]によって 1 つの音源から発生する音と映像の輝度変化の同期性を用いた音源位置推定手法が提案されたのに続き，これまでにもいくつかの手法が提案されてきた．Smaragdis ら[16]は音と映像の両方の信号を要素とするベクトルを考え，まず主成分分析でベクトルの次元数を落とし，そして独立成分分析によりベクトル空間中の独立成分を見つけることによって音源を検出した．Darrell ら[6]は，音と映像の相互情報量最大化に基づく手法を提案している．具体的には，音と映像信号を別々の低次元空間内のベクトルに射影する際に，その 2 つのベクトル間の相互情報量を最大化する射影ベクトルを求め，射影ベクトルの高いところを音源位置として抽出した．Kidron ら[10]は *Canonical Correlation Analysis*（CCA）を用いた手法を提案している．通常 CCA による解析では大量のデータが必要となるが，Kidron らはノイズを含む少量のデータでは CCA により得られる相関が必ず最大となることに着目し，これを制約条件とした．そして，CCA 変換ベクトルの $L1$ ノルムを最小化することにより効率的に音源位置を推定することを実現した．さらに，Monaci ら[14]は音と映像の相関を考える場合に輝度変化よりも物体の動きの変化の方が音情報とより強い相関を示すと考え，*Matching Pursuit*（MP）と呼ばれる方法により局所画像特徴の移動と回転を計算し，この動き情報と音情報との相関に基づいて音源位置を推定した．

これらの既存の手法では，すべて各画素もしくは各局所画像特徴ごとの独立な処理に基づくため，断片化された音源位置の推定結果しか得ることができないという共通の問題が存在した．断片化された音源位置推定結果でも十分な場合も考えられるが，たとえばテレビ会議などのアプリケーションで映像中の話者の位置を知りたい場合，断片化された形ではなく話

†1 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo
†2 ソニー株式会社情報技術研究所
Information Technologies Laboratories, Sony Corporation

をしている人物に対応する部分を 1 つの領域として得られることが望ましい．

これに対し，本研究では画像分割と音と映像の相関分析とを統合することにより，領域の断片化を抑制しつつ複雑な背景からでも精度良く音源，特に話者の領域を抽出可能な手法を提案する．音と映像の相関分析による音源推定に関する研究が報告されたのが比較的最近であるのに対し，画像の領域分割は非常に長い歴史を持ち，これまでにも数多くの手法が提案されてきた．(たとえば文献 2)，9)，18) など)．この中でも，近年 Boykov ら[2] によって提案されたグラフカット最適化に基づく画像分割手法はその性能の高さから注目されている．この手法では，画像分割の問題を各画素をノードとしたグラフのラベリング問題としてとらえることにより，データ項と平滑化項と呼ばれる 2 つの項の和として定義されるエネルギー関数をグラフカットにより最小化することによって入力画像が前景と背景とに分割される．ここでデータ項は各画素が前景と背景のいずれかに属するとした場合の尤度に相当し，ユーザにより手動で指定された前景と背景のサンプルから得られるモデル（輝度分布など）により計算される．また，平滑化項は隣接画素間の輝度差や距離などにより定義され，隣接画素間におけるラベルの変化，すなわち分割された領域の断片化を抑制するのに寄与する項となっている．

良いセグメンテーションの性能を得る一方，ユーザの手動指定が必要なことが Boykov ら[2] の手法の応用範囲をかなり限定してしまう．それをなくすため，いくつの手法が提案された．たとえば，Kolmogorov ら[11] はステレオマッチングで得た距離情報を用いて前景を分割した．Yu ら[17] は検出された顔およびその下の領域の色分布を学習し，ビデオの中で存在する人物を分割した．

本研究では，グラフカットにおけるエネルギー関数のデータ項を音と映像の相関に基づいて定義することにより，Boykov らの手法のように手作業で前景と背景のモデルを与えることを不要とし，通常のグラフカットによる画像分割ではうまく働かないような複雑背景を含む映像であっても話者領域をより安定に抽出することを可能としている．また，エネルギー関数の平滑化項に関しても，隣接画素として同一フレーム内の近傍に加え時間的に隣接するフレーム内の近傍も考慮することにより，連続する画像フレーム間で話者の画像領域が大きく変動することを効果的に抑制している．

以下，本論文の構成は次のとおりである．まず 2 章において提案手法におけるグラフカットによる画像分割について述べたのち，3 章で音と映像の間の相関を表す指標について説明する．4 章で提案手法を用いた実験の結果について報告したのち，5 章で結論と今後の課題について述べる．

2. グラフカットによる画像分割

本章では提案手法におけるグラフカットによる画像分割の概要について説明する．

グラフカットによる画像分割では式 (1) のようにデータ項 $D_p(f_p)$ と平滑化項 S_{pq} を合計したコスト関数 E を最小化することにより全画素への最適なラベル付けが計算される．提案手法では，映像中の連続フレームをひとまとめとして考え，グループ化されたフレーム中のすべての画素 p に対して前景もしくは背景のラベル f_p を割り当てる．ここで前景と背景のラベルはそれぞれで $f_p = 1$ と $f_p = 0$ とする．

$$E(f) = \lambda \cdot \sum_{p} D_p(f_p) + \sum_{\{p,q\} \in neighbor} S_{pq}(f_p, f_q) \tag{1}$$

このコスト関数において，各画素が音源であるかどうかを評価するデータ項 $D_p(f_p)$ は音と映像の相関度 $AVC(p)$ に基づき式 (2) のように定義される．なお，この相関度 $AVC(p)$ の詳細は 3 章で説明するが，この相関度は相関の高さに応じて 0 から 1 までの値をとるように定義される．

$$D_p(f_p) = \begin{cases} AVC(p) & f_p = 0 \\ 1 - AVC(p) & f_p = 1 \end{cases} \tag{2}$$

これにより，画素が p が音源でないとする場合（$f_p = 0$）は，その画素における音と映像の相関度 $AVC(p)$ が大きいほどコストが高くなり，逆に画素が p が音源であるとする場合（$f_p = 1$）は相関が小さいほどコストが高くなる．すなわち，画素 p において音と映像の相関が高ければ高いほど，その画素が前景として抽出される可能性が高くなる．

得られる領域の断片化を抑制するための平滑化項 S_{pq} では**図 1** に示すように時空間的に隣接する 2 画素の組 pq を考慮する．すなわち，フレーム t における画素 p に注目する場合，同一フレーム t および前後フレームの 8 連結近傍に位置する 26 画素を隣接画素とする．

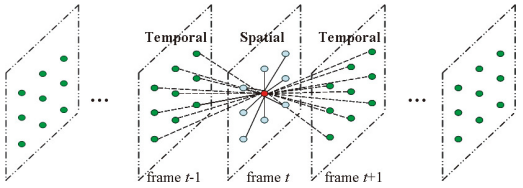


図 1　時空間隣接関係
Fig. 1　A demonstration of the temporal and spatial neighbors.

このような隣接画素の組 pq に対して，平滑化項 S_{pq} を式 (3) のように定義する．

$$S_{pq}(f_p, f_q) = e^{-\frac{(I_p - I_q)^2}{2\sigma^2}} \cdot \frac{1}{dist(p,q)} \cdot T[f_p \neq f_q] \tag{3}$$

ここで I_p と I_q は隣接画素 p と q の輝度，σ は推定された画像のノイズの分散であり，本論文では経験的に決めた値 3 を用いた．$dist(p,q)$ は p と q の三次元グリッドにおけるユークリッド距離であり，本論文で扱う隣接関係では 1 と $\sqrt{2}$ と $\sqrt{3}$ のいずれかとなる．$T[\cdot]$ は 0 または 1 を返すブール関数であり，平滑化項は画素 p と q のラベル f_p と f_q が異なる場合にコストが増加する．

隣接画素 p と q の間にエッジが存在する場合には画素値 I_p と I_q の差が大きくなり，これら 2 つの画素に異なるラベルが割り当てられても平滑化項のコストが増大しない．逆に，隣接画素の輝度値が近い場合に異なるラベルが割り当てるとコストが増大する．このように，この平滑化項は画像のエッジを反映した分割とする効果がある．最後に，λ はデータ項と平滑化項のバランスを調整する係数であり，本研究における実験ではすべて $\lambda = 0.1$ に設定した．

このようにして定義されたエネルギー関数 E[12] をもとに，ひとまとめにした連続フレーム中の全画素をノードとするグラフを構築したのち，Max-Flow アルゴリズムにより E を最小化するようにグラフを分割[3] することにより連続フレーム中の全画素を前景と背景とに分割する[2]．

3. 音と映像の相関度 $AVC(p)$

本章では，本研究で用いる音特徴と映像特徴について述べたのち，各画素 p について音と映像の相関の高さを表す相関度 $AVC(p)$ の定義について説明する．

3.1 音　特　徴

本研究では単一の音入力の大きさに基づいて音特徴を計算している．そのため，ステレオ音声付きの映像など複数の音入力がある場合には，いずれかの入力のみを用いる，あるいか全入力の平均を用いるなどして単一の音入力とする．

音入力と映像入力を考えた場合，一般に音入力のサンプリングレートが映像入力のそれよりも大幅に高い．そのため，音と映像の相関を考えるために，次に述べるように音のサンプルを複数まとめて 1 フレームとした．その様子を**図 2** に示す．ここで，フレーム長さ T_a は映像の 1 フレーム分の長さ（30 ms など）であり，フレーム t の範囲は映像の t フレーム目に相当する時間枠を示している．

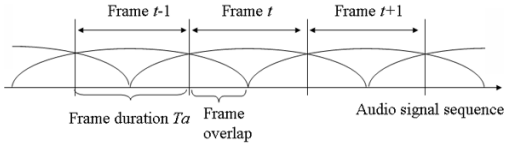


図 2　音信号のフレーム化
Fig. 2　The frame division of the audio signals.

フレーム間の連続性と保ちつつ，映像の t フレーム目に相当する音の大きさを計算するために半フレーム分（$T_a/2$）の重複を持たせたハミング窓による平滑化を用いた．実際には次式のような標準的なハミング窓を利用している．

$$h(n) = 0.54 - 0.46\cos(\frac{\pi n}{N-1}) \tag{4}$$

ここで $h(n)$（$n = 0 \ldots N-1$）は n 番目のサンプルに対する係数を表し，N は 1 フレーム中に含まれる音入力のサンプル数である．たとえば，ビデオ映像が 10 fps で撮影され，音信号が 8 kHz で記録されているとすると，フレームの長さと重複はそれぞれ $1000/10 = 100$ ms と $100/2 = 50$ ms となり，各フレームのサンプル数 N は $(100 + 50 \cdot 2) \cdot 8 = 1600$ となる．各フレームについて，音入力 $s_a(n)$ にハミング窓係数 $h(n)$ の掛け合わせた値の 2 乗平均により得られる音エネルギーの対数をとったものを音特徴とする．

$$\log\left[\frac{1}{N}\sum_{n=0}^{N-1}\{s_a(n)h(n)\}^2\right] \tag{5}$$

3.2 映 像 特 徴

Monaci ら[14] が示したように音と映像の相関を考える場合，画像の輝度変化よりも対象物体の動きの方が音情報とより強い相関を示す．そこで，本研究では連続する 2 フレームの画像から計算されるオプティカルフローをもとに映像特徴を定義することにした．入力映像としてはグレースケール画像もしくはカラー画像の輝度成分を考え，Lucas-Kanade 法[13] によりオプティカルフローを計算する．本研究の実験ではウィンドウサイズは 7×7 とした．さらに，テクスチャの少ない領域では安定にオプティカルフローを求めることが難しいため，ウィンドウ中の画素濃淡値の分散があらかじめ設定された閾値よりも小さければ，この画素におけるオプティカルフローは安定でないと判断し，フローの両成分ともに 0 とした．

さらに，本研究では映像中における話者領域の切り出しを目的とするため，音源の主な動きとしては話者の発話にともなう動作，たとえば口や目の動きや頷きなどの動きなどが想定

される．そこで，ここではオプティカルフローの鉛直方向成分を映像特徴として利用することとした．

3.3　音と映像の相関度

先行研究 8)，14) で示されたように，音源の音特徴と映像特徴との間において時間的な同期性が顕著に現れる傾向が強い．**図 3** は本研究で用いる音特徴と映像特徴の時間的変化の一例であるが，ここでも音特徴と映像特徴との間の同期性が明らかに見てとれる．音特徴と音源（ここでは話者の口近くの画素）から抽出された映像特徴とは増減のタイミングがよく一致しているのに対し，音源でない部分（ここでは背景の動物体）における映像特徴ではこのような同期性がほとんど見受けられない．

ここで，音源からの映像特徴であっても音特徴と増減自体傾向は一致しないということに注意しなければならない．たとえば，口周辺部における動きを考えた場合，口を開きながら発話した際に上唇と下唇で鉛直方向の動きの反対となる．そのため，音特徴と映像特徴との同期性を考慮する場合には，オプティカルフローの鉛直方向成分の増減と音の相関ではなく，その相関の絶対値を用いる方が望ましい．同様のことが目の動きや頷きによる顔の動きに関してもいえる．このような理由により，フレーム t 内の画素 p における音特徴と映像特徴の間の相関度 $AVC(p)$ を式 (6) に示すように前後 5 フレームを用いて計算された正規化相関の絶対値とした．なお，映像の無音部分は音と映像の相関を求めるのに有効でないため，この相関度の計算に先立ち入力映像において音特徴の値が一定の閾値よりも小さいフレームをすべて除外した．

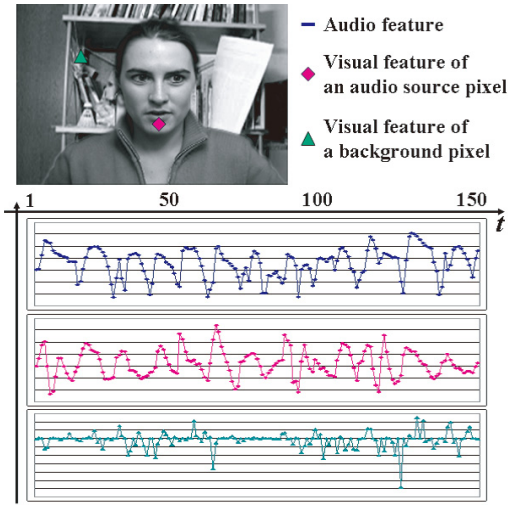


図 3　音特徴と映像特徴の時間変化の例：グラフ上段（青），中段（ピンク），下段（緑）はそれぞれ音特徴，音源付近の画素における映像特徴，背景中の動物体付近の画素における映像特徴の時間変化を示す

Fig. 3　Examples of temporal changes of the audio and visual features: The top curve in blue shows the temporal variation of the audio feature. The middle curve in pink corresponds to the temporal variation of the visual feature from an audio source pixel, while the bottom one in green shows that from a pixel on a moving background object.

$$
\begin{aligned}
&AVC(p) \\
&= \left| \sum_{i=-2}^{2} (A_{t+i} - \overline{A}_t)(V_{t+i}(p) - \overline{V}_t(p)) \right| \\
&\Big/ \left(\sqrt{\sum_{i=-2}^{2} (A_{t+i} - \overline{A}_t)^2} \cdot \sqrt{\sum_{i=-2}^{2} (V_{t+i}(p) - \overline{V}_t(p))^2} \right) \qquad (6)
\end{aligned}
$$

ここで A_t と $V_t(p)$ はそれぞれフレーム t における音特徴と画素 p での映像特徴であり，$\overline{A}_t$ と $\overline{V}_t(p)$ はその 5 フレームの平均である．なお，$AVC(p)$ はグラフカットにおけるデータ項として利用されることから，フレーム t における画素 p の時空間的近傍のデータのみから計算される必要がある．そこため，ここでは十分短い時間として $AVC(p)$ の計算に 5 フレームを用いることとした．

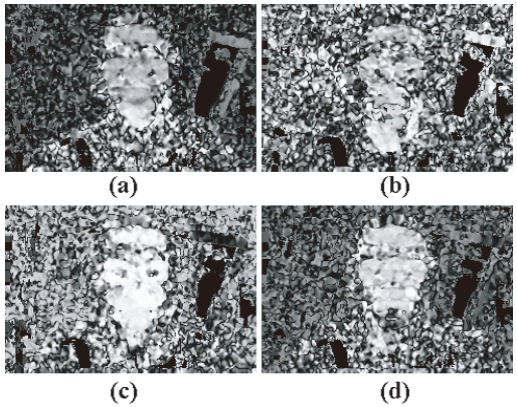


図 4　4 つのフレームの AVC 相関値：相関値高い輝度は高い相関値に対応する

Fig. 4　AVC value of four frames: Pixels with lighter intensities correspond to higher AVC values.

図 3 のビデオに対して計算された相関度の例を**図 4** に示す．フレームによっては話者の口周辺の画素でも相関度が低くなることがあるが，全体としては話者の顔部分で背景よりも高い値が得られていることが分かる．

4. 評価実験

CUAVE データベース[15] を用いて提案手法による話者領域切り出しに関する実験を行った．利用した映像は 1 人の人物が緑色の背景の前で英語の数字を話す様子を記録したものとなっており，映像のフレームレートは 29.97 fps，音声は 44 kHz のステレオ音声となっている．本論文で報告する実験ではすべてステレオ音声の左チャネルのみを利用した．**図 5** (a) に映像の一例を示す．実験では元の画像サイズの 720×480 画素を 240×160 画素に縮小したものを利用した．

複雑な背景からの話者切り出しを実験するために，CUAVE データベースの映像の人物部分をクロマキーの要領で背景色（ここでは緑）をもとに切り出し，別途デジタルカメラ（SONY DSC-F717）撮影された背景映像に重ねこんだ映像を準備した．図 5 (b)〜(d) にこのようにして準備された実験用映像を示す．図 5 (b) では人物の左右両方に位置する物体が揺れている様子が記録されている．図 5 (c) の背景にはカメラをパンさせながら屋外を撮影した様子が記録されており，図 5 (d) は背景の動きが鉛直方向である場合をテストするために背景が図 5 (c) の背景映像を 90 度回転させたものとなっている．

まず最初に，提案手法において何フレーム分をまとめて処理するのがよいかを調べるために，1 グループとしてまとめるフレーム数の変化が話者領域の切り出しに与える影響について調べた．**図 6** にそれぞれ 10 フレーム，20 フレーム，40 フレームをまとめて処理した結果を示す．10 フレームをまとめて処理した場合よりも 20 フレームをまとめて処理した場合の方が良い結果が得られていることが分かる．一方，20 フレームと 40 フレームでは大きな違いが見受けられなかった．他の映像を利用した実験においても同様の傾向が見受けられた．まとめて処理をするフレーム数が多くなるにつれ分割処理に要するメモリ量と計算時間が大幅に増加してしまうことからも，ある程度十分なフレーム数をまとめて処理すれば十分であると考え，本論文における実験ではまとめて処理をするフレーム数を 40 とした．なお，40 フレームをまとめて処理するのに Intel Core2Duo 1.83 G/1 G RAM の PC を用いて 14 秒程度の時間がかかった．

次に，無音部分を除いた 40 フレームを 1 グループとし，図 5 の映像に提案手法を適用した結果を**図 7** に示す．提案手法によって，さまざまな複雑な背景を含む映像であっても断

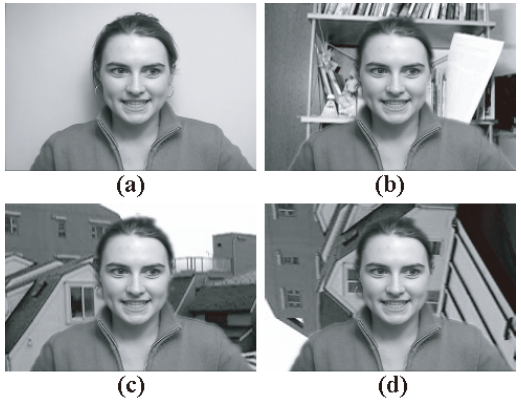


図 5　実験に用いた映像：(a) は CUAVE からのオリジナルの映像．(b)〜(d) は複雑背景に人物を重ねこんだ映像
Fig. 5　Videos of experimental data: (a) is the original data of CUAVE. (b), (c) and (d) are combined with our taken data.

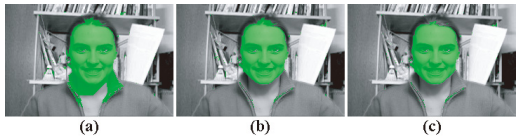


図 6　グループ化されたフレームの数とセグメンテーションの結果：(a) は 10 フレームでまとめた場合の結果である．(b) と (c) は 20 と 40 フレームの結果である
Fig. 6　The segmentation results for different numbers of frames grouped together for segmentation: (a) is the segmentation result using ten video frames. (b) and (c) are the results of twenty and forty frames, respectively.

片化されることなしに話者領域を抽出できていることが見てとれる．比較のために，同じ 40 フレームの時空間画像に対して Boykov らの手法[2] を適用した結果を**図 8** に示す[⋆1]．ここでは，あらかじめ手作業で準備した前景と背景のマスク（図 8）をもとに全フレームから前景と背景の画素値の分布を学習した．Boykov らの手法はユーザによるインタラクティブな操作を前提に提案されており，マスクを手動で修正することによりセグメンテーション結

⋆1 図 7 (d) は (c) の背景を 90 度回転したものであるため，図 8 では図 7 (d) の映像に対する切り出し結果は省略した．

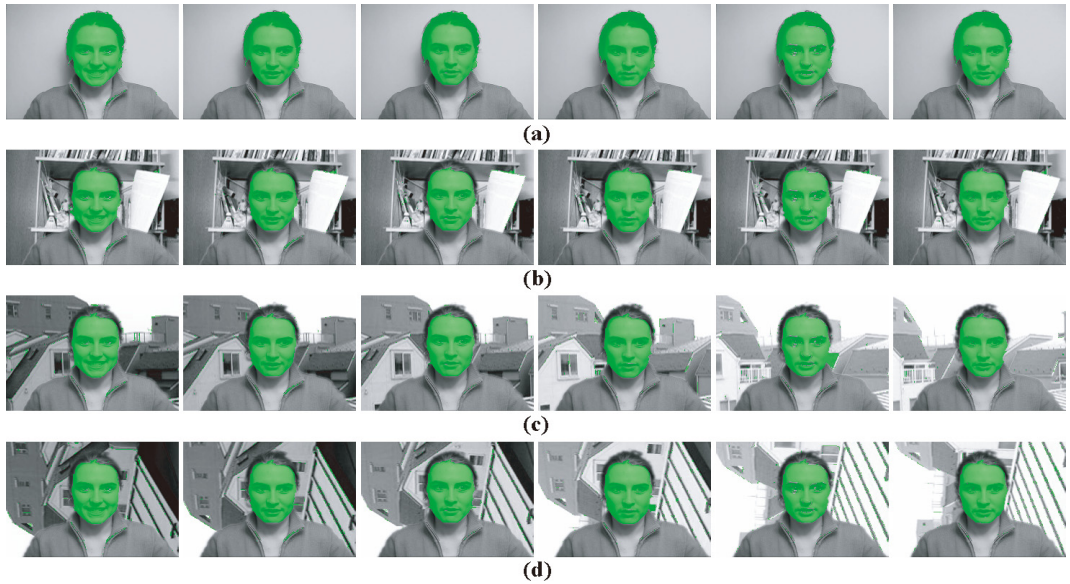


図 7　提案手法による話者領域切り出しの結果：40 フレームをまとめて処理した結果となっており，その中の 6 フレームを示す
Fig. 7　Segmentation results of the four video sequences. Forty frames are grouped to process. Six frames of their results are displayed.

果を改善することが可能となる．そのため，両者の結果を単純には比較できないものの，通常のグラフカットによる画像分割ではうまく抽出できないような複雑背景中の人物領域であっても，音と映像の相関を考慮することにより人物領域をより正確に切り出すことができていることが分かる．

提案手法による話者領域検出の精度を定量的に評価するために，図 7 の実験で用いた映像 40 フレームから 4 フレームを選び，人手により話者領域を選び正解値とした（**図 9**）．なお，どこまでを話者領域とするかを客観的に決めることは難しいが，ここでは顔のうち髪を含まない部分を話者領域としている．この正解値を用いて図 7 (a)〜(c) の切り出し結果の検出率（detection rate）と誤検出率（false positive rate）を計算した結果を**表 1** に示す．この結果から，背景によらず高い検出率が得られていることが分かる．また，背景が単純な (a) で誤検出率が高くなったのは髪の部分が話者領域に含まれてしまったことが主な要因となっている．

次に，提案手法におけるパラメータの選択が切り出し結果に与える影響について調べるため，式 (1) 中の λ と式 (3) 中の σ を変えながら図 7 (a) 中の話者領域の切り出しを行った．その際の検出率と誤検出率を**表 2** に示す．この結果から，パラメータの値を大きく変えても切り出しの精度は大きく変化しないことが分かる．

また，**図 10** に人物の位置が移動する映像に対して提案手法を適用した結果を示す．提案手法では対象人物の動きを考慮していないものの，フレーム間で画像上に話者の中心の移動量が数画素に限られる場合には話者領域をうまく抽出できていることが分かる．

CUAVE データベースに含まれる別の映像に対し，同じパラメータを用いて提案手法を適用した結果を**図 11** と**図 12** に示す．図 11 から別の人物の映像であってもパラメータを調整することなしに話者領域がうまく抽出できていることが分かる．また，図 12 のように，映像中に複数の人物が含まれている場合においても，話者以外の領域が一部誤検出されてしまうこともあるものの，提案手法を用いることにより話者領域をおおむね良好に切り出せることが確認された．

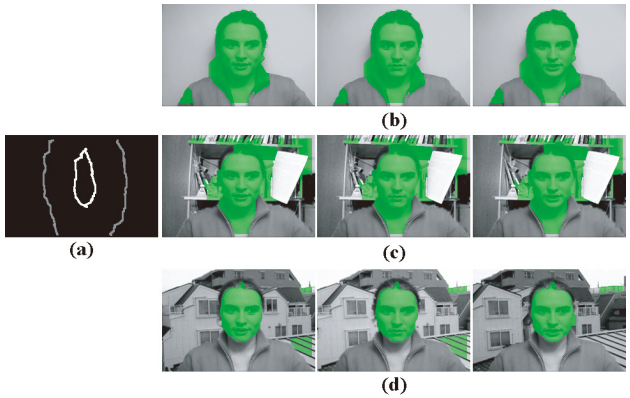


図 8　手法 2) による切り出しの結果：(a) は手で与えた前景と背景のマスク（白：前景，グレー：背景）を示す．(b)〜(d) はそれぞれ図 7 の (a)〜(c) の切り出し結果の例（40 フレーム中の 3 フレーム）を示す

Fig. 8　Experimental results by the method of 2): (a) is the manually labeled mask. White pixels belong to the foreground, and gray pixels correspond to the background. (b), (c) and (d) are the segmentation results of video (a), (b) and (c) in Fig. 7, respectively.

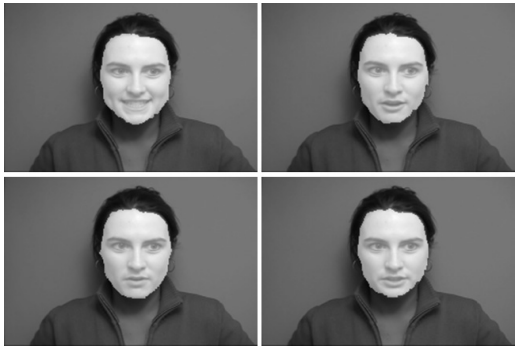


図 9　手動で与えた話者領域の正解値：図 7 で用いた 40 フレームから選んだ 4 フレームに対し，話者領域を白で示している

Fig. 9　Manually labeled ground truth: White pixels correspond to the speaker regions manually assigned to four frames selected from 40 frames used in the experiments in Fig. 7.

表 1　図 7 の切り出し結果における検出率（detection rate）と誤検出率（false positive rate）

Table 1　Detection rates and false positive rates of segmentation in Fig. 7.

	Detection rate (%)	False positive (%)
Video (a)	98.9	5.7
Video (b)	98.5	1.3
Video (c)	98.1	1.6

表 2　さまざまなパラメータの値で図 7 (a) 中の話者を切り出した結果：DR は検出率，FP は誤検出率を示す

Table 2　Detection rates (DR) and false positive rates (FP) of the segmentation of Fig. 7 (a) with different parameter values.

DR(%) / FP(%)	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.05$
$\sigma = 1$	90.8 / 5.5	91.7 / 4.7	91.8 / 4.6
$\sigma = 3$	98.7 / 5.3	98.9 / 5.7	98.9 / 5.7
$\sigma = 5$	98.8 / 5.5	98.8 / 5.5	98.7 / 5.7

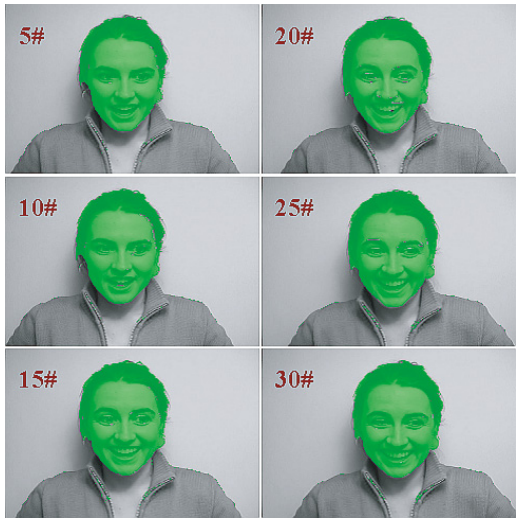


図 10　動いている話者に対する適用結果

Fig. 10　Segmentation result of a non-stationary speaker.

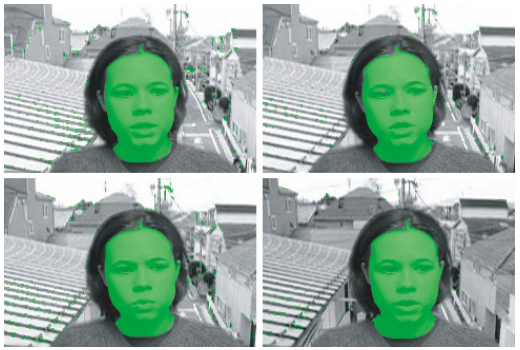


図 11　別の人物の映像へ適用した結果
Fig. 11　Segmentation result of another speaker.

図 12　複数人物が映っている映像に適用した結果
Fig. 12　Results of segmentation applied to video clips of multiple people.

5. おわりに

本論文では，音と映像の相関を用いることにより複雑背景から話者領域を自動的に切り出す手法を提案した．音と映像の統合による音源推定に関する従来手法では推定される音源が断片化されてしまうという問題があったのに対し，提案手法では，連続フレームをまとめた時空間画像に対して，時空間的な近傍で計算される音と映像の相関に基づきグラフカット最適化による分割処理を行うことにより，領域の断片化を抑制しつつ複雑背景から話者領域を切り出すことを実現した．今後は，今回利用したものとは別の音特徴と映像特徴の利用を検討していく．具体的には，音に関してはエネルギーだけでなく周波数特性を考慮した特徴量への拡張を考えている．また，提案手法では映像特徴としてオプティカルフローの鉛直成分を経験的に用いたが，一定方向に限定せずにオプティカルフロー自体の利用を検討していく．さらに，カラー画像への拡張や音と映像の関係として相関以外の統計量への拡張も考えたい．

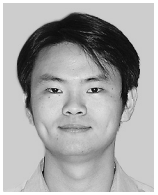
参 考 文 献

1) Boykov, Y. and Funka-Lea, G.: Graph Cuts and Efficient N-D Image Segmentation, *Int'l J. of Computer Vision*, Vol.70, No.2, pp.109–131 (2006).
2) Boykov, Y. and Jolly, M.P.: Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images, *Proc. Int'l Conf. on Computer Vision* (*ICCV2001*), Vol.1, pp.105–112 (2001).
3) Boykov, Y. and Kolmogorov, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.26, No.9, pp.1124–1137 (2004).
4) Boykov, Y., Veksler, O. and Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.23, No.11, pp.1222–1239 (2001).
5) Casanovas, A.L.: Blind audiovisual source separation using sparse redundant representations, Master thesis, Signal Processing Institute, EPFL (2006).
6) Darrell, T. and Fisher III, J.W.: Speaker association with signal-level audiovisual fusion, *IEEE Trans. Multimedia*, Vol.6, No.3, pp.406–413 (2004).
7) Driver, J.: Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading, *Nature*, Vol.381, pp.66–68 (1996).
8) Hershey, J. and Movellan, J.R.: Audio vision: Using audiovisual synchrony to locate sounds, *NIPS*, pp.813–819, The MIT Press (1999).
9) Kass, M., Witkin, A. and Terzolpoulos, D.: Snakes: Active contour models, *Int'l J. of Computer Vision*, Vol.1, No.4, pp.321–331 (1988).
10) Kidron, E., Schechner, Y.Y. and Elad, M.: Pixels that sound, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR2005*), pp.88–95 (2005).
11) Kolmogorov, V., Criminisi, A., Blake, A., Cross, G. and Rother, C.: Bi-layer segmentation of binocular stereo video, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR2005*), Vol.2, pp.1186–1194 (2005).
12) Kolmogorov, V. and Zabih, R.: What Energy Functions can be Minimized via Graph Cuts?, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.26, No.2, pp.147–159 (2004).
13) Lucas, B. and Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision, *Proc. 7th Int'l Joint Conf. on Artificial Intelligence*

(*IJCAI1981*), pp.674–679 (1981).
14) Monaci, G., Escoda, O.D. and Vander-gheynst, P.: Analysis of multimodal signals using redundant representations, *Proc. Int'l Conf. on Image Processing* (*ICIP2005*), pp.145–148 (2005).
15) Patterson, E.K., Gurbuz, S., Tufekci, Z. and Gowdy, J.N.: Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus, *EURASIP J. on Applied Signal Processing*, Vol.2002, No.11, pp.1189–1201 (2002).
16) Smaragdis, P. and Casey, M.: Audio/visual independent components, *Proc. Int'l Symposium on Independent Component Analysis and Blind Source Separation* (*ICA2003*), pp.709–714 (2003).
17) Yu, T., Zhang, C., Cohen, M., Rui, Y. and Wu, Y.: Monocular video foreground/background segmentation by tracking spatial-color Gaussian mixture models, *Proc. IEEE Workshop on Motion and Video Computing* (*WMVC2007*), pp.55–63 (2007).
18) Zhu, S.C. and Yuille, A.: Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.18, No.9, pp.884–900 (1996).

(平成 19 年 9 月 25 日受付)
(平成 20 年 3 月 10 日採録)

(担当編集委員　鷲見 和彦)



劉　玉宇

2000 年北京郵電大学通信工学科卒業．2003 年清華大学大学院電子工学研究科修士課程修了．同年ソニー株式会社に入社．2006 年より東京大学大学院情報理工研究科の博士課程．音と映像の相関の分析に関する研究に従事．

佐藤　洋一（正会員）

1990 年東京大学工学部機械工学科卒業．1997 年カーネギーメロン大学大学院計算機科学部ロボティクス学科博士課程修了．Ph.D. in Robotics. 同年より東京大学生産技術研究所研究機関研究員，講師，助教授を経て，現在，同大学大学院情報学環准教授．コンピュータビジョン，ヒューマン・コンピュータ・インタラクション，コンピュータグラフィックスに関する研究に従事．2008 年電子情報通信学会論文賞，2006 年電子情報通信学会論文賞，1999 年情報処理学会山下記念研究賞，1999 年日本バーチャルリアリティ学会論文賞等を受賞．電子情報通信学会，日本バーチャルリアリティ学会，ACM，IEEE 各会員．