情報処理学会論文誌　Vol. 50　No. 4　1421–1430 (Apr. 2009)

カメラワークを含むショット・リバースショット区間の検出

吉高淳夫[†1]　尾崎昂[†2]　平嶋宗[†2]

映像への効率良いアクセスを実現するためには映像の構造化が不可欠であり，また構造化を行うためにはカット検出だけでなくシーン境界検出等，相互に関係するショットをまとめる処理が必要である．映画やドラマにおいては，2 種類の類似したショットが交互に繰り返される，ショット・リバースショットが会話の場面等で多用されるという特徴がある．ショット・リバースショットはカメラ操作をともなわない静的なショットから構成されることが比較的多いため，キーフレームの色領域に関する空間関係の類似度評価等，画像の静的特徴により検出するものが多かった．しかし，パン，チルト等のカメラ操作が施される場合もあり，そのような映像では全体が変化するために従来の静的な類似性判定では検出することは困難である．このようなショット区間の検出は映画のシーン境界検出やインデクシング，検索に必要であり，カメラワークを含むものも検出できることが精度向上のために望まれる．本論文ではカメラワーク成分をキャンセルする処理を施すことで一様な動きをした背景部分の除去を行い，残された主体部分の類似性判定に基づきショットどうしの類似性を判定することによって，カメラワークを含むショット・リバースショットの場面を抽出する手法を提案する．そして，カメラワークキャンセルを考慮しない従来検出法との比較により，提案手法の効果について評価した．

Extraction of Shot/Reverse-shot Including Camera Work

Atsuo Yoshitaka,[†1] Takashi Osaki[†2] and Tsukasa Hirashima[†2]

Video parsing is one of the important technique to offer efficient way to access video contents, and therefore scene boundary detection is fundamental process for video parsing. Scene boundary detection, that is based on shot similarity, has been studied for movie structure analysis. According to the film grammar, shots that consist of the alternation of two group of similar shots are called 'shot/reverse shot'. Previous studies that extract shot/reverse shot are based on comparison of color histogram or color based region clusters that are extracted by static key-frame analysis. Therefore, the extraction of a scene composed of shot/reverse-shot including camera work is an open issue, since static shot similarity evaluation cannot cope with such case effectively. In this paper, we proposed a method of detecting shot/reverse-shots under the influence of camera work based on camera work cancellation and dominant object similarity. We evaluated the proposed method by comparing static based method.

1. はじめに

インターネットによる映像配信やディジタル放送の普及にともなう多チャネル化により，大量の映像が提供されている．また大容量の HDD を搭載した PC やビデオレコーダの普及にともない，ユーザが膨大な量の映像を蓄積・管理し，視聴することが可能となった．蓄積される映像が大容量になるに従って，コンテンツの始めから終わりまでの単純な再生や，早送りと再生の繰返しによる操作だけでなく，内容に基づくインデクスからブラウジングする等，多くの映像の中から効率良く映像の内容を把握する仕組みへの要求が高まっている．

映像の内容に基づくインデクスを提供するためには，映像を意味的にまとまりのある単位に分割し，内容に基づいてそれらを関連付けるといった，映像の構造化が不可欠となる．構造化を行うためには，最初に映像を特定の単位に分割する必要があり，最も基本的な単位として映像をショットに分割することが考えられる．検出されるショット境界は映像が編集等により切り替わる点であり，ショット長は通常数秒〜数十秒程度であるので，内容把握を目的とする場合はこれらをより大きな単位にまとめて構造化することが必要である．そこでショットどうしの意味的なまとまりを検出し，それらのショットを統合することによって内容に基づいたシーンを検出すること等が行われている．

シーン検出手法の 1 つとして，特定のショットの編集パターンによって構成されるショット列を映像から抽出することが考えられる．ニュースやスポーツ映像を対象としてシーン検出を行う研究として，文献 1)，2) があげられる．これらの研究では，特定ジャンルの映像の特徴として，各シーンの先頭で共通した特徴的なパターンが含まれるショットが出現する等，シーン間で現れるショットの並びのパターンに関する規則性に着目している．しかし，映画やドラマといった映像においては，ニュースやスポーツ映像と同様な固定的で特徴的な

†1 北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science and Technology
†2 広島大学大学院工学研究科
Graduate School of Engineering, Hiroshima University

パターンがシーン内に出現しないため，他のジャンルの映像と同様な手法ではシーンの検出が困難である．

そこで映画におけるシーンを検出するために，類似したショットが繰り返し出現する構造に着目した手法が研究されている．映画の撮影・編集規則について述べられている「映画の文法」[9]によると，2種類のショットを交互に繰り返すことにより構成されるショットシーケンスは「ショット・リバースショット」に分類される．このようなショットの繰返しにより構成されるシーンのうち，最も頻出するシーンは，1名ずつ撮影された計2名の人物をスチル（カメラを静止させた状態）撮影したショットが交互に繰り返されるシーンであり，会話の場面等で多用される．映像の類似性に着目してこのような構造を検出する研究では，色相成分に基づくクラスタリング[3),4)]や色相ヒストグラム[5),6)]に基づく手法が提案されている．

会話の場面等でスチル撮影されたショットにおいては，人物の姿勢の変化等がある場合はあるもののショット内の映像変化がほとんどないため，静的な色相領域のクラスタリングによって求められる領域やヒストグラムの類似性によって類似ショットの繰返し構造を検出するのは有効な手法だといえる．しかし，被写体の大きな動作やカメラワークのある映像により構成されるショット・リバースショットシーケンスの場合は映像変化が大きいため上にあげた静的な類似性評価手法では検出が困難である．

シーン内ではBGMが途切れずに連続することが多いという特徴から，映像の類似性に加えてBGMの連続性も考慮したシーン検出手法[7),8)]等も提案されているものの，シーン境界を多少前後してBGMが挿入される演出もしばしば見られることから，映像の類似性に基づくシーン境界の判定を補強する位置付けである．また，ショット・リバースショットの場面は1シーンを構成する部分要素，すなわちより粒度の小さい単位であり，一方BGMの開始，終了はシーン単位でなされることが多いため本研究で目標とするショット・リバースショット区間の境界検出にBGM検出は直接的には寄与しないと考えられる．これらのことから，ショット・リバースショットシーケンスの検出精度をより高めるためにはカメラワークを含むショット・リバースショットシーケンスへの対応を直接的に考慮する必要があると考えられる．しかしながら，それを検出することの重要性，いい換えれば実際の映画等の映像においてカメラワークをともなったショット・リバースショットシーケンスがどの程度用いられているかは明らかにされておらず，またそれを検出するという課題は未解決であった．

本論文では，従来手法では未解決であったカメラワークを含むショット・リバースショットの場面を抽出する手法を提案する．ショット・リバースショットからなるショットシーケンスの区間は，2名の被写体の撮影に各々1台ずつカメラを割り当てて撮影し，後の編集で2台のカメラで撮影された映像が交互になるよう編集されるのが基本である．したがって，スチル撮影されている場合，ある人物を撮影した編集前の映像のフレーム間類似性は高いといえる．一方カメラワークが施された場面については，つねに主体を追跡するように撮影されるため，背景がそれにともなって変化する一方で主体部分には大きな変化が生じにくいという特徴があると考えられる．

提案手法ではこのような特徴に着目し，カメラワークによって生じる背景部分の変化をキャンセルし，主体部分を抽出したうえでその類似性に基づいてショットの類似性を判定する．主体部分の類似性判定のためにはカメラワークが施されている映像における主体の検出が必要となる．その手法としては文献11)–13)があげられるが，これらの手法では，映画映像によく見られる雷やカメラフラッシュ等の閃光，あるいはスポットライトの通過のような短時間に輝度が激しく変化する場面の検出が困難である[11),13)]ことや，主体（主となる被写体）の出現位置はつねに映像フレーム中央であるという前提を置いた検出処理で，出現位置が映像フレームの右端あるいは左端付近であるような場合が考慮されていない[12)]といった問題がある．そのため本研究ではカメラワークを検出し，それによる映像変化をキャンセルして一様に変化する背景部分を除去することにより主体部分を検出する．一様に変化する背景部分に相当するカメラワークパラメータについては時空間投影画像を用いて算出する．時空間投影画像とは，フレーム内の一定位置の直線を各フレームから抽出し，それらを時間方向に並べて得られる画像である．時空間投影画像では，背景の移動が時間軸に対して傾きを持つ線として表される．そのため，短時間に輝度が激しく変化する映像に対しては一定の傾きを持つ線分の補間によりその影響を抑えることが比較的容易であり，また被写体自体の動きを含む映像に対しては複数線分の傾きの共通性を考慮することによりカメラワークによる映像変化量を被写体自体の動きによるものと区別してとらえるのが比較的容易であることがあげられる．また，主体が映ることの少ない部分である映像フレームの端付近に時空間投影画像生成のための参照部分を設定することにより，主体の位置変化や大きさにも影響されにくいカメラワークパラメータの抽出が可能となる．そして，カメラワークをキャンセルする処理を施すことで，一様な動きをした背景部分の除去を行い出現位置に依存しない主体の検出を行う．そして，検出された主体部分に対してショット間での類似性を判定し，類似ショットどうしを統合することによって，カメラワークを含むショット・リバースショットの場面を抽出する．

本論文では，2章でショット・リバースショットの場面について，映画における使用頻度とカメラワークを含むショット・リバースショットの特徴について述べる．3章では，カメ

ラワークを含むショット・リバースショットの抽出に必要な主体の検出手法について述べ，4 章では，ショット・リバースショットの場面を抽出する手法について説明する．5 章で提案手法の評価実験とその結果について述べたあとで，6 章で本論文をまとめる．

2. ショット・リバースショットの演出場面

「映画の文法」[9] によると，ショット・リバースショットとは，2 名以上の人物等の主体どうしのやりとりを撮影，編集する場合に，それぞれの主体が映るショットが交互になるよう編集することで，人物同士の対話やその緊迫感を盛り上げる編集上の演出のことである．ショット・リバースショットの例を**図 1** [⋆1] に示す．

2.1 映画におけるショット・リバースショット

ショット・リバースショットのうち，カメラワークを含むショット・リバースショットがどの程度含まれているのかについて，映画全体におけるショット・リバースショットの割合とともに調査を行った．**表 1** の映画における調査結果を**表 2** に示す．また，ショット・リバースショットにおいて使用されたカメラワークの種類とその構成比についての調査結果を**図 2** に示す．

調査した映画中のショット・リバースショットのうち，カメラワークを含む区間の割合はショット数を単位として約 35%と無視できるほど少なくはない．したがって，ショット・リバースショットシーケンスを検出し，シーン検出等の構造化やそれに基づくインタラクションをより的確に行うためにはカメラワークを含んだものにも対応する必要があるといえる．単一の処理ですべてのカメラワークを検出できるわけではないため，各々のカメラワークの種類に応じた検出手法が必要となる．このため本研究では，図 2 におけるカメラワークのうち約 75%と最も多くを占めており，ほぼ同様の手法で検出可能である，一定の位置でカメラを水平方向に旋回する操作であるパン，垂直方向に旋回する操作であるチルト，水平方向または垂直方向にカメラを移動する操作であるトラックのいずれかのカメラワークが含まれるショット・リバースショットを対象とした．

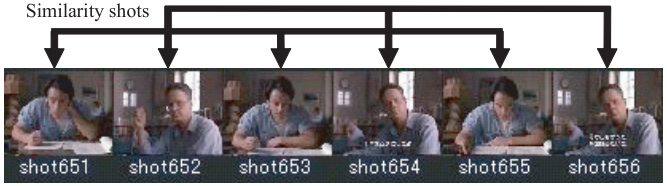

図 1　ショット・リバースショットの例[⋆1]
Fig. 1　An example of shot/reverse-shot.

⋆1 The Shawshank Redemption, Frank Darabont (Dir.), Warner Bros. (1994)

2.2 カメラワークを含むショット・リバースショット

カメラワークを含むショット・リバースショットではその区間内で映像全体が変化し，カメラが固定された状態で撮影された区間に比べて映像フレーム全体にわたってショット間で類似するフレームがほぼ存在しない．そのため，フレーム全体を対象とした映像の類似性等に基づく従来のシーン検出手法ではショットの類似性判定が困難である．ここで，ショット・リバースショットにおけるパン，チルト，トラック操作が施されている区間ではつねに

表 1　調査を行った映画
Table 1　Movies analyzed.

作品名	監督	制作年	ジャンル	時間	配給会社
Speed	Jan de Bont	1994	アクション	115 分	20th Century Fox
The Shining	Stanley Kubrick	1980	ホラー	143 分	Warner Bros.

表 2　映画 2 本のショット・リバースショット（S・RS）の出現頻度
Table 2　Frequency of appearance of shot/reverse-shot in two movies.

	区間数と映画全体に対する割合			S・RS に対する割合	時間と映画全体に対する割合	
	区間	Shot	%	%	時間	%
S・RS	100	758	28.8	—	85 m 57 s	33.8
カメラワークを含む S・RS	33	268	10.2	35.4	32 m 36 s	12.8

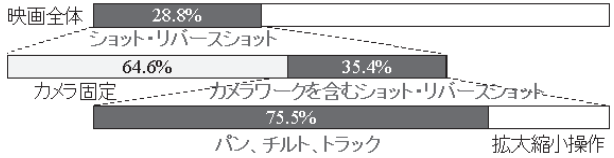

図 2　ショット・リバースショット中のカメラワーク構成比
Fig. 2　Camera work ratio in shot/reverse-shots.

主体となる人物を追跡して撮影されるという点に着目する．その際，**図 3** [*1]のように映像の背景部分はカメラワークによって全体が一様に変化するが，主体部分は背景部分と異なり人物の姿勢程度の変化しかないことが多い．したがって，主体部分を構成する色相の特徴量は大きく変化しないと考えられるため，背景の一様な変化とは異なり，人物の姿勢の変化は色相のヒストグラムによる類似性判定においては無視できると考えられる．

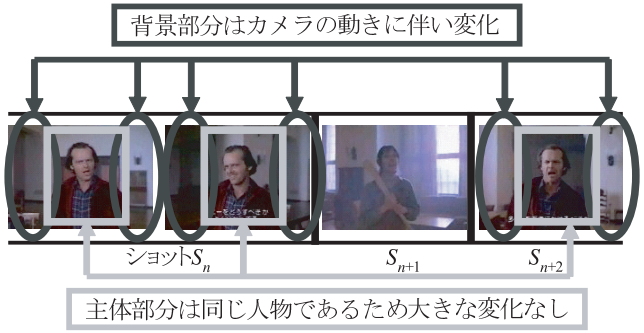

図 3　カメラワークを含むショット・リバースショットの例[*1]
Fig. 3　Example of shot/reverse-shot including camera work.

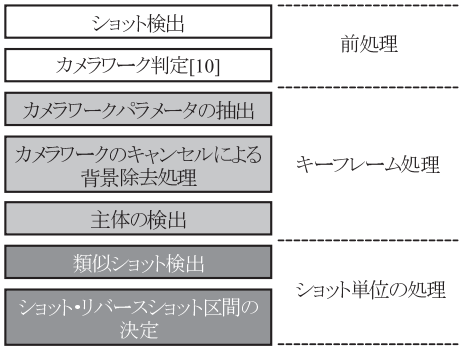

図 4　提案手法の概要
Fig. 4　Overview of the proposed method.

⋆1 The Shawshank Redemption, Frank Darabont (Dir.), Warner Bros. (1994)

提案手法ではカメラワークパラメータを求め，背景移動量相当の映像の動きをキャンセルする．背景部分はカメラワークによる変化に加えて，ショット間で同じカメラ位置から撮影していないこともあるためにショット間においても大きく異なる場合がある．そこで，背景部分の除去を行い，主体となる人物を検出し，主体部分の色ヒストグラムの類似性判定に基づき，ショット・リバースショットの場面を構成する類似ショット対であるか否かの判定を行う．本手法の概要を**図 4** に示す．

3. カメラワーク映像における主体の検出

本研究では時空間投影画像によりカメラ操作量を求め，カメラ操作にともなって一様な動きをした背景部分を検出，除去することで主体を検出する．カメラワーク区間とその種類の判定は時空間投影画像に基づく手法[10] により行った．

3.1 カメラワークパラメータの抽出

時空間投影画像とは，フレーム内の一定位置の直線（以後定線とよぶ）を各フレームから抽出し，それらを時間軸に沿って並べて得られる画像である（**図 5**）．時空間投影画像の横軸は時間 f [frame]，縦軸は定線の長さである．時空間投影画像内のエッジにはカメラワー

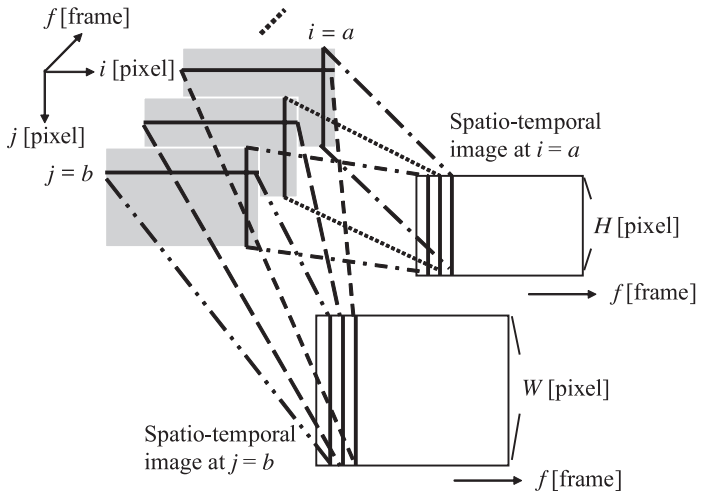

図 5　時空間投影画像
Fig. 5　Spatiotemporal tomographic image.

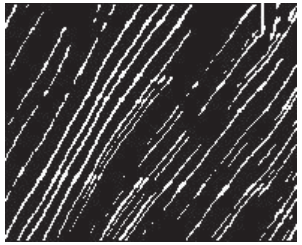

図 6　パン，チルト，トラックにおける時空間投影画像の例
Fig. 6　Spatiotemporal tomographic image in pan, tilt and track.

クの種類やその操作量に従って異なる特徴が現れるため，本研究では時空間投影画像に現れるエッジの特徴抽出により，一様な動きをした背景部分の移動量を求める．映画やドラマの映像でのカメラと被写体，背景との位置関係においては，カメラの回転操作であるパン，チルトとカメラの平行移動操作であるトラックによって現れる時空間投影画像上の特徴は同様であると見なせる場合が多い．フレームサイズ $W \times H$ [pixel] の映像中，パン，チルト，トラックのうち，水平方向のカメラワークでは直線 $j = b$（$0 \leq b \leq H-1$），垂直方向のカメラワークでは直線 $i = a$（$0 \leq a \leq W-1$）を定線とした時空間投影画像において，カメラ操作の結果として背景の各画素が時間経過にともなって定線の軸上を移動するため，時間軸に対して斜めに傾いたエッジ群が現れるという特徴がある（**図 6**）．このとき主体の動きの影響を受けにくいフレームの端付近に定線をとり，以下の処理を行う．カメラワーク区間内の時空間投影画像の直線エッジを検出し，全エッジ e 本の角度 θ_n（$-80° \leq \theta_n \leq -5°$，$5° \leq \theta_n \leq 80°$）を求める．差分をとるフレーム間隔を dif [frame]，θ_n（$n = 1, \ldots, e$）の最頻角を θ_{mod} とし，背景移動量 mov [pixel] を次の式 (1) により求める．

$$mov = dif \times \tan\theta_{mod} \tag{1}$$

式 (1) において mov の値が大きい場合，後述の背景除去処理でフレーム間の差分画像が生成可能な範囲が狭くなってしまう．また mov の値があるシーン内のショット間で大きく異なる場合は，背景として除去されない領域の現れ方に違いが生じ，カメラ操作量に依存しない主体検出が困難になる．そのため dif の値は $\tan\theta_{mod}$ の値により動的に決定する．

3.2　背景除去処理

検出したカメラワークパラメータに基づき，カメラワークによる変化をキャンセルした差分画像を生成する．本論文では左旋回のパン操作を例に説明する．5×5 [pixel] のブロック B_{kl}（$k = 1, \ldots, \lfloor (W - mov)/5 \rfloor$，$l = 1, \ldots, \lfloor H/5 \rfloor$）で式 (2) の条件を満たすとき，つまり**図 7**[⋆1]において主体検出を行うフレーム f_n の座標 (i, j) における輝度 $V_n(i, j)$ と，フレーム f_{n+dif} の座標 $(i + mov, j)$ における輝度 $V_{n+dif}(i + mov, j)$ との，ブロック内の各画素の輝度の差の合計が閾値 Thv 以下のとき，フレーム f_n のブロック B_{kl} を背景部分と判断して除去する．

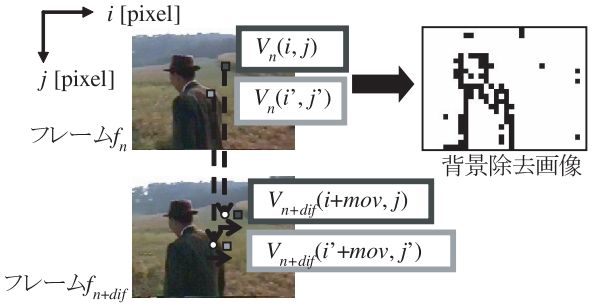

図 7　背景除去処理
Fig. 7　Removal of background region.

$$\sum_{i=5(k-1)}^{5k-1} \sum_{j=5(l-1)}^{5l-1} |V_n(i, j) - V_{n+dif}(i + mov, j)| \leq Thv \tag{2}$$

なお，背景除去処理は透視投影ゆがみにより背景が大きく変化する場合は，ゆがみに起因するフレーム間の差異も検出されてしまうという問題がある．しかし，人物等の主体を追うカメラワークの場面では主体と背景とのコントラストを強調するために背景部分は見かけの変化が大きくない場合が多く，また，カメラワーク区間では 0.5 秒ごとのフレームを比較しているので問題にならない場合が多いと考えられる．しかし，主体の移動速度が速く，背景もカメラから比較的近い場合等はその影響が大きく現れると考えられるので，そのような状況への対策は今後検討する必要がある．

3.3　主体の検出

主体領域内の輝度には大きな変化がないため，背景除去処理の過程で主体の内部も除去されてしまうことになる．いい換えれば，背景除去処理によって主体の境界に相当する領域が背景除去処理の結果得られる．主体領域を検出するために，当該映像フレームを色相に基づ

⋆1 The Shawshank Redemption, Frank Darabont (Dir.), Warner Bros. (1994)

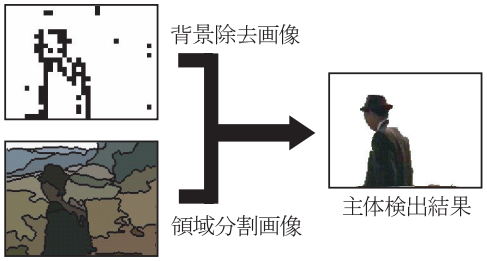

図 8　主体の検出
Fig. 8　Detection of a dominant object.

いて領域分割し，それによって得られた領域と背景除去処理によって得られた主体の境界を統合する．映像フレームの領域分割では，K-means 法により各フレームに対して色特徴量が類似した領域ごとにクラスタリングする．まず，映像フレームの横を 5 等分，縦を 4 等分して計 20 の初期クラスタ領域を設定する．次に各クラスタ領域における画素の特徴量の平均値を算出する．画素の特徴量 S は RGB 表色系による色成分とクラスタ内の画素の位置 (i, j) とする．すなわち，

$$S = \sqrt{k_c \left\{ \left(R - \overline{R}\right)^2 - \left(G - \overline{G}\right)^2 - \left(B - \overline{B}\right)^2 \right\} + k_p \left\{ \left(i - \overline{i}\right)^2 + \left(j - \overline{j}\right)^2 \right\}} \quad (3)$$

ここで，$\overline{R}$，$\overline{G}$，$\overline{B}$ はそれぞれ該当クラスタ内における R，G，B の平均値，$(\overline{i}, \overline{j})$ は該当クラスタ内における画素位置の平均値とする．また，k_c，k_p は重み係数であり，$k_c = 2$，$k_p = 1$ とした．フレーム内の各画素に対してその特徴量がどのクラスタの特徴量の平均値と最も近いかを判定し，最も近いクラスタに統合することでクラスタの再構成をする．この処理を各クラスタを構成する画素が変化しなくなるまで繰り返す．このようにして領域分割を行うが，その結果大きな領域の周囲に小さな領域ができることがあるため，これに対して 5×5（ピクセル）の大きさの多数決フィルタを適用してそのような領域を近傍の大きな領域に統合する．さらに面積がフレームの面積の 1/384 以下の小領域は，隣接する領域のうち最も面積の大きい領域に統合する．このようにして映像フレーム全体を領域分割する．

次に，カメラワークパラメータをキャンセルすることで得られた，主体と背景との境界にあたる領域（**図 8** [⋆1]左上）を領域分割結果と重ね合わせ，主体領域に相当する領域であるかを判定する．抽出された各領域に対して，背景除去処理により抽出されたブロックの面積の割合が 30%以上でかつ領域の境界に属するブロックの面積の割合が 40%以上という条件を満たす領域を主体領域として抽出する．この条件は予備実験により決定した．

⋆1 The Shawshank Redemption, Frank Darabont (Dir.), Warner Bros. (1994)

4. カメラワークを含むショット・リバースショットの抽出

類似ショットが交互に現れる区間の検出により，ショット・リバースショットの場面を抽出する．このときショット・リバースショットにおけるカメラワーク区間とスチル撮影の区間では，ショット間において類似している領域が異なるため，カメラワークの存在の有無に応じた類似ショット対の判定を行う．

4.1 類似ショット検出

類似ショット対を抽出するために前後のショット間の類似度を算出する．ショットの類似度はキーフレーム間の色相類似度によって求める．

ショット内から先頭フレームと最終フレーム，さらに先頭から 30 フレーム（約 1 秒）ごとのフレームをキーフレームとして抽出する．このとき，カメラワーク区間については映像変化が大きいため 15 フレーム（約 0.5 秒）ごととする．

比較対象の両方のショットにパン，チルト，トラック操作が施されている場合，主体部分どうしでキーフレーム間の類似性を判定する．また，比較を行う一方のショットにパン，チルト，トラックが施され，他方のショットはスチル撮影である場合，カメラワークの施されたショットにおける主体領域が他方のショット中のキーフレームに含まれるか否かといった主体の存在判定を行い，主体部分に関してキーフレーム間の類似性を判定する．映画，ドラマ等の映像ではショット・リバースショット間において主体が映る位置は大きく変わらないことが多い傾向があることから，カメラワーク区間のフレームにおいて検出された主体領域と同じ位置の領域をスチルショット区間における主体候補領域として比較判定を行う．比較を行う両方のショットともスチル撮影の場合は，従来のシーン検出手法と同様に，フレーム全体を対象としてキーフレームどうしの類似性判定を行う．

キーフレーム k_1，k_2 間の類似度 $H_n(k_1, k_2)$ は，HSV 色特徴量のヒストグラムを用いて式 (3) によって評価する．HSV 色特徴量は人間の視覚特性に応じて H = 6，S = 2，V = 3，計 36 個の部分空間に分割した．これは人の視覚特性として，色相の変化と比較して彩度の変化には反応しにくいことに基づいている．式 (3) の $h_{k_1 R(rx,ry)}(c_z)$ は，キーフレーム k_1 内の ry 行 rx 列目の部分領域 $R(rx, ry)$（$rx = 1, \ldots, N$，$ry = 1, \ldots, N$）に対して，z 番目の部分空間 cz（$z = 1, \ldots, 36$）に含まれる画素数である．N の値については，2 つのキー

フレームのどちらか一方でも主体を検出している場合は主体領域どうしのみの比較を行うため $N = 1$ とし，それ以外の場合は背景部分の空間関係を考慮するため $N = 8$ として 64 の領域に分割した．

$$H_n(k_1, k_2) = \frac{1}{N^2} \sum_{rx=1}^{N} \sum_{ry=1}^{N} \left(1 - \frac{1}{\frac{W}{N} \times \frac{H}{N}} \sum_{z=1}^{36} |h_{k_1 R(rx,ry)}(c_z) - h_{k_2 R(rx,ry)}(c_z)| \right) \tag{4}$$

ショット間の類似度は，ショット内のキーフレーム間類似度 $H_n(k_1, k_2)$ が閾値以上であるフレーム数の割合によって評価し，ショットどうしが類似ショット対であるかどうかの判定を行う．

4.2　ショット・リバースショット区間の決定

ショット・リバースショットでは 2 種類の映像を交互に配置するように編集するのが基本であるので，これを式 (5) のように定義し，その区間の抽出を行う．すなわち，類似ショット a，b（$a \neq b$）が交互に出現し，それらのショットの総数が 4 ショット以上である区間を，ショット・リバースショットとする．ただし，区間内に a，b どちらとも類似しないショット c，d の存在を許容する．これは類似ショット a，b が交互に編集されるほか，途中で回想ショット等類似性の点で独立したものがしばしば挿入されることを考慮したものである．

$$S = (axby)_n,\ n \geq 2 \tag{5}$$
$$x = \{Null, c\}$$
$$y = \{Null, d\}$$

5. 評 価 実 験

カメラワークを含むショット・リバースショットの抽出精度を評価するために，カメラワーク以外のシーンの誤検出要因を排除した映像，シーンの誤検出が生じうる要因を変化させた映像，実映画映像の 3 種類の映像に対して実験を行った．実映画以外の映像では，カメラワーク以外でシーンの誤検出となりうる要因を排除して評価するため，主体と背景の色特徴量差が十分あり，カメラワークの速度やフレーム内における主体の大きさやその位置が映画映像と同程度になるように配慮した，以下の 2 種類のショット・リバースショットを撮影，編集して用意した．

(a) 2 名の動く人物をパンによって追跡したショットが，それぞれ交互になるように編集した映像

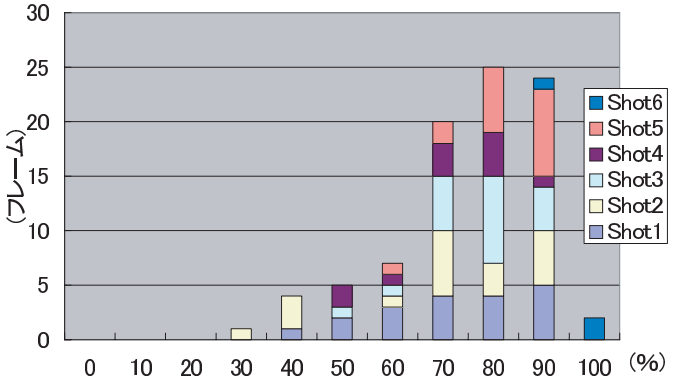

図 9　主体の検出精度（$Precision_r$）
Fig. 9　Result of dominant object detection ($Precision_r$).

(b) 動く人物をパンによって追跡したショットと，静止した別の人物をカメラを固定した状態で撮影したショットが，それぞれ交互になるように編集した映像

用いた映像形式は，フレームサイズ 160 × 120 [pixel]，フレームレート 30 [fps]，24 ビットカラーである．(a)，(b) の映像はそれぞれ約 1 分，6 ショットで構成されており，また (b) の映像に対するカメラワークのショットの割合は映画での割合と同じく半分程度としている．なお，検出結果には影響しないため，式 (5) における非反復ショット c，d は含めていない．

5.1　主体検出精度の検証実験

(a)，(b) の映像におけるカメラワーク区間のキーフレームに対して主体検出処理を行った．主体検出精度の評価は対象領域の面積単位で行い，検出された領域に対する実際の主体領域の割合を $Precision_r$，実際の主体領域に対する検出された領域の割合を $Recall_r$ とする．$Precision_r$ を**図 9**，$Recall_r$ を**図 10** に示す．主体の検出精度について，全体としては $Precision_r$ の平均が約 75%，$Recall_r$ の平均が約 84%となっており，過剰に抽出される領域がやや多いものの，主体領域自体はほぼ抽出されているといえ，比較ショット間における主体どうしの類似度を求める際にはあまり影響しないものと考えられる．

5.2　本手法と従来手法の比較実験

本実験では (a)，(b) の映像（※ 1）について，誤検出が生じると考えられる要因を変化させ，本手法と従来のシーン検出手法における抽出精度の比較を行った．映像の類似性に基づくシーン検出手法は色相ヒストグラム[5),6),8)] またはクラスタリング[3),4)] によるものが一

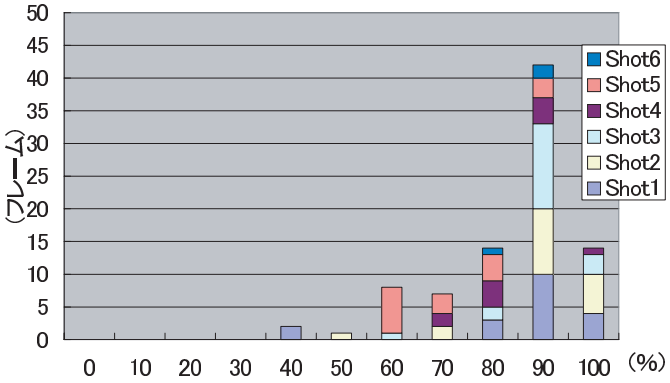

図 10　主体の検出精度（$Recall_r$）
Fig. 10　lt of dominant object detection ($Recall_r$).

般的であるので，ここでは比較手法として文献 5)，6) と同様な手法の 1 つである類似画像判定アルゴリズム[8] を使用した．比較手法にカメラワークのキャンセル処理を加えたものが提案手法となる．変化させた要因は以下のとおりである．

- 背景の複雑さ（単純・複雑）（※ 2）
- 主体のフレーム内での大きさ（小・中・大）（※ 3）
- カメラワークの速度（遅・中・速）（※ 4）

ショット・リバースショット区間内のすべてのショットを繰返し構造を持つ一連のショットシーケンスであると判定できた場合を正しい判定とし，2（※ 1）× 2（※ 2）× 3（※ 3）× 3（※ 4）= 36 通りの映像に対する抽出実験結果を**表 3** に示す．表中の数字はショット・リバースショットの映像数である．

実験結果より，背景が単純な場合は本手法（表中，静的比較+CW キャンセルと表記）と比較手法（表中，静的比較と表記）の抽出精度に顕著な差が現れなかった．これは背景が単純であれば，背景の動きが色相ヒストグラム上の差として現れないことから，カメラワークを考慮しない手法でもショットの類似性が判定可能であったことによる．しかしながら，背景が複雑な場合は，比較手法の抽出精度は約 56%であったことに対し，本手法の抽出精度（Recall）は約 94%であった．したがって，従来のシーン検出手法では抽出されないショット・リバースショットが本手法により抽出可能であることが分かる．

また，主体の大きさやカメラ操作の速度といった要因を変化させたところ，要因の変化にかかわらずショット・リバースショットの抽出精度（Recall）が約 92%から 100%であったことから，本手法は主体の大きさやカメラ操作の速度にも影響されにくい手法であるといえる．しかし，実際の映画と同様の条件での撮影が困難であるため本実験では確認できなかった要因である，照明変化等により抽出精度が低下する可能性がある．

表 3　カメラワークを含むショット・リバースショットの抽出精度
Table 3　Result of shot/reverse-shot extraction including camera work.

	背景が単純	背景が複雑	全体
静的比較+CW キャンセル	1.00 (18/18)	0.94 (17/18)	0.97 (35/36)
静的比較	0.89 (16/18)	0.56 (10/18)	0.72 (26/36)
	主体大	主体中	主体小
静的比較+CW キャンセル	0.92 (11/12)	1.00 (12/12)	1.00 (12/12)
静的比較	0.67 (8/12)	0.50 (6/12)	1.00 (12/12)
	速度大	速度中	速度小
静的比較+CW キャンセル	0.92 (11/12)	1.00 (12/12)	1.00 (12/12)
静的比較	0.67 (8/12)	0.75 (9/12)	0.67 (8/12)

表 4　映画におけるショット・リバースショットの抽出精度（$Recall_d$）
Table 4　Result of shot/reverse-shot extraction in movie.

		$Recall_d$ (Nc/Na)
カメラワークを含む S・RS	本手法	0.40 (6/15)
	比較手法	0.20 (3/15)
映画全体における S・RS	本手法	0.80 (40/50)
	比較手法	0.74 (37/50)

5.3　映画映像に適用した実験の結果

映画 “The Shining（Stanley Kubrick (Dir.), Warner Bros., 1980)” について，カメラワークを含むショット・リバースショット区間の抽出精度を求める実験を行った．前節で述べた実験と同様に提案手法と比較手法の抽出精度を求め，さらに映画全体におけるショット・リバースショットの抽出精度も含めた結果を**表 4** に示す．ここで，Nd，Na，Nc をそれぞれ，Nd：提案手法によって抽出された区間数，Nc：正解と一致する区間数，Na：主観により求めた正解区間数とする．正解区間は映画の文法を理解している被験者 1 名の主観判断により決定し，抽出された区間と正解区間との間の前後 1 ショットのずれはシーン検索等の応用においては許容できる誤差であると考え，正判定とした．

実験の結果，カメラワークを含むショット・リバースショット区間のみで比較すると $Recall_d$

は比較手法で 0.2 であるのに対し，提案手法では 0.4 となり，精度が 20%向上した．スチルショットのみからなるショット・リバースショット区間を含めた映画全体で見ると約 6%の精度向上が見られた．スチルショット区間のみでの $Recall_d$ は提案手法，比較手法ともに約 0.97（34/35）とほぼ十分な精度で検出できているといえるため，現状ではまだ十分ではないものの，カメラワークを含むショット・リバースショットの検出が検出精度向上のためには必要であることが分かる．通常 Recall と Precision はトレードオフの関係にあるため Recall が向上している場合それと同時に Precision が低下することが考えられるが，映画全体で比較した際の Precision（= Nc/Nd）は比較手法で 0.88（37/42），提案手法で 0.87（40/46）となり，Recall の向上度合いに対して Precision の低下を小さく抑えることができた．よって，提案手法は Precision，Recall を総合的に見てショット・リバースショットの検出精度向上に寄与しているといえ，映画映像に対するシーン等へ応用した場合の検出精度の向上に寄与すると考えられる．

5.4　映画において検出もれが多い原因

前節に示した，実映画映像を対象とした実験結果ではショット・リバースショットの検出もれが多かったため，本手法により検出されなかった主原因を調査した．その結果，主体と背景の色特徴量が類似している状況下で主体が正しく検出されずにショットどうしが類似していると判定されない映像が 4 区間あった．また，照明の変化によりカメラワークのキャンセルが困難であった映像が 3 区間，また，カメラワークの検出もれにより，カメラワークを含むショットの類似性を判定できない映像が 2 区間あった．そのほか，色ヒストグラムの類似性判定においてショットを過剰に統合してしまったものが 2 区間あった．これらの多くはカメラワーク以外の要因によるものであったため，実映画への適用のためには色相判別処理等カメラワーク検出処理以外の部分を改善する必要があると考えられる．

6. おわりに

本研究では，映像の静的な類似性に基づいた従来のショット・リバースショット区間検出では困難であった，カメラワークを含むショット・リバースショットの場面の抽出手法を提案した．ショット・リバースショット区間では 2 名の被写体各々を 1 台ずつのカメラで撮影して編集することが一般的であり，また，カメラワークが施されている場合は主体を追跡するように撮影することから，カメラワークにより背景部分が変化しても主体部分の変化は小さいという特徴に着目した．提案手法では，カメラワークの種類とその操作量を時空間投影画像により求め，カメラワークによる映像の変化をキャンセルした映像から主体領域を抽出し，主体領域の類似性をショット間で評価することによりカメラワークを含むショット・リバースショットの関係を検出した．実験により，カメラワーク以外の映像変化要因のない理想的な状態の映像における提案手法の効果が明らかになったといえ，また，現状では 1 本の映画全体に対する評価結果でしかないが，他の要因が影響するものの実映像でも精度が向上するという結果が得られた．本手法はシーン検出の際に，映像全体の変化が大きいため誤ってシーン境界と判定してしまうことが多いカメラワークを含むシーンに対して，より正しいシーン境界の判定や検出精度の向上に寄与すると考えられる．今後の課題として，映画に対して 5.4 節であげた検出ミスとなる要因に関する処理の改善の検討等があげられる．

参　考　文　献

1) 青木　恒：映像対話検出によるテレビ番組コーナ構成高速解析システム，信学論（D-II），Vol.J88-D-II, No.1, pp.17–27 (2005).
2) 椋木雅之，寺尾元宏，池田克夫：カット構成の規則性を利用したスポーツ映像のプレイ単位の分割，信学論（D-II），Vol.J85-D-II, No.6, pp.1016–1024 (2002).
3) 岡本啓嗣，八杉将伸，馬場口登：固定長の時空間画像に基づく映像シーンのクラスタリング，信学論（D-II），Vol.J86-D-II, No.6, pp.877–885 (2003).
4) Lehane, B., O'Connor, N. and Murphy, N.: Dialog Scene Detection in Movies Using Low and Mid-level Visual Features, *International Conference on Image and Video Retrieval*, LNCS 3569, pp.286–296 (2005).
5) 村野井亮治，早坂里奈，太田浩二，趙　継英，松下　温：映画におけるシーンの抽出を利用した階層的なビデオブラウザの構築，情報処理学会研究報告，Vol.1997, No.89, 1997-IM-032, pp.29–34 (1997).
6) 青木　恒，堀　修：繰返しショットの統合による階層化アイコンを用いたビデオ・インタフェース，情報処理学会論文誌，Vol.39, No.5, pp.1317–1324 (1998).
7) Pfeiffer, S., Lienhart, R. and Elsberg, W.E.: Scene Determination based on Video and Audio Features, *IEEE Multimedia Computing and Systems*, Vol.1, pp.685–690 (1999).
8) Yoshitaka, A. and Miyake, M.: Scene Detection by Audio-Visual Features, *IEEE International Conference on Multimedia and Exposition*, pp.49–52 (2001).
9) ダニエル・アリホン（著），岩本憲児，出口丈人（訳）：映画の文法，紀伊国屋書店 (1980).
10) 吉高淳夫，松井亮治，平嶋　宗：カメラワークを利用した感性情報の抽出，情報処理学会論文誌，Vol.47, No.6, pp.1696–1707 (2006).
11) Yuan, C., Ma, Y.-F. and Zhang, H.-J.: Extracting Video Object's Motion Trajectory by Velocity Voting, *IEEE Paci.c-Rim Conference on Multimedia*, pp.452–454 (2003).

12) 奥村真澄，高木真一，小舘亮之，富永英義：動き補償と色情報を組み合わせた MPEG 映像からの人物領域抽出，情報処理学会研究報告，Vol.2002, No.25, 2002-AVM-36, pp.31–36 (2002).
13) Oh, J.-H. and Sankuratri, P.: AUTOMATIC DISTINCTION OF CAMERA AND OBJECT MOTIONS IN VIDEO SEQUENCE, *IEEE International Conference on Multimedia and Exposition*, pp.81–84 (2002).

(平成 20 年 8 月 6 日受付)
(平成 21 年 1 月 7 日採録)

吉高　淳夫（正会員）

1989 年広島大学工学部第 2 類（電気系）卒業，1991 年同大学大学院博士課程前期修了，1994 年同博士課程後期単位取得退学．現在，北陸先端科学技術大学院大学情報科学研究科准教授．博士（工学）．マルチメディアデータ検索，感性情報等に基づいた動画像処理，映像を利用したインタラクティブシステムが主な研究分野．映像情報メディア学会，IEEE Computer Society 各会員．

尾崎　　昂

2005 年広島大学工学部第 2 類（電気・電子・システム・情報系）卒業．2007 年同大学大学院博士課程前期修了．修士（工学）．画像処理，映像構造化に興味を持つ．

平嶋　　宗（正会員）

1986 年大阪大学工学部応用物理学科卒業，1991 年同大学大学院博士課程修了，同年同大学産業科学研究所助手．同講師，九州工業大学情報工学科助教授を経て，2004 年より広島大学大学院工学研究科教授．人間を系に含んだ計算機システムの高度化に関する研究に従事．工学博士．ED-MEDIA95，ICCE2001，ICCE2002 で優秀論文賞．2008 年度教育システム情報学会論文賞．人工知能学会，電子情報通信学会，教育システム情報学会，教育工学会，日本教育心理学会，IAIED，APSCE，AACE 各会員．