Vol.2014-MUS-103 No.54
2014/5/25

歌手映像と歌声の解析に基づく音楽動画中の歌唱シーン検出手法の検討

平井 辰典[1,a)]　中野 倫靖[2,b)]　後藤 真孝[2,c)]　森島 繁生[3,d)]

概要：本稿では，ライヴ動画や PV などに代表される音楽動画において，歌手が歌っているシーンである「歌唱シーン」を検出する手法について検討する．音楽において歌手は最も主要な役割を担っており，音楽動画における歌唱シーンも同様に動画のハイライトの一つであると言える．歌唱シーンは動画サムネイル生成や，大量の音楽動画の短時間ブラウジングなどにおいて有用である．歌唱シーンを検出するためには，歌手の顔認識，楽曲中の歌声区間検出といった要素手法及びそれらを組み合わせる手法についての検討が必要である．本稿では，顔認識を用いた映像解析，歌声区間検出を用いた音響解析，それらを複合した Audio-visual 解析のそれぞれについて比較・検討しながら歌唱シーン検出の実現可能性について議論する．

1. はじめに

音楽動画は，動画共有サービスにおいて最も人気がある動画ジャンルである．Videotrine[1] によると，YouTube における再生回数歴代上位 30 動画のうち，音楽動画が 29 作品を占めている．なかでも，再生回数歴代一位の PSY による “GANGNAM STYLE” の Music clip は 2014 年 4 月現在，19.5 億再生という驚くべき再生回数を示している．

このように動画視聴サイトで再生回数が多いようなポピュラー音楽動画において歌手は中心的な役割を担っている．このことは，ソロとして活躍するアーティストには特定の楽器奏者よりも歌手の方が多いことからも推察できる．前述の 29 作品のうち 26 作品は特定のソロ歌手を中心とした Music clip またはライヴ動画である．また，残りの 3 作品についても歌手を含んだ音楽グループの Music clip であり，依然として歌手は作品中で重要な役割を担っている．このような音楽動画作品を鑑賞する際には，視聴者は歌手に注目して動画を検索，鑑賞することが多いと考えられる．本研究では音楽動画の中でも歌手に注目し，歌手が映像中で歌い，対応する歌声が聴取可能なシーンである「歌唱シーン」を検出する手法について検討する．

過去から現在までに存在する音楽動画の数は単調増加を続けており，ユーザが興味のある音楽動画すべてを視聴することは困難である．視聴困難なほど多くの音楽動画が存在する一方，ユーザは限られた時間の中でその内容を試聴しながら自分が見たい動画を探さなければいけない．そんな状況にも関わらず，多くの動画視聴サイトにおいてユーザが動画を視聴するかを決めるための判断材料は動画内の画像 1 枚からなるサムネイル画像と，動画に付加されている限られた情報（メタデータ）のみであり，動画の内容そのものは実際に視聴しなければわからない．このような背景から，音楽動画の代表的な箇所を抽出する動画要約やサムネイル生成といった技術は，ユーザがたくさんの音楽動画の要点を試し見し，視聴する動画を決める上で有用である．特に，ポピュラー音楽によく見られる，歌手にスポットを当てた音楽動画では，歌唱シーンの検出をすることは音楽動画の要約及びサムネイル生成の一つの手法となりうる．

本研究では歌唱シーンの検出手法について検討し，歌唱シーンに特化した音楽動画のサムネイル生成や，大量の音楽動画を短時間でブラウジングするための音楽動画中の歌唱シーンインデクシングの実現を目指す．本稿では，音楽動画中の歌唱シーン検出手法の検討と位置付けて，歌手映像を解析することによる動画像処理からのアプローチ，楽曲中の歌声を解析することによる音響信号処理からのアプローチ，それらを統合することによる Audio-visual 解析のアプローチのそれぞれの方面からその実現可能性について実験を交えて議論する．

[1] 早稲田大学
Waseda University
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
[3] 早稲田大学理工学術院総合研究所 / JST
Waseda Research Institute for Science and Engineering / Japan Science and Technology Agency
[a)] tatsunori_hirai@asagi.waseda.jp
[b)] t.nakano@aist.go.jp
[c)] m.goto@aist.go.jp
[d)] shigeo@waseda.jp

2. 関連研究

音楽動画の配信サイトや音楽ランキング番組などでは，音楽動画のサムネイルが作られ，多くの音楽動画の代表的な箇所を短時間で試聴できるようになっている．ポピュラー音楽動画における歌手の重要性は前述の通りであり，歌唱シーン検出はそのような音楽動画のサムネイル生成に応用可能な技術である．

楽曲のサムネイル生成に関する研究はいくつかあるが [2], [3], [4]，音楽動画を鑑賞するという目的において音響信号処理のみに基づいた動画サムネイル生成は最良の方法とは言い難い．例えば，音楽動画においてサビのみを抽出するようなサムネイルを生成したとしても，歌手の顔が映っていないようなサビを抽出してしまう場合など，音楽的には代表箇所であっても，映像も同様に代表箇所であるとは限らない．また一方で，動画の要約に関する研究も多くなされており [5] ，多くの動画分野に適用されている．しかし，音楽動画は最も人気のある動画ジャンルであるにも関わらず，音楽動画要約及びサムネイル生成についての研究事例は少ない [6], [7], [8]．特に，著者らが知る限りでは歌手の顔に注目した研究事例はない．Agnihotri らの手法では，映像解析，音響解析に加え，歌詞などを含むメタデータを使った音楽動画の要約手法を提案している [6]．この手法では要約の一要素として，映像中の顔の有無について検討しているが，歌手の顔であるかどうかについては注目していない．Xu らの研究でも同様に，顔が登場しているか否かをサムネイル生成の要素の一つとして導入しているが，登場する顔が歌手の顔であるとは限らない [7]．歌手の顔に注目する場合，顔認識手法に基づいて顔認証レベルで歌手の顔を認識するアプローチも考えられるが，映像のみによるアプローチでは抽出した顔が歌唱しているかどうかの判断までは難しい．一方で中村らは，音楽動画に付与されたコメントから分析した視聴者の反応と音響解析に基づくサビ検出手法を組み合わせた音楽動画のサムネイル生成手法を提案しているが [8]，映像を考慮していないため，歌手が登場するかどうかを判定することは難しい．このように，音楽動画のサムネイル生成という観点で歌唱シーンの考慮やそれに相当するような手法は提案されていない．

また，本研究は音楽動画への歌唱シーンに関するアノテーションの自動付与研究と捉えることができる．動画へのアノテーションの自動付与に関する研究は，動画の内容理解，動画検索技術や動画のイベント認識に関連して盛んに行われている．音楽動画中の歌唱シーンの検出はその中でも動画中の特定イベント認識として位置付けられる．

このような動画解析に関する研究は，動画像処理に基づく映像解析以外のアプローチでも行われている．直接映像を解析せずに動画の内容に関するアノテーションを自動付与する例として，動画に付与されたソーシャルアノテーションを利用する手法が挙げられる．佃らは，視聴者によって動画像の時系列に同期して付与されるコメント情報に基づき，動画の登場人物毎の盛り上がり箇所等を推定し，動画に対する時系列アノテーションとして検索に活用している [9]．Nakamura らはコメント情報から動画の印象を推定し，動画の検索に応用する手法を提案している [10]．これらの研究で示されるように，動画の内容を理解するためにソーシャルアノテーションなどの間接的な情報を活用する手法は有効である．しかし，ソーシャルアノテーションには主観的な情報も含まれており，必ずしも動画の内容を反映しているとは限らない．そこで，動画そのものを直接解析して内容を理解することで，ソーシャルアノテーションによる情報を客観的に補うことが有効である．

動画そのものを解析することによる動画の内容理解に関する研究は多く行われており，TRECVID (TREC Video Retrieval Evaluation) と呼ばれる動画検索技術に関する国際的な評価ワークショップで大々的に取り上げられている [11]．TRECVID では，動画の内容解析やイベント認識を初めとした動画検索に関するタスクで，参加チームがそれぞれのシステムの性能を競い合う．TRECVID の動画内容理解タスクにおいて，“Singing scene” というラベルが含まれた動画があり，そのようなタスクに対応するため，参加チームは映像特徴量に加え，MFCC などの音響特徴量を用いた動画解析手法などを提案している [12]．TRECVID では動画を汎用的に分類，ラベリングするためのタスクについて議論されており，歌唱シーンの検出に特化した解析のような特定イベント認識はその範囲外である．音楽動画のように適用範囲を限定する場合には，一般的なイベント認識手法よりも特定イベント認識手法の方が，タスクに特化している分その認識精度が上がりやすい．

音響解析と映像解析を組み合わせた Audio-Visual 解析によって動画中の特定イベントを認識する研究も多くなされている [13], [14], [15], [16]．Hrybyk らは，ギター演奏動画において映像中の指の位置の認識と音響信号処理を組み合わせることで，ギターコードの認識精度が向上することを確認している [14]．Petridis らは，顔認識と MFCC やピッチを用いた音響信号処理の組み合わせによる話し声と笑い声の識別手法について検討し，特徴量の組み合わせ方による精度を比較している [15]．この研究では，映像情報と音響情報を有効に組み合わせることにより，単一のモーダルで認識した時よりも高精度で認識可能であることが示されている．

本研究では，上述の Audio-Visual 解析で扱われていた単一モーダルでも認識可能なタスクとは違い，映像中で歌手が歌っており，なおかつ楽曲においても歌手の歌声が聞こえる歌唱シーンの検出タスクに取り組む．単一モーダルレベルでは，顔認識手法や歌声区間検出手法などに関する研究が行われてきているが，歌唱シーンはそのいずれかだけが

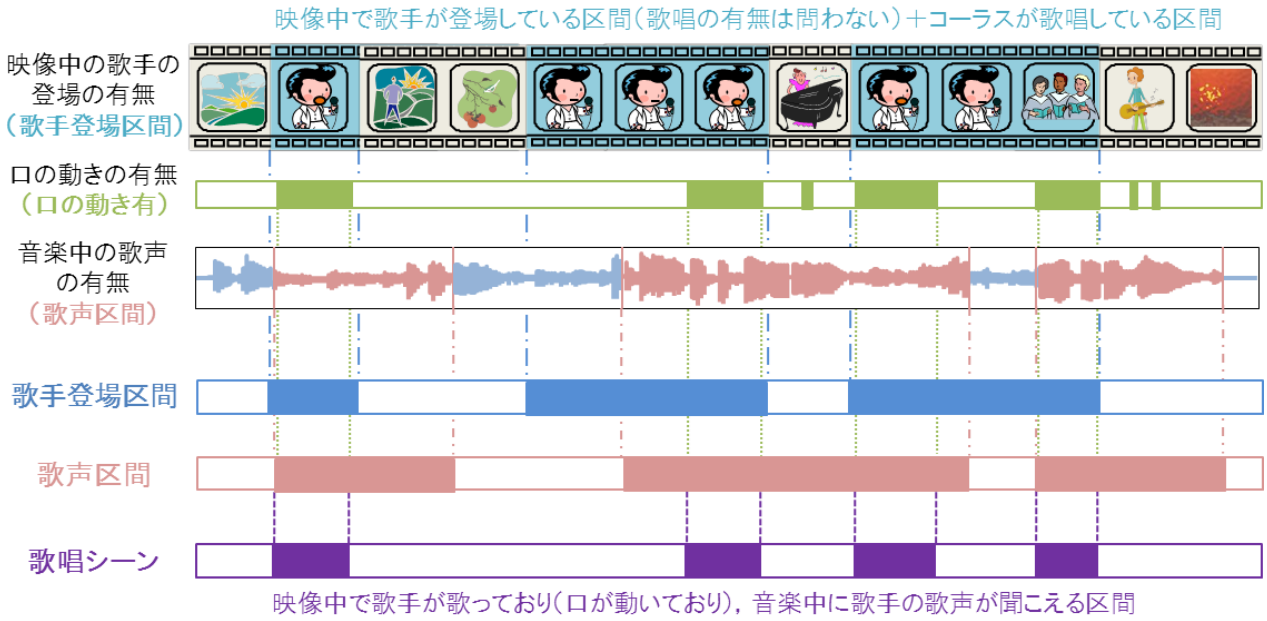

図 1　歌手登場区間，歌声区間，歌唱シーンの定義
Fig. 1　Definitions of singer appearing scene, vocal part, and singing scene.

完璧な精度であっても十分ではなく，両者を組み合わせなければ認識できない．本稿では，映像中の顔認識手法に基づく口の動き検出（MAD : Mouth Aperture Detection）と混合音中の歌声区間検出（VAD : Vocal Activity Detection）を組み合わせ，それぞれの検出手法での歌唱シーン検出精度と両者を組み合わせた際の歌唱シーン検出精度を比較・検討する．

3.　歌唱シーン検出手法の検討

本研究では，映像解析のみによる検出手法，音響解析のみによる検出手法，両者を複合した検出手法のそれぞれでどのような検出が可能で，歌唱シーン検出に対してどの程度の精度を示すかを比較・検討する．検討にあたり，本研究で扱う音楽動画における「歌手登場区間」，「歌声区間」，「歌唱シーン」という 3 つの概念を**図 1** 及び以下に示すように定義する．

(1) **歌手登場区間**：映像中で，歌唱の有無に関わらず歌手（ボーカル）が登場している区間．メインの歌手以外の歌手（コーラス等）の場合，歌唱している区間のみ（i.e. 映像中にボーカルが登場する区間＋コーラスが歌唱している区間）．
(2) **歌声区間**：音楽中に歌声が含まれている区間．メインの歌手以外の歌唱（コーラス）区間も含む．
(3) **歌唱シーン**：音楽動画において，映像中で歌手が歌っており（口が動いており），なおかつ楽曲においても歌手の歌声が聞こえるシーン．メインの歌手以外の歌手（コーラス等）の歌唱も含む．

歌唱シーンについては音楽と映像の両方から決定されるが，歌声区間は音楽のみから，歌手登場区間は映像のみから判断可能である．本研究では，これら 3 つのうちの「歌唱シーン」の検出を目指す．

本稿を通して音楽動画 10 作品を用いて実験を行いながら映像解析による手法，音響解析による手法，両者を複合した手法のそれぞれの精度比較をする．実験に使用した 10 作品とそれぞれの動画に歌唱シーンが含まれている割合，歌声区間の割合，歌手登場区間の割合を**表 1** に示す．10 作品のうち，動画番号 5 の Let it be のみがバンドによる演奏動画であり，他 9 作品は Music clip である．Music clip では，歌唱シーンの他にも音楽と直接関係のないストーリーを持った映像がしばしば付加されている．動画番号 8 の作品のように歌手登場区間が作品のほとんどを占めているものもあり，この Music clip は全編が演奏シーンで構成されている．表 1 で示した歌唱シーン，歌声区間，歌手登場区間はそれぞれ手動でラベリングすることで付与した正解値である．

歌唱シーンのラベリングは，動画を 1 フレーム毎に鑑賞しながら，歌手の口が動いており，それに対応する歌声が聴取可能な動画フレームを歌唱シーンとしてラベリングした．ただし，口が映っていないフレームや，歌手が映っていても後姿などで本当に歌っているのかが映像からは判別できないフレームについては，文脈からの予測が可能であっても歌唱シーンとはしなかった．逆に，歌いながら一瞬下を向くなど，楽曲中の休符の長さに満たない間だけ口が映らなかった場合，前後フレームが歌唱シーンであった場合に限って歌唱シーンであるとした．

前述の定義の通り，歌手登場区間は動画の全編を通して

表 1　実験に使用した音楽動画と各シーンの割合
Table 1　Music clips used in experiments and its ratio of each scene.

動画番号	タイトル（アーティスト名）	歌唱シーンの割合 [%]	歌声区間の割合 [%]	歌手登場区間の割合 [%]
1	Almost Human（Kimonos）	28.1	62.1	42.4
2	Baby ft. Ludacris（Justin Bieber）	38.0	89.7	71.4
3	First Love（宇多田ヒカル）	44.5	75.5	74.4
4	Island in the sun（Weezer）	31.1	66.5	42.0
5	Let it be（The Beatles）	41.0	66.7	54.6
6	SMILE（おとぎ話）	37.4	70.3	62.7
7	Winter, again（GLAY）	66.8	66.8	87.0
8	はあとぶれいく（向井秀徳アコースティック＆エレクトリック）	64.8	64.8	95.9
9	ふたつのハート（サニーデイ・サービス）	39.5	59.3	51.5
10	魔法のバスに乗って（曽我部恵一 BAND）	56.1	70.4	75.1

歌唱している人物が登場するシーン全体を表している．歌唱シーンと歌手登場区間の割合を比較することで必ずしもその人物が歌っている区間であるとは限らないことがわかる．

4. 歌手映像の分析による歌唱シーン検出

4.1 映像中の顔検出手法

顔認識技術は日々進歩してきているが，現在でも極端な顔向きや，装飾品などによるオクルージョン，極端な表情変化などがある顔画像に対してはその認識精度は高いとは言えない．特に，実験環境で撮影されたものではない実動画における精度はまだ汎用的なものではない．音楽動画における顔を認識することはさらに困難である．曲に合わせてアーティストが動くことが多いため，その顔向き変化は激しく，歌うことによる表情変化もあり，メイクや装飾品によるオクルージョンや，極端な照明条件など，多くの困難な課題に同時に対応できなければならない．

本研究では，それらの困難な条件での顔認識に取り組む代わりに，現状の顔認識手法で検出可能な顔が映っているフレームの情報を基に，映像フレームの時間的な連続性を生かして，顔認識結果を時間方向に伝播する．これによって，従来の手法では顔を検出することが困難であった映像フレームにおいても顔を検出することを可能としている．本手法の詳細については，[17], [18] に著者らによる先行手法としてまとめている．本稿ではこの手法における顔検出及び顔トラッキング手法のみを適用して顔を検出している．また本稿では，顔がトラッキングしきれなかったフレームについても，映像フレームの時間的連続性の観点から同一ショット内で顔が検出されてさえいれば顔登場フレームである可能性が高いと仮定し，顔検出フレームに加える．ただし，顔そのものを検出できているわけではないため，この後の口の動き検出等は実際に顔を検出できたフレームに対してのみ適用する．

表 2 に映像中の顔検出手法単体での，3 章で定義した歌手登場区間と歌唱シーンそれぞれの検出精度を示す．ここで，適合率とは，検出結果の総フレーム数に対して検出できた正解フレーム数の割合を表し，検出の正確さを示す．再現率とは，総正解フレーム数に対して検出できた正解フレーム数の割合を表し，正解をどれだけ網羅できたかを示している．どちらか一方が高くても精度が高いとは言えず，F 値が両者を考慮した精度の指標となっている．

表 2　顔検出による歌手登場区間と歌唱シーンの検出精度
Table 2　The accuracy of singer appearance detection and singing scene detection with face detection.

動画	歌手登場区間検出			歌唱シーン検出		
番号	適合率	再現率	F 値	適合率	再現率	F 値
1	0.463	0.887	0.608	0.344	0.997	0.512
2	0824	0.758	0.790	0.523	0.903	0.663
3	0.867	0.699	0.774	0.665	0.895	0.763
4	0.465	0.881	0.608	0.345	0.883	0.496
5	0.630	0.775	0.695	0.447	0.733	0.555
6	0.771	0.596	0.672	0.648	0.840	0.732
7	0.914	0.999	0.955	0.701	0.998	0.824
8	1.000	1.000	1.000	0.676	1.000	0.807
9	0.669	0.376	0.482	0.626	0.459	0.530
10	0.882	0.882	0.882	0.729	0.976	0.834
平均値	0.680	0.714	0.679	0.570	0.869	0.672

表 2 の歌手登場区間検出の適合率については，歌手登場区間の割合が低い動画について低くなっており，検出した顔が歌手の顔であるかどうかが歌手が登場する割合にある程度比例した結果となっている．歌手登場区間検出の再現率は，歌手についてのみの数字ではあるが，本手法での顔検出技術の性能を表している．例えば動画 9 では，全歌手登場フレームのうちの 45.9 %のフレームでのみ歌手の顔が検出できたという結果になっている．この動画は，歌手が登場しているシーンがすべて斜め，または横から撮影されており顔検出が比較的困難な動画である．

歌手登場区間検出と歌唱シーン検出では，歌唱シーン検出の方が再現率が高い．これは，歌手が登場している区間

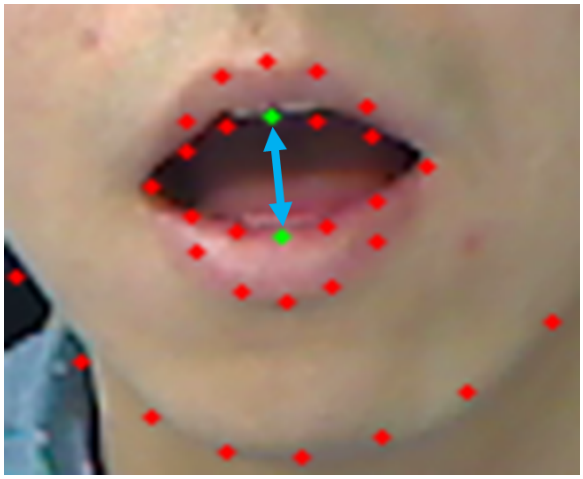

図 2　唇間の距離の抽出
Fig. 2　An extraction of the degree of mouth aperture.

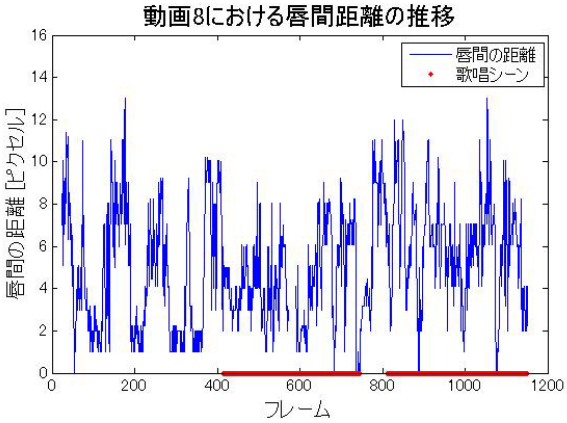

図 3　動画 8 における唇間距離の推移の様子
Fig. 3　A transition of lip distance in video no.8.

のうち，歌唱している箇所の方が顔が検出しやすいということを示している．実際に動画を鑑賞してみると，歌手が登場しながらも歌唱していないシーンは，歌手が移動しているシーンやダンスをしているシーンなど，顔が大きく動いているような顔検出に不向きなシーンであることが多かった．一方歌唱シーンでは，顔の動きは比較的小さく，顔がしっかりと撮影されているケースが目立った．このことから，音楽動画において歌唱シーンははっきりと撮影される傾向にあると予想できるが，より多くの音楽動画を基に検証する必要がある．歌唱シーン検出の方が適合率が低い理由は，表 1 における歌手登場区間の割合と歌唱シーンの割合の比較からも明らかで，歌手が登場しているシーンが歌唱シーンであるとは限らないことによる．

ここで，顔検出で注目するのは映像中に顔が映っているかどうかだけであり，それが歌手であるか否かは判定できない．そのため，本研究では検出した顔が歌手であるかを判定するために，口の動き検出を行う．

4.2　口の動き検出手法

上述した手法によって検出した顔における口の動きを検出する．本手法では Irie らの顔器官検出手法 [19] を用いているため，口の位置（特徴点の位置）については，[17] の手法における顔検出及び顔トラッキングが成功したフレームにおいて検出可能である．検出した口領域を基にその人物が歌っているかどうかを判定する．

歌唱をする際，口は開閉を繰り返す．それにより，歌唱区間では非歌唱区間に比べて唇の開閉の頻度が多くなる．そこで，本稿では口の動き検出手法として，唇間の距離を用いる．唇間の距離は**図 2** に示すように上唇の中央下端と下唇の中央上端によって測る．

唇間の距離をそのまま用いると，顔の大きさや顔向きによって値が変わってしまう．そこで，唇間の距離を算出するにあたって，顔の向きと大きさを正規化する．具体的には，Irie らの手法 [19] によって推定できる 3 次元の顔向き情報を基に，両目間の長さが 50 ピクセルとなるように 3 次元アフィン変換を行う．ここで，3 次元アフィン変換を行う際に顔の奥行きを推定することはできないため，顔を 3 次元空間中の平面の板とみなして変換をすることになる．しかし，上唇と下唇の奥行きは大きく変わらないため，唇間の距離はこの平面近似の影響をあまり受けない．

この正規化によって，唇間の距離は 0〜40 ピクセル程度の範囲の値となる．唇間の距離が変動する頻度の高い箇所は歌唱シーンであると予想できるため，映像をショット毎に分割し，ショット内の唇間距離の標準偏差を特徴量として口の動き検出を行う．ショットとは，映像においてシーンやカメラの切り替わりがなく，フレームが連続に繋がっている区間のことであり，[17] に示した映像フレームのヒストグラムを用いた手法で自動検出する．

ショット内の唇間距離の標準偏差が n ピクセル以上である時，該当ショット全体を歌唱シーンであるとする．本稿では，n の値を変えて検証した結果，実験的に $n = 3$ としている．**図 3** に動画 8 の冒頭 1200 フレームにおける唇間距離の推移の様子を示す．このように唇間距離は，非歌唱区間においても顔特徴点検出の不安定さに起因するノイズが大きく乗ってしまう．そのため唇間距離を直接用いることは効果的ではなく，本稿ではその標準偏差に注目した．

表 3 に口の動き検出手法単体による歌唱シーンの検出精度を示す．顔検出のみによる歌唱シーンの検出精度と比較すると，わずかではあるが精度が向上している．特に適合率が 0.570 から 0.609 に向上しており，口の動きを考慮することで，口の動きを考慮しない場合に比べて歌唱シーンをより的確に検出できることがわかる．ただし，再現率については 0.869 から 0.823 に低下しており，実際の歌唱シーンの一部を非歌唱区間であると推定してしまっている．

本研究では，唇間距離を直接特徴量としてショット毎に閾値を定め，唇間距離が閾値以上となるフレームを開口フレームとし，開口フレームの間隔によって歌唱シーンを検

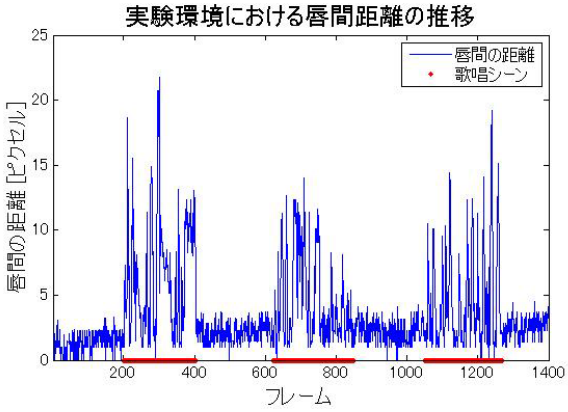

図 4　実験環境における唇間距離の推移の様子
Fig. 4　A transition of lip distance in laboratory environment.

表 3　口の動き検出による歌唱シーンの検出精度
Table 3　The accuracy of singing scene detection with MAD.

動画番号	タイトル	適合率	再現率	F 値
1	Almost Human	0.333	0.851	0.479
2	Baby ft. Ludacris	0.555	0.891	0.684
3	First Love	0.809	0.809	0.809
4	Island in the sun	0.409	0.842	0.551
5	Let it be	0.453	0.733	0.560
6	SMILE	0.706	0.809	0.754
7	Winter, again	0.701	0.998	0.824
8	はあとぶれいく	0.676	1.000	0.807
9	ふたつのハート	0.679	0.331	0.445
10	魔法のバスに乗って	0.767	0.965	0.855
	平均値	0.609	0.823	0.677

出する手法や，口の左右の距離を考慮した口の面積に基づく検出手法についても実験をした．

それらの結果については本稿では省略するが，顔特徴点の検出精度不足から，唇間距離や口の面積の値には多くのノイズが乗ってしまい，フィルタ処理等を行っても歌唱シーン検出の精度が低かった．ショット内の標準偏差を導入することによって，ノイズをある程度吸収することができたが，一方でフレーム単位での検出ができないという欠点もある．唇間距離を直接用いた開口フレームの検出は，Web カメラの前に立って口を動かすような実験環境下では比較的良好な検出精度であった．実際に実験環境下で歌唱と非歌唱を数秒おきに繰り返した際の唇間距離の推移を**図 4** に示す．このように，実験環境下では唇間距離を直接用いるだけでも口の動きが検出可能である．ただし，母音の「い」や「う」のように口を縦に開かないような発音の際には唇間距離は大きくならないため，縦の距離以外にも注目する等の改良の余地はある．図 3 及び図 4 の結果を比較することで，音楽動画のように顔検出そのものが困難な動画に対しては唇間距離を直接用いることは効果的でないことがわかる．

表 4　歌声区間検出による歌声区間と歌唱シーンの検出精度
Table 4　The accuracy of vocal part detection and singing scene detection with VAD.

動画	歌声区間検出			歌唱シーン検出		
番号	適合率	再現率	F 値	適合率	再現率	F 値
1	0.836	0.641	0.726	0.407	0.690	0.512
2	1.000	0.537	0.699	0.508	0.644	0.568
3	0.987	0.722	0.834	0.599	0.742	0.663
4	0.998	0.632	0.774	0.449	0.609	0.517
5	0.990	0.814	0.893	0.634	0.849	0.726
6	0.933	0.653	0.769	0.517	0.680	0.587
7	0.995	0.831	0.906	0.995	0.831	0.906
8	0.880	0.787	0.831	0.880	0.787	0.830
9	0.962	0.660	0.783	0595	0.613	0.604
10	0.991	0.941	0.965	0738	0.879	0.803
平均値	0.957	0.722	0.818	0.632	0.732	0.672

5. 歌声区間検出手法

歌声区間は 3 章で定義したように音響信号のみから判断されるものである．本研究では混合音中の歌声区間検出手法として，Fujihara らによる HMM ベースの手法を用いた [20]．この手法では，歌声区間と非歌声区間のそれぞれを GMM で学習し，それぞれの状態を行き来する HMM によって楽曲を表現することで歌声区間を推定している．また，推定の過程で混合音の各フレームに対して歌声と非歌声の対数尤度がそれぞれ算出できるため，検出結果の尤もらしさを考慮することもできる．

表 4 に歌声区間検出手法単体による歌声区間の検出精度と歌唱シーンの検出精度を示す．本手法では，歌声区間，非歌声区間のモデリングを学習ベースで行っているため，歌声区間検出の結果は安定しており，高い適合率を示している．しかし，歌唱シーンは映像の考慮が不可欠で歌声区間のみからは推定できないため，歌声区間検出に比べ歌唱シーン検出の精度が確かに下がっていることが確認できる．

6. 歌手映像と歌声の複合分析

上述した映像のみを用いた口の動き検出手法と，歌声のみを用いた歌声区間検出手法を組み合わせることによる歌唱シーン検出手法を検討する．本研究において歌唱シーンは，3 章の定義及び図 1 に示したように，「歌手が映像中で実際に歌っており対応する歌声が聴取可能なシーン」として定義している．そのため，映像中の歌手の口の動きと混合音中の歌声区間が完璧に検出できているような理想的な状況では，口の動き検出の結果（口の動きがある区間が 1，それ以外が 0）と歌声区間検出の結果（歌声区間が 1，非歌声区間が 0）の論理積を取った結果が最適な歌唱シーン検出の結果となる．ここで，論理積を取るとは，両検出結果が 1 であった場合に歌唱シーンを表す 1 を取り，それ以外

表 5　口の動き検出結果と歌声区間検出結果の論理積及び論理和による歌唱シーンの検出精度（10 動画の平均）

Table 5　The accuracy of singing scene detection with a combination of MAD and VAD using logical conjunction and disjunction (Average of 10 clips).

複合方法	適合率	再現率	F 値
論理積（10 動画の平均）	0.755	0.604	0.654
論理和（10 動画の平均）	0.545	0.951	0.683

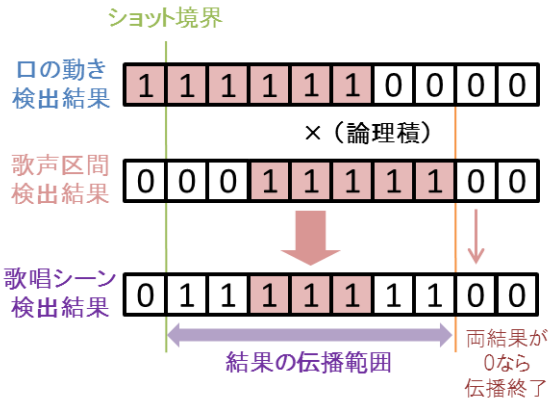

図 5　検出結果の複合方法

Fig. 5　A combination of detection results.

の場合には非歌唱シーンを表す 0 を取るような論理演算を表す．このような結果の複合では，歌手以外の人物が音楽とは関係のない文脈で口を動かすような場合には誤った結果を返してしまうが，本稿ではそのような例外的なシーンは考慮しない．

各手法で理想的な検出精度が実現できている場合にはこの論理積による結果の複合が有効だと考えられるが，実際の検出精度は理想とは遠い．現実の検出結果に対してこのような単純な複合を行うと，各手法が共に歌唱シーンであると推定している分適合率の向上は期待できるが，再現率の上限が各検出結果の低い方の値となってしまう．

一方で，各検出結果の論理和を取ることを考えると，再現率の向上は見込めるが，各手法の誤検出結果までも吸収してしまうこととなり，適合率が低下してしまう．各検出結果の論理積，論理和を取ることによる歌唱シーンの検出結果を**表 5** に示す．この結果と表 3（口の動き検出による歌唱シーン検出），表 4（歌声区間検出による歌唱シーン検出）を比較すると，論理積を取ることによる適合率の向上，論理和を取ることによる再現率の向上が確認できる．しかし，理想的な状況下では最適な検出精度となるべき論理積を取ることによる結果は，各検出手法単体と比べても F 値が下がってしまっている．これは各検出手法で検出可能な箇所が異なることでそれぞれの検出結果がうまく生かされていないことによる．

そこで，両検出結果の論理積を取ることで歌唱シーンであると推定されたフレームを種として，動画の時間的な連

表 6　口の動き検出と歌声区間検出の複合による歌唱シーンの検出精度

Table 6　The accuracy of singing scene detection with a combination of MAD and VAD.

動画番号	タイトル	適合率	再現率	F 値
1	Almost Human	0.464	0.766	0.578
2	Baby ft. Ludacris	0.585	0.719	0.645
3	First Love	0.855	0.809	0.831
4	Island in the sun	0.575	0.619	0.597
5	Let it be	0.488	0.713	0.579
6	SMILE	0.821	0.670	0.738
7	Winter, again	0.701	0.998	0.824
8	はあとぶれいく	0.676	1.000	0.807
9	ふたつのハート	0.679	0.331	0.445
10	魔法のバスに乗って	0.773	0.965	0.858
	平均値	0.662	0.759	0.690

続性に基づいて，**図 5** に示すように検出結果を時間方向に伝播させる．検出結果の伝播は動画の時間的な連続性が保証される映像ショット単位で行う．歌声区間として検出され，なおかつ口の動きも検出されたフレームを歌唱シーンであるとして，その前後フレームが同一ショットであれば，該当フレームも同様に歌唱シーンであると推定する．ただし，結果の伝播は同一ショット内の歌声区間もしくは口の動きがある区間のみとして，非歌声区間であり口の動きも検出されないフレームがあればそこで結果の伝播は終了する．以上のようにして推定した歌唱シーンの検出精度を**表 6** に示す．

両検出結果を複合することによって検出した際の F 値の平均は 0.690 となっており，各検出手法単体に比べて歌唱シーンの検出精度がわずかではあるが向上していることが確認できる（表 3，表 4，表 6）．本手法の検出精度向上は，各検出手法の精度向上に依存するところが大きいが，各手法の複合方法にも改善の余地があると考えている．

今回は歌唱シーンであるか否かという 0 か 1 の結果のみに注目したが，歌唱シーンであるかどうかの推定の確信度も算出可能である．例えば，論理積を取って歌唱シーンであると推定された箇所は確信度の高い歌唱シーンであり，論理和によって推定された箇所は確信度は低いが歌唱シーンの可能性があるということや，歌声区間検出手法の歌声区間の尤度なども用いることで算出できる．今後，単純に両者の結果を複合するだけではなく，歌唱シーンらしさの定量化を初めとした複合手法の改善に取り組んでいきたい．

7.　まとめと今後の方針

本研究では，顔認識を用いた映像解析，歌声区間検出を用いた音響解析，それらを複合した Audio-visual 解析のそれぞれについて比較・検討しながら歌唱シーン検出の可能性について述べた．

歌唱シーンの検出が実現することによって，音楽動画の

サムネイル生成や，歌手が歌っているシーンに基づいた音楽動画の検索等が可能となる．さらに，歌唱シーンの検出は，混合音からの歌声のモデリングや歌手の顔認識の精度向上にも寄与できると考えている．例えば，歌声区間検出では実際に歌声がない箇所を歌声区間と誤検出してしまうことがある．歌声のモデリングをする際など，少ないフレーム数であっても歌声区間である確信度が高いフレームが抽出できればいいような場合には，歌唱シーン検出によってより確実に歌声が含まれているフレームが抽出できる．また，歌手の顔認識の場合には，歌唱シーン検出によってその人物がボーカルであるかどうかを判断できるため，より確信度の高い歌手のクラスタリングなどに貢献できる．

本稿の結果では，映像解析と音響解析の複合による精度の向上はわずかであり，手法の検討に留まったが，今後，各モーダルの情報を統合する手法，統合すべき情報そのものについてさらに検討していきたい．それによって，歌唱シーン検出の手法を確立していきたい．特に現状では口の動き検出の効果が小さいため，唇間の距離以外の特徴に注目することや，データの学習による口の動き検出手法などについて検討していきたい．

本稿では精度の比較のみを扱ったが，映像解析，音響解析のそれぞれに関する別の側面からの評価も必要であると考えている．例えば，処理速度の観点から各手法を比較すると，映像解析に比べて音響解析の方が高速に処理できる．精度よりも速度が求められるような状況では映像解析を複合することは適切でない可能性もある．さらに，ソーシャルアノテーションなどのテキスト解析では，音響解析よりも高速に処理可能である上にコンテンツそのものがなくても解析可能であるといったメリットがある．そのような様々な側面を考慮した上で各モーダルの情報を統合することについて検討していきたい．

今後，歌唱シーン検出手法を拡張していくことで歌手に限らず，楽器の奏者の検出手法と楽器の物体認識手法を組み合わせた，音楽動画中のミュージシャンの役割認識などが実現できる可能性がある．本研究を通じて，音楽動画のインデクシングなどに有用な手法として研究の対象を広げていきたい．

謝辞 本研究の一部は，日本学術振興会特別研究員（DC1）科学研究費及び，JST CREST「コンテンツ共生社会のための類似度を可知化する情報環境の実現」の支援を受けて実施された．

参考文献

[1] Videotrine: http://en.videotrine.com/
[2] Cooper, M., Foote, J.: Summarizing popular music via structural similarity analysis, *Proc. of IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, pp.127–130 (2003).
[3] Chai, W., and Vercoe, B.: Music Thumbnailing via Structural Analysis, *Proc. of ACMMM2003*, pp.223–226 (2003).
[4] Bartsch, M., and Wakefield, G.: Audio Thumbnailing of Popular Music Using Chroma-Based Representations, *IEEE Trans. on Multimedia* Vol.7, pp.96–104 (2005).
[5] Money, A., and Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art, *Journal of Visual Communication and Image Representation*, Vol.19, pp.121–143 (2008).
[6] Agnihotri, L., Dimitrova, N., and Kender, J.: Design and Evaluation of a Music Video Summarization System, *Proc. of ICME2004*, pp.1943–1946 (2004).
[7] Xu, C., Shao, X., Maddage, N., and Kankanhalli, M.: Automatic Music Video Summarization Based on Audio-Visual-Text Analysis and Alignment, *Proc. of SIGIR2005*, pp.361–368 (2005).
[8] 中村聡史，山本岳洋，後藤真孝，濱崎雅弘：視聴者反応と音響特徴量に基づくサムネイル動画の生成手法，情処学論データベース (TOD)，Vol.6，No.3，pp.148–158 (2013)．
[9] 佃 洸摂，中村聡史，山本岳洋，田中克己：映像に付与されたコメントを用いた登場人物が注目されるシーンの推定，情処学論，Vol.52，No.12，pp.3471–3482 (2011)．
[10] Nakamura, S., and Tanaka, K.: Video Search by Impression Extracted from Social Annotation, *Proc. of WISE2009*, pp.401–414 (2009).
[11] Smeaton, A., Over, P., and Kraaij, W.: Evaluation campaigns and TRECVid, *Proc. of MIR'06*, pp.321–330 (2006).
[12] Muhling, M., Ewerth, R., Zhou, J., and Freisleben, B.: Multimodal Video Concept Detection via Bag of Auditory Words and Multiple Kernel Learning, , Advances in Multimedia Modeling, Vol. 7131, pp,40–50 (2012).
[13] Potamianos, G. , Neti, C., Gravier, G. , Garg, A., and Senior, A.: Recent Advances in the Automatic Recognition of Audio-Visual Speech, *Proc. of IEEE*, Vol.91, pp. 1306–1326 (2003).
[14] Hrybyk, A., and Kim, Y.: Combined Audio and Video Analysis for Guitar Chord Identification, *Proc. of ISMIR2010*, pp.159–164 (2010).
[15] Petridis, S., and Pantic, M.: Audiovisual Discrimination Between Speech and Laughter: Why and When Visual Information Might Help, *IEEE Trans. on Multimedia*, Vol.13, pp.216–234 (2011).
[16] Eyben, F., Petridis, S., Schuller, B. and Pantic, M.: Audiovisual Vocal Outburst Classification in Noisy Acoustic Conditions, *Proc. of ICASSP2012*, pp.5097–5100 (2012).
[17] 平井辰典，中野倫靖，後藤真孝，森島繁生：シーンの連続性と顔類似度に基づく動画コンテンツ中の同一人物登場シーンの同定，映情学誌，Vol.66，No.7，pp.J251–J259 (2012)．
[18] 平井辰典，中野倫靖，後藤真孝，森島繁生：音楽動画コンテンツ中のアーティスト名とその登場シーンの同定手法，情処研報音楽情報科学，2012-MUS-94-24，pp.1–8 (2012)．
[19] Irie, A., Takagiwa, M., Moriyama, K., and Yamashita, T.: Improvements to Facial Contour Detection by Hierarchical Fitting and Regression, *Proc. of ACPR2011*, pp.273–277 (2011)．
[20] Fujihara, H., Goto, M., Ogata, J., Okuno, H.: LyricSynchronizer: Automatic Synchronization System Between Musical Audio Signals and Lyrics, *IEEE Journal of Selected Topics in Signal Processing*, Vol.5, pp.1252–1261 (2011)．