ちょっとした出来事の自動映像編集：映画文法に基づき撮影されたターゲット映像の参照による理解しやすい映像の生成
Automatic Video Editing of a Minor Event: Generation of Easily Comprehensible Image by Reference to Target Image Shot on the Basis of Film Grammar

古川 智裕[†]，金谷 友樹[†]，榎津 秀次[‡]
FURUKAWA Chihiro, KANAYA Yuki, ENOKIZU Hideji

[†] 芝浦工業大学大学院工学研究科，[‡] 芝浦工業大学工学部
[†] Graduate School of Engineering, Shibaura Institute of Technology [‡] Shibaura Institute of Technology
[†] m11138@shibautra-it.ac.jp, [‡] enokizu@sic.shibautra-it.ac.jp

Abstract

In the present study, we proposed the mechanism of automatic video editing that uses target image on the basis of the film grammar to generate the easily comprehensible image of a minor everyday event. Several minor everyday events were shot by eight digital video cameras set around the shooing space. On the other hand, we have prepared the target image of each everyday event shot on the basis of the film grammar previously. Two stages are primarily needed to generate easily comprehensible image. In the first stage, an image, which was shot from the most appropriate camera position, is selected by comparing eight images with the target image. Then, in the second stage, the selected image is cropped and zoomed by reference to the target image. Some images were generated by automatic video editing system that implemented a series of picture processing involved with these two stages. Generated images were similar to each target image appeared to be easily comprehensible. However, we found some problems, for example, estimating appropriate vale of the target image and segmenting the target image, to be overcome.

Keywords — Automatic Video Editing, Film Grammar, Everyday Event, Easily Comprehensible image

1. 研究目的

近年，ビデオカメラや映像編集ソフトの低価格化・高性能化などにより手軽に映像を撮影・編集することができるようになった．しかし，それにより作成された映像とテレビや映画などの映像では理解のしやすさに明らかな違いがある．これには様々な理由が挙げられるが，その中でも特に大きな問題として編集の問題がある．テレビや映画の編集を行っている人たちは長い年月をかけて自分たちが培ってきた知識や技術を用いて編集を行っている．しかし我々がその編集の知識や技術を身につけようと思うと多くの時間や資金などが必要となり困難である．そのため，映像を自動的に撮影し編集する様々な研究が進められている．先行研究[1][2][3][4]では，出力された映像がわかりやすい映像なのかどうかの判断を主観評価などによって求めていた．そこで本研究では映画文法[5]と呼ばれる映画編集のルールに基づき製作された映像を元に，その映像に近い映像をシステムにより撮影・編集することで，視聴者にとって理解しやすい映像を制作することを目的にする．

2. ターゲット映像

ターゲット映像とは，映画文法と呼ばれるルールに則って撮影・編集された映像のことである(図 2.1)．映画文法とは，映画制作関係者などが経験則的に培われた知識を自然言語で表現したもので，視聴者が映像を見たときに意識しなくても映像を理解できるような編集方法が記されたものである．映画のあるワンシーンをショットという映像の単位に分け，図 2.2 のように，人物の動きやカメラの配置，カメラで撮影した映像の代表的なものを描いた図とともに，ショットサイズや，カメラの操作方法などが自然言語で記されている．シーンとは，ある定められた空間の中で起こった出来事（イベント）の流れであるとする．ショットとは，時間的・空間的な切れ目なしに連続して撮影された映像の単一断片を指すものである．また，ショットサイズとは人物の体のどの部分でカ

メラを切るか(カッティング)の高さのことを指し，バストショットやミディアムショットなど複数の種類がある．なお，本研究で使用するショットサイズは人物の肩から上が映るクロースショット(CS)，人物の腰から上が映るミディアムショット(MS)，人物の全身が映るフルショット(FS)の 3 種類とした．これは，見た目として違いがわかりやすく，カッティングの場所が比較的明確であるからである．

図 2.1　ターゲット映像

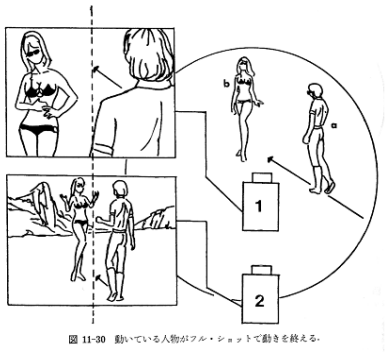

図 2.2　映画文法内の図

3.　イベント

イベントとはイベント情報より求められる，ある定められた空間の中で起こったカメラの切り替えに関わる動作を指す．本研究は移動・向きの変化・姿勢の変化・発話の 4 つをイベントとして扱っている．この 4 つのイベントの組み合わせを 4 次元配列で表記する．たとえば，移動と発話が同時に起こった場合は．(5,0,0,1)のように表す．なお，この移動の値は移動の向きを表している．

4.　システム構成

本システムでは，図 5 に示すようにイベント導出部，カメラ決定部，ショット映像生成部で構成されている．イベント導出部では，撮影空間で撮影された映像からイベント情報を導出し，イベントを決定する．カメラ決定部ではイベント導出部で導出されたイベントやイベント情報などを使い，ターゲット映像にもっとも近いカメラ映像を求め，カメラ番号を出力する．ショット映像生成部では，イベント導出部，カメラ決定部で得られた情報よりカメラ映像をターゲット映像に近づけ，音声を合成し動画として出力する．

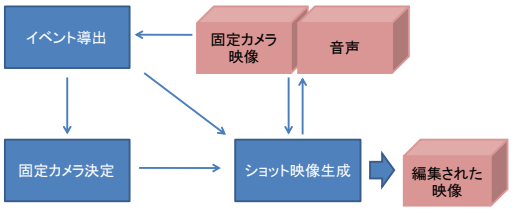

図 4　システム全体の流れ

5.　撮影空間

図 5 のように撮影空間とは，縦横に 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る．この撮影空間のまわりにカメラ 8 台用意し，高さ約 1.6m，カメラの中心が撮影空間の中心(5,5)を映すように設置する．左下のカメラ番号をカメラ 1 番とし，時計回りに 2 番，3 番…とする．このカメラの高さは映画での一般的なアイレベルを参考に人物の目線の高さに合わせた．その理由としては，アイレベルで撮影された映像は視聴者が普段見ている世界との見え方が同じなため安定感や安心感が得られるからである．

ターゲット映像に対し，撮影空間の固定カメラで撮影された映像のことを固定カメラ映像とする．

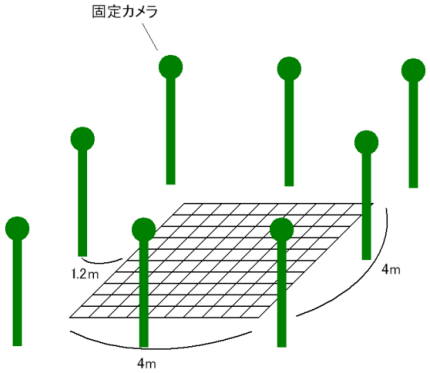

図 5　撮影空間

6. イベント情報

イベントを決定するために必要な情報のことで，撮影空間内に写っている人数，撮影空間内で人物のいる座標，人物の正面を検出したカメラ，人物の姿勢，発話の有無のことを指す．人物のいる座標が変わった場合移動を検出，そのときに正面を向いているカメラ番号をイベントに格納する．人物の向きが 1 秒以内に 90 度以上変わった場合にはイベントの向きの変化の値を 1 に，姿勢が 1 フレームの間に 30 ピクセル以上減少した場合人物の姿勢の値を 2(座る)にし，30 ピクセル以上増加した場合は人物の姿勢の値を 1(立ち)にする．また，人物の発話があった場合には発話の値を 1 にする．

7. 人物領域情報

人物領域情報はイベント情報とは異なり，撮影された固定カメラ映像の各フレームの中で人物がどのように映っているかを表す画像上の情報である．ここでカメラごとに抽出する情報を以下に示す．

1) 人物の重心点
2) 人物領域の右端，左端，上端，下端の各座標
3) 顔検出の中心点，半径の大きさ

これらの情報はイベント情報を出力する際に同時に出力できるため，人物領域情報の出力のためには新たな画像処理をすることはない．なお，ここでの座標はカメラの画像上の座標のことで，左上を始点としたピクセル数のことである．

8. イベント導出部

イベント導出部でのイベント情報導出方法について記述する．まず，人物の位置については各固定カメラから取得した映像を 0.2 秒ごとに静止画として保存したものを利用し，解析をする．固定カメラごとに取得した背景のみの画像(背景画像)を読み込み，フレーム単位での解析を行う．そして，各固定カメラから撮影空間内の人物に対して直線を引き交点を求めることで，人物の位置を検出する．解析手順を以下に示す．

1) 入力画像と背景画像のグレースケール化
2) 入力画像と背景画像の平滑化
3) 入力画像と背景画像の差分を取る
4) 差分画像の二値化
5) ノイズ除去
6) ラベリングをし，人物領域の重心を算出
7) 固定カメラからの角度計算
8) 各固定カメラから直線を引き交点を求める

それにより求められた交点に一番近い座標に人物がいるとし，人物の位置を導出している．なお，人物の位置は(0,0)～(10,10)で表現する．

人物の向きは撮影空間のまわりにある 8 台のカメラをすべて使って検出する．まず，8 台のカメラ全てで顔検出をし，検出された場合，その人物は検出されたカメラの方向を向いていると判断させた．なお，この顔検出には精度を高めるために色相による制限を行っている．

人物の姿勢に関しては，撮影空間のまわりにあるカメラを使い検出する．人物の位置を求めるために使用した背景画像との差分の情報とラベリングにより得られる情報より人物の姿勢を決定する．本研究では人物は立っている状態と座っている状態の 2 種類の姿勢を扱うものとし，人物が撮影空間に入ってきた場合無条件で立っているものとして扱っている．1 フレームである 0.2 秒の間に，人物領域の縦幅の値が 30 ピクセル減少した場合は座っている状態に姿勢の変化が起きる．なお，この 30 ピクセルという値は実際に姿勢の変化のある映像を解析して得られた結果である．また 1 フレームの間に人物領域の縦幅の値が 30 ピクセル増大した場合人物が立っている状態に姿勢の変化が起きたとしている．

音声に関しては，登場人物に装着したワイヤレスヘッドセットマイクによって拾われた音声を解析することによって導出する．録音した音声を 0.2 秒ごとに区切り，それを一つの単位とした．登場人物が発話をした場合，音声データの振幅が大きくなる．それを利用して，振幅がある閾値を超えた回数が 0.2 秒間に 50 回以上だった場合人物の発話として検出する．

こうして求められたイベント情報をもとに，イベ

ントを決定していく．まず，移動に関しては求めた座標に変化が 1 秒以上あった場合，その間で移動のイベントを検出する．イベントを検出した場合，同時に移動している固定カメラの番号を移動の方向として出力する．向きの変化は，得られた顔の向きが 1 秒以内に 90 度以上変化した場合その最初のフレームから 1 秒間を人物の向きの変化として検出する．姿勢の変化は，人物の姿勢が切り替わったときに，その最初のフレームから前後 1 秒間を姿勢の変化とする．発話は，音声処理により求めた人物の発話があった場合に発話を検出する．

9. ターゲット映像記述情報

ターゲット映像記述情報とは，ターゲット映像をカメラ映像との類似度を比較できるように必要な情報をテキスト形式でまとめたものである．1 フレームごとに抜き出す情報を以下に示す．

1) ターゲット映像のフレーム番号
2) 人物番号
3) 人物の位置
4) 人物の向き
5) 人物のショットサイズ
6) イベント

ターゲット映像を 0.2 秒ごとの静止画に分割し，分割した最初の画像を 0 フレームとしてフレーム番号を決定する．人物の位置は背景画像より背景差分を用いて求められた領域の重心点と画像左端の間の距離を位置とする．人物の向きとイベントの移動の向きに関しては分割された画像の顔の向きを図 9 の 8 方向から選択する．移動の向きは前後のフレームを見て移動している方向を選択する．ショットサイズは映画文法にあるショットサイズの中から人物の全体が映るフルショット(FS)腰から上が映るミディアムショット(MS)人物の顔が中心に映るクロースショット(CS)の 3 種類より選択する．イベントは，イベント導出と同じく移動・向きの変化・姿勢の変化・発話の 4 つをイベントとして扱う．この 4 つのイベントを 4 次元配列で表記する．

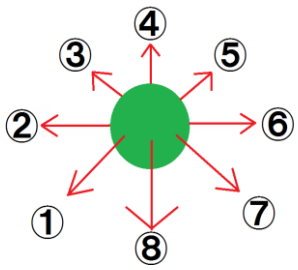

図 9　人物の向き

10. カメラ映像記述情報

イベント導出部より得られた情報より，撮影空間上のカメラの映像をテキスト形式で表す．フレーム単位で表す情報を以下に示す．

1) カメラ映像のフレーム番号
2) 人物番号
3) 人物の位置
4) 人物の向き
5) イベント

この値はイベント導出部で得られるイベント情報とイベント，人物領域情報と対応している．フレーム番号と人物番号はイベント情報と共通．人物の位置は人物領域情報の人物の重心の座標の値が入る．向きはイベント情報にある向きと対応しており，イベントはイベントが格納される．なお，イベント情報で与えられている向きは撮影空間上のカメラの番号となっているが，カメラ映像記述情報ではターゲット映像記述情報の向きと合わせるため，前もって変換してある．

11. カメラ決定部

カメラ決定部では，イベント導出部より得られた情報をもとに，カメラ映像記述情報を生成する．それと前もって作成しておいたターゲット映像記述情報と比較することで，どのカメラの映像が最もターゲット映像に近いのかを決定する．

まず，ターゲット映像の最初のイベントを見て，それと一致するイベントをもつカメラ映像記述情報をフレーム単位で全て抜き出す．さらに，その中からターゲット映像の人物向きが一致するカメラ映像を全て抜き出す．この抜き出されたカメラ映像記述情報のフレーム番号が連続している部分

を 1 つのショットとして扱う．この時のフレーム番号と最適カメラ番号をショット映像生成部に送る．また，次の処理で画像処理であるトリミングを行うかどうか，ターゲット映像記述情報のショットサイズと上端，下端，右端，左端の 4 つの端点の情報をトリミング情報として求める．
トリミングを行うかの判断は，ターゲット映像のショットサイズがフルショットでない場合はトリミングを行うとする．
これをターゲット映像全てのフレームに関して行う．

12. ショット映像生成部

ショット映像生成部ではカメラ決定部で決定した最適カメラからターゲット映像記述情報，イベント導出部の情報より最適カメラの画像に画像処理を行い，よりターゲット映像に近い映像を生成する．具体的には，ターゲット映像記述情報のショットサイズの情報をもとに最適カメラの画像をトリミングし，ショットサイズを一致するように画像処理を行う．前処理であるイベント導出部とカメラ映像決定部から，人物領域情報とイベント情報より人物領域の上端，下端，右端，左端の座標情報と人物領域の重心点の座標．さらに，カメラ映像決定部により決定した最適カメラの番号とトリミング情報を受け取る．その情報からカメラ映像をトリミングする．その切り取りの座標を決定する手順を以下に示す．

1) 切り取りを行う始点の決定
2) 横幅(width)の決定
3) 縦幅(height)の決定
4) 得られた値をもとに画像を切り取る

なお，縦幅と横幅の決定にはショットサイズごとのアルゴリズムによって決定する．次にショットサイズごとのアルゴリズムを示す．

まずクロースショット(CS)の場合は，人物領域情報の上端の y 座標と人物の重心点の座標から顔の中心点を求める．次に人物の顔の領域を円と見立てて中心点の座標と人物の上端の座標から顔の大きさの半径を求める．これにより求まった顔領域の半径と人物の重心から縦幅 height を決定する．横幅と縦幅の比はカメラ映像の比と同じく 4:3 としているため，縦幅が決まることで横幅 width も決定する．これより，顔の中心点が中心になるように始点を決定し，切り出しを行う．

次に，ミディアムショット(MS)の場合はトリミング情報中の人物領域の左端と右端の値より画像を 3 分割して比を求める．次に固定カメラの画像に移り，人物領域情報より人物の左端と右端の値が得られる．その差を先ほどの 3 分割した比の中央の値と対応させ，左右の幅のピクセル数を決定する．すべてのピクセル数の値を足した数値が横幅 width となる．また，縦幅 height の値も同時に求まる．次に始点を求める．ミディアムショットは腰の高さを下限とするため，人物の重心点の y 座標を切り出しの下点となるよう始点を決定した．

最後にフルショット(FS)の場合は，本研究の撮影空間で撮影された映像のショットサイズは人物の全身が移るフルショットであるため，フルショットに関しては画像処理によるショットサイズの変更を行わず，固定カメラ映像を使用した．

これにより得られた画像と音声を合成し，生成動画を生成する．なお，フレームレートはカメラ映像を静止画に分割したときと同じく 5fps とした．

13. 結果

結果として，ある一定の条件下での撮影空間上での人物の認識から最適カメラの決定．また，ターゲット映像と同じショットサイズへの変更を行うことができた．ただし，固定カメラで映る人物の背景差分より得られる人物領域が重なってしまった場合，領域がうまく検出できなかった．これは人物領域が重なることで 1 つの領域として認識されてしまったことが原因だと考えられる．また，複数の人物に対しての人物番号の割り当てがターゲット映像記述情報では撮影空間に入ってきた順番なのに対し，カメラ映像記述情報では領域の x 座標が小さい順番で行っていたため，ターゲット映像記述情報の値を検出しやすいように変更する

必要があった.

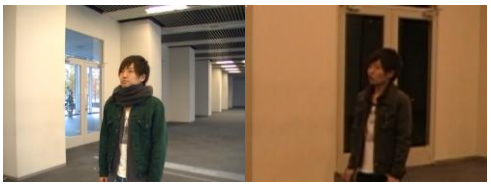

図 13.1　ターゲット映像(左)と出力映像(右)

図 13.2　人物領域が重なってしまう場合

14. 考察

これにより, 撮影空間上での動作をターゲット映像に近い形で出力することができた．しかし，今回はターゲット映像のパターンが少なく，またターゲット映像中のイベントと撮影空間上のイベントとの順番，内容が一致していないとうまく出力することができなかった．そのため，ターゲット映像となる映像をより多くのパターン用意する必要がある．また，ターゲット映像のイベントを 1 つ，または 2 つ程度とし，ターゲット映像を組み合わせることでカメラ映像のイベントと対応させていくなどの工夫が必要になると考えられる．トリミングにより出力された映像は画質が粗く見にくいものになってしまったが，今後カメラの画質や画素数が上がるにつれて画像の拡大を行っても視聴に耐える映像を作れると考えられる.

図 14　画像の劣化
(左:ターゲット映像 右:トリミング後の映像)

参考文献

[1] 金谷 友樹, 梶山 大介, 榎津 秀次, “撮影空間におけるイベントの流れの自動撮影・編集―映画文法に基づくショット選択ルールの適用―” 電子情報通信学会技術研究報告, Vol.110, No.33, pp.125-130(2010)

[2] 尾形 涼, 中村 裕一, 大田 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.

[3] 西崎 隆志, 尾形 涼, 中村 祐一, 大田 友一, (2006) “会話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.

[4] 足立 順, 滝口 哲也, 有木 康雄, (2007) “固定カメラ映像からの音声・画像情報を用いた映像コンテンツの生成”, 画像の認識・理解シンポジウム.

[5] Arijon,D.（著）, 岩本 憲児, 出口 文人（訳）(1980) “映画の文法”, 紀伊国屋書店.