イベント映像自動生成
映画のショット解析から導出される構図ルールとその適用
Automatic generation of event image: on the basis of composition rules derived from shot analysis of some films

野田 佑帆[†]，金谷 友樹[†]，杉本 徹[†‡]，榎津 秀次[†‡]
Yuho Noda, Yuki Kanaya, Toru Sugimoto, Hideji Enokidzu

[†]芝浦工業大学大学院工学研究科，[‡]NECソフト株式会社，[†‡]芝浦工業大学工学部
[†]Graduate School of Engineering, Shibaura Institute of Technology, [‡]NEC soft Corporation,
[†‡]Shibaura Institute of Technology
[†]ma12081@shibaura-it.ac.jp, [†‡]sugimoto@sic.shibautra-it.ac.jp, [†‡]enokizu@sic.shibautra-it.ac.jp

Abstract

The present study is intended to construct automatic editing system for recorded video of the scene that was performed on the basis of a short scenario. On the basis of the shot analysis of two films, we generated the event-driven composition rules. The event-driven composition rules were applied to event information to extract composition information. Composition information involved the layout of characters and the shot size. Our system was able to make up some scene videos. However, several problems are still left that need to be addressed to improve our system.

Keywords — automatic editing, shot analysis, composition rules.

1. はじめに

老若男女問わず，映画を観ることで，作品の中に何かを感じ取り，また自分が作品の登場人物であるように感じることもできる．なぜこんなにも映画というものが親しみやすいものであるのか．まず映画には，映像制作者が長年積み上げてきた知識を映像編集に用いることで，映像に没入感を生ませている．また，Gibson の著書によると，「映画を観ることは，現実生活で普通に生ずる出来事を観察することと類似している」[1]と述べている．つまり，我々が知覚している世界は，映画と同じような映像の世界であるということである．また，Zack の論文によると，「人間の知覚は，“椅子”や“犬”といったモノと，“車を購入する”や“ケーキを切る”といったイベントといった単位で出来事を認知している」[2]と言及している．これには映画の構成と関係がある．

映画を構成している最小単位にショットと呼ばれるものがある．ショットとは，切れ目なしに連続して撮影された映像のことであり，モノを映していたりや登場人物が歩いているといった単一の出来事を映していることが多い．このショットを編集によって繋ぎ合せることで映画が作られている．つまり，人の知覚がイベント単位で認知されているとすれば，ショットで構成されている映画は認知しやすいものであると言える．

そこで本研究では，映画が人の認知に適しているという仮定をもとに，日常的な出来事を自動で映像生成するシステムを提案する．なお映像化する際のルールは映画の解析をもとに作成する．

2. システムの構成

本システムは，図１のようにイベント導出，構図ルール，固定カメラ選定，ショット映像生成で構成されている．

イベント導出では，撮影空間と呼ばれる 4.0m ×4.0m 空間に設置した固定カメラによって撮影された映像を処理することによって人物の位置，人物の向き，人物の領域を導出する．また登場人物に装着したマイクから拾われた音声を処理することによって発話の有無を導出する．これらの情報から起きている出来事を判定する．

固定カメラ選定では起きている出来事から，映画の解析結果から求めた構図ルールを元に，最も映画的に撮られているカメラを決定する．

ショット映像生成では，決定したカメラ映像にトリミング処理を行い，ショットサイズを変更し

た後，それらを繋ぎ，動画として出力する．

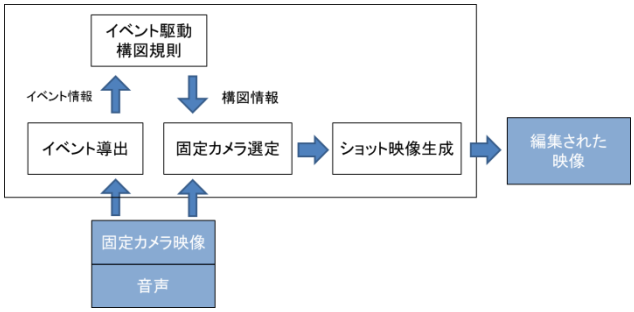


図 1　システムの全体図

3. 撮影空間

撮影空間は縦横 4.0m，床は 0.4m 間隔で 10×10 マスの格子状になるように区切られている．この撮影空間の周りに固定カメラを設置する．

固定カメラは，図 2 のように座標(0,0)にあたる部分から，時計回りに 1 から 8 まで番号を振ってある．固定カメラの高さは約 1.6m で，撮影空間の中心から縦横に 3.2m，対角線上に 4.53m の距離で設置する．

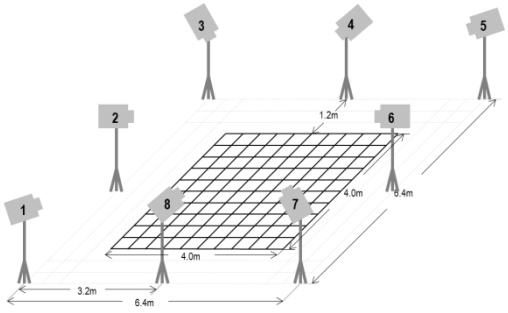


図 2　撮影空間

4. イベント導出

図 3　イベント導出の流れ

イベント導出は図 3 のように画像処理，音声処理，イベント検出で構成されている．

4.1. 画像処理

画像処理では，固定カメラの映像を 0.2 秒ごとにサンプリングし，静止画として保存しておく．各画像のピクセルの輝度値の推移から各ピクセルを動状態(白)，静状態(グレー)，背景状態(黒)の 3 つの状態に分類する．例を挙げると，図 4 では(a)は元画像，(b)は背景画像で，各ピクセルの輝度値の推移を見ることで(c)のような 3 値画像が得られる．この画像のノイズを処理した画像が(d)になる．この 3 値画像から「人物領域」，「顔の向き」を検出する．

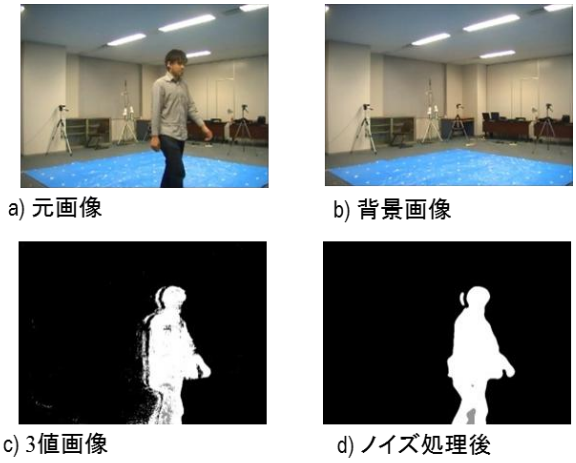


図 4　3 値画像の生成

4.1.1. 人物領域の決定

ピクセルの状態によって 3 値化された画像から人物が画面上のどの位置にいるのかを検出する．画像の左上を原点(0,0)として，水平方向を x 座標，垂直方向を y 座標としたとき，画像の原点から x 軸方向に 1 ピクセルずつ見ていき，隣のピクセルと状態が同じところを記憶していく．横の列を上からすべて見ていったとき，一番小さな座標が人物の左側の座標，一番大きな座標が人物の右側の座標となる．これと同じ動作を y 軸に沿って行い，人物の上端の座標と下端の座標が決まる．この結果，図 5 のように領域の上下左右の四端点と重心を求める．

図 5　人物領域

4.1.2. 顔の向き検出

8 台のカメラ全ての静止画に顔認識を行い，正面の顔を検出したカメラ番号をもとに人物の向きを計算する．人物の向きは撮影されたカメラから見て図 6 左のどの方向を向いているかを番号で示している．図 6 右は 5 番のカメラから撮影された映像で正面の顔が検出された場合，5 番のカメラの人物の向きは 8 となる．

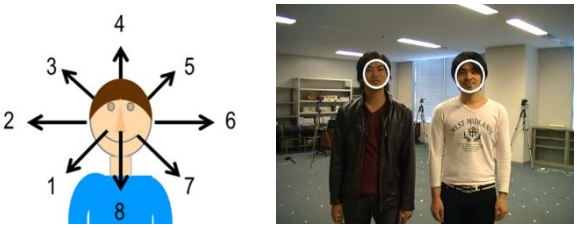


図 6 顔の向き

4.2. 音声処理

ワイヤレスマイクによって録音された音声を wave 形式で保存し，波形情報を抽出し処理を行っていく．wave ファイルはファイルの先頭部分にヘッダ情報(チャンネル数・サンプリング周波数など)が付加されており，ヘッダ情報の後に音声の波形データが記録されているものである．なお，サンプリング周波数は 44100[Hz]である．図 7 のように，抽出した波形データから発話と大声閾値を決め，それ以上の振幅が現れたフレームを記憶していく．発話の閾値は 8000，大声の閾値は 20000 としている．

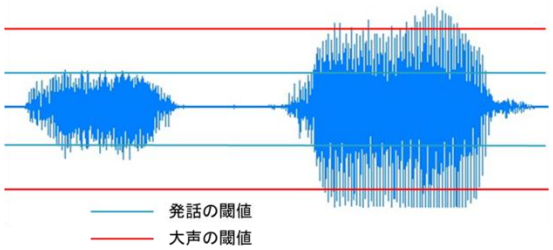


図 7 音声波形

4.3. イベントの検出

本研究内で扱うイベントを決定する際に，映像編集に関する著書である映画文法[3]と映画の解析結果を参考にした．

映画文法に記されているイベントは，以下の 3 つに分類することができる．

① 「人物が静止している」
② 「人物が移動している」
③ 「人物が会話(発話)している」

これらを踏まえたうえで，映画“ゴーストバスターズ(1984)”をショットに分割し，人物の動作を抜き出す解析を行った．その結果表れた上位 7 つの動作を表に示す．

表 1 解析結果

動作	回数
発話	777
移動	635
見ている	298
振り向く	151
聞く	89
立つ	35
座る	35

全 3175 回現れた動作のうち上位 7 つで約 63%を示している．この動作のうち，人物が静止しているに該当するものが「振り向く」，「立つ」，「座る」であり，人物が移動しているに該当するものが「移動」であり，人物が会話しているに該当するものが「発話」である．よって本研究では「発話」，「人物の移動」，「向きの変化」，「姿勢の変化」の 4 つをイベントと定義した．

なお，「見ている」，「聞く」は動作として認識し辛いため，今回はイベントとして扱わないこととした．

4.3.5. 人物の移動

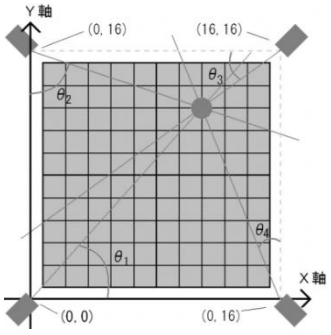


図 8 人物の座標

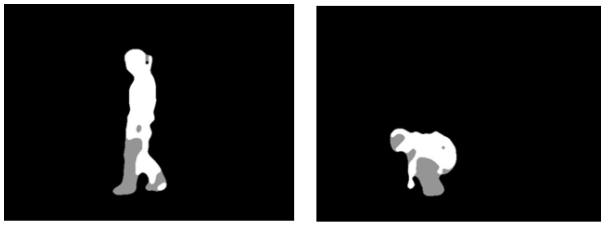
人物領域の中心点とカメラの画角の関係から，図 8 のように直線を引き，その交点から人物の空間上における座標を求める．この座標が 15 フレームごとに変化した場合，人物の移動が起きていると判定する．

4.3.2. 向きの変化

顔認識で人物の正面から移したカメラの番号を向きとして，30 フレーム以内に向きが 90 度以上変化した場合に向きの変化を検出する．

4.3.3. 姿勢の変化

姿勢の変化は，人物領域の上下左右の四端点のうち上下の点の座標を用いて計算する．上下間の幅が 10 フレームで 50 ピクセル増減した場合，姿勢の変化を検出する．図 9 では，(a)から(b)の状態に変化した場合，イベントは「座る」という動作になる．



a) 立っている状態　b) 座っている状態

図 9　姿勢の変化

4.3.4. 発話の有無

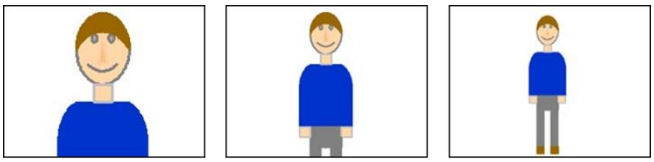
音声処理で発話と大声それぞれの閾値を超えたフレームを検出しているので，そのフレームがどれだけ多く出現するかによって発話と大声があるかどうか判定する．大声の場合，閾値を超えた値が 6 フレームの間に 50 回表れた場合大声を検出する．また，発話の場合は大声の閾値には満たないが発話の閾値を超える値が 6 フレームの間に 50 回表れた場合発話を検出する．

5. ショット解析

映画をショット毎に分けていき，必要な情報を抜き出していく作業を本研究ではショット解析と呼ぶ．ショット解析の目的としては，起きている出来事に対して，人がイメージする最も典型的な構図は何であるかを調べるためである．

映画における構図は映画制作者が長年積み上げてきた知識をもとに，視聴者に対して最も起きている出来事が伝わりやすい構図を使用していることを仮定として，ショット解析を行っていく．

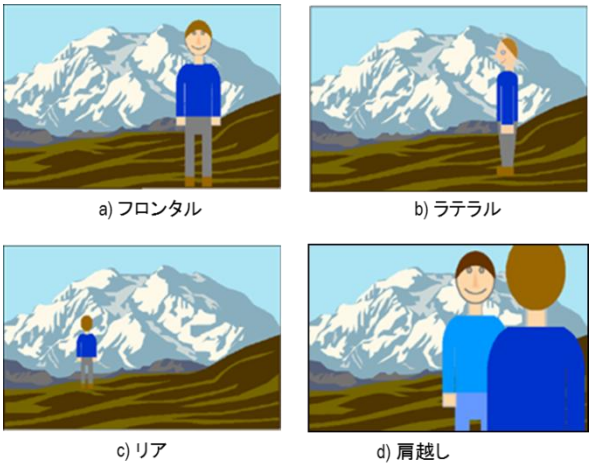
ショット解析により抜き出す情報は，出来事(人物の動作)に対するショットサイズ，コンポジションの情報である．ショットサイズは被写体の大きさを示しており，図 10 のように全身を映したフルショット，腰から上を映したミディアムショット，胸から上を映したクロースショットがある．



a) Close Shot　b) Medium Shot　c) Full Shot

図 10　ショットサイズ

コンポジションは人物の映り方の立ち位置を示したものである図 11 のように(a)のような正面を映したフロンタル，(b)のような横から映したラテラル，(c)のような後ろから映したリア，(d)のような人物の肩越しに映した肩越しがある．



a) フロンタル　b) ラテラル　c) リア　d) 肩越し

図 11　コンポジション

5.1. 構図ルールの構築

ショット解析では，動作の記述が自然言語で自由に表記されている．そこで動作を以下のパラメータからなるイベントとして分類する．各パラメータは整数で表す．

① 人物の移動

人物の移動が無い場合は0とする. 人物が移動する場合，画面の手前を8とし時計回りに1から番号を振り，その方向を記述する.

② 向きの変化

人物の向きが90度以上変化した場合に1とする. 90度以上変化しない場合は0とする.

③ 姿勢の変化

立っている状態を0とする．立っている状態から座った状態に変化した場合は1とする.

座っている状態から立った状態に変化した場合は2とする.

④ 発話

登場人物のセリフが入った場合を1とする．さらに，人物が叫ぶなどの大きな声を出している場合は大声として値を2にする．セリフも大声もない場合は0とする.

6. カメラの選定

カメラの選定では，まず映画編集の原則の 1 つであるイマジナリーラインの原則によってカメラを制限する．この原則はカメラの切り替えを行う際に，二人の人物を結ぶ線を超えた反対側のカメラを使用すると視聴者が理解しづらい映像になってしまうことを防ぐためである．図 12 は撮影空間上でのイマジナリーラインの例を示している.

その後，映画のショット解析をもとに作成した構図ルールを利用し，コンポジションの向きと人物の向きの一致したカメラを決定する.

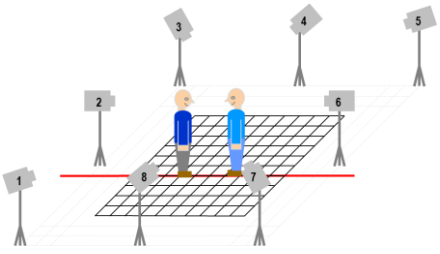


図 12 撮影空間上のイマジナリーライン

7. ショット映像生成

ショット映像生成では，イベント導出から人物領域，固定カメラ選定から使用するカメラ番号とショットサイズの情報を受け取る．受け取った情報を基に，1フレームごとにトリミングと拡大をする．図13はショットサイズがミディアムショットの場合のトリミング処理を施した結果である．その後，トリミング処理を行った各静止画を繋ぎ合せて動画として出力する.

a) カメラの元映像　　b) トリミング後の映像

図 13 カメラ映像のトリミング

8. 結果

縦横 4m の空間内で起こる人物の行動を周りに設置した 8 台のカメラで撮影し，得られた 8 台分のカメラの映像を本システムに入力して解析することで，自動編集システムの有用性を確かめた．イベント導出では，二人の人物が重なってしまった場合でも人物ごとの領域情報を求めることができた．これに伴い，人物の座標も誤検出が減り，イベントをより高い精度で導出できるようになった．固定カメラ選定では，人数，イベントを入力とすることで，コンポジションに基づいたカメラ選択を行えるようになった．システムの出力としては，特に違和感の映像が出力できた.

9. 考察

イベント導出では，カメラ切り替えに必要な情報を出力することに成功したが，一部想定外の数値をとることがあった．これは，ピクセル輝度状態の判定が上手くできておらず，図14のように本来であれば動状態と判定されるべきピクセルが別の状態で判定され，結果として人物として認識されていないというものであった．このことから，判定式の見直しや閾値の変更などによって精度を上げる必要があると考えられる．顔検出においても人物の顔を検出できないという例が見られた．これは，人物の重なりによって顔の正面をとらえることができない状況が

起きてしまうことが問題であった．人物が正面で会話をする場合しばしばこのような状況が起きるため，正面の顔検出のみではなく横顔の検出も同時に行うことで解決できるのではないかと考えられる．また，今回輝度値の変化から映像を見ていったが，同じ被写体を映した8台のカメラ映像の輝度変化には，共通の変化を表す箇所が見られた．これを利用し，これまで手動で行ってきたカメラ映像の同期が自動で行えるのではないかと考えられる．

図 14　誤検出

ショット解析において，今回，出来事に対して最も典型的な構図として出てきたものが，本当に人のイメージする 構図であるのか検証を行っていないため，今後検証を行う必要がある．

固定カメラ選定では，構図ルールの構築の際に，人数とイベントの組み合わせによっては，出現回数が1回やまったく出てこない人数とイベントの組み合わせもあった．これは，ショット解析のサンプル数が少ないのが原因だと思われるため，ショット解析のサンプル数を増やす必要があると考えられる．サンプル数を増やすことで，今回出現しなかった人数とイベントの組み合わせや，出現回数が同じになってしまい，決定できなかった人数とイベントの組み合わせを1つに決定できるのではないかと考える．さらに，csv形式で記述する場合，二人の人物が同時に行動を起こした場合の記述方法を考える必要がある．

ショット映像生成では，トリミングをし，動画として出力することができた．しかし，人物が二人の場合にコンポジションやショットサイズ通りにトリミングすることができない場合が多かった．そのため，固定カメラ選定では，トリミングをする範囲を考慮したカメラの選択をする必要があると考える．

10. 結論

ショット解析から求めた構図ルールを利用して自動編集するシステムを構築することができた．

しかし，出来上がった動画の評価実験や，ショット解析結果と人のイメージの検証などを行っていく必要がある．

また，ピクセル状態分析などの各処理の精度の向上や，映像の同期の自動化，さらには，構図ルールに利用するショット解析のサンプル数をさらに増やしていくなど，多くの問題を改善していく必要がある．

参考文献

[1] J. J. Gibson, (1985) “生態学的視覚論”, 古崎敬訳, 株式会社サイエンス社, pp. 315.

[2] J. M. Zacks, (2008) “Event perception”, Scholarpedia, 3, 3837.

[3] D. Arijon, (1980) “映画の文法”, 岩本憲次・出口丈人訳, 紀伊国屋書店

[4] 尾形 涼, 中村 裕一, 太田 友一, (2004) “制約充足と最適化による映像編集モデル, ” 電子情報通信学会論文誌, vol.J87-D-II, no.12, pp.2221－2230.

[5] 西崎 隆志, 尾形 涼, 中村 祐一, 太田 友一, (2006) “会議シーンを対象とした自動撮影・編集システム, ”電子情報通信学会論文誌, vol.J89-D, no.7, pp.1557-1567.

[6] 井上 亮文, 吉田 竜二, 平石 絢子, 重野 寛, 岡田 謙一, 松下 温, (2004) “映画の映像理論に基づく対面会議シーンの自動撮影手法, ” 情報処理学会論文誌, vol.45, no.1, pp.212-221.

[7] 村上 正行, 西口 敏司, 亀田 能成, 角所 考, 美濃 導彦, (2005) “京都大学での実践に基づく講義アーカイブの調査分析, ” 日本教育工学会論文誌, pp.253-262.

[8] 冨野 由悠季, (2011) ”映像の原則 改訂版”, キネマ旬報社.

[9] 西 貴行, 藤吉弘亘, 梅崎太造, (2004) ”ピクセル状態分析による固定監視映像の圧縮”,第 10 回画像センシングシンポジウム