J-058

映画の構造化のためのショットサイズ抽出とシーン解析

Shot Size Extraction and Scene Parsing for Movie Structurizing

伊藤 学†　Uwe Kowalik†　**青木 輝勝**†　**安田 浩**†
Manabu Ito　Uwe Kowalik　Terumasa Aoki　Hiroshi Yasuda

１．はじめに

現在，蓄積された番組の検索に加え，見たいシーンの検索や要約など，映像の構造化を基としたあらゆる機能が必要とされ，それらに関する技術が一層注目されている．従来の報告は，スポーツ（特に野球）やニュース番組など，決まった画面構造を有した構造化しやすい映像を対象としたものが多いが，汎用的な映画に関する研究は少ない．そこで，映画コンテンツに焦点をあて，その構造化を可能とする要素技術として，顔のアップや全身など，人物のショットサイズ（以降 SS）抽出手法と，それを用いた感動的なシーンや緊迫するシーンなど，１つの映画における重要（印象的）シーンの解析法について述べる．重要シーンを解析することは，その映画の象徴的シーンとして記憶に残りやすいといえるため，映画検索においてその内容を確認する要約映像の一部としての使用が期待できる．ここでシーンとは，映画中のあるシチュエイションでの出来事を意味し，アクションの“起因～経過～帰結”となる一連の繋がりを持つものである．

視聴者に感動や緊張を与えるためには，自分と感情を同居させている登場人物の表情を的確に見せることが重要と考えている．つまり，１つの映画において感動的／緊張感のある重要なシーンは，フェイスやネックアップなどの大きな SS が多用されていると考える．

２．提案手法

図１にシステムのフローを示す．入力されたシーンに対し，顔検出として映像中に映っている人物の顔と複数の特徴点を検出し，SS 抽出の基準となる区間（FL: Face Length）を算出し取得する．次に，抽出された FL とフレームサイズとの割合を算出し，独自のモデルにマッピングし，ショットサイズを抽出，さらにすべての SS を対象に解析して重要度を算出，重要シーンとして判定する．

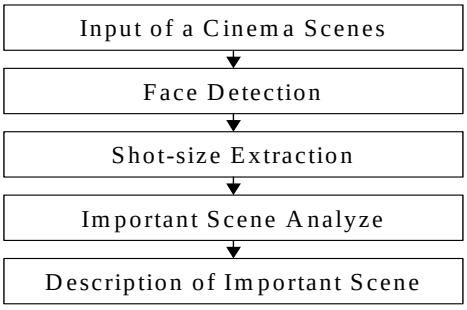


図１　提案手法のフロー

†東京大学　先端科学技術研究センター
The University of TOKYO, RCAST

2.1　顔領域検出

図２に顔検出システム[1]により取得可能な顔の特徴点と，SS 判定の基準となる区間 FL の位置を示す．入力映像は，顔検出モジュールに送られる．この顔検出モジュールとは入力した映像のうちどの部分が顔であるかを周波数領域で識別するためのモジュールである．SS 抽出の基準となる FL は，表情や口の動きなど位置変化の影響を受けにくい目と目の中間で鼻の根元の鼻根点（FPa：Face Position a）と，唇の上部に位置する上唇点（FPb）の２つの特徴点座礁を用い，その距離を FL とした．ここで，[2]にて紹介されている９段階の SS のうち，人物に焦点をあてた“ボディ”から“フェイス”までの６段階を用いた．一般的に，顔領域の解析では矩形を用いることが多いが，ここでは画面の縦方向の長さのみを用いている．なぜなら，SS が人物の高さ方向の長さを基準としていること，表示デバイスの形体が，３：４，９：１６さらにシネマサイズなど様々であり，面積を基に解析するとそれに影響されてしまうためである．

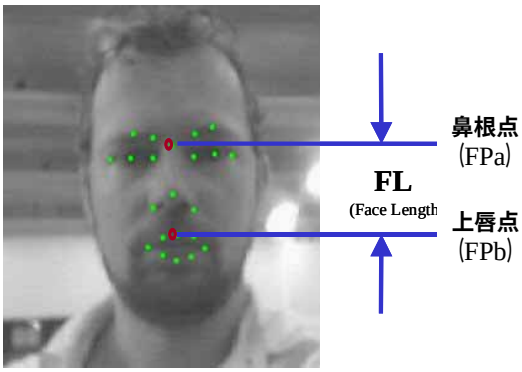


図２　顔の特徴点と判定基準区間 FL

2.2　ショットサイズ判定

本手法では，顔の“大きさ（SS_{ratio}_L）”，“位置（SS_{ratio}_P）”とを考慮した SS 判定を採用している．SS 抽出の定義として，フレームの最上部（Ha）に頭の最上部が接している状態を基準とし，最下部（Hb）に映る体の位置で判断する．

フレームの高さを H とすると，SS_{ratio}_L は，

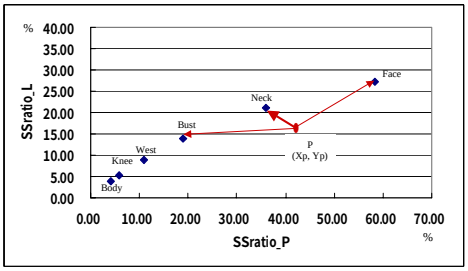
$$SS_{ratio}\_L = \frac{FL}{H} \times 100 = \frac{|FPa - FPb|}{|Ha - Hb|} \times 100 \qquad (1)$$

顔の位置も大きさ同様に，FL がフレームの最下部を Hb から鼻根点 FPa までの長さ（H_{FL}）に対して占める割合を用い SS_{ratio}_P は，

$$SS_{ratio}\_P = \frac{FL}{H_{FL}} \times 100 = \frac{|FPa - FPb|}{|FPa - Hb|} \times 100 \qquad (2)$$

次に，人体寸法[3]を参考に上記 2 つの値が示す SS の判定対象を構築する．参照したデータは，18 歳以上 30 歳未満の男子 217 名，女子 204 名の合計 421 名で，計測値は男女全ての平均を採用した．厳密には，男子と女子とではサイズに差があるため区別して検討しなければならないが，現在の顔検出では，画像から男女を見分けることは困難であるため全ての平均を用いた．

それぞれの SS が占める FL と長さを抜き出し，それぞれの値（SS_{ratio}_L および SS_{ratio}_P）を算出し，図３に判定モデルを構築した．X軸に SS_{ratio}_P，Y軸に SS_{ratio}_L とし，その交点が各 SS の基準点 Qss である．任意のフレームより得られた値P(Xp, Yp)は，各 SS の基準点 Qss(Xq, Yq)までの距離が最小となる SS（ボディ～フェイス）が付与される．ここで，付与される SS は，後述する重要シーン解析に用いるため図３のような，整数値を与える．



形式	特徴量
Body	1
Knee-up	2
West-up	3
Bust-up	4
Neck-up	5
Face	6

図３　SS 判定モデル

2.3　重要シーン解析

登場人物がどのようなアクションをするかを見極めるため，登場人物の SS を用いて，シーンの経過に対する“SS 経過（$SS_{transition}$）”とシーン中に用いられている“SS 傾向（$SS_{average}$）”，2 つのパラメータを採用し重要シーンを判定する．

$SS_{transition}$ は，ｙ軸に SS，ｘ軸に時間を与え，両軸ともに０～１に範囲に正規化して得た線形近似直線の傾きを用いる．次に，$SS_{average}$ は，シーン中の SS の平均を用いる．最後に，$SS_{transition}$ および $SS_{average}$ を用いて，重要シーン値(ISV: Important Scene Value)を算出する．ISV は，$SS_{transition}$ を傾き，$SS_{average}$ を切片とした直線 SS L(x)により，ｘ軸と囲まれる面積と定義する．式（３）に示す．また，図４にて，その詳細を示す．

$$ISV(x) = \int_0^1 SSL(x)dx = \int_0^1 (SS_{transition}x + SS_{average})dx \qquad (3)$$

ここで，本方式は，顔検出システムより検出されたフレーム数のみを用いて解析している．本来であれば，顔検出システムが 100%の検出精度が必要となるが，現在，自然画からの顔検出技術ではこれを満たすものは存在しない．特に，小さすぎる顔や後向き，さらにマスクなどで顔を覆っているものなどは，検出できない．しかしながら，映像の構造化を図るための自動化は必須であるこ

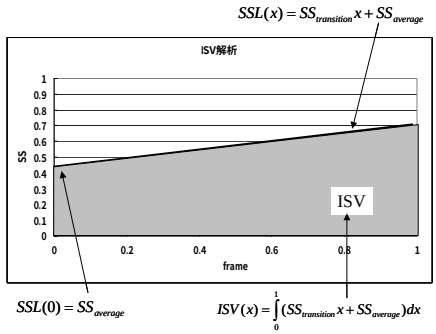


図４　重要シーン判定モデル

から，ここでは現在の検出制度でできる範囲を前提として，後述する実験と主観評価にて検討している．また，映画ではシチュエイションを伝えるために風景などのショットがわずかではあるが含まれている．そもそも，このショットは人物を含んでいない場合が多く，これも検出漏れと判断されてしまう．しかしながら，実験で用いた映画を視聴し手動で人物の含まれていないフレーム数を抽出した結果，全フレーム数は 209574 フレーム，そのうち風景は 2470 フレーム，つまり風景フレームは割合にして 1.2%しかないため，無視する．

3．実験

１つの映画中より５つのシーン（Scene 1～5）を抜き出し，提案手法と被験者による主観評価実験を行った．主観評価では，Scene 2 に対して 16 人中 9 人が 1 位，5 人が 2 位と判断し評価順位は 1 位となり，続いて Scene 1 が，1 位 4 人，2 位 10 人となり評価順位は 2 位となった．ISV の結果と比較すると，重要シーンの 1 位と 2 位の 2 つにおいては正解率 100%という良好な結果が得られた．重要シーンは，映画における最も緊張や感動を得るものを意味するため，Scene 2 が提案手法および評価において同義と解析されたことは，本手法の有効性を証明したといえる．

4．まとめと今後の課題

本稿では，映画コンテンツの構造化を図るべく，登場人物の SS に着目し，その抽出手法とそれを用いた重要シーン解析を述べ，良好な結果を得た．制作者側の意図を読むには，シナリオを用い構文解析する手段もあるが，入手は困難である．一方で，本手法のように映画ストリームそのものから解析する手段として，BGM の有無と音声レベルを用いる方法もあるであろう．今後は，重要シーンと音との関連を追及し，精度向上を目指していく．また，多くのジャンルでの実験など，システムの精度向上に取り組む．

【参考文献】

[1] 純丘：“エンターテイメント映画の文法　ヒットを約束する脚本からカメラワークまで”，三秀舎（2005）
[2] U.Kowalik, T.Aoki, H.Yasuda：“BROAFERENCE - A Next Generation Multimedia Terminal Providing Direct Feedback on Audiences Satisfaction Level”，INTERACT 2005, pp. 962-965,（2005）
[3] 河内，持丸：“2005　AIST 人体寸法データベース”，産業技術総合研究所，H16，PRO 287