# 機械学習による映像文法の自動獲得

吉岡　希† 山本　岳洋†† 日置　淳也†† 大島　裕明†† 荘司　慶行††† 田中　克己†

† 福知山公立大学 情報学部　〒 620 – 0886 京都府 福知山市 字堀
†† 兵庫県立大学 情報科学研究科　〒 651 – 2197 兵庫県 神戸市 西区 学園西町
††† 青山学院大学 理工学部　〒 252 – 5258 神奈川県 相模原市 中央区 淵野辺
E-mail: †{32045104,tanaka-katsumi}@fukuchiyama.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp,
†††ad22w056@gsis.u-hyogo.ac.jp, ††††ohshima@ai.u-hyogo.ac.jp, †††††shoji@it.aoyama.ac.jp

**あらまし**　本論文では，映画映像を機械学習して映像文法を自動獲得する手法とシステムについて述べる．映像文法は映像の撮り方に関する多くの規範であり，例えば，ショットサイズ，アングル，映像の対称性（シンメトリ），構図（開いた・閉じた構図），POV ショット，ショット・リバースショットなど多様な映像文法が知られている．本研究では，いくつかの映像文法に限定して，映画映像を機械学習して新たな映画映像の映像文法を獲得する仕組みを提案する．

**キーワード**　映像文法，映画，機械学習

## 1　はじめに

映像文法とは，ショットの撮影やショットのつなぎ方に関する基本的な規則である [1]．映像文法は映像の撮り方やつなぎ方に関する多くの規範であり，例えば，ショットサイズ，アングル，映像の対称性（シンメトリ），構図（開いた・閉じた構図），POV ショット，ショット・リバースショットなど多様な映像文法が良く知られている．

本論文では，映画映像を機械学習して映像文法を自動獲得する手法について述べる．本研究では，いくつかの映像文法に限定して，映画映像を機械学習して新たな映画映像の映像文法を獲得する仕組みを提案する．具体的には，映像文法としてバストショット，クロースアップ，シンメトリの 3 つの映像文法に限定し，与えられたフレーム画像から文法を予測する問題に取り組む．

第 2 節では，関連研究，特に，映像文法に関する研究を紹介し，本研究の位置づけを示す．第 3 節では本研究における提案手法を提示し，実際に行った実験に関して概要を示す．第 4 節では，映画映像を機械学習して新たな映画映像の映像文法を獲得する仕組みを提案する．第 5 節では，本研究で行った結果について評価・考察し，第 6 節ではショット系列で判別する映像文法の獲得について提案し，第 7 節では，結論と今後の課題について述べる．

## 2　関連研究

映像文法とは，ショットの撮影やショットのつなぎ方に関する基本的な規則である [1]．映画の文法については，ダニエル・アリホン（Daniel Arijon）の解説 [2]，ジェニファー ヴァン シルの解説 [3] などがある．

また，国内での，映像文法の解説については今泉らの著作 [4] [5] が代表的なものである．これらの中で，ショットサイズ，アングル，映像の対称性（シンメトリ），構図（開いた・閉じた構図），POV ショット，ショット・リバースショットなど多様な映像文法が解説されている．図 1 は，シンメトリが用いられている映画の例 [4] [5] である．また，図 2 は，POV ショットが用いられている映画のイラスト例 [5] である．最後に，図 3 は，エスタブリッシングショットが用いられている映画のイラスト例 [5] である．


映画「初恋のきた道」


映画「暖流」

図 1　シンメトリ（映像文法）の映画イラスト例 [4] [5]

映画映像のショット分類を機械学習で行う研究 [6] も開始されており，この研究では，映画映像のショット分類（ショットサイズの分類）を ResNet50 を用いて行っている．

## 3　提案アーキテクチャ

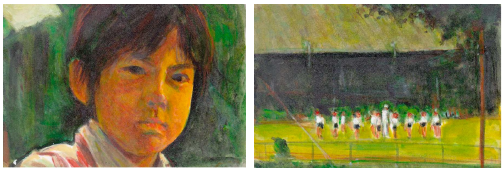
機械学習によって獲得する映像文法を以下の 2 種類に分類し，それぞれの映像文法を獲得することを第 4 節，第 6 節で述べる．本研究では特にショット単体で判別できる映像文法について，ResNet18 を用いて分類を行う．また，もう 1 つの実験として Mask R-CNN を用いて背景除去した画像を入力データとして ResNet18 により分類を行う．これら 2 つの実験に関して差分を調査する．

（1）　ショット単体で判別できる映像文法

- ショットサイズ（ロング/バスト/クロースアップ等）

映画「Shall We Dance?」

映画「幸福の黄色いハンカチ」



映画「砂の器」

図 2　POV ショット（映像文法）の映画イラスト例 [5]

映画「メリー・ポピンズ」

図 3　エスタブリッシングショット（映像文法）の映画イラスト例 [5]

- ハイアングル/ローアングル
- シンメトリ/アシンメトリ（映像の左右対称性/非対称性）
- 閉じた/開いた構図（映像の構図の開閉）

（2） ショット系列で判別できる映像文法

- POV（Point of View）ショット

人物の眼のショット＋その人物から見える映像

- ショット・リバースショット

相対する複数の人物の，前面からのショット＋背面からのショット

- エスタブリッシングショット

シーンの冒頭などで，場所の状況や出演者の位置関係を認識させるためのショット群のこと．通常は，超ロングショット/ロングショットから始まり，パンショットが続き，最後にズームアップショットとなることが多い．

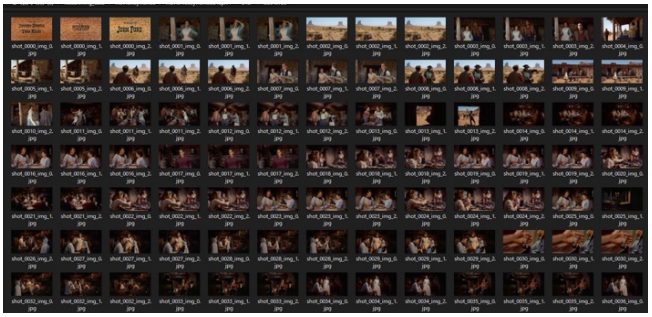


図 4　映画のショット毎のキー画像系列
（MovieNet データセット）

表 1　映画データの内訳

| | |
|---|---|
| train | 40 |
| validation | 8 |
| test | 8 |

| | train | validation | test |
|---|---|---|---|
| bust-shot | 252 | 223 | 300 |
| close-up | 244 | 300 | 300 |
| symmetry | 296 | 148 | 228 |

図 5　学習データの枚数内訳

## 4　ショット単体で判別できる映像文法予測

### 4.1　データセットの構築

本研究では，学習に用いる映画データセットとして，MovieNet [1]データセットの中の，Movie per-shot keyframes（240p）（161GB）（映画毎に，各ショットのキー画像（静止画像）を系列化したもの）を用いた [7]．図 4 に，Movie per-shot keyframes 中の 1 つの映画のキー画像系列を示す．MovieNet は映画理解のための総合的なデータセットを提供する組織であり，Huawei 社の有志らによって管理されている．Movie per-shot keyframes は MovieNet にアクセス後，OpenDataLab のページに移動しアカウントを登録した上でユーザサービス契約とプライバシーポリシーに基づいてデータをダウンロードする．データセット内は全て tar ファイルで管理されており，これらを解凍することで映画のキー画像系列を得る．tar ファイルは合計で 56 個あり，各ファイルが 1 つの映画の画像系列である．train に 40 個，validation に 8 個，test に 8 個の映画を割り当てアノテーションを行った（表 1）．Movie per-shot keyframes には合計で 246,047 枚のフレーム画像が存在する．tar ファイル 1 つに対し，1 本の映画が対応しているため，Movie per-shot keyframes は合計で 56 本の映画によって構成されるフレーム画像のデータセットである．

本研究では，Movie per-shot keyframes から顔が映っている画像のみをアノテーションし，学習に用いる．これは後述する

1：MovieNet: https://movienet.github.io

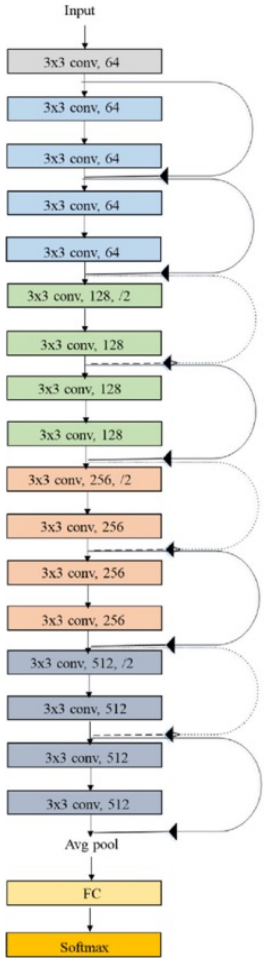



図 6　ResNet18 による残差学習によるショットの自動分類

図 7　バストショットとアノテーションした画像

Mask R-CNN により画像加工したデータを学習した場合との差分を検証するためである．アノテーションは今泉らの著書 [4] を参考に行う．図 5 にアノテーションした画像の枚数とその内訳を示す．

### 4.2　ResNet18 を用いた映像文法予測

畳み込みニューラルネット（Convolutional Neural Network, CNN）の拡張である残差学習（Residual Learning）ResNet18 [8] [9] を用いて，ショット単体で判別できる映像文法による画像分類を行った．ResNet18 の学習の仕組みを図 6 に示す．

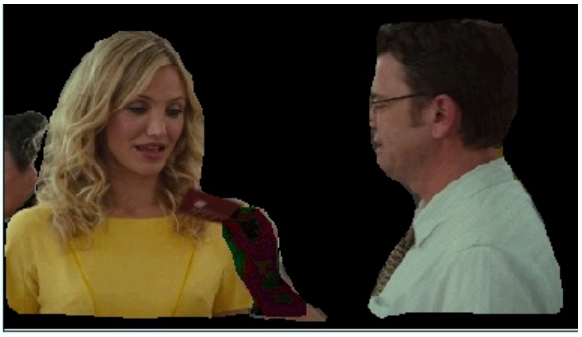


図 8　図 7 に対して Mask R-CNN により人物の部分と認識された部分以外を黒でマスクする加工を行った画像

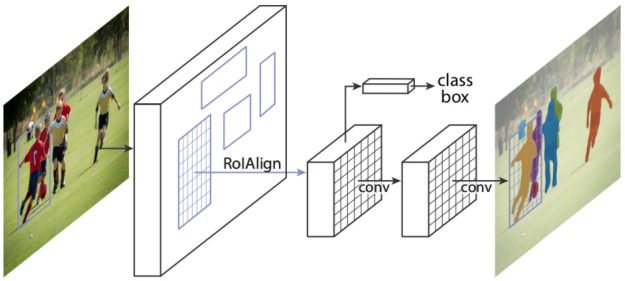



図 9　Mask R-CNN のフレームワーク [10]

表 2　ハイパーパラメータの設定

| | |
|---|---|
| learning rate | 2e-6 |
| epochs | 100 |
| patience | 10 |

### 4.3　Mask R-CNN による画像データ加工

Faster R-CNN を拡張した Mask R-CNN [10] を用いて，アノテーションした画像データに対し，人物領域を抽出し背景を取り除いた画像データを用意した．Mask R-CNN では物体検出やセグメンテーションを実現する手法であり，人物領域のみの抽出も可能である．Mask R-CNN のフレームワークを図 9 に示す．実際に Mask R-CNN を用いてアノテーションした画像を加工した例を図 7，図 8 に示す．

### 4.4　比 較 手 法

MovieNet より入手したデータセットをアノテーションした画像データを用いて機械学習を行った場合と，アノテーション後の画像を Mask R-CNN により人物領域のみを抽出した画像データを用いて機械学習を行った場合にどの程度正解率に差が出るのか比較を行った．以降これらはそれぞれ「背景除去なし」と「背景除去あり」と記述する．Mask R-CNN を利用するのは，判別する映像文法が，バストショット，クロースアップ，シンメトリであり，いずれも人物領域のみを抽出した画像で十分で有り，さらに，背景などを取り除くことで，認識性能も改善が見込めると考えたためである．

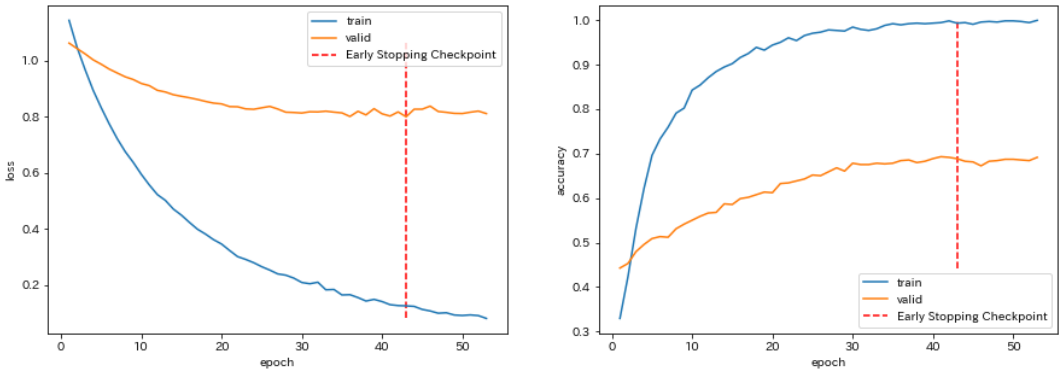



図 10　背景除去なしにおける学習曲線

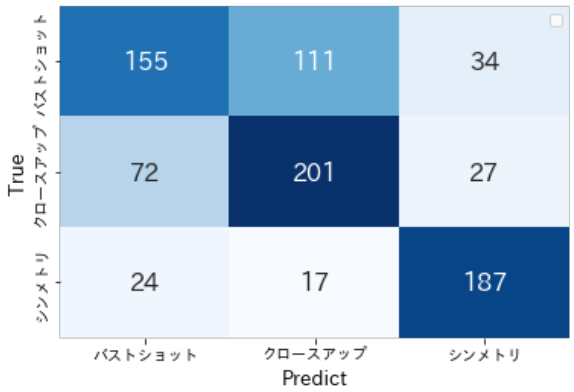



図 11　背景除去なしにおける混同行列

表 3　学習を止めた epoch 数

| | |
|---|---|
| 背景除去なし | 43 |
| 背景除去あり | 69 |

表 4　各手法における正解率

| | |
|---|---|
| 背景除去なし | 0.655 |
| 背景除去あり | 0.576 |

## 5　実　　験

### 5.1　実験設定

本研究では第 3 節及び，第 4 節で記述したショット単体で判別できる映像文法について実験を行う．アノテーション後の画像データをそれぞれ図 5 のように分け，バストショット，クロースアップ，シンメトリの分類実験を行った．エポック数はいずれの場合も 100 とし，学習率の調整によって検証データの損失が下がらなくなり十分に学習し，過学習が起こるタイミングがわかるようなグラフになるよう調整した．ResNet18 の実装は PyTorch を使用し，入力画像データは（224*244）にリサイズしたものを使用した．

### 5.2　結　　果

ここでは，ショット単位で判別できる映像文法として，バストショット，クロースアップ，シンメトリを取り上げ，この映像文法をアノテーションとして有する画像を ResNet18 で機械学習し，新たな映像ショットを三値分類（バストショット，クロースアップ，シンメトリ）するシステムを構築した．その分類結果を図 10，図 11，図 12，図 13 に示す．混同行列ではテストデータに対する正解率を表示している．この結果はハイパーパラメータを表 2 のように設定して得たものである．図 10，図 12 は左のグラフは epoch と損失の関係性を示し，右のグラフは epoch と正解率の関係性を示している．共に横軸が epoch となっており，縦軸が損失，正解率となっている．背景除去なしと背景除去ありの正解率を表 4 に示し，各手法において学習を止めたときの epoch 数を表 3 に示す．

図 14，図 15 では実際に背景除去なし，背景除去ありにおいて正しくバストショットと判別できた画像を示す．どちらの手法でも正しく判別できた例がこれである．図 16，図 17 を用いて背景除去なしと背景除去ありの差分について示す．本来この画像はバストショットであると判別されることが正しい画像であり，背景除去なしの場合には正しく判別できている．しかしながら，背景除去ありでは誤ってシンメトリと判別してしまっている．また，図 18 はシンメトリであるが，誤ってバストショットと判別してしまった例である．実際にはバストショットでもありシンメトリでもある．これはバストショットであるにも関わらず，図 18 をシンメトリとアノテーションしたことが原因であり，マルチラベル分類を行いたいという展望でもある．

### 5.3　考　　察

図 11，図 12 からどちらの提案手法の場合であってもシンメトリと予測することはできているが，バストショットとクロースアップの判別ができていないケースが多く見受けられる．また，背景除去ありの方がバストショットとクロースアップの判別ができていないが，これは背景がないことによって遠近感の判定が難しくなったことが原因であると考えられる．また，元の画像を (224*224) にリサイズすることによって文法の判別ができなくなることも考えられる．図 16，図 17 において背景除去ありが誤って文法の判別をしてしまったのは，シンメトリの場合人物領域のみならず，背景にある様々なオブジェクトがシンメトリであると判別するのに必要な情報であったのではないかと考えられる．

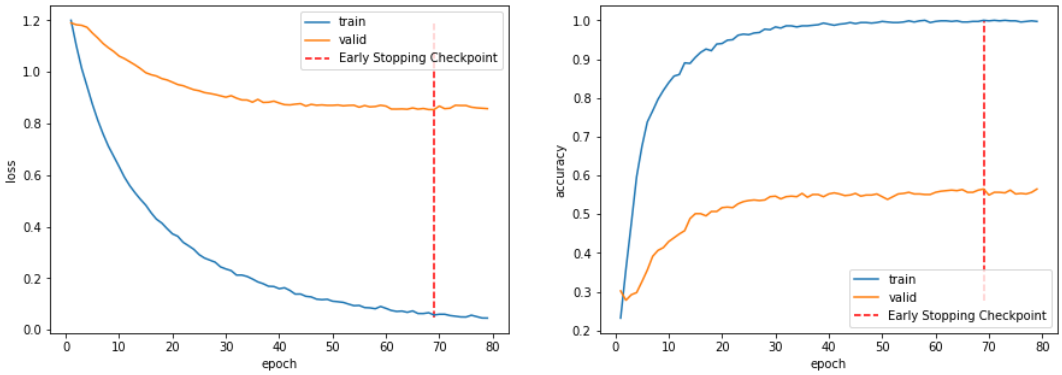



図 12　背景除去ありにおける学習曲線

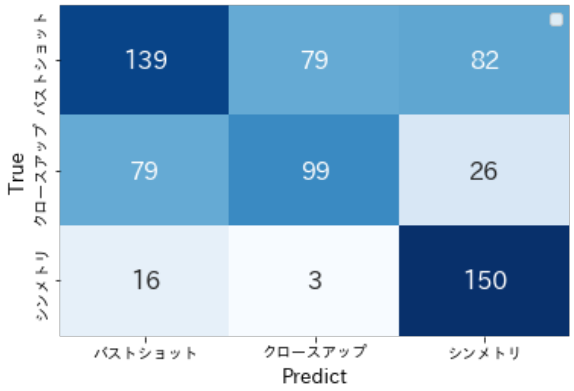



図 13　背景除去ありにおける混同行列

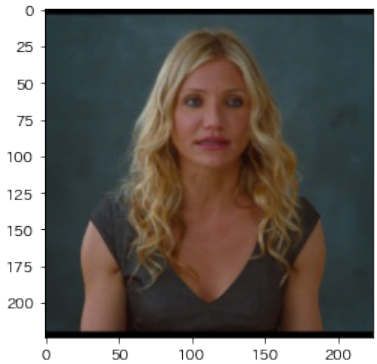


図 14　背景除去なしにおいてバストショットと判別した画像

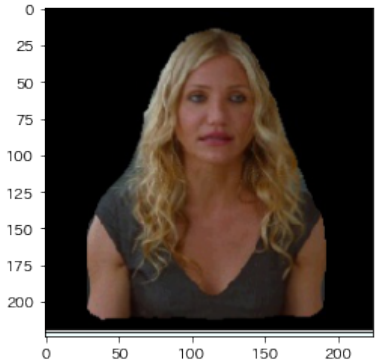


図 15　背景除去ありにおいてバストショットと判別した画像

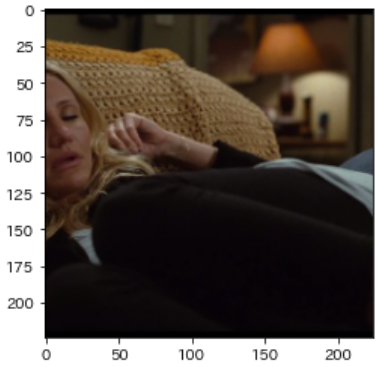


図 16　背景除去なしにおいて正しくバストショットと判別した画像

## 6　ショット系列で判別できる映像文法予測

畳み込みニューラルネット (Convolutional Neural Network, CNN) の拡張である残差学習（Residual Learning）ResNet18 [8] [9] を複数個用いてショット系列で判別できる映像文法による画像分類を行う仕組みを図 19 に示す．ショット系列で判別できる映像文法の場合，ショット単体で判別できる映像文法とは違い入力として系列になった画像を与える必要がある．それを実現するために，入力画像 1 枚に対して 1 つの ResNet18 を適応し，各入力画像から得られたベクトルを基に映像文法を判別するものである．図 19 では入力画像が連続する 2 枚の画像である場合の学習の仕組みとして示している．

## 7　ま　と　め

本論文では，機械学習によって獲得する映像文法を以下の 2 種類に分類し，それぞれの映像文法を獲得する手法を提案した．

（1）　ショット単体で判別できる映像文法

- ショットサイズ（ロング/バスト/クロースアップ等）
- ハイアングル/ローアングル
- シンメトリ/アシンメトリ（映像の左右対称性/非対称性）
- 閉じた/開いた構図（映像の構図の開閉）

（2）　ショット系列で判別できる映像文法

- POV（Point of View）ショット

人物の眼のショット＋その人物から見える映像

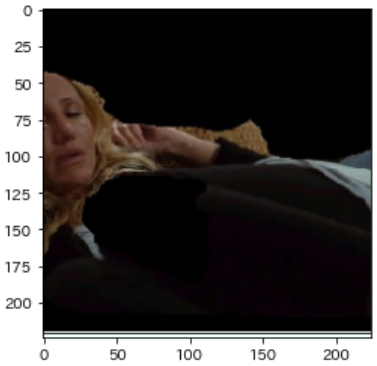


図 17　背景除去ありにおいて誤ってシンメトリと判別した画像

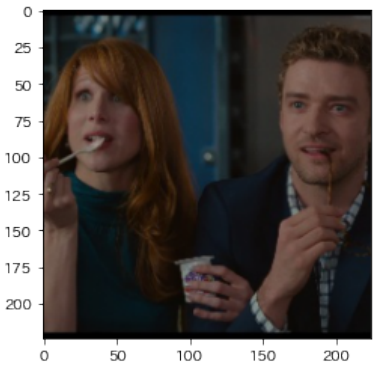


図 18　シンメトリであるがバストショットと分類した画像

- ショット・リバースショット

相対する複数の人物の，前面からのショット＋背面からのショット

- エスタブリッシングショット

シーンの冒頭などで，場所の状況や出演者の位置関係を認識させるためのショット群のこと．通常は，超ロングショット/ロングショットから始まり，パンショットが続き，最後にズームアップショットとなることが多い．

ショット単体で判別できる文法であるバストショット，クロースアップ，シンメトリのついて三値分類を ResNet18 を用いて分類した場合と，Mask R-CNN によって学習データの人物領域のみを抽出して ResNet18 を用いた場合について正解率にどの程度の差が確認されるのか実験した．

今後の課題は以下の通りである．

（1） 提案アーキテクチャの改良

（2） アノテーションの効率化

（3） バストアップ, クロースアップ，シンメトリ以外の映像文法抽出

（4） ショット系列で判別できる映像文法の抽出

## 謝　　辞

本研究は JSPS 科研費 JP21H03774，JP21H03775，JP22H03905 の助成を受けたものです．ここに記して謝意を表します．

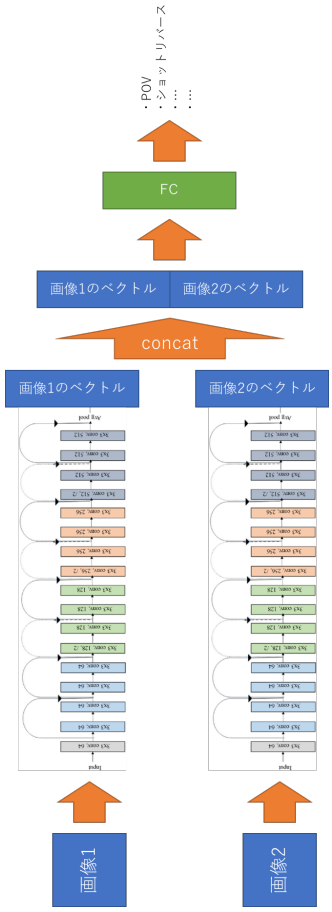



図 19　ResNet18 対残差学習によるショット系列の自動分類

## 文　　献

[1] 嶌田聡. 知っておきたいキーワード　映像文法. 映像情報メディア学会誌, Vol. 62, No. 9, pp. 1403–1405, 2008.

[2] ダニエルアリホン (著), 岩本憲児 (訳)，出口丈人 (訳). 映画の文法 -実作品にみる撮影と編集の技法-. 紀伊國屋書店, 1980.

[3] ジェニファーヴァンシル (著), 吉田俊太郎 (訳)．映画表現の教科書 —名シーンに学ぶ決定的テクニック 100. フィルムアート社, 2012.

[4] 今泉容子. 映画の文法 改訂増補 -日本映画のショット分析-. 彩流社, 2019.

[5] 今泉容子 (著), 田中克己 (編著), 黒橋禎夫 (編著). 映像のデザイン（第 7 章），情報デザイン-京都大学デザインスクール・テキストシリーズ-. 共立出版, 2018.

[6] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A unified framework for shot type classification based on subject centric lens. In *The European Conference on Computer Vision (ECCV)*, 2020.

[7] MovieNet. Movienet data set (movie per-shot keyframes (240p). *https://movienet.github.io*.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385v1 [cs.CV]]*, 2015.

[9] DeepAge. Residual network(resnet) の理解とチューニングのベストプラクティス. DeepAge 人工知能の今と一歩先を発信するメディア, `https://deepage.net/deep_learning/2016/11/30/resnet.html`, 2016.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.