HOKKAIDO UNIVERSITY

Title	複数特徴を利用した映像意味解析とリランキングの導入による検索の高精度化に関する研究
Author(s)	吉田, 壮
Citation	北海道大学. 博士(情報科学) 甲第12197号
Issue Date	2016-03-24
DOI	10.14943/doctoral.k12197
Doc URL	http://hdl.handle.net/2115/61840
Type	theses (doctoral)
File Information	Soh_Yoshida.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

博士論文

複数特徴を利用した映像意味解析とリランキングの導入による検索の高精度化に関する研究

吉田 壮

2016年3月

北海道大学 大学院情報科学研究科
メディアネットワーク専攻

目 次

第 1 章　序論　1

1.1　本研究の背景 1

1.2　本研究の目的 3

1.3　論文の構成 6

第 2 章　映像意味解析とリランキングに関する従来研究　8

2.1　はじめに 8

2.2　映像のシーン分割手法 9

2.2.1　ショットへの分割 10

2.2.2　シーンへの分割 11

2.3　物体識別手法 14

2.3.1　学習用コーパス 15

2.3.2　画像特徴 17

2.3.3　識別モデル 21

2.4　検索結果のリランキング手法 22

2.5　本研究で解決すべき課題と解決法 24

2.6　まとめ 25

第 3 章　複数特徴の利用による映像の構造に注目したシーン分割手法　27

3.1　はじめに 27

3.2　歌謡番組映像の構造に注目したシーン分割法 28

3.2.1　色特徴および出演者情報に基づいた有向グラフの生成 . . 29

3.2.2 音響特徴に基づいた有向グラフの生成 32
3.2.3 有向グラフの統合とシーン境界の算出 33
3.2.3.1 有向グラフの統合とシーン境界の算出 33
3.2.3.2 音響特徴に基づいた司会区間の事前判定 34
3.3 実験 35
3.3.1 映像の構造に注目したシーン分割の精度評価 35
3.3.2 顔認識のシーン分割精度へ与える影響に関する考察 41
3.4 まとめ 44

第 4 章 撮像対象の自動抽出処理を導入した物体識別手法 45
4.1 はじめに 45
4.2 Support Vector Machine 46
4.3 撮像対象の自動抽出処理を導入した SVM に基づく識別器の高精度化 47
4.3.1 正例ブロックの検出 49
4.3.2 正例ブロックを用いた識別超平面の更新 51
4.3.3 多クラス識別 53
4.4 実験 57
4.4.1 実験データセット 57
4.4.2 評価実験 58
4.5 まとめ 59

第 5 章 映像特徴とタグ情報の類似性に注目したリランキング手法 62
5.1 はじめに 62
5.2 映像特徴とタグ情報を用いた異種グラフの構築 63
5.2.1 映像グラフの構築 63
5.2.1.1 低次特徴抽出 63
5.2.1.2 重み関数の算出 64

5.2.2 タググラフの構築 64
5.2.2.1 タグフィルタリング 65
5.2.2.2 重み関数の算出 68
5.2.3 異種グラフの構築と重みの伝播 68
5.3 グラフ構造に基づく検索結果のリランキング法 70
5.3.1 グラフ構造を考慮したリランキングスコアの推定 70
5.3.2 尤度算出 72
5.3.3 リランキングスコアの最適化 73
5.4 実験 74
5.4.1 実験条件 75
5.4.2 評価方法 75
5.4.3 精度評価 76
5.5 まとめ 78

第 6 章 結論 **80**
6.1 本論文の成果 80
6.2 今後の課題 83

謝辞 **84**

著者の研究業績 **93**

図 目 次

1.1 A generic framework for video indexing and retrieval. 5

2.1 A video search framework with video semantic analysis and reranking. 9
2.2 Example of scene segmentation in [1]. 12
2.3 Example of scene segmentation using SVM training scheme in [2]. . . 14

3.1 Overview of scene structures in music programs. 28
3.2 Overview of the proposed method. 29
3.3 Scene segmentation results obtained by applying the proposed method and the other methods to Video A. Note that shadow areas in (a) represent singer introduction scenes, gray areas in (b) represent singer introduction scenes which are obtained in the proposed method, red lines in (c)-(h) represent positions where scene boundaries are correctly obtained in each method, and blue lines represent positions where scene boundaries are excluded from the detection. 38
3.4 Quantitative evaluation of the proposed method and the comparative methods based on recall and precision. 39
3.5 Verifying the Effect of Face Recognition on Scene Segmentation Accuracy based on recall and precision. 43

4.1 Overview of Support Vector Machine. 48
4.2 An example of clipping local-blocks $g_{i,s,m}^{\mathrm{pos}}$ from the positive example F_i^{pos}. 49

4.3 An example of positive-blocks $b_{i,s}^{\mathrm{pos}}$ detected from the positive example F_i^{pos} 52
4.4 An example of "the relationship between s and $f^{\mathrm{svm}}(\mathbf{x}_{i,s}^{\mathrm{pos}})$" and "the detection of s_i^{opt}". 53
4.5 Example images of Caltech 101 and Caltech 256 datasets. 55
4.6 An examples of target object detection from Caltech 101 and Caltech 256. Categories "ant", "ibis", "car-side", "helicopter", "cougar-body", "crocodile" are included in Caltech 101. All categories shown in this figure are included in Caltech 256. This means Caltech 256 includes Caltech 101. 56
4.7 Relationship between the number of training examples per class and the mean recognition rate per class. 61

5.1 MAP Comparison of nearest-neighbor number for local learning regularization. 78
5.2 MNDCG Comparison of the Video Search Performance. 79

表 目 次

3.1 Dataset used in the experiments. 36

3.2 Performance comparison of the proposed method and the machine learning-based traditional method. 42

4.1 Four kinds of groups $A_s (s = 1, 2, 3, 4)$, whose areas are $(0.8)^s A_0$ (pixels), used in our method, where $A_0 = 65536$. 50

4.2 Aspect ratios (vertical:horizontal) used for clipping local blocks of each group A_s. 50

4.3 Classification accuracy (%) comparison on Caltech 101 and Caltech 256. 58

5.1 10 Event Queries used in the experiments. 74

5.2 MAP Comparison of each Regularization. 77

第1章　序論

1.1　本研究の背景

高速・大容量な通信回線の普及に伴い，インターネットによる動画共有サービスの利用が増加している．特に，YouTube[1] に代表される動画共有サービスの拡大は著しく，膨大な量の映像が日々アップロードされ，多くのユーザがこれらの映像を視聴している．具体的な数値としては，YouTube が 2015 年に公表した統計データ [2] で，映像を視聴する利用者数はひと月平均で 10 億人を超え，利用者の視聴時間のひと月の総計は 60 億時間にのぼることが報告されている．このような状況の中で，ユーザが望む映像を効率的に検索し閲覧するための技術開発が急務とされている．

インターネットの世界では，Web ページの増加に合わせて，大規模なテキストデータを対象とする検索技術が急速に発展してきた．この検索技術の発展が，Web コンテンツの驚異的な増大とその有用性を一層高めたとも言える．テキストデータと同じように大量の映像の中から望む映像を素早く検索するため，映像検索の結果は一般に，予め映像に必要なテキスト情報（以降，メタデータ）を付与し，それとクエリとして与えるキーワードとの適合度に基づく順位付けに従い提示される [3]．このとき，Web 上の映像は極めて膨大であることから，手動によるメタデータの付与は現実的でなく，自動化が必須である．現状の研究では，1) 映像を同一意味内容のまとまりを持つ区間に分割した後，2) 分割された各区間に何が映っているのかを識別することでメタデータの付与を可能とする．これに

[1] http://www.youtube.com

[2] http://www.youtube.com/yt/press/statistics.html

より自動で得られたメタデータとクエリを比較して適合度を算出し，順位付けすることで検索結果が提示される．しかしながら，特に，2) において行われるメタデータの自動付与に誤りが生じると，自ずと順位付けに影響を与え，ユーザが望む映像が下位に配置されてしまう．

この問題を解決するため，3) 順位付けにおいて下位に存在する映像の中でユーザが望む映像の順位が高くなるように並べ替えるリランキング手法が提案されている [4]．先に示す 2) の手法の精度がある程度高い場合は，このリランキング手法の適用により，望む映像の順位を上げることが可能であり，メタデータ付与の誤りによる検索の精度劣化の低減が期待できる．

このような背景の下，1) および 2)，さらに 3) について，様々な研究が行われている．以下に，各々の従来研究とその問題点を示す．まず，1) 映像をシーンに分割する従来手法について説明する．シーンは「同一もしくは類似する内容を表すショットが連続する区間」として，ショットは「1 台のカメラにより切れ間なく撮像される区間」として定義され，映像のシーン分割は，隣接するショットの意味内容が同一もしくは類似するか否かを判定して統合することで可能となる．シーンの意味的なまとまりに関する定義は映像のジャンル毎に異なり，各々の手法で個別に設定される場合が多い [3]．代表的な従来研究に，映像の構造に注目したシーン分割手法 [1] が高精度なものとして挙げられる．文献 [1] では，映像のシーン分割問題をグラフの分割問題に置き換えることで，シーン分割を実現している．具体的に，ショットから抽出した画像特徴の類似性を基に，ショットのクラスタリングを行い，得られる各クラスタをグラフのノードとし，隣接する 2 つのショットが同一クラスタ内に存在しないときに，クラスタ間に辺を与えることで有向グラフを取得する．さらに，このグラフを複数のサブグラフに分割し，各サブグラフを 1 つのシーンとみなすことでシーン分割を実現している．しかしながら，シーン境界毎に特徴の変化が大きいジャンルの映像では，画像，音響という特徴を利用するだけでは，映像が有する繰り返し構造の検出は困難となる問題が存在した．

次に，2) 映像の各シーンの代表画像であるキーフレームに対して，機械学習

を利用し，各々の撮像物体の識別を行うことで，メタデータの付与を可能とする．従来，共通の識別フレームワークが用いられ，改良手法が様々提案されている [5]．具体的に，まず撮像物体毎に学習用画像を収集し，学習用画像から局所特徴量などの画像特徴を抽出する．それに基づき Bag of Features (BoF) [6] などの画像表現を作成し，Support Vector Machine (SVM) [7] などで学習ならびに識別を行う．しかしながら，背景の多様性や同時に撮像されている他の物体の影響から，撮像物体の識別を行う上で雑音となる特徴が含まれ，識別精度が低下する問題が存在した．

最後に，3) 検索結果のリランキング手法は，キーワード検索で得られた映像間の類似性に基づいてその実現を可能とする．具体的に，キーワード検索で下位に配置された望む映像の検出を行うため，上位の映像と類似する映像をクラスタリングで検出する手法 [8]，それらを学習データとした識別モデルの構築を行う手法 [9–11]，映像間の類似性を表現するグラフを構築し，重要度の高い映像を検出する手法 [12, 13] が提案されている．一方，動画共有サービスから提供される映像には，映像毎に関連リンクや付随するタグなど，映像の特徴以外にも映像間の関係を把握する上で有益な情報が含まれている．しかしながら，これらを活用してリランキングの精度を改善する手法はこれまで提案されていない．

1.2　本研究の目的

本論文では，映像検索における課題である，メタデータ付与の自動化のための映像意味解析技術について述べる．映像意味解析については，映像のシーン分割と撮像物体識別に大別し，それぞれの高精度化のため，複数特徴を利用した映像の構造に注目したシーン分割手法と多様な背景や他の物体による影響を抑制する物体識別手法を提案する．さらに，ユーザが与えるキーワードとメタデータに基づいて行われる検索において，望む映像を上位に配置するような検索結果を提示するためのリランキング手法の導入について述べる．以上のように，映像意味解析の各手法の精度向上とリランキング手法の新たな提案により，映像検索の高精

度化を実現することを本研究の目的としている．

具体的に，まず，1) の問題の解決を目指し，映像の複数特徴から得られるシーン境界の判定結果を協調的に用いるシーン分割手法の提案を行っている．提案手法では，各ジャンルの中でも映像構造が明確に存在する歌謡番組に焦点を当て，例えば,「司会区間」の次に「歌唱区間」が続き，それらが番組内で繰り返し現れるという歌謡番組固有の構造を利用する．この構造に基づいてショットの階層クラスタリングを行い，その結果から取得される有向グラフのカット辺を探索する．なお，ショットの階層クラスタリングを行う際には，歌謡番組に注目した画像，音響特徴を導入する．画像特徴については，シーンが遷移するときに，同時に各シーンの中心となる出演者も変化することを考慮し，新たに出演者の変化を検出する特徴を顔認識法の結果から定義する．各々の特徴から得られる結果を統合する協調処理により歌謡番組固有の繰り返し構造を検出することで，シーン分割の高精度化を実現する．

次に，2) の高精度化を目指し， 撮像対象の自動抽出処理を導入した物体識別手法の提案を行っている．具体的に，SVM を用いた識別超平面を算出し，特徴空間における特徴ベクトルと超平面との距離に基づいて，撮像物体の存在する領域を内包する矩形を自動検出する処理を提案する．提案手法では，背景領域に存在する特徴を除去することで，撮像物体の領域に限定した特徴抽出を可能とし，新たな識別超平面を算出する．これにより，背景や他の物体による悪影響を抑制した物体識別を実現する．

最後に，3) を実現する新たな手法として，映像特徴とタグ情報の類似性に注目したリランキング手法の提案を行っている．提案手法では，まず映像のリンク関係を基に，映像とそれに付随するタグそれぞれの類似性を表現可能なグラフを構築し，そのグラフ上において，映像とタグ，各々の類似性を伝播させる．さらに，キーワード検索による順位が，類似性の高い映像ノード間で近くなるようにエネルギー関数を与え，その最小化問題を解くことで，グラフの整合性を保つようなリランキングが可能となり，望む映像が上位に配置される映像検索を実現する．

以上について，一般的な映像検索フレームワークの概略と本論文の流れを図 1.1

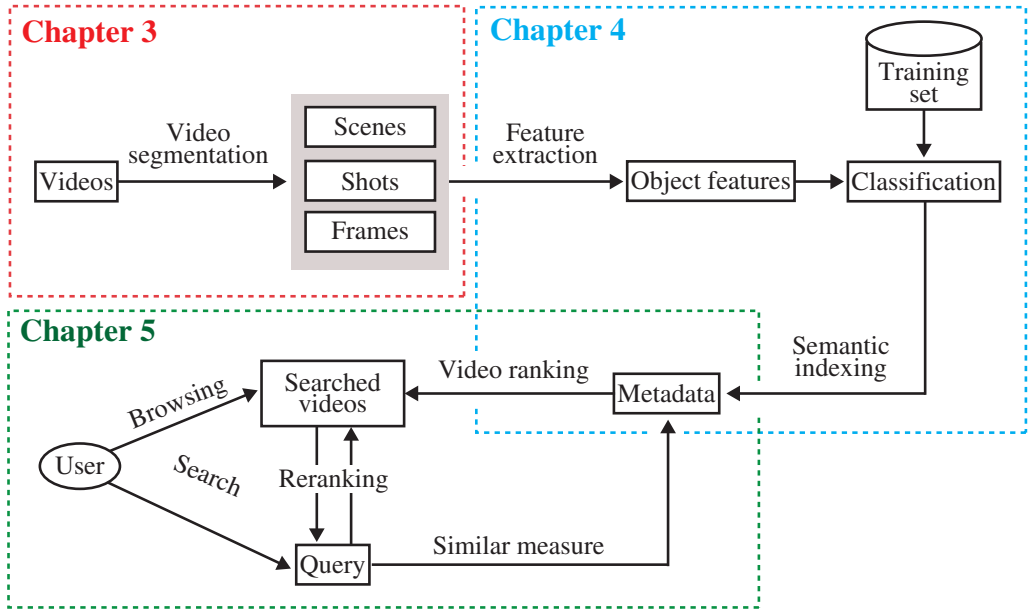


図 1.1: A generic framework for video indexing and retrieval.

にまとめる．図 1.1 に示すように，本論文で構成される研究は，映像を起点としてユーザがその映像を検索し閲覧するまでの各種の映像意味解析と検索結果の提示技術に対応する．映像意味解析として，複数特徴を利用した映像の構造に注目したシーン分割手法と多様な背景や他の物体による影響を抑制する物体識別手法の提案し，メタデータ付与の自動化へ貢献する．さらに，望む映像を上位に配置するような検索結果を提示するため，映像特徴とタグ情報の両者を用いたリランキング手法を導入する．これらの構成によって，メタデータ付与の自動化および映像検索の高精度化のための技術開発を行うことで，本研究は，ユーザが望む映像の効率的な検索と閲覧の実現に貢献する．

1.3　論文の構成

本論文は 6 章からなり，以下に各章の概要を述べる．

- 第 1 章では，研究の背景および目的について述べた．
- 第 2 章では，映像検索の高精度化を目的とした，映像意味解析とリランキングに関する従来研究について説明し，現状の課題と本論文の解決法について明確にする．
- 第 3 章では，複数特徴の利用による映像の構造に注目したシーン分割手法の提案を行う．まず，歌謡番組固有の構造を明確にするとともに，その構造を検出するのに適した画像，音響特徴を説明する．次に，特徴毎に階層クラスタリングを行い，繰り返し構造を検出し，その結果を統合することで，シーン境界判定を行う手法を提案する．実験では，複数特徴を用いることで，歌謡番組固有の構造を高精度に検出可能であることと，その構造に従うことでシーン分割の精度向上が可能となることを確認する．
- 第 4 章では，撮像対象の自動抽出処理を導入した物体識別手法の提案を行う．まず，提案手法で用いる識別器である SVM とその識別超平面の性質について説明する．次に，その性質を利用した撮像対象の自動抽出法を提案し，特徴ベクトルと識別器の更新法をそれぞれ示す．実験では，撮像物体の領域を自動で選定可能であることと，選定した領域のみから抽出した特徴に基づいて更新した識別器を用いて識別精度の向上が可能となることを確認する．
- 第 5 章では，映像特徴とタグ情報の類似性に注目したグラフ構造に基づくリランキング手法の提案を行う．まず，映像のリンク関係を基に，映像とタグそれぞれの類似性を表現可能なグラフの構築方法を示し，そのグラフ上において，映像とタグ，各々の類似性を伝播させる手法を説明する．次に，キーワード検索による順位が，類似性の高い映像ノード間で近くなるようにエネルギー関数を導出する．最後に，エネルギー関数の最小化方法

を説明する．実験では，映像特徴とタグ情報の類似性の両者の類似性を用いることの有効性を示す．

- 第 6 章では，本論文の結論を示し，今後の課題，考察を行う．

第2章　映像意味解析とリランキングに関する従来研究

2.1　はじめに

本章では，映像検索の高精度化を目的とした，映像意味解析とリランキングに関する従来研究について述べる．映像は，写真などの静止画像と異なり時間軸を持つことから，どの時刻に何が映っているかを記述したメタデータの付与が必要である．自動でメタデータ付与を行う現状の研究では，映像をシーンに分割した後，各シーンに何が撮像されているか識別することが行われ，本論文における映像意味解析は，映像のシーン分割と，撮像物体の識別に大別する．一方，メタデータの自動付与に誤りが生じると自ずと順位付けに影響を与え，ユーザが望む映像が下位に配置されてしまう．この問題を解決するための研究として，検索結果のリランキング手法が提案されている．撮像物体の識別精度がある程度高い場合は，このリランキング手法の適用により，望む映像の順位を上げることが可能であり，メタデータ付与の誤りによる検索の精度劣化の低減が期待できる．なお，図 2.1 に，上の映像意味解析とリランキングを導入した映像検索フレームワークの概要を示す．

本章の構成は以下の通りである．2.2 では，映像のシーン分割に関する従来研究について説明する．2.3 では，撮像物体識別に関する従来研究について説明する．2.4 では，検索結果のリランキングに関する従来研究について説明する．2.5 では，各々の従来研究が抱える課題と本論文の解決法について検討する．

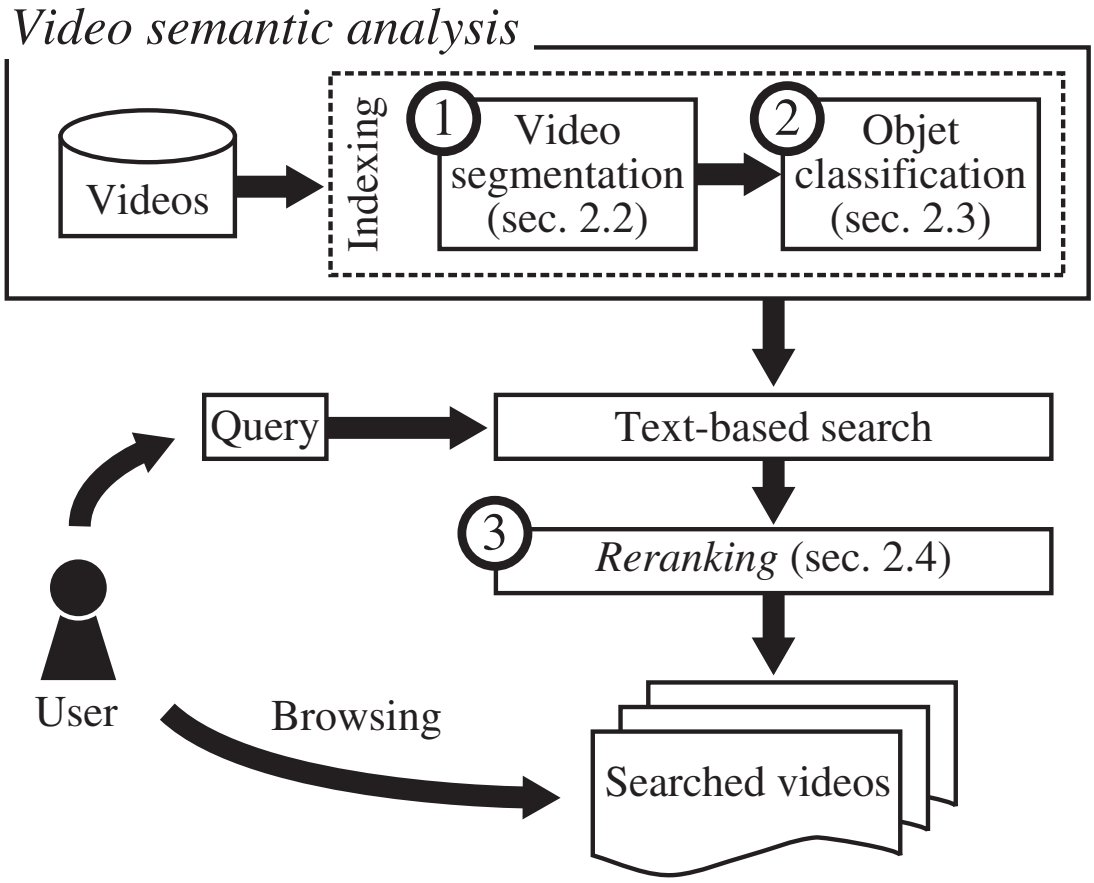


図 2.1: A video search framework with video semantic analysis and reranking.

2.2　映像のシーン分割手法

映像の分割単位には，シーン，ショットおよびフレームがある．シーンは幾つかのショットを含み，1 つのショットの中に，複数枚のフレームがある．シーン単位に分割した映像区間に，自動的にメタデータを付与できれば，検索効率は非常に高まる．また，一般に，映像のシーン境界は，映像のショット境界から選出されるため，高精度なショット分割手法が必要とされる．

本節では，準備として，映像のシーン，ショットおよびフレームの定義について述べ，これらの分割単位間の関係を説明する．シーンは，同一もしくは類似する内容を表すショットが連続する区間であり，ショットは，1 台のカメラにより

切れ間なく撮像される区間として定義されている．なお，フレームは静止画像であり，映像の最小分割単位である．映像のフレーム，ショットおよびシーンの定義では，複数枚のフレームがショットに含まれ，さらに，シーンの中に 1 つあるいは複数個のショットがある．シーンはショットより上位にある映像の分割単位であり，ショットはフレームの上位にある映像分割単位でもある．以降，映像のショット分割とシーン分割に関する従来手法についてそれぞれ説明する．

2.2.1　ショットへの分割

ショット境界検出技術は，一つのカメラで途切れることなく撮影された映像の区間であるショットを取り出すために，フレーム間の映像データの連続性を評価し，連続性が途切れた境界をショット境界とする方法を基本としている [14]．一般に，ショット境界では，前後の隣接するフレームの画像特徴が大きく変化する．したがって，連続性の評価には，隣接するフレーム間の特徴が変化するか否かを捉えることが有効であり，隣接するフレーム間の特徴量間の変化に注目するヒストグラム比較法がよく用いられる．ヒストグラム比較法については，フレーム間の RGB や HSV などの色ヒストグラムの差分を χ^2 検定を用いて比較し，画面全体ではなく 16 個程度に分割したブロック内を処理対象とする分割 χ^2 検定法 [15] が高精度なショット分割を実現している．ただし，ヒストグラム比較法は，類似度の高いショット間で切り替えがあった場合に検出できないことや，動きの速いショットで誤検出をすることがある．このため，文献 [16] では，ショット境界検出のための新しい特徴として動き方向ヒストグラム特徴を提案した．この手法では，画面全体の動きが激しいショットとカット点において動きベクトルの方向分布が異なることに注目し，動きベクトルの方向分布の違いをヒストグラムの絶対差分和により表現することで，ショットの誤検出を防ぐことを可能としている．

2.2.2　シーンへの分割

シーンの境界は，意味的な境界でもあるため，映像の内容や見る人，利用用途によっても異なる点で分割されることがあり，自動分割は困難な課題とされている．現在提案されているシーン分割の手法では，映像内の色の変化や，音響信号の連続性や構造など，比較的表層的な情報を用いて映像内容の連続性を判断し，その切れ目をシーンの切れ目として扱うことが多い [3]．そのため，初期の研究では，ショットに属するフレームから算出される色特徴や，ショット全体の音響信号のボリューム，零交差率などに基づいてグループ化する手法が提案されている [17]．各々が高精度なシーン分割を目指しているが，特徴量の改良だけでは精度向上に限界がある．現在では，映像の構造解析に基づく手法 [1] やその改良手法 [18] が提案されている．また，映像が画像，音響などの複数の要素から成り立っていることから，それらを統合して用いる処理を導入することが一般的となっている [2,17,19–21]．その中でも，色配置，動き，そして音などの特徴量を算出し，シーンが変化するショット境界を SVM に基づいて検出することで，シーン分割を実現する手法 [2] が提案されている．以降，現在主流である映像の構造解析に基づくシーン分割手法と機械学習に基づくシーン分割手法の各ベースライン手法について，それぞれの説明を行う．

映像の構造解析に基づくシーン分割

文献 [1] では，ショットの明確な繰り返し構造を有する映像を対象とし，それらの階層クラスタリング結果に基づいて作成される有向グラフの分割問題を解くことでシーン分割を可能とする．文献 [1] ではまず，ショットがあらかじめ得られていると仮定し，画像特徴に基づいてショットのクラスタリングを行う（以降，得られるクラスタをショットクラスタと呼ぶ）．さらに，ショットクラスタをノードとした有向グラフを作成し，得られた有向グラフに対してカット辺を探索することで，シーン分割を行う．以下では，具体的な処理について説明を行う．

今，N_s 個のショット $s_i (i = 1, 2, \cdots, N_s)$ が与えられたとき，文献 [1] では，階

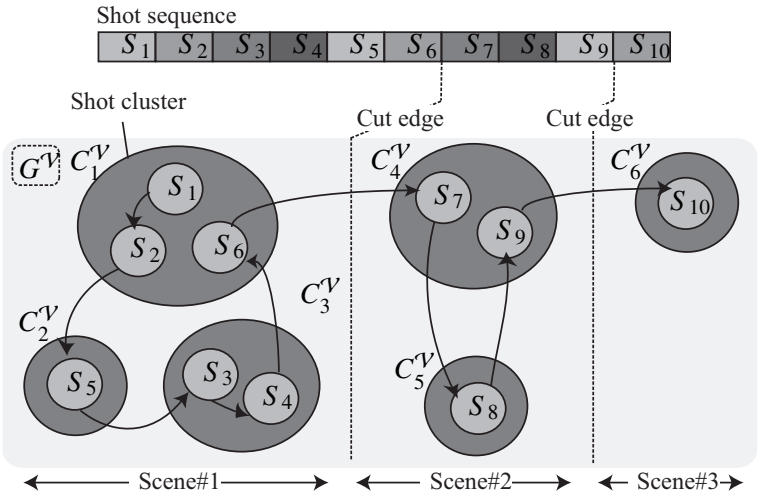


図 2.2: Example of scene segmentation in [1].

層的クラスタリングの一手法である完全連結法によってショットクラスタを生成する．具体的にはまず，N_s 個のショット s_i を構成要素とする N_s 個のクラスタ $C_1 = \{s_1\}, C_2 = \{s_2\}, \cdots, C_{N_s} = \{s_{N_s}\}$ を作成する．次に，異なる二つのクラスタ C_u, C_v 間の類似度 $Sim_C^{\mathcal{V}}(C_u, C_v)$ を次式のように定め，最も類似度の高い二つのクラスタを結合して一つのクラスタとする．

$$Sim_C^{\mathcal{V}}(C_u, C_v) = \min_{s_i \in C_u, s_j \in C_v} Sim_s^{\mathcal{V}}(s_i, s_j) \tag{2.1}$$

ここで，$Sim_s^{\mathcal{V}}(s_i, s_j)$ は異なる二つのショット s_i, s_j 間の類似度であり，ショット s_i および s_j に含まれるフレーム f_l および f_m 間の色ヒストグラムに基づいて次式のように算出される．

$$Sim_s^{\mathcal{V}}(s_i, s_j) = \begin{cases} \max\limits_{f_l \in s_i, f_m \in s_j} Sim_f(f_l, f_m) & \text{if } D_T(s_i, s_j) \leq Th_T \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

ただし，$Sim_f(f_l, f_m)$ はフレーム f_l および f_m 間の色ヒストグラムの類似度を表しており，ヒストグラムの重なり率で定義される．なお，$Sim_f(f_l, f_m)$ はヒストグ

ラムの重なり率の最大値で除することにより，正規化されているものとする．また，$D_T(s_i, s_j)$ は二つのショット s_i, s_j 間の時間的距離を表しており，ショット s_i と s_j の間に存在するフレームの数で与えられる．さらに，Th_T はあらかじめ設定される閾値である．文献 [1] では，全てのクラスタ間について，式 (2.1) で算出される類似度 $Sim_C^{\mathcal{V}}(\cdot,\cdot)$ が $Th_{C^{\mathcal{V}}}$ よりも小さくなるまで，もしくはクラスタ数が 1 になるまでクラスタの結合を繰り返し，最終的に得られるクラスタをショットクラスタ $C_k^{\mathcal{V}}$（$k = 1, \cdots, N_{C^{\mathcal{V}}}; N_{C^{\mathcal{V}}}$ はショットクラスタの総数）とする．

次に，生成されたショットクラスタを用いて有向グラフ $G^{\mathcal{V}}$ を作成する．この有向グラフ $G^{\mathcal{V}}$ では，それぞれのショットクラスタをノードとし，隣接するショット s_i, s_{i+1} がそれぞれ異なるショットクラスタ $C_u^{\mathcal{V}}, C_v^{\mathcal{V}}$ に含まれている場合に $C_u^{\mathcal{V}}$ から $C_v^{\mathcal{V}}$ へ辺が作成される．図 2.2 に有向グラフ $G^{\mathcal{V}}$ の例を示す．図 2.2 では，例えば隣接するショット s_2 および s_3 が，それぞれショットクラスタ $C_1^{\mathcal{V}}$ および $C_2^{\mathcal{V}}$ に含まれるため，$C_1^{\mathcal{V}}$ から $C_2^{\mathcal{V}}$ へ辺が作成されている．また，$C_1^{\mathcal{V}} \rightarrow C_2^{\mathcal{V}} \rightarrow C_3^{\mathcal{V}} \rightarrow C_1^{\mathcal{V}}$ および $C_4^{\mathcal{V}} \rightarrow C_5^{\mathcal{V}} \rightarrow C_4^{\mathcal{V}}$ においてショットの繰り返しが観察できる．

最後に，得られた繰り返し構造に基づき映像をシーンに分割する．このシーン分割は，作成された有向グラフ $G^{\mathcal{V}}$ において，取り除くと二つの接続されていないグラフに分かれる辺（カット辺）を探索することで実現される．図 2.2 では，ショット s_6，s_7 間および s_9，s_{10} 間において有向グラフ $G^{\mathcal{V}}$ のカット辺が存在するため，それらをシーンの境界と判定することでシーン分割が行なわれる．以上の処理により，文献 [1] では，構造解析を用いたシーン分割を実現する．

機械学習に基づくシーン分割手法

次に，文献 [2] で提案されている機械学習に基づくシーン分割手法について説明する．具体的には，分割対象映像をショットに分割し，個々のショットから抽出した特徴量をもとにシーン境界識別器を構築することにより，シーン境界であるショット境界を検出する手法である．以下では，具体的な処理について説明する．

今，N_s 個のショット $s_i (i = 1, 2, \cdots, N_s)$ が与えられたとき，文献 [2] では，各ショットより，特徴量を抽出する．具体的にはまず，ショット s_i から，音，動き，

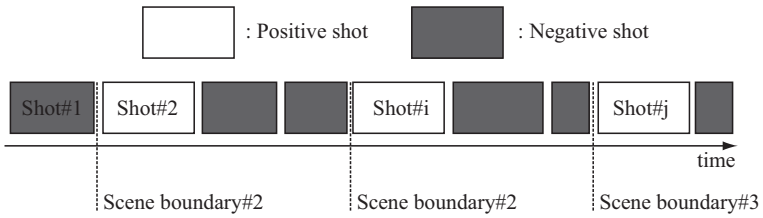


図 2.3: Example of scene segmentation using SVM training scheme in [2].

色配置，そして，時系列の 4 種類に大別した計 51 次元の特徴ベクトルを算出する．次に，算出した特徴ベクトルを利用し，シーン境界であるショット境界を抽出するための識別器を，SVM で構築することにより，シーン分割を行う．識別器を構築するために，学習データ，すなわち，予めシーン境界点が付与されているショットを用いる．そして，シーン境界出現ショットを正例，それ以外のショットを負例として，前述した 51 次元の特徴ベクトルに基づいて，SVM の学習を行う．学習処理における正例及び負例の設定方法の概要を図 2.3 に示す．

続いて，学習の結果得られた識別器を利用し，シーン境界が未知の映像に対するシーン分割を行う．まず，分割対象の映像をショットに分割し，対象映像のジャンルに基づき個々のショットから特徴を算出する．そして，各ショット全ての特徴を識別器に入力し，シーン境界が含まれれるショットを抽出する．最後に，分析対象映像のシーン境界を，上記処理によって抽出された全ショットの開始点に設定することにより，文献 [2] では，最終的なシーン分割結果を得る．

2.3　物体識別手法

通常，映像検索の際に，最も一般的な検索キーワードは，何が映っているのかという被写体の名称で,「人」「鳥」などの名詞で与えることが多い．これらの被写体の名称を，前述のシーン単位に分割した映像区間に，自動的に付与できれば，検索効率が高まる．

自動メタデータ付与は，被写体などを記述するメタデータが付与されていない映像に対して複数のメタデータを自動的に付与する手法を指す．また，映像に対しては，シーンの代表フレームを決定することで，画像認識技術が適用される．特に，画像認識では物体のカテゴリを識別する一般物体認識 [22] の研究の成果がそのまま使われることが多い．そのため，本節では，一般物体認識を用いたメタデータ付与に関する従来研究について述べる．メタデータ付与の枠組みは，実データの多様性を十分に反映したコーパスを整備した上で，画像の特徴抽出，機械学習を用いたメタデータの識別モデルの構築，モデルを用いた画像のメタデータ予測となっている．つまり，学習用コーパス，画像特徴，識別モデルが画像・映像のメタデータ予測性能に貢献することから，以下では，それぞれを取り上げ説明する．

2.3.1　学習用コーパス

機械学習を用いたメタデータ付与には，学習のためのコーパスが必要である．初期の研究では，最大数百程度のカテゴリに分類された数千から数万程度の画像による学習用コーパスを構築し，これに基づく画像の識別の研究が進められてきた．現在では，実用的な応用を考え，カテゴリ数を数千からそれ以上に増やした大規模コーパスの整備が進められている．これには，マルチメディア関連機器の低廉化および大容量化，そして，インターネットのさらなる進展などにより，画像・映像などのマルチメディア情報が大量に収集するのことが容易になってきていることが背景にある．以下では，前者を小規模コーパス，そして，後者を大規模コーパスとして，それぞれについて代表的なコーパスを説明する．

小規模コーパス

標準的に利用されているデータセットは Corel5K [23] である．これは 5000 枚の画像から構成され，4500 枚の訓練画像と 500 枚のテスト画像に分かれている．

各訓練画像にはあらかじめ 1 から 4 つのタグが付与されており，訓練画像には 371 種類のタグとテスト画像には 260 種類のタグが用いられている．また，約 2 万枚の画像を含む IAPR-TC12 [24] がある．これは各画像に付与されたキャプションから一般的な名詞を抽出して利用され，約 300 種のタグから各画像に平均 5 つ程度割り当てる．

上記は 1 枚の画像に複数タグが付与されたデータセットであるが，物体カテゴリ識別用として背景画像を含む 102 カテゴリ，約 1 万枚の画像を含む Caltech101 [25] が最も有名である．この他に 256 カテゴリ，約 3 万枚の画像を含む Caltech256 [26]，約 20 カテゴリ，約 1 万枚の画像を扱う Pascal VOC[1] がある．

大規模コーパス

代表的な一般画像認識用の大規模コーパスとして ImageNet [27] が挙げられる．ImageNet は WordNet [28] の階層的構造を有しており，現在までに 11,231,732 枚の画像と 15,589 カテゴリが収集されている．ただし，1 カテゴリあたり 500 から 1000 枚の画像が含まれている．画像はインターネット上から画像検索エンジンを用いて収集し，Amazon Mechanical Turk[2] を通じ，人手によりカテゴリに属する適切な画像を選別しており，検索エンジンのエラーによるノイズは存在しない．

また，映像解析のワークショップとして TRECVID [29] が存在する．TRECVID とは，映像意味解析および検索の高度化を目的とした競争型国際プロジェクトであり，National Institute of Standards and Technology (NIST) の主催の下，2003 年から毎年開催されている．いくつかのタスクが実施されているが，特に映像意味解析に直接関係するのが Semantic Indexing (SIN) タスクである．SIN タスクは，映像のショットから Flower や Bus などの物体を検出することを目的としており，2011 年には学習用 400 時間，評価用 200 時間の映像が用意された．

最近では YouTube などの映像共有サイトのデータを大規模に収集する例が出てきている [30]．映像共有サイトにアップロードされる映像にはユーザが付与し

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[2] https://aws.amazon.com/jp/mturk/

たタグが存在するが，タグが十分に付与されず映像内容を表現できていない問題が存在するため，適切な学習データが不足している．また，アップロードされた動画の質の低さや扱うテーマの幅の広さから困難な課題と言える．

2.3.2　画像特徴

まず画像特徴を局所特徴と大域特徴に分類する．次に，局所特徴から一つの画像特徴を得るには多数存在する局所特徴をまとめる枠組みが必要となる．局所特徴から大域的な特徴を生成するフレームワークとして Bag of Features（BoF）を説明する．

局所特徴と大域特徴

画像特徴は，局所特徴と大域特徴に分類できる．よく用いられる局所特徴として，輝度勾配ヒストグラムを基盤とした SIFT [31]，HOG [32]，SURF [33]，ある着目領域と周囲の類似度を表現した Self Similarity [34]，着目点からの距離に応じて形状を曖昧に表現し，着目点と形状の曖昧な表現との関係を記述する Geometric Blur [35] などがある．大域特徴として，HLAC [36]，GIST [37]，カラーヒストグラムなどがある．特徴点の検出方法にもよるが局所特徴は一枚の画像あたり数百から数万点得られる．そのため特徴点検出手法や特徴点における記述子の計算コストは，データセットが大規模になれば大きな負担となる．しかしながら各特徴点の検出や記述は他の特徴点と独立に計算できる場合が多く，SiftGPU[3] や CUDASURF[4] のように近年盛んに開発されている GPU を用いて高速化が実現されている．

[3]http://www.cs.unc.edu/ccwu/siftgpu/

[4]http://www.mis.tu-darmstadt.de/surf

Bag of Features

画像識別で標準的に利用されている枠組みが Bag of Features (BoF) [6] である．BoF に基づいた画像識別は 1) 局所特徴抽出 (feature extraction)：画像の各特徴点から局所特徴を抽出する．2) コーディング (coding)：codebook を生成・各 codeword に局所特徴を割り当てる．3) プーリング (pooling)：各 codeword に割り当てられた局所特徴を統合し，一つの画像特徴として生成する．そして最後に得られた特徴から SVM などの識別器を構築することにより識別が行わる．この処理の流れは *feature extraction-coding-pooling* pipeline と呼ばれ，近年報告されている画像認識手法はこのフレームワークに沿ったものが大多数を占める [38–40]．このフレームワークの中で，コーディングを行う際に特徴を符号化することから分類に有効な情報が損失されてしまっていると，2008 年に Boiman らにより指摘された [41]．従来より用いられていたに BoF は，一つの局所特徴をただ一つの codeword に割り当てる．これはベクトル量子化 (Vector Quantization，VQ) であり，簡便な処理である一方で，このような hard-assignment coding を行った場合，その制約は強力であり，量子化誤差の発生は避けられない．そのため，文献 [41] の発表以降，様々な局所特徴のコーディングおよびプーリングの高精度化が活発に研究され，今もその注目度は高い．また，それと同時に画像の分類精度は大幅に上昇した．以降では，コーディングおよびプーリングについて代表的な手法を紹介する．なお，説明のため，$\mathbf{b}_i$ ($\mathbf{b}_i \in \mathbb{R}^d$) を合計 n 個の codeword とする．ただし，d は局所特徴の次元を表す．行列 $\mathbf{B}_{d\times n} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n)$ を codebook とする．$\mathbf{x}_i(\mathbf{x}_i \in \mathbb{R}^d)$ を，1 枚の画像から算出された i 番目の局所特徴とし，$\mathbf{u}_i(\mathbf{u}_i \in \mathbb{R}^n)$ を $\mathbf{x}_i$ が属する codeword とする．

1) Hard-assignment coding

   前述した通り，一般的な BoF では，各局所特徴をただ一つのコードワードに割り当てる．具体的に，二つの局所特徴点間は次の様にユークリッド距

離で定義される．

$$u_{ij} = \begin{cases} 1 & \text{if } j = \arg\min_{j=1,\cdots,n} \|\mathbf{x}_i - \mathbf{b}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

2) Soft-assignment coding

VQ による量子化誤差を軽減させる方法として，最も簡単なのは，局所特徴を一つの codeword ではなく，複数の codeword に割り当てることである．文献 [42] では，次式に基づき局所特徴を複数の codeword に割り当てる．

$$u_{ij} = \frac{\exp(-\beta\|\mathbf{x}_i - \mathbf{b}_j\|_2^2)}{\sum_{k=1}^{n} \exp(-\beta\|\mathbf{x}_i - \mathbf{b}_k\|_2^2)} \tag{2.4}$$

上式において，β はコーディングの柔軟さを制御するパラメータである．また，文献 [42] では，n 個全ての codeword を用いて u_{ij} を算出している．

3) Sparse Coding

Soft-assignment coding により局所特徴の柔軟なコーディングが可能となるが，上記で示したように，n 個全ての codeword を用いて u_{ij} を算出しているため，局所特徴の粗い近似となる．これに対して，文献 [38] では，codeword を用いてより正確に局所特徴を表現するために，$\mathbf{u}$ に L1 正則化の制約を与え，$\mathbf{u}$ の少数の要素が非ゼロとなるようにした．これは，局所特徴が少数の codeword に属することを意味する．これにより，VQ は次式で示すように Sparse Coding (SC) の問題として捉えることが可能となる．

$$\mathbf{u}_i = \arg\min_{\mathbf{u}\in\mathbb{R}^n} \|\mathbf{x}_i - \mathbf{B}\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1 \tag{2.5}$$

ただし，λ は正則化パラメータである．

4) Locality-constrained Linear Coding

SC と異なり，Locality-constrained Linear Coding (LLC) [39] はスパース性よ

りも局所性を優先させる．これにより，$\mathbf{x}_i$ に対して遠くに位置する局所特徴に制約を設ける．LLC では次式による最適化によりコーディングが行われる．

$$\begin{aligned} \mathbf{u}_i &= \arg\min_{\mathbf{u}\in\mathbb{R}^n} \|\mathbf{x}_i - \mathbf{B}\mathbf{u}\|_2^2 + \lambda\|\mathbf{d}_i \odot \mathbf{u}\|_2^2 \\ s.t.\ \mathbf{1}^{\mathrm{T}}\mathbf{u}_i &= 1 \end{aligned} \tag{2.6}$$

ただし，$\mathbf{d}_i = \exp(\mathrm{dist}(\mathbf{x}_i, \mathbf{B})/\sigma)$，$\mathrm{dist}(\mathbf{x}_i, \mathbf{B}) = \mathrm{dist}(\mathbf{x}_i, \mathbf{b}_1), \mathrm{dist}(\mathbf{x}_i, \mathbf{b}_2), \cdots, \mathrm{dist}(\mathbf{x}_i, \mathbf{b}_n)$ は各 $\mathbf{b}_j$ と $\mathbf{x}_i$ 間のユークリッド距離により定義される．また，σ は重みベクトル $\mathbf{d}_i$ を制御するパラメータである．さらに，文献 [39] では，高速化のため，式の第二項を無視し，$\mathbf{x}_i$ の k 近傍を直接選択することにより，第一項目の最小化を行なっている．

コーディングにより得られた $\mathbf{u}_i$ を画像特徴として利用するためには，$\mathbf{u}_i$ を統合して一つの特徴ベクトルして表現するプーリングを行う必要がある．プーリングには以下に示す二つの処理が存在する [43].

1) Sum-pooling / average-pooling
   sum-pooling では，$\mathbf{p}$ の i 番目の成分を，$p_j = \sum_{i=1}^{l} u_{ij}$ と定義する．ただし，l は一つの画像から得られる局所特徴数を表す．続いて，average-pooling では，p_j の要素数 l で平均を取ったものと定義する．

2) Max-pooling
   max-pooling では，$\mathbf{p}$ の i 番目の成分を，$p_j = \max_i u_{ij} (i = 1, 2, \cdots, l)$ と定義する．max-pooling は sum- / average-pooling より良い分類性能を示すことが実験的に示されている．

以上までに説明したコーディングおよびプーリングは量子化誤差による情報損失を如何に抑えるかに注目した手法である．これに加えて BoF では，画像空間情報が失われてしまうという問題が存在する．このような問題を解決するため，

画像空間情報を表現する手法として Spatial Pyramid Matching (SPM) が提案されている [40]．これは，画像を一定間隔のグリッドに分割し，分割されたセル内で BoF のヒストグラムを求める．そして，分割のレベルを変えて得られた全てのセルのヒストグラムをつなげて一つの特徴ベクトルとする．コーディングおよびプーリングに基づいた画像分類では，SPM を用いることが一般的となっている．このとき，コーディング法の違いにより，SPM の様々な拡張が行われている．簡単に説明すると，例えば，VQ により得られた結果を用いた手法 [40]，SC を導入した手法 [38]，そして，文献 [44] では，Laplacian matrix を用いた正則化を式 (2.6) に第 3 項目として追加し，精度向上を実現している．上記はそれぞれ，KSPM，ScSPM，LScSPM と呼ぶ．以上のように，SPM は画像分類に有効な処理として用いられている．

2.3.3　識別モデル

識別対象のクラスが増えると低次元の特徴ベクトルでは十分な識別性能が得られなくなるため特徴ベクトルの次元は増大させる必要がある．実際に S'anchez らによって，高い識別率を維持するためには特徴ベクトルの高次元化が重要であることを実験的に示している [45]．特徴ベクトルが高次元であっても性能を示すモデルが必要となるため，次元の呪いにかからないと言われている SVM が一般に用いられる．なお，SVM の詳細については，第 4 章で後述する．クラス数が増えた場合，通常の 2 クラス識別機である SVM を多クラスに拡張する必要があるが，ほとんどの場合 1-vs-all SVM が利用される．1-vs-all SVM は各クラス識別機の出力値の大小関係が正規化されていないために多クラス識別問題において適切である保証はないが，多くの物体識別で十分に高い性能が得られることが報告されている．

2012 年に，Neural Network [46] を多階層化した Deep Learning と呼ばれる手法が一般物体認識に適用され，これまでの認識精度を大幅に向上させるとともに，認識に有効な特徴量を自動的に学習可能となった [47]．Neural Network やそれを

多段化する技術は古くから知られていたが，近年のコンピューターの性能の進歩や，インターネット上の大量の映像や画像の利用と，それに適した学習手法の開発により，良好な性能を得ることができるようになった．一方，学習に時間を要することや，大量のデータが必要であること，パラメーターの調整が複雑であることなどから，実用化には時間を要する．

2.4　検索結果のリランキング手法

リランキングは，映像特徴の類似性を考慮して検索順位（ランキング）を並べ替え，望映像が上位の結果として与えられる手法であり，キーワード検索の高精度化のための一手法として，盛んに研究が行われている [4,8–13]．従来のリランキング手法は，機械学習を用いて検索上位の映像と類似する映像の識別モデルを構築する手法 [9–11]，映像特徴に基づいた映像のクラスタリングから類似映像を検出する手法 [8]，Web 映像のリンク関係などで定義されるグラフ構造を解析して重要度の高い映像を検出する手法 [12,13] などが提案されている．このように，リランキング手法は，モデルベース，クラスタリングベース，そして，グラフベースに大別される．その中でも，グラフ構造を解析するリランキングは，映像共有サイトで定義されるリンク関係を利用可能であることから，効率的に類似映像を検出可能であることが知られている [4]．そこで，本節では，グラフ構造に基づいたリランキングに注目して説明する．

具体的に，文献 [12] では，画像特徴間の類似度により重み付けられたネットワークに対し，PageRank [48] を適用して重要度の高い画像をランキングする手法である VisualRank を提案した．さらに，文献 [13] では，映像間の類似度を表すネットワークとキーワード検索によるランキングの両者を考慮して定式化されるエネルギー関数の最小化問題を解き，最適なランキングを推定する．以降，文献 [12] で提案された VisualRank について説明し，文献 [13] については，第 5 章で後述する．

VisualRank

VisualRank は PageRank [48] を画像に適用したものである．PageRank は Web ページのランク付けを行うためのアルゴリズムであり，Web ページ間のリンク構造を行列で表現する．各ページのランクについては，反復計算によってその値が求められる．したがって，多くの重要なページからリンクされているページが上位にランク付けされる．具体的に，VisualRank は，リンク構造を表現する行列の代わりに，画像の類似度行列を用いられ，次式に示す更新式を反復することで求める．

$$R = S \times R \tag{2.7}$$

ただし，R は各画像の VisualRank 値を並べた列ベクトル，S は類似度行列の各列を正規化した行列をそれぞれ表す．また，S の各列は値の合計が 1 になるように正規化されているものとする．VisualRank の前に重要な画像が予想可能な場合は，重要と予想される画像の VisualRank 値の初期値を大きくすることが可能である．式 (2.7) は，反復計算によって R は S の絶対値最大固有値に対応する固有ベクトルへと収束する．類似度行列の大きさは画像数の 2 乗となるため，計算量を抑えるためにはランク付けの対象となる画像の数を減らす必要がある．そこで，対象画像を絞り込むためには，高速なテキストベースな手法が用いられ，画像検索での上位画像を対象画像とし，それらの再ランク付けに VisualRank を用いるようにすることで，計算量を抑えることを可能としている．また，ランク値の初期値を画像によって変え，Jing らは，画像検索での上位 1000 枚を，VisualRank を用いて再ランク付けしている．式 (2.7) に示した更新式に，PageRank での補正ベクトル P を加えたものが次式で表される．

$$R = \alpha(S \times R) + (1 - \alpha)P \tag{2.8}$$

ただし，$0 \leq \alpha \leq 1$ はパラメータである．VisualRank では，式 (2.8) の更新で得られた R の値に従い，最終的なランキングを提示可能とする．

2.5　本研究で解決すべき課題と解決法

本節では，映像検索の高度化を実現するために，本論文で解決すべき課題について明確にする．これまでの節で示した従来研究の課題と本論文の解決法を以下にまとめる．

1) **映像のシーン分割に関する研究**

映像シーン分割に関する従来研究の問題は，映像から算出される画像，音響などの複数特徴を効果的に用いておらず，映像が有する繰り返し構造の検出が困難であることである．そこで，本論文では，映像の複数特徴から得られるシーン境界の判定結果を協調的に用いるシーン分割手法を提案する．提案手法では，各ジャンルの中でも映像構造が明確に存在する歌謡番組に焦点を当て，例えば,「司会区間」の次に「歌唱区間」が続き，それらが番組内で繰り返し現れるという歌謡番組固有の構造を利用する．この構造に基づいてショットの階層クラスタリングを行い，その結果から取得される有向グラフのカット辺を探索する．なお，ショットの階層クラスタリングを行う際には，歌謡番組に注目した画像，音響特徴を導入する．画像特徴については，シーンが遷移するときに，同時に各シーンの中心となる出演者も変化することを考慮し，新たに出演者の変化を検出する特徴を顔認識法の結果から定義する．各々の特徴から得られる結果を統合する協調処理により歌謡番組固有の繰り返し構造を検出することで，シーン分割の高精度化を試みる（第 3 章）．

2) **撮像物体の識別に関する研究**

撮像物体識別に関する従来研究の問題は，対象物体を含む画像全体から抽出される局所特徴などを用いて識別モデルを構築するため，背景の多様性

や同時に撮像されている他の物体の影響から，撮像物体の識別を行う上で雑音となる特徴が含まれ，識別精度が低下することである．そこで，本論文では，撮像対象の自動抽出処理を導入した識別器の提案する．具体的に，SVM を用いた識別超平面を算出し，特徴空間における特徴ベクトルと超平面との距離に基づいて，撮像物体の存在する領域を内包する矩形を自動検出する処理を提案する．提案手法では，背景領域に存在する特徴を除去することで，撮像物体の領域に限定した特徴抽出を可能とし，新たな識別超平面を算出する．これにより，背景や他の物体による悪影響を抑制した物体識別を試みる（第 4 章）．

3) **検索結果のリランキングに関する研究**

検索結果のリランキングに関する従来研究の問題は，キーワード検索で得られた映像間の類似性に基づいてその実現を可能とする手法が提案されてきた一方で，動画配信サービスから提供される映像に付随する，関連リンクやタグなどの情報を活用してリランキングの精度を改善する手法が提案されていなかったことである．従って，本論文では，映像特徴とタグ情報の類似性に注目したリランキング手法を新たに提案する．提案手法では，まず映像のリンク関係を基に，映像とタグそれぞれの類似性を表現可能なネットワークを構築し，そのネットワーク上において，映像とタグ，各々の類似性を伝播させる．さらに，キーワード検索による順位が，類似性の高い映像ノード間で近くなるようにエネルギー関数を与え，その最小化問題を解くことで，グラフの整合性を保つようなリランキングが可能となり，望む映像が上位に配置される映像検索を試みる（第 5 章）．

2.6　まとめ

本章では，映像意味解析とリランキングに関する従来研究について説明した．また，従来手法における問題点について述べ，映像意味解析とリランキングを導

入した検索を実現するために，本論文で解決すべき課題とその解決法ついて明確にした.

第3章　複数特徴の利用による映像の構造に注目したシーン分割手法

3.1　はじめに

本章では，複数特徴の利用による映像の構造に注目したシーン分割手法を提案する．まず，歌謡番組固有の映像構造を説明する．歌謡番組では，司会区間が次の曲を紹介するなど番組進行を行うためのシーンとなり，番組内で繰り返し現れる．そして，司会区間の次には曲紹介を受けて歌手による歌唱区間が続く．つまり，提案手法では，図 3.1 に示す様に，歌謡番組において司会者と歌手が繰り返し出演する構造を有することに注目する．したがって，提案手法では，この構造に基づいてショットの階層クラスタリングを行い，その結果から取得される有向グラフのカット辺を探索することで，歌謡番組映像のシーン分割を可能とする．さらに，提案手法では，ショットの階層クラスタリングを行う際に歌謡番組に注目した画像，音響特徴を導入する．まず，画像特徴については，事前に収集された各出演者の顔画像を学習データとした核非線形相互部分空間法 (Kernel Mutual Subspace Method，KMSM）[49] を用いることで得られる類似度を利用する．また，音響特徴には，無音，音楽，音声，雑音の音響信号の分類手法により得られる結果を用いる．提案手法では，これらの特徴から得られる結果を統合した協調処理により，従来手法では困難であった歌謡番組におけるシーン分割を実現する．

本章の構成は以下の通りである．3.2 では，歌謡番組映像の構造に基づいたシー

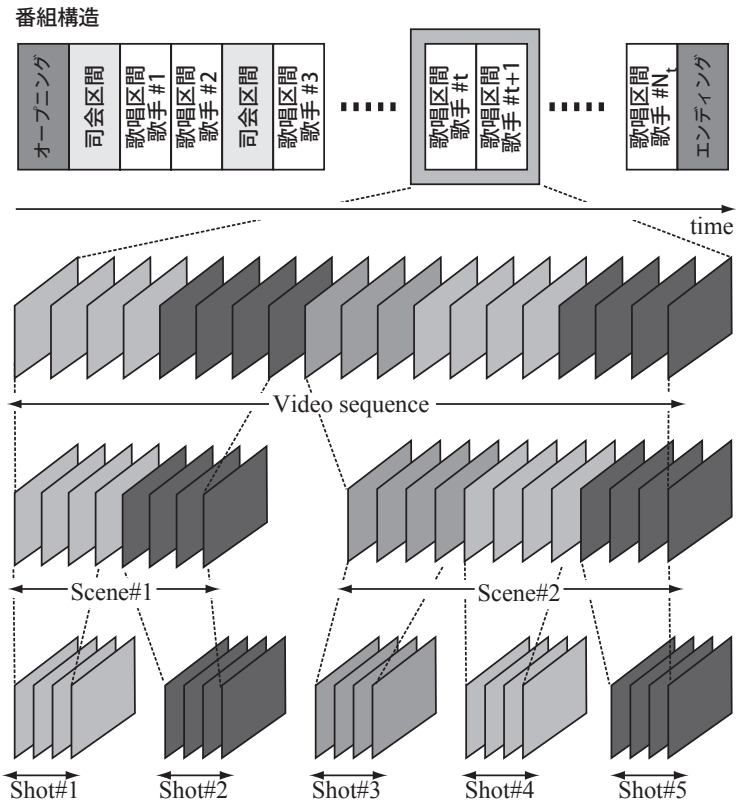


図 3.1: Overview of scene structures in music programs.

ン分割手法を提案する．3.3 では，提案手法の有効性を確認するため，評価実験の結果を示す．

3.2　歌謡番組映像の構造に注目したシーン分割法

本節では，歌謡番組映像の構造に注目したシーン分割法を提案する．提案手法では，上記のシーン分割を行うため，映像中のショットから画像特徴量および音響特徴量を抽出し，各々の特徴量から 2 種類の有向グラフを生成する．さらに，生成した 2 種類の有向グラフのシーン分割結果を統合する協調処理を導入することで歌謡番組が有する構造を表現することが可能となる．このとき，シーン境界

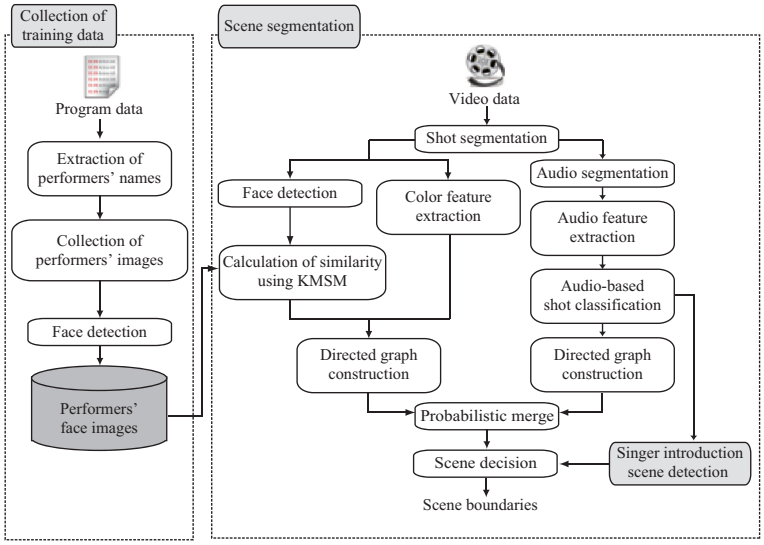


図 3.2: Overview of the proposed method.

検出の前処理として，誤検出の抑制のため，司会区間の判定処理を導入する．以上により，提案手法では，高精度なシーン分割を実現する．なお，図 3.2 に，提案手法の概略図を示す．

以降，3.2.1 では，画像特徴として色特徴および出演者情報に基づいた有向グラフの生成，3.2.2 では，音響特徴に基づいた有向グラフの生成，そして，3.2.3 では，司会区間の判定を事前処理に加え，二つの有向グラフの統合結果を用いたシーン境界の決定方法について説明する．

3.2.1　色特徴および出演者情報に基づいた有向グラフの生成

提案手法では，対象の映像から取得される HSV 色ヒストグラムおよび検出された顔領域の KMSM による各出演者との類似度に基づいて有向グラフの生成を行う．歌謡番組において，出演者の変化を検出することは有効であるが，出演者が全てのショットに存在しているとは限らない．そこで，そのような場合，一つのシーンでは同一のステージ（照明，セット）が用いられ，画像特徴として色特

徴が類似することが予想されるため，提案手法ではその利用を検討する．

まず，2.2.2 と同様に HSV 色ヒストグラムに基づいてショット s_i, s_j 間の類似度 $Sim_s^{\mathcal{V}}(s_i, s_j)$ を算出する．次に，色特徴に加えて，各ショットに登場する出演者の顔領域を検出し，各出演者の顔画像を学習データとした KMSM に基づき算出される顔画像間の類似度を新たな特徴量として導入する．これにより，隣接ショット間における出演者の変化の検出が期待できる．このとき，学習による顔の識別処理ではなく，番組内の全顔画像を画像特徴でクラスタリングするなどの方法も考えられるが，番組内に出現する顔画像は向きが一定ではなく，画像特徴のクラスタリングにより同一人物を一つのクラスタにまとめることは困難である．よって，一人の人物につき様々な向きの顔画像を学習した識別処理を採用する．さらに，グループを扱う場合には顔の分類だけではなく，グループのメンバーであることを判断する必要があり，識別処理の必要が生じる．

KMSM は，相互部分空間法（Mutual Subspace Method，MSM）[50] にカーネル非線形主成分分析（Kernel Principle Component Analysis，KPCA）[51] を適用して非線形拡張したものであり，顔画像に対して高い識別性能を示すことが知られている．KMSM では，MSM と同様，辞書側と入力側の部分空間のなす角度に基づいて識別を行う．提案手法ではまず，各出演者の顔画像を学習データとした KMSM に基づき，辞書側部分空間を作成する．また，ショット s_i の先頭フレームから AdaBoost [52] を用いて顔領域を検出し，この検出位置を初期位置として，Mean-shift アルゴリズム [53] を用いることで，ショット s_i 中の各フレームに対して顔領域の追跡を行う．これにより，ショット s_i 中に存在する出演者の顔画像集合が収集され，それらから入力側部分空間が作成される．

以上の結果から，提案手法では，ショット s_i 中から検出された顔領域と既知の出演者との類似度を次式で求める．

$$\phi_u(s_i) = \frac{1}{N_p^{s_i}} \sum_{h=1}^{N_p^{s_i}} S_{h,u}^{s_i} \tag{3.1}$$

ただし，$N_p^{s_i}$ はショット s_i 中から検出された出演者数，$S_{h,u}^{s_i}$ は，KMSM を用いて

ショット s_i 中から h 番目に検出された出演者の顔領域の画像集合を入力側部分空間とした場合の u 番目の辞書側部分空間との類似度を示す．なお，KMSM を用いた類似度の算出方法の詳細については文献 [49] に示されている．ここで，提案手法では，出演者が属する歌手グループを考慮するため，ショット s_i 中に出演者 u が属する歌手グループ $V^g(g = 1, 2, \cdots, N_G$；$N_G$ は対象映像に登場する歌手グループ数) が存在する確率 $\psi_g(s_i)$ を次式で定義する．

$$\psi_g(s_i) = \max_{u \in V^g} \phi_u(s_i) \tag{3.2}$$

ただし，$\psi_g(s_i)$ は $\sum_{g=1}^{N_G} \psi_g(s_i) = 1$ となるように正規化されているものとする．以上のようにして得られる類似度により，提案手法では出演者情報に基づいた二つのショット s_i と s_j 間の類似度を χ^2 距離を用いて次式で定義する．

$$Sim_s^{\mathcal{VP}}(s_i, s_j) = \begin{cases} 1 - \sqrt{\sum_{g=1}^{N_G} \dfrac{(\psi_g(s_i) - \psi_g(s_j))^2}{\psi_g(s_i) + \psi_g(s_j)}} & \text{if } D_T(s_i, s_j) \leq Th_T \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

最後に提案手法では，異なる二つのクラスタ C_u, C_v 間の類似度 $Sim_C^{\mathcal{VP}}(C_u, C_v)$ を次式のように定義し，最も類似度の高い二つのクラスタを結合して一つのクラスタとする．

$$\begin{aligned} Sim_C^{\mathcal{VP}}(C_u, C_v) \quad = \quad & \min_{s_i \in C_u, s_j \in C_v} \alpha_{i,j} Sim_s^{\mathcal{V}}(s_i, s_j) \\ & + (1 - \alpha_{i,j}) Sim_s^{\mathcal{VP}}(s_i, s_j) \end{aligned} \tag{3.4}$$

ただし，$\alpha_{i,j}$ はショット s_i, s_j 共に出演者が検出されている場合 0，少なくともどちらか一方のショットにおいて出演者が検出されていない場合 1 となる．このように定義された類似度を用いてショットのクラスタリングを行うことで，新たに出演者の変化を検出することが可能となる．完全連結法におけるクラスタの統合は，全て

のクラスタ間の類似度が$Th_{C^{\mathcal{VP}}}$よりも小さくなるまで，もしくはクラスタ数が 1 になるまで繰り返され，最終的に得られるクラスタを$C_k^{\mathcal{VP}}$ ($k = 1, 2, \cdots, N_{C^{\mathcal{VP}}}; N_{C^{\mathcal{VP}}}$はクラスタの総数) とする．さらに，得られたクラスタを用いて，2.2.2 と同様にして，有向グラフ$G^{\mathcal{VP}}$を生成する．

3.2.2　音響特徴に基づいた有向グラフの生成

提案手法では，対象映像に対して音響特徴に基づいたショット分類法 [54] を適用する．文献 [54] では，ショットs_i中の音響信号に対して，四つのオーディオクラスに属する確率（以降，帰属確率）$\phi_{i,c}$（$c = 1, \cdots, 4$; それぞれ無音, 音楽, 音声, 雑音に対応）の算出を可能とする．提案手法では，各区間の音響信号が属するオーディオクラスの帰属確率を音響特徴量とすることで，音楽の有無の差を明確に見出し，司会区間の繰り返し構造の検出を可能とする．ここで，異なるショットs_iおよびs_j間の類似度$Sim_s^{\mathcal{A}}(s_i, s_j)$を次式のように定義する．

$$Sim_s^{\mathcal{A}}(s_i, s_j) = \begin{cases} \sum_{c=1}^{4} \min(\phi_{i,c}, \phi_{j,c}) & \text{if } D_T(s_i, s_j) \leq Th_T \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

ただし，Th_Tはあらかじめ設定される閾値である．このように定義された類似度を用いてショットのクラスタリングを行うことで，オーディオクラスの帰属確率に注目したクラスタリング結果が得られる．なお，ショットのクラスタリングには，文献 [1] と同様に完全連結法を用いる．このとき，完全連結法におけるクラスタの統合は，全てのクラスタ間の類似度が$Th_{C^{\mathcal{A}}}$よりも小さくなるまで，もしくはクラスタ数が 1 になるまで繰り返され，最終的に得られるクラスタを$C_k^{\mathcal{A}}$ ($k = 1, 2, \cdots, N_{C^{\mathcal{A}}}; N_{C^{\mathcal{A}}}$はクラスタの総数) とする．さらに，得られたクラスタを用いて，2.2.2 と同様にして，有向グラフ$G^{\mathcal{A}}$を生成する．

3.2.3　有向グラフの統合とシーン境界の算出

本節では，3.2.1 および 3.2.2 で生成された有向グラフからカット辺を探索し，その時刻がシーン境界かどうかを判定することでシーン分割を行う．ここで，提案手法には，あらかじめ設定する必要がある閾値が複数存在し，それらの値を変動させることでシーン分割の精度は大きく変化する．しかしながら，それらの最適な値を見つけ出し，設定することは容易ではない．そこで，我々は，これらの閾値を変更してシーン分割を複数回行い，得られた複数の結果を統合することで最終的な結果を取得する．これにより，提案手法では事前の閾値設定が不要となる．さらに，我々は，音響特徴を用いて，あらかじめ司会区間の判定を行う事前処理を導入する．これにより，誤検出となり得るシーン境界の候補をあらかじめ除去することが可能となる．以降，3.2.3.1 では，有向グラフの統合とシーン境界の算出法について，3.2.3.2 では，音響特徴を用いたシーン境界判定の事前処理について説明する．

3.2.3.1　有向グラフの統合とシーン境界の算出

今，n 回目の試行におけるシーン分割結果を $I_i^{(n)}$ （$i = 1, 2, \cdots, N_s - 1$）とし，$s_i$ および s_{i+1} 間にカット辺が存在する場合に $I_i^{(n)} = 1$，存在しない場合には $I_i^{(n)} = 0$ とする．すなわち，$I_i^{(n)}$ は次式のように定義される．

$$I_i^{(n)} = \begin{cases} 1 & \text{if a cut edge exists between } s_i \text{ and } s_{i+1} \text{ in } n\text{-th result} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

ここで，提案手法では 2 種類の有向グラフを生成しているため，Th_T と $Th_{C^{\mathcal{VP}}}$ の値を変更することで得られるシーン分割結果 $I_i^{(n^{\mathcal{VP}})}$ $(n^{\mathcal{VP}} = 1, 2, \cdots, N^{\mathcal{VP}})$，および Th_T と $Th_{C^{\mathcal{A}}}$ の値を変更することで得られるシーン分割結果 $I_i^{(n^{\mathcal{A}})}$ $(n^{\mathcal{A}} = 1, 2, \cdots, N^{\mathcal{A}})$ の 2 種類の結果が得られる．そこで，提案手法ではこれらの結果をそれぞれ以下の式で統合し，s_i および s_{i+1} 間にシーン境界が存在するスコアを

$P_i^{\mathcal{VP}}$ および $P_i^{\mathcal{A}}$ と定義する．

$$P_i^{\mathcal{VP}} = \frac{1}{N^{\mathcal{VP}}} \sum_{n^{\mathcal{VP}}=1}^{N^{\mathcal{VP}}} I_i^{(n^{\mathcal{VP}})} \tag{3.7}$$

$$P_i^{\mathcal{A}} = \frac{1}{N^{\mathcal{A}}} \sum_{n^{\mathcal{A}}=1}^{N^{\mathcal{A}}} I_i^{(n^{\mathcal{A}})} \tag{3.8}$$

最後に，提案手法では，最終的なシーン分割結果を以下の式で取得する．

$$P_i = w_i^{\mathcal{VP}} P_i^{\mathcal{VP}} + w_i^{\mathcal{A}} P_i^{\mathcal{A}} \tag{3.9}$$

ただし，$w_i^{\mathcal{VP}}$ 及び $w_i^{\mathcal{A}}$(ただし $w_i^{\mathcal{VP}} + w_i^{\mathcal{A}} = 1$) は二つの値を制御するパラメータである．最終的に，対象映像において，P_i が閾値 T_s よりも高い値を示す境界をシーン境界として分割する．

以上により，提案手法では，映像中の色特徴と出演者情報および音響特徴に注目して得られるそれぞれのシーン分割結果を統合することで，歌謡番組における司会区間と歌唱区間の効果的な分割が可能となる．さらに，異なるシーン間における出演者の変化を有効に用いることが可能となり，高精度なシーン分割が実現される．

3.2.3.2　音響特徴に基づいた司会区間の事前判定

前節で説明したように，提案手法では，文献 [54] に基づいて算出される帰属確率を音響特徴として用いる．提案手法では，算出された音響特徴から得られる有向グラフによりシーン境界の判定を行う，しかしながら，対象映像を歌謡番組にした場合，より簡便な処理により，司会区間の検出が可能になる．このように，歌謡番組の映像構造を利用し，あらかじめ，異なる歌謡番組においても共通して存在する司会区間を検出することで，シーン境界判定の誤検出の抑制が可能となる．

具体的に，歌謡番組における司会区間の検出では，音楽の有無の差を明確に見出すことが重要となる．かつ司会区間では，4 つの帰属確率の中で音声が顕著に現れると考えられることから，音楽 $(\phi_{i,3})$ と音声 $(\phi_{i,2})$ に対応する帰属確率の比を r_i として次式のように算出する．

$$r_i = \frac{\phi_{i,3}}{\phi_{i,2}} \tag{3.10}$$

ここで算出した r_i が閾値 Th_r を超えるとき，提案手法ではショット s_i を司会区間と判定する．ただし，F フレーム以内で隣接する区間については連続する区間として検出する．最後に，提案手法では，判定された司会区間に存在するショット境界をあらかじめシーン境界の判定から除外する．なお，司会区間と判定された区間の内，極端に短い時間長の区間は無視する．本論文では，このときの閾値を経験的に設定する（10 秒以内）．

3.3　実験

本節では，提案手法の有効性を確認するため，3.3.1 で映像の構造に注目したシーン分割の精度評価を行う．さらに，3.3.2 では，顔認識のシーン分割精度に与える影響について実験を行い考察する．実験には，実際にテレビで放映された歌謡番組映像を用いた．各フレームのサイズは 352 × 240 画素，24bit 階調カラー，フレームレートは 30fps である．これらの映像に対して提案手法を適用する前に，文献 [15] で提案されている分割 χ^2 検定法を適用し，ショット分割を行う．

3.3.1　映像の構造に注目したシーン分割の精度評価

本節では，提案手法の有効性を評価するため，実験を行い，その結果を考察する．本実験では，司会区間の判定に用いる閾値を 1800(= F) とし，事前に学習

表 3.1: Dataset used in the experiments.

		出演者数	映像長 (min.)
番組 I	**映像 A**	25	62
	映像 B	24	58
	映像 C	22	58
	映像 D	19	62
	映像 E	18	53
番組 II	**映像 F**	8	41
	映像 G	9	42
	映像 H	8	41

データとして，対象映像に含まれる出演者の顔画像をテスト映像各出演者毎に 600 枚ずつ収集した．なお，各出演者の顔画像は次の手順により収集した．

1) 評価用映像とは異なる複数の歌謡番組映像から対象の出演者が登場する映像区間を手動で選択する．このとき，各出演者の顔画像の枚数が 600 枚となるまで，繰り返し歌謡番組映像からの収集を行う．

2) 選択した映像区間の各フレームより AdaBoost [52] を用いて顔領域を検出する．ただし，誤検出や欠損が存在する画像は手動で除外する．

このとき，本実験で用いる顔検出法は，Intel 社がオープンソースで公開しているコンピュータビジョン関連のライブラリ (Open Source Computer Vision Library,OpenCV) [55] で実装されているアルゴリズムに従う．以上のように収集した顔画像集合に対し，出演者名およびグループ名をラベルとして付与したものを，実験では学習データ

として用いる．なお，学習データの自動収集については今後の課題である．表 3.1 には，提案手法で用いるデータセットの詳細について示す．本実験では，これら 8 本の映像に対して提案手法を適用することでシーン分割を行う[1]．このとき，評価データの映像 (**A**〜**H**) 毎に登場する出演者数分の KMSM の辞書側部分空間を算出する．例えば，映像 A ならば登場する 25 人分の出演者の学習データを準備する．また，これらの映像における正解シーンは手動で与えた．

映像 A に対してシーン分割を行った結果を図 3.3 に示す．図 3.3 では，(a) に手動で与えた正解のシーン境界を示す．(b) には，事前処理により得られた司会区間を示し (c) には提案手法により得られたシーン分割を示す．また，提案手法を適用する際に得られた他の有向グラフを用いた場合のシーン分割結果として，(d) に色特徴と音響特徴によるシーン分割結果，(e) に色特徴と出演者情報によるシーン分割，(f) に色特徴のみによるシーン分割結果，(g) に音響特徴のみによるシーン分割結果をそれぞれ示す．そして，比較のため，(h) に文献 [1] によるシーン分割結果を示す[2]．なお，$P_i^{\mathcal{V}}$ は，3.2.3.1 と同様の手順により色特徴のみから得られたスコア，そして，$P'_i = w_i^{\mathcal{V}} P_i^{\mathcal{V}} + w_i^{\mathcal{A}} P_i^{\mathcal{A}}$ としてそれぞれ定義される．図 3.3 (c) – (g) では，図 3.3(a) に示す対象映像の正解シーン境界数が 31 であることから同様に，各特徴毎に得られた P_i の上位 31 件のシーンの検出結果を示している．図 3.3 より，提案手法により検出された正解シーン境界数が 24 に対して，比較手法 (d) – (h) により検出された正解シーン境界の数はそれぞれ，21，18，13，15 および 8 であることが確認できる．これにより，提案手法は比較手法において発生している過分割および未分割を抑制していることも確認できる．

[1] このとき，$Th_{C^{\mathcal{VP}}} = 0.02, 0.04, \cdots, 0.98, Th_{C^{\mathcal{A}}} = 0.01, 0.02, \cdots, 0.49, Th_T = 100, 200, \cdots, 4900$ とする．すなわち，$N^{\mathcal{VP}} = N^{\mathcal{A}} = 2401$ と設定した．

[2] 閾値 Th_T および $Th_{C^{\mathcal{V}}}$ は適切な値を実験的に与え，それぞれ $Th_T = 0.58$ および $Th_{C^{\mathcal{V}}} = 3000$ と設定した．

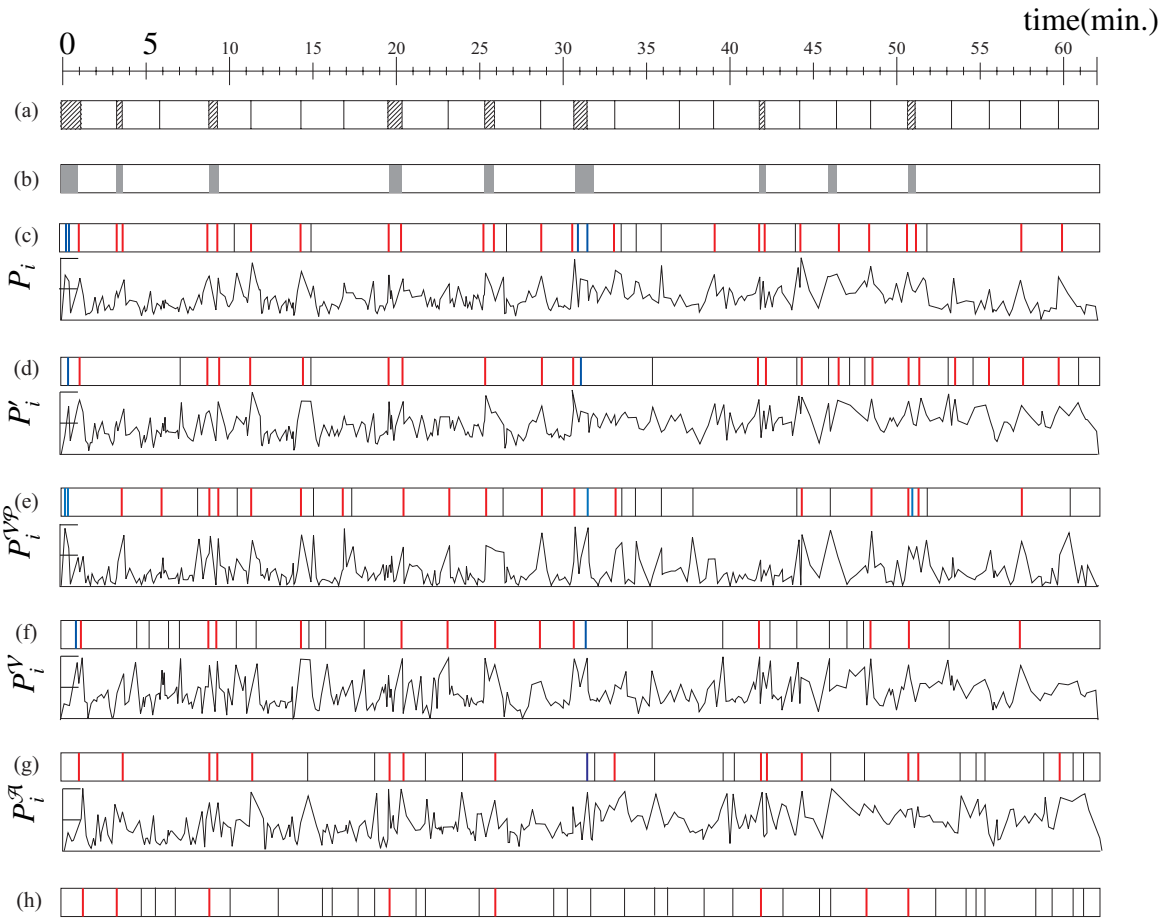


図 3.3: Scene segmentation results obtained by applying the proposed method and the other methods to Video A. Note that shadow areas in (a) represent singer introduction scenes, gray areas in (b) represent singer introduction scenes which are obtained in the proposed method, red lines in (c)-(h) represent positions where scene boundaries are correctly obtained in each method, and blue lines represent positions where scene boundaries are excluded from the detection.

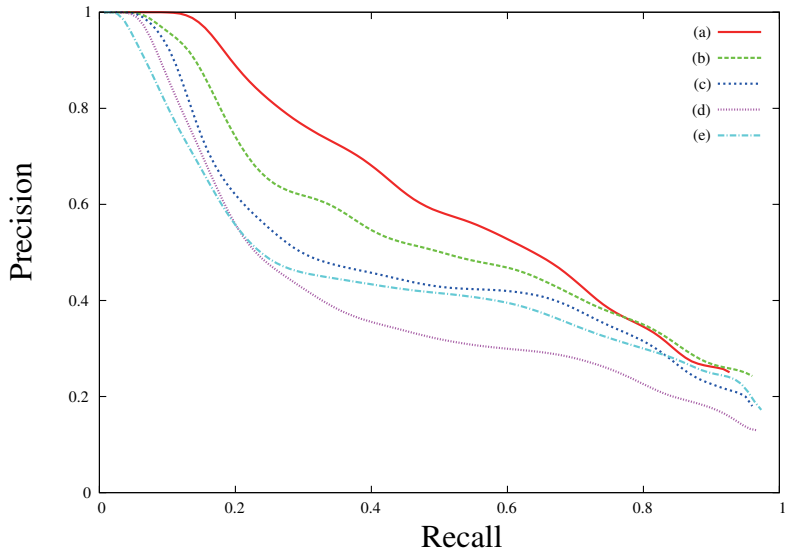


図 3.4: Quantitative evaluation of the proposed method and the comparative methods based on recall and precision.

次に，シーン分割精度を定量的に評価するため，映像 **A** – **H** 全てに対して提案手法と比較手法を適用し，以下に示す Recall および Precision を算出する．

$$\text{Recall} = \frac{\text{検出された正解シーン境界数}}{\text{正解シーン境界の数}} \tag{3.11}$$

$$\text{Precision} = \frac{\text{検出された正解シーン境界数}}{\text{検出されたシーン境界の数}} \tag{3.12}$$

このとき，閾値 Th_s を変更させることにより，図 3.4 に示す Recall-Precision カーブを得る．図 3.4 では，(a) に提案手法によるシーン分割の結果，(b) に色特徴と音響特徴によるシーン分割結果，(c) に色特徴と出演者情報によるシーン分割結果，(d) に色特徴のみによるシーン分割結果，そして，(e) に音響特徴のみによるシーン分割結果を示す．図 3.4 より，Recall および Precision ともに提案手法が比較手法の結果を上回っており，提案手法が高精度なシーン分割を実現していることが確認できる．

ところで，図 3.3 の結果には，シーン境界の過検出と未検出が存在する．以下では，それらについて各特徴毎に注目し分析する．まず，過検出は，色特徴のみを用いた場合に多数発生する．歌謡番組では，歌唱中の照明による色の変化が大きい．このような場合にシーン境界を過剰に検出している．また，提案手法においても，一つのショットもしくはシーンに複数の出演者が同時に映し出される場面において，出演者の推定精度が低下し，出演者の変化を過剰に検出する．次に，未検出は，音響特徴のみを用いた場合に発生する．音響特徴を用いた場合，司会区間の境界の漏れは少ないものの，司会区間を挟まずに次の歌手による歌唱区間が連続する場面において，音響特徴による特徴の差の検出が困難となるため，シーン境界においてもスコア $P_i^{\mathcal{A}}$ の値が低くなる．これらの問題は，色特徴および音響特徴の二つの特徴を同時に用いることで，双方の過検出および未検出の抑制を実現している．このことから，歌謡番組への協調処理の導入の有効性が確認できる．しかしながら，歌謡番組では，色特徴および音響特徴の変化が必ずしもシーン境界における特徴の差にならない．これに対し，出演者情報を用いた提案手法では，これらの特徴を用いた場合よりも，安定して歌謡番組のシーン境界を検出可能であることが図 3.3 からも確認できる．なお，提案手法における，以上の過検出の問題は今後の検討課題である．

最後に，機械学習ベースのシーン分割法との定性的比較を行い，提案手法の有効性を評価する．一般的な機械学習ベースの手法では，学習データを用意して，シーン境界の識別を行うことから，同じような編集方法の番組では，学習データが十分に与えられれば，高精度なシーン境界の検出が可能であると考えられる．文献 [2] では，あらゆるジャンルのシーン境界を検出可能な特徴量を用いることで汎用な識別器の構築を実現している．しかしながら，歌謡番組には様々な歌や演出が存在し，それに応じてシーン境界における様々な特徴のパターンが存在する．以上の文献 [2] の手法の有効性に対して，提案手法は，他ジャンルの映像に対して汎用的にシーン分割を行うことは困難であるが，歌謡番組が有する固有の性質に注目した特徴およびそれに基づく映像の構造解析を行うことで，番組毎にシーン境界の特徴が変化する歌謡番組に対して，高精度なシーン検出が可能にな

ると考えられる．実際に実験により検証を行った結果を以下に示す．本実験では，上記で示す多様な特徴の変化がある歌謡番組を対象とした場合，機械学習ベースの手法では高精度な識別が困難であることを確認する．そのため，機械学習ベースのシーン分割法として文献 [2] に従った SVM に基づいた手法との比較を行う．文献 [2] では，正例をシーン境界直後のショット，それ以外のショットを負例として学習を行った SVM 識別器を構築する．SVM は，LIBSVM [56] により実現し，また，カーネルとして RBF カーネルを用いている．実験データには，映像 **A–H** を用いる．ただし，簡単な比較のため，各ショットから提案手法で用いる HSV 色ヒストグラム（V）および文献 [54] に基づく音響特徴（A）を特徴量として用いて識別器を構築する．提案手法および SVM に基づいた手法ともに，同じ特徴を用いている．そして，評価方法として，leave-one-out 法により，Recall, Precision，そして F 値を算出する．F 値の算出式は以下の通りである．

$$\text{F 値} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3.13}$$

実験結果を表 3.2 に示す．表 3.2 により，提案手法が比較手法の結果を上回っている．以上から，一般的な機械学習ベースの手法では歌謡番組に対しシーン分割を行うことが困難であることが確認できる．

3.3.2　顔認識のシーン分割精度へ与える影響に関する考察

本節では，提案手法における顔認識の精度がシーン分割精度へ与える影響について考察する．本実験では，KMSM で顔画像の類似度算出を行う際に，シーン分割対象である映像から取得した顔領域の画像を学習データとして用いる．さらにその画像数を増減させることで，次式で定義される識別率を算出し，複数の顔認識精度に対応する類似度算出を行う．

$$\text{識別率} = \frac{\text{正しく識別された出演者数}}{\text{全ショットで検出された出演者数}} \times 100 \tag{3.14}$$

表 3.2: Performance comparison of the proposed method and the machine learning-based traditional method.

		Recall	Precision	F-measure
SVM	V	0.181	0.118	0.143
	A	0.322	0.245	0.278
	V+A	0.381	0.325	0.351
提案手法	V	0.68	0.29	0.406
	A	0.62	0.397	0.484
	V+A	0.627	0.468	0.536

具体的に，本実験では，映像 A を評価対象とし，学習データを，映像 A の各フレームから収集された顔画像（①）と，3.3.1 の (1)，(2) で収集された顔画像（②）を用意し，その割合を変化させることで，複数の異なる識別精度となるような結果を得た．例えば，②の画像の割合を①の画像に比べて多くすれば，顔認識精度の低下が実現される．評価は，3.3.1 と同様に，式 (3.11)，(3.12) から Recall および Precision を算出する．このとき，閾値 Th_s を変更させることにより，図 3.5 に示す Recall-Precision カーブを得る．図 3.5 では，顔認識の識別率が，(a) 100% である場合のシーン分割結果，(b) 73% である場合のシーン分割結果，(c) 61% である場合のシーン分割結果，(d) 52% である場合のシーン分割結果，(e) 41% である場合のシーン分割結果，(f) 33% である場合のシーン分割結果，また，比較のため，(g) 色特徴と音響特徴を用いた場合のシーン分割結果を示す．ただし，図 3.5(a) は，ショット s_i に存在する出演者を手動で判定し，式（3.2) で算出される類似度 $\psi_g(s_i)$ における g 番目の出演者との類似度を最大値 1 もしくは最小値 0 で与えることにより算出した．図 3.5(b)-(g) については，先に述べた通り，学習

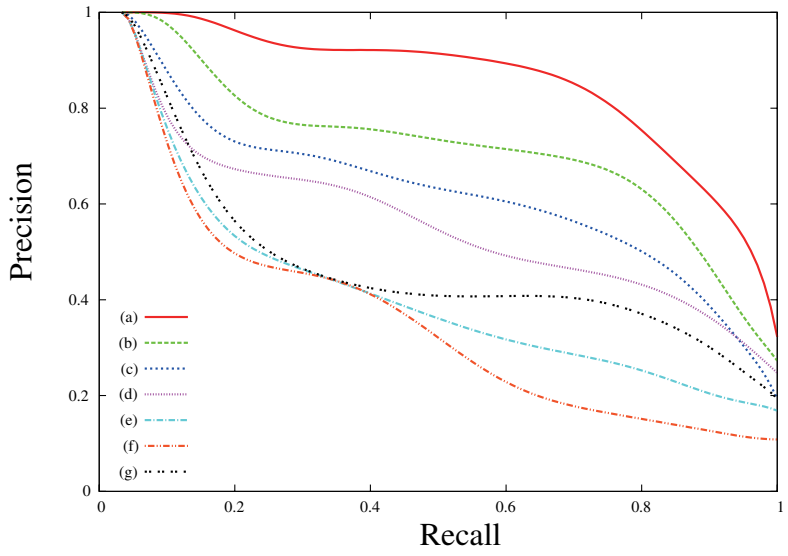


図 3.5: Verifying the Effect of Face Recognition on Scene Segmentation Accuracy based on recall and precision.

データ内の映像 A の各フレームから収集された顔画像と 3.3.1 の (1)，(2) で収集された顔画像との割合を変更することで，様々な識別精度となる結果を求めている．図 3.5 より，顔認識の精度に従い，シーン分割精度もそれに応じて最終的なシーン分割の精度が変化していることが確認できる．これより，提案手法の精度は，顔認識の精度の向上に従っていることが確認できる．なお，4.1 の実験条件で行った場合，映像 **A** における顔認識の識別率は 52% となる．また，識別率が 52% のときに図 3.5(g) の色特徴と音響特徴を用いた場合の結果を上回る．このことから，顔認識を用いた画像特徴は，一定の顔認識の誤りが生じた場合においても，シーン分割の高精度化を実現可能であることが分かる．これは，提案手法で新たに定義した歌謡番組の映像構造を利用し，有向グラフに基づいたシーン境界判定を行うことが有効に作用しているためと考えられる．

これらの結果より，提案手法では，最終的なシーン分割精度は顔認識の精度に依存する一方で，歌謡番組の映像構造を利用することで，従来よりも高精度なシーン分割を実現可能となっていることが確認できる．

3.4　まとめ

本章では，複数特徴の利用による映像の構造に注目したシーン分割手法を提案した．提案手法では，まず，画像特徴として，従来の色特徴に加えて，映像中に登場する各出演者の顔画像を学習データとした KMSM により算出される類似度を求めた．そして，音響特徴として，音響信号が属するオーディオクラスを判定した結果を特徴量として用いた．これにより，従来手法では困難であった歌謡番組映像のシーン境界の検出が可能となった．また，各々の特徴量から 2 種類の有向グラフを生成し，得られた結果を統合することで，高精度なシーン分割を行うことが可能となり，実験によりそれを確認した．

第4章　撮像対象の自動抽出処理を導入した物体識別手法

4.1　はじめに

本章では，映像中に撮像されている物体の識別のため，撮像対象の自動抽出処理を導入した物体識別手法を提案する．提案手法では，学習画像から対象物体の検出し，その検出結果から新たな識別超平面を算出することで画像識別の精度向上を試みる．具体的に，提案手法では，Support Vector Machine (SVM) [7] を用いた識別超平面を算出し，特徴空間における特徴ベクトルと超平面との距離に基づいて，物体領域を内包する矩形を検出する．これにより，背景領域に存在する特徴を除去することで，特徴ベクトルの更新が可能となる．さらに，この特徴ベクトルを用いることで，新たな識別超平面を算出する．このとき，物体領域以外に存在する特徴量は，学習の過程に影響を及ぼさない．以上により，提案手法では，物体領域に限定した特徴抽出が可能となるため，背景や他の物体による影響を抑制した高精度な物体識別が実現される．

本章では，まず 4.2 で，SVM の概要について説明する．次に，4.3 では，撮像対象の自動抽出処理を導入した SVM に基づく識別器の高精度化法を説明する．そして，4.4 で，提案手法の有効性を確認するための実験結果を示す．

4.2 Support Vector Machine

本節では，提案手法で用いる Support Vector Machine (SVM) [7] について説明を行う．SVM は学習データとなる特徴ベクトル $\mathbf{x}_i (i = 1, 2, \cdots, N; N$ は学習データの総数) と，これに割り当てられたラベル y_i から，未知のデータ $\mathbf{x}$ のラベルを識別するためのアルゴリズムである．このとき，図 4.1 に示すように，マージンが最大となるよう識別超平面を算出することで，高い汎化性を持つことが一般に知られている．SVM では，次式で表される識別超平面における $\mathbf{w}$ と b をマージンが最大となるように設定する．

$$f^{svm}(\mathbf{x}) = \mathbf{w}^T \cdot \mathbf{x} + b \tag{4.1}$$

このとき，マージンは次式で与えられる．

$$Margin = \frac{y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b)}{\|\mathbf{w}\|} \tag{4.2}$$

したがって，マージンを最大とするような $\mathbf{w}$，b を求める問題は，以下の制約付き最適化問題を解くことに帰着される．

$$\begin{aligned} \arg\min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ s.t. \quad & y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, \cdots, N \end{aligned} \tag{4.3}$$

さらに，ラグランジュ乗数 $\mathbf{a} = (a_1, \cdots, a_N)^T$ を定義しラグランジュの未定乗数法を用いると，式 (4.3) は以下の目的関数を $\mathbf{a}$ について最大化する問題に帰着される．

$$\tilde{L}(\mathbf{a}) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j \tag{4.4}$$

ただし，**a** は以下の制約条件を満たす．

$$\sum_{i=1}^{N} a_i y_i = 0 \tag{4.5}$$

$$0 \leq a_i, \ \ i = 1, \cdots, N \tag{4.6}$$

以上より，式 (4.4) で得られた **a** を用いて，識別超平面 $f^{svm}(\mathbf{x})$ は次式で表される．

$$f^{svm}(\mathbf{x}) = \sum_{i=1}^{N} a_i y_i \mathbf{x}_i \cdot \mathbf{x} + b \tag{4.7}$$

さらに，特徴ベクトル $\mathbf{x}_i$ を非線形の写像 $\Phi(\mathbf{x}_i)$ により変換し，その空間上で SVM を適用することで次式により非線形分離が可能となる．

$$f^{svm}(\mathbf{x}) = \sum_{i=1}^{N} a_i y_i K^{svm}(\mathbf{x}_i, \mathbf{x}) + b \tag{4.8}$$

ただし，カーネル関数 $K^{svm}(\mathbf{x}_i, \mathbf{x}) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})$ で表される．

ここで，式 (4.1) で定義される SVM の出力値は，$\Phi(\mathbf{x})$ と識別超平面の距離に基づいて算出される．したがって，入力の特徴ベクトルが正例である可能性を示す度合を表現可能となる [57]．以上の性質により，提案手法では，物体領域の選定を実現するため，SVM 出力値 $f^{\mathrm{svm}}(\mathbf{x})$ に注目する．

4.3　撮像対象の自動抽出処理を導入した SVM に基づく識別器の高精度化

本節では，撮像対象の自動抽出処理を導入うした SVM に基づく識別器の高精度化法を説明する．今，与えられた学習画像から，対象カテゴリーの画像を正例，

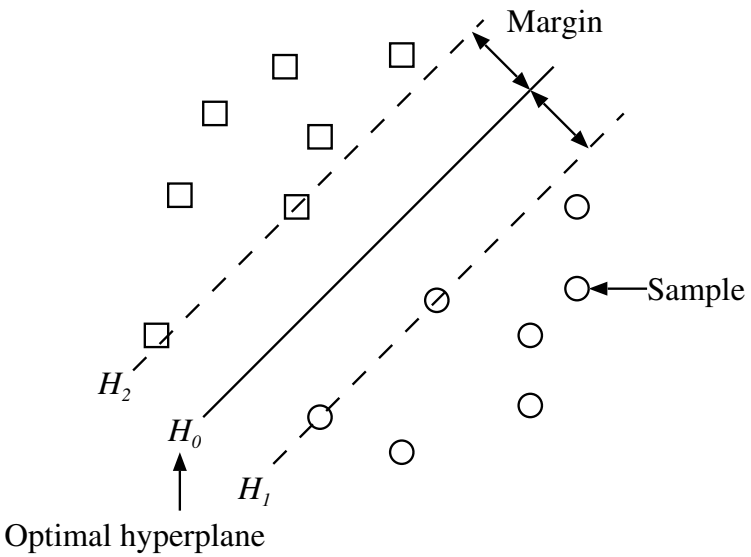


図 4.1: Overview of Support Vector Machine.

それ以外の画像を負例として，前節で示したように予め SVM の学習により識別超平面を算出する．提案手法では，高精度な識別器の構築を実現するため，識別器の学習過程に物体領域を検出する機能を導入する．この機能は以下の 2 つの処理により実現される．

処理 1) SVM による超平面間の距離に基づく撮像対象を含む領域の検出

処理 2) 新たな正例の選択およびそれを用いた識別超平面の更新

提案手法では, まず**処理 1** で，学習画像から SVM により算出される識別超平面を用いて，正例から物体領域を含むブロック (以降, 正例ブロック) を検出する．さらに**処理 2** で，検出された正例ブロックを新たな正例として用いて新たな識別超平面を算出する．以上の処理により，物体領域のみを含む正例が新たに生成されるため，背景領域に存在する特徴量による影響の抑制が可能となる．したがって，提案手法を用いることで，物体領域に限定した特徴抽出が可能となり，識別器の高精度化が実現される．

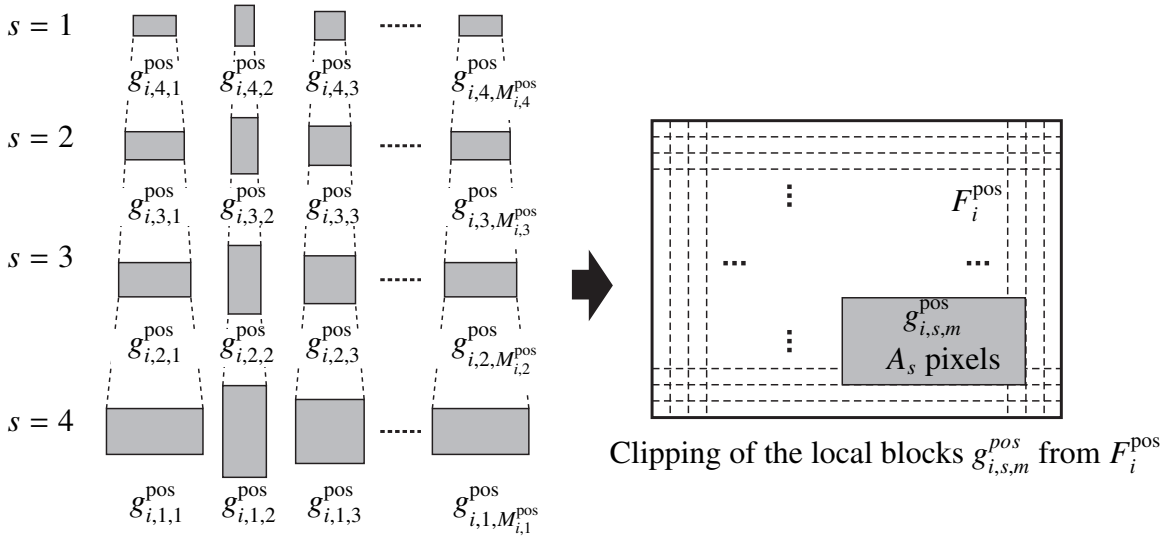


図 4.2: An example of clipping local-blocks $g^{pos}_{i,s,m}$ from the positive example F^{pos}_i.

本節では，まず，4.3.1 では，SVM に基づく正例ブロックの検出法について説明する（**処理 1**）．次に，4.3.2 では，正例ブロックの検出結果を用いた SVM による識別超平面の更新について説明する（**処理 2**）．最後に，4.3.3 で，提案手法における多クラス識別手法について説明する．

4.3.1　正例ブロックの検出

本節では，正例ブロックの検出法について説明する．今，予め対象物を含む画像として，N 枚の正例 F^{pos}_i $(i = 1, 2, \cdots, N)$ が与えられているものとする．提案手法では，まず，正例ブロックを検出するため，正例 F^{pos}_i に対し，局所ブロック $g^{pos}_{i,s,m}(s = 1, 2, \cdots, S, m = 1, 2, \cdots, M^{pos}_{i,s})$ を互いに重なりを許すように切り出す．以下では，局所ブロック $g^{pos}_{i,s,m}$ の詳細につい説明する．

まず，s は同一のサイズを持つ局所ブロックのグループを表すインデックスを表す．提案手法では，様々な縦横比およびサイズを持つ局所初ブロックを正例 F^{pos}_i から切り出す．このとき，図 4.2 で示すように，インデックス s を用いることで，

表 4.1: Four kinds of groups $A_s(s = 1, 2, 3, 4)$, whose areas are $(0.8)^s A_0$ (pixels), used in our method, where $A_0 = 65536$.

A_1	A_2	A_3	A_4
$0.8A_0$	$(0.8)^2A_0$	$(0.8)^3A_0$	$(0.8)^4A_0$

表 4.2: Aspect ratios (vertical:horizontal) used for clipping local blocks of each group A_s.

4 : 1	3 : 1	2 : 1	3 : 2	1 : 1	2 : 3	1 : 2	1 : 3	1 : 4

同一のサイズを持つ局所ブロックのグループを A_s と表す．提案手法では，A_s の総数を 4 とし，表 4.1 で示す 4 種類の異なる面積の集合 $A_s(s = 1, 2, 3, 4)$ を用いる．各集合 A_s に対し，正例 F_i^{pos} から様々な縦横比の局所ブロックを切り出す．具体的に，我々は，表 4.2 で示すように 9 種類の縦横比の局所窓を用い，これらの局所窓から局所ブロック $g_{i,s,m}^{\mathrm{pos}}$ を切り出す．但し，$M_{i,s}^{\mathrm{pos}}$ は，正例 F_i^{pos} における，面積 $A_s(s = 1, 2, \cdots, S)$ の局所ブロックの総数を表す．以上のように，提案手法では，$M_{i,s}^{\mathrm{pos}}$ 個の面積 A_s で様々な縦横比を持つ局所ブロック，つまり，計 $\sum_{s=1}^{S} M_{i,s}^{\mathrm{pos}}$ 個の局所ブロックを正例 F_i^{pos} から抽出する．

さらに，提案手法では，各局所ブロック $g_{i,s,m}^{\mathrm{pos}}$ から，特徴ベクトル $\mathbf{x}_{i,s,m}^{\mathrm{pos}}$ を算出する．このとき，正例 F_i^{pos} から，S 種類の正例ブロック $b_{i,s}^{\mathrm{pos}}(s = 1, 2, \cdots, S)$ が算出される．具体的に，各集合 $A_s(s = 1, 2, \cdots, S)$ に対し，それに対応する正例ブ

ロックを次のように取得する．

$$b_{i,s}^{\mathrm{pos}} = g_{i,s,m_{i,s}^{\mathrm{opt}}}^{\mathrm{pos}} \tag{4.9}$$

但し，

$$m_{i,s}^{\mathrm{opt}} = \arg \max_{m=1,2,\cdots,M_{i,s}^{\mathrm{pos}}} f^{\mathrm{svm}}(\mathbf{x}_{i,s,m}^{\mathrm{pos}}) \tag{4.10}$$

上式において，$f^{\mathrm{svm}}(\mathbf{x}_{i,s,m}^{\mathrm{pos}})$ は式 (4.1) で示す SVM の出力値を表す．但し，SVM における識別超平面は，正例 F_i^{pos} $(i = 1, 2, \cdots, N)$，および，負例から算出されているものとする．また，提案手法では，式 (4.9) を用いて，各集合 $A_s (s = 1, 2, \cdots, S)$ に対する，正例ブロック $b_{i,s}^{\mathrm{pos}}$ を選定する．以上により，図に示すように，各正例 F_i^{pos} から S 種類の正例ブロックが得られる．

前節で示したように，SVM の出力値 $f^{\mathrm{svm}}(\mathbf{x}_{i,s,m}^{\mathrm{pos}})$ は，高次元特徴空間において特徴ベクトル $\Phi(\mathbf{x}_{i,s,m}^{\mathrm{pos}})$ と超平面間の距離に基づいて算出される．さらに，出力値 $f^{\mathrm{svm}}(\mathbf{x}_{i,s,m}^{\mathrm{pos}})$ は，入力の特徴ベクトル $\mathbf{x}_{i,s,m}^{\mathrm{pos}}$ が，正例における対象物から得られた可能性を示す割合を表現している．したがって，SVM の出力値が最大となる局所ブロックを選定することで，対象の物体領域を含む正例ブロックの検出が実現される．

4.3.2　正例ブロックを用いた識別超平面の更新

本節では，前節で得られた新たな正例ブロック $b_{i,s}^{\mathrm{pos}} (i = 1, 2, \cdots, N, s = 1, 2, \cdots, S)$ を用いた SVM 識別超平面の更新手法について説明する．図 4.3 に示すように，正例ブロック $b_{i,s}^{\mathrm{pos}} (s = 1, 2, \cdots, S)$ は，各正例 F_i^{pos} から検出される．これらの正例ブロックは，物体領域を含む一方で，異なる面積を持つ．そのため，提案手法では，SVM 識別超平面の更新のため，各正例ブロック $b_{i,1}^{\mathrm{pos}}, b_{i,2}^{\mathrm{pos}}, \cdots, b_{i,S}^{\mathrm{pos}}$ から，最適な正例ブロックの選択を行う．この処理の導入により，物体領域を含む最小のブロッ

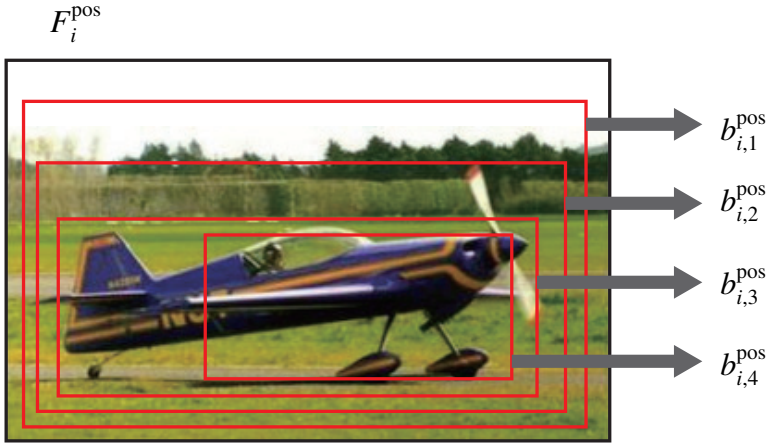


図 4.3: An example of positive-blocks $b_{i,s}^{\mathrm{pos}}$ detected from the positive example F_i^{pos}.

クの選択が可能となる．このとき，我々は各正例ブロックにおける SVM の出力値 f^{svm} の集合 A_s に対する変化に着目する．A_s を変化させたときの $b_{s,i}^{\mathrm{pos}}$ は，ある面積において物体領域を包含する最小の局所ブロックとなり，この場合の f^{svm} は他の正例ブロックよりも高くなる傾向があると考えられる．一方，ブロックの面積が物体領域よりも小さくなるとき，対象物から得られる特徴の一部が失われてしまうため，f^{svm} は対象物を含む場合であっても小さい値となる可能性が大きくなる．そこで，提案手法では図 4.4 に示すように，次式を満たす最小の s を s_i^{opt} として，最適な正例ブロック $b_{i,s_i^{\mathrm{opt}}}^{\mathrm{pos}}$ を検出する．

$$f^{\mathrm{svm}}(\mathbf{x}_{i,s}^{\mathrm{pos}}) - f^{\mathrm{svm}}(\mathbf{x}_{i,s+1}^{\mathrm{pos}}) > 0 \tag{4.11}$$

但し，$\mathbf{x}_{i,s}^{\mathrm{pos}}$ は，$b_{i,s}^{\mathrm{pos}}$ から算出される特徴ベクトルである．

以上により，提案手法では，各正例 F_i^{pos} から最適な正例ブロック $b_{i,s_i^{\mathrm{opt}}}^{\mathrm{pos}}$ の算出を可能とする．我々は，このようにして得られた物体領域を正確に内包する正例ブロック $b_{i,s_i^{\mathrm{opt}}}^{\mathrm{pos}}(i = 1, 2, \cdots, N)$ を用いることで，SVM 識別超平面を更新する．このとき，特徴ベクトルは $b_{i,s_i^{\mathrm{opt}}}^{\mathrm{pos}}$ から得られるため，更新した SVM 識別超平面を用いることで，識別精度の向上が期待できる．

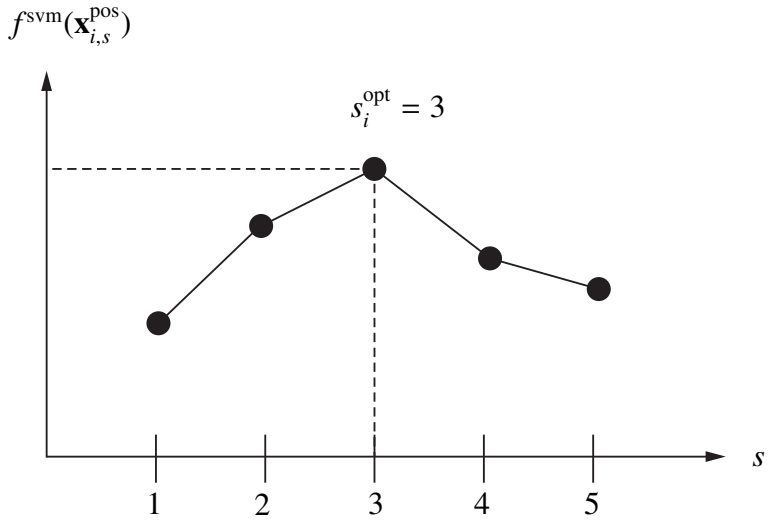


図 4.4: An example of “the relationship between s and $f^{\mathrm{svm}}(\mathbf{x}_{i,s}^{\mathrm{pos}})$” and “the detection of s_i^{opt}”.

4.3.3　多クラス識別

本節では，提案手法に基づいた多クラス識別手法について説明する．従来，SVM に基づいた多クラス識別手法が提案されており [40]，我々もそれに従う．

前節で示した処理に処理により，提案手法では，正例ブロック $b^{\mathrm{pos}}_{i,s_i^{\mathrm{opt}}}$ を用いた識別器の構築を可能とする．一般に，2 クラス問題を含む多クラス識別手法では，各クラスの学習画像を事前に用意する．まず，各カテゴリ k ($k = 1, 2, \cdots, K$; K は物体のクラスの総数) において，クラスに含まれる学習画像 F_i^k ($i = 1, 2, \cdots, N^k$; N^k はクラス k の学習画像の総数) を定義する．

まず，各クラス k に対し，我々は，SVM 識別超平面 H^k を正例 $F_i^k (i = 1, 2, \cdots, N^k)$ および負例 $F_i^{k'} (k' = \{1, 2, \cdots, K | k' \neq k\}, i = 1, 2, \cdots, N^{k'})$ を用いて算出する．次に，得られた識別超平面 H^k を用いて，最適な正例ブロック $b^k_{i,s_i^{\mathrm{opt}}}$ を算出し，各クラス k に対し，識別超平面 $\hat{H}^k$ を更新する．但し，更新した識別超平面は，正例として $b^k_{i,s_i^{\mathrm{opt}}}$，そして，負例として $F_i^{k'}$ および $b^{k'}_{i,s_i^{\mathrm{opt}}}$ $(k' = \{1, 2, \cdots, K | k' \neq k\}, i = 1, 2, \cdots, N^{k'})$ を用いることで得られる．このとき，提案手法では，全てのクラス

$k(k = 1, 2, \cdots, K)$ に対する識別器が得られる．

最終的に，提案手法では，新たなテスト画像が与えられた際に，得られた識別超平面 $\hat{H}^k(k = 1, 2, \cdots, K)$ を用いて識別を行う．具体的に，未知のクラスのテスト画像から得られた特徴ベクトル $\mathbf{x}$ が与えられた際，提案手法では，次式を満たす $\hat{k}$ を識別結果とする．

$$\hat{k} = \arg \max_{k=1,2,\cdots,K} f_k^{\mathrm{svm}}(\mathbf{x}) \tag{4.12}$$

但し，$f_k^{\mathrm{svm}}(\cdot)$ は，クラス k における識別関数を表す．したがって，式 (4.12) を用いることで，画像識別を実現する．

図 4.5: Example images of Caltech 101 and Caltech 256 datasets.

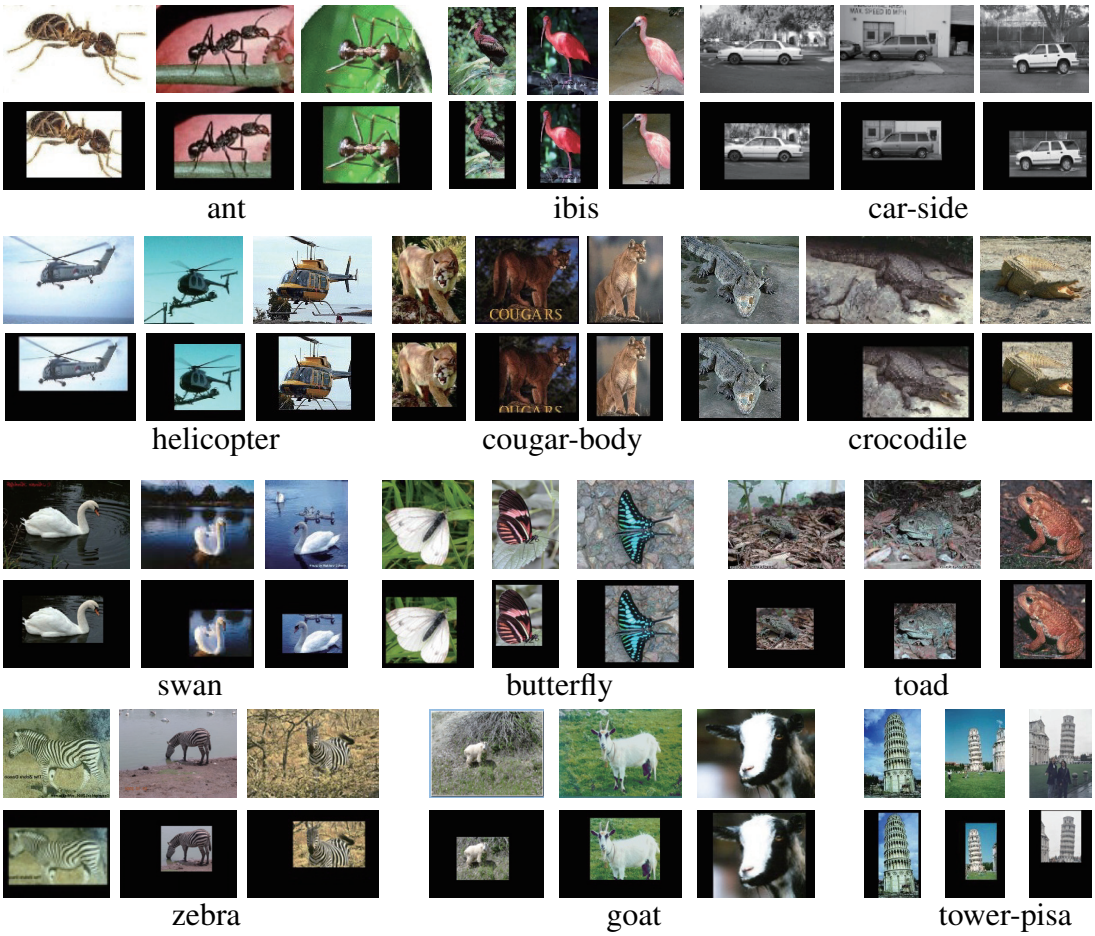


図 4.6: An examples of target object detection from Caltech 101 and Caltech 256. Categories “ant”, “ibis”, “car-side”, “helicopter”, “cougar-body”, “crocodile” are included in Caltech 101. All categories shown in this figure are included in Caltech 256. This means Caltech 256 includes Caltech 101.

4.4　実験

本節では，提案手法の有効性を確認するため，実画像を用いた評価実験の結果を示す．実験では，Caltech 101 [25] および Caltech 256 データセット [26] に含まれる画像を用いる．

4.4.1　実験データセット

本節では，評価実験で用いる Caltech 101 および Caltech 256 データセットの詳細について説明する．各データセットの詳細は次の通りである．

(i) Caltech 101

Caltech 101 は，Fei-Fei ら [25] によって収集された画像データセットである．本データセットは，101 種類の物体カテゴリーと背景画像からなる 102 種類のカテゴリーを持つ画像で構成され，主に Google Image Search を用いて人手により収集した 9,144 枚の画像が含まれている．カテゴリー毎に 31 枚から 800 枚の画像を含んでいる．多くの画像は 300×300 枚程度の中解像度となっている．Caltech 101 データベースは，色，姿勢そして照明が画像ごとに変化するといった特徴を持つ．

(ii) Caltech 256

Caltech 256 は，Griffin ら [26] によって収集された画像データセットであり，Caltech 101 の拡張として公開された．本データベースは物体カテゴリー数が 256 種類であり，総画像数は 30,607 枚となっている．Caltech 256 では，最小カテゴリーでも 80 枚の画像が含まれている．また，Caltech 101 データセットとは異なり，向きや大きさをある程度統一する操作を行っていないため，より分類が難しい画像データセットとされている．

図 4.5 に，各データセットの例を示す．Caltech 101 と比較すると，Caltech 256 データセットは対象物体の向きや大きさが多様であることが確認できる．

表 4.3: Classification accuracy (%) comparison on Caltech 101 and Caltech 256.

Methods	Datasets	Conventional method	Proposed method
SVM [7]	Caltech 101	44.8 ± 0.7	**46.2 ± 0.9**
	Caltech 256	11.6 ± 0.5	**13.2 ± 0.9**
SVDD [58]	Caltech 101	38.8 ± 0.6	**39.9 ± 0.8**
	Caltech 256	10.1 ± 0.7	**11.1 ± 0.5**
SVM-KNN [59]	Caltech 101	64.6 ± 0.3	**65.3 ± 0.5**
	Caltech 256	32.3 ± 0.5	**33.1 ± 0.8**
Lazebnik SPM [40]	Caltech 101	64.2 ± 0.3	**65.6 ± 0.7**
	Caltech 256	31.1 ± 0.8	**32.5 ± 0.6**
Bosch SVM [60]	Caltech 101	71.2 ± 0.3	**72.8 ± 0.5**
	Caltech 256	36.3 ± 0.5	**37.6 ± 0.7**

4.4.2　評価実験

本節では，まず，提案手法における物体領域検出の有効性を評価する．図 4.6 に，正例ブロックの検出結果を示す．図 4.6 より，提案手法により，対象の物体領域が検出されていることが確認できる．

次に，提案手法の画像分類精度を Caltech 101 および Caltech 256 データセットそれぞれを用いて評価する．本実験では，文献 [61,62] に従い，各カテゴリー毎にランダムに学習画像を 30 枚，そして，残りの画像をテスト画像として選択する．このとき，テスト画像の枚数を Caltech 101 において 50 枚，そして，Caltech 256 において 25 枚を上限とした．表 4.3 に，二つのデータセットそれぞれの分類結果を示す．ただし，表 4.3 は，以上に示した学習画像およびテスト画像の選択

を 3 回行い，それらの平均と標準偏差を示している．

表 4.3 に示す通り，本実験では，SVM に基づいた複数の従来手法に対して提案手法を適用した．ただし，本実験では，表 4.3 における SVM [7] および Support Vector Data Description(SVDD) [58] に対しては，SIFT を適用する．なお，SIFT の詳細は文献 [60] と同様である．次に， SVM-KNN [59], Lazebnik SPM [40]， そして， Bosch SVM [60] の実験条件についても，各文献に従う．また，提案手法における正例ブロックの検出には，文献 [60] で提案されている SIFT および HOG に基づく，Pyramid Histogram Of visual Words (PHOW) および Pyramid Histograms Of Gradients (PHOG) の二つの特徴量を用いた．ただし，パラメータ設定については，文献 [60] と同様である．本実験では，HOG の勾配方向数は 40 とし，1×1, 2×2，そして， 4×4 の 3 レベルの SPM カーネル [40] とした．以上の特徴量を用いて検出した正例ブロックを用いて，上記の識別器を構築した．

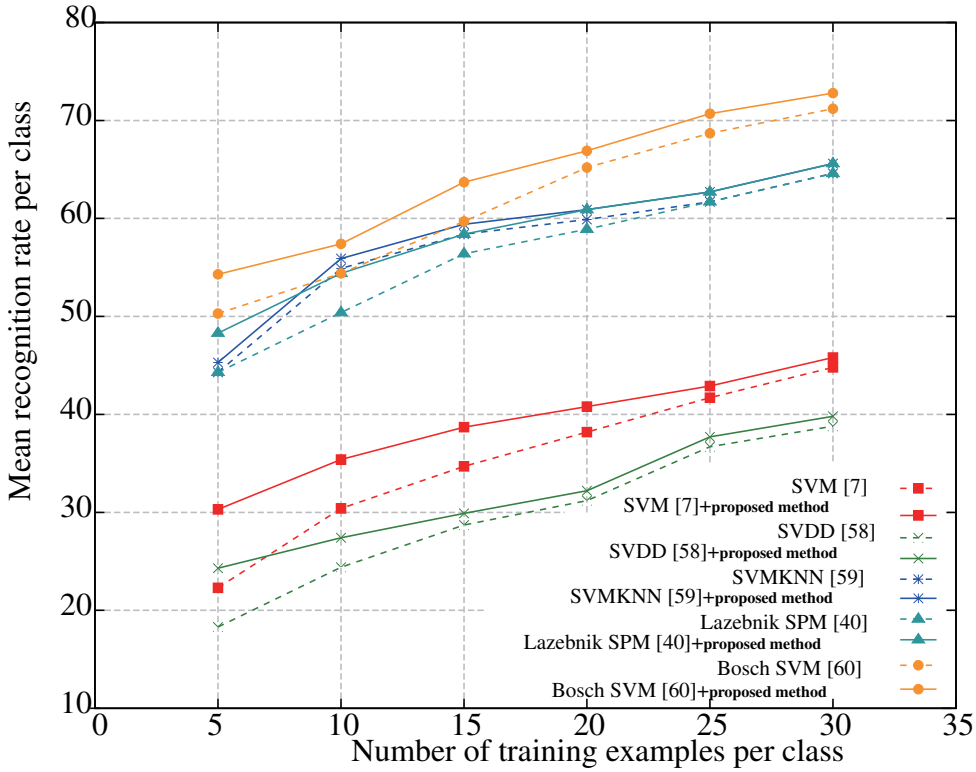
表 4.3 の結果より，提案手法の適用により各識別器のみを用いる場合と比較して，精度向上を実現していることが確認できる．さらに，本実験では，提案手法の学習画像数の変化の影響を検証するため，二つのデータセットそれぞれに対し，ランダムに $5, 10, 15, 20, 25, 30$ 枚の画像を学習画像として選択し，残りの画像をテスト画像とした精度評価を行う．本検証は，従来より広く行われており，本実験においてもそれに従う [26, 40, 59]．図 4.7 に，結果を示す．図 4.7 では，提案手法および各従来手法の結果を比較として示す．図 4.7 からも，提案手法は，比較手法を上回っており，提案手法の導入により精度向上が実現できることが確認できる．

以上の二つのデータセットを用いた実験結果から，提案手法の有効性が確認され，SVM に基づく識別器の高精度化を実現した．

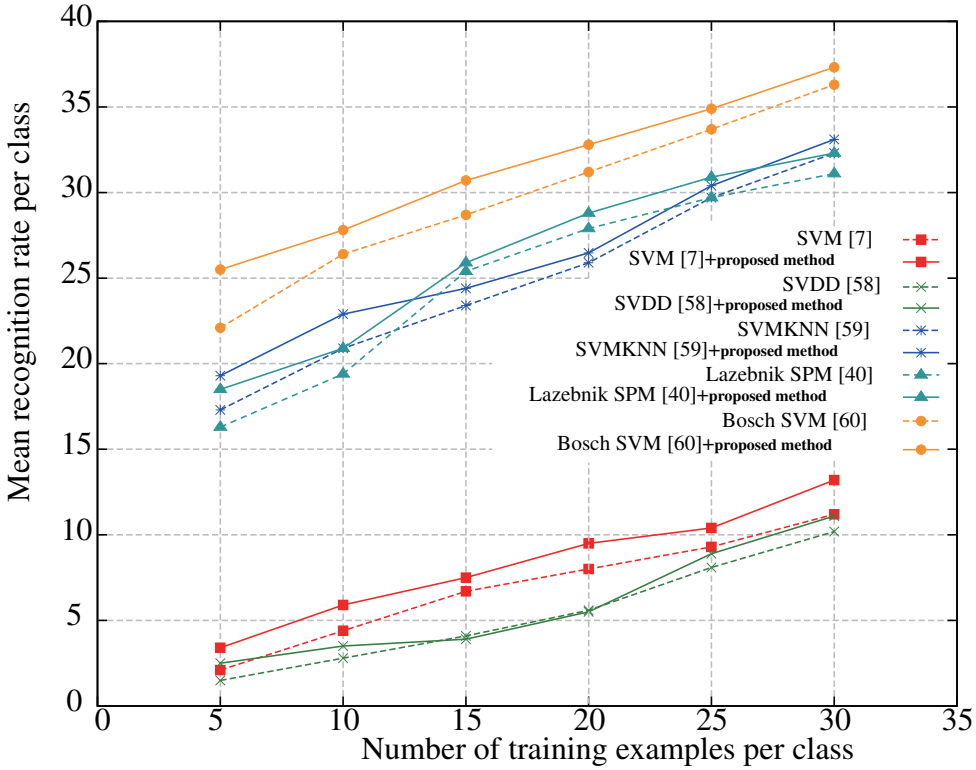
4.5　まとめ

本章では，撮像対象の自動抽出処理を導入した物体識別手法を提案した．提案手法では，学習画像から対象物体を検出し，その検出結果から新たな識別超平

面を算出することで物体識別の精度向上を実現した．具体的に，提案手法では，SVM を用いた識別超平面を算出し，特徴空間における特徴ベクトルと超平面との距離に基づいて，物体領域を内包する矩形を検出した．これにより，背景領域に存在する特徴を除去することで，特徴ベクトルの更新を可能とした．さらに，この特徴ベクトルを用いることで，新たな識別超平面を算出した．以上により，提案手法では，撮像対象の物体領域に限定した特徴抽出を可能とし，実験により，背景や他の物体による影響を抑制した高精度な物体識別が実現されることを確認した．



(a) Caltech 101.



(b) Caltech 256.

図 4.7: Relationship between the number of training examples per class and the mean recognition rate per class.

第5章　映像特徴とタグ情報の類似性に注目したリランキング手法

5.1　はじめに

本章では，映像特徴とタグ情報の類似性に注目したリランキング手法を提案する．提案手法ではまず，映像のリンク関係を基に，映像とタグそれぞれをノードとするサブグラフを構築し，二つのサブグラフを統合したグラフ（以降，異種グラフと呼ぶ）を構築する．このとき，異種グラフのエッジには，映像特徴とタグ情報の類似性に基づき重み付けを施す．さらに，異種グラフのエッジを介して，タググラフの重みを映像グラフに伝播させる．次に，提案手法では，キーワード検索の順位を基に算出するスコアを，グラフ近傍に位置する類似映像間で平滑化するため，ノード毎に局所モデルを構築する．この局所モデルは，各ノードの近傍ノード集合の映像特徴からスコアを推定する回帰モデルで表される．したがって，提案手法では，推定されるスコアと局所モデルとの誤差を表す正則化項を加えたエネルギー関数の最小化問題を解くことで，近傍の類似映像間のスコアを平滑化可能とし，グラフの整合性を保つリランキングを可能とする．以上により，映像特徴とタグ情報の類似性の両者を考慮したリランキングが可能となり，従来手法と比較して高精度な映像検索が実現される．

本章の構成は以下の通りである．5.2 では，映像特徴とタグ情報を用いた異種グラフの構築方法を説明する．5.3 では，構築した異種グラフに基づく検索結果のリランキング法を説明する．5.4 では，提案手法の有効性を確認するため，評価実験の結果を示す．

5.2　映像特徴とタグ情報を用いた異種グラフの構築

提案手法では，映像特徴とタグ情報の両者を考慮したリランキングを実現するため，映像およびタグそれぞれをノードとするサブグラフを構築する．さらに，それらの関係性に基づいてサブグラフを統合した異種グラフを構築する．今，説明のため映像とタグで構成される二つのグラフをそれぞれ G^k $(k = \{M, T\})$ と表す．ただし，$V^k = \{v_1^k, v_2^k, \ldots, v_{N_k}^k\}$ はノード集合，N^k はノード数，そして，E^k はエッジ集合を表す．このとき，エッジは重み関数 w^k により重み付けを施す．ただし，重み関数 $w^k(i, j)$ はノード v_i^k，v_j^k 間の重みを表す．また，$w^k(i, j)$ を (i, j) 成分の要素として持つ重み行列を $\mathbf{W}^k$ と表す．なお，映像およびタグのクローリング方法の詳細については，5.4.1 で後述する．

以降では，映像グラフ，タググラフ，そして，2 種のサブグラフを統合した異種グラフの構築方法をそれぞれ説明する．

5.2.1　映像グラフの構築

提案手法では，映像共有サイトにおける関連映像検索から得られるリンクを基に，映像グラフ G^M を構築する．なお，本文では，映像共有サイト YouTube を利用し，説明を進める．本節ではまず，YouTube を利用したキーワード検索から得られた検索上位 N_M 件の映像を収集，そして，そのリンク関係を解析することで，ノード集合 $V^M = \{v_1^M, v_2^M, \ldots, v_{N_M}^M\}$ を取得する．次に映像グラフのノード間の関係性を表現するため，各映像から低次特徴を抽出する．最後に，ノード v_i^M および v_j^M 間の類似度を算出することで重み関数 w^M を定義する．

5.2.1.1　低次特徴抽出

我々はまず，映像 v_i^M から以下に示す 3 種類の低次特徴を抽出する．ただし，提案手法では，文献 [63] を利用し，映像 v_i^M を m_i 個のショット $\mathbf{s}_i = \{s_i^1, s_i^2, \ldots, s_i^{m_i}\}$ に分割する．文献 [63] では，フレーム間の非類似性を表すスコアを算出し，閾

値に基づきショット境界を検出する．そこで，提案手法では，ショット境界以外のスコアの上位 3 件に対応するフレームを各ショットのキーフレームとする．

HSV 色ヒストグラム HSV 色ヒストグラムをショットのキーフレームから抽出する．このとき，配色の偏りに配慮するため，キーフレームを 5×5 に分割し，各面について $4^3 = 64$ 次元を，キーフレーム 1 枚について計 1600 次元の特徴ベクトルを得る．

SIFT SIFT 特徴量をキーフレームに対し，グリッド幅 6 の dense sampling に基づき抽出する．さらに，Bag-of-words モデル [39] を適用することで，各キーフレームに対し 1000 次元の特徴ベクトルを得る．

MFCC MFCC を各ショットの区間全体から抽出する．提案手法では，先行研究に倣い MFCC12 次元の他，ΔMFCC，ΔΔMFCC，対数パワー，Δ 対数パワー，そして，ΔΔ 対数パワーを抽出した計 39 次元の特徴ベクトルを得る．ただし，Δ は時間微分を表す．

5.2.1.2　重み関数の算出

抽出された低次特徴を用いてノード v_i^M，v_j^M 間の重み関数 $w^M(i,j)$ を次式で定義する．

$$w^M(i,j) = \sum_{l \in \mathcal{L}} \alpha_l \exp\left\{ -\frac{\min_{u \in \mathbf{s}_i, v \in \mathbf{s}_j} \mathrm{ED}(\mathbf{f}_u^l, \mathbf{f}_v^l)}{2\sigma_l^2} \right\} \tag{5.1}$$

ただし，$\mathcal{L} = \{SIFT, HSV, MFCC\}$, $\mathrm{ED}(\mathbf{f}_u^l, \mathbf{f}_v^l)$ は特徴ベクトル $\mathbf{f}_u^l$,$\mathbf{f}_v^l$ 間のユークリッド距離, そして，α_l $(\sum_l \alpha_l = 1)$ は各特徴の重要度を決定する重みである．また，σ_l はパラメータを表し，全てのユークリッド距離のメディアンを用いて算出する．

5.2.2　タググラフの構築

提案手法では，映像に付与されているタグを N_T 個収集する．さらに，同一映像に付与されているタグ間にエッジを張ることでタググラフ G^T を構築する．本

節では，収集したタグを要素とするノード集合を $V^T = \{v_1^T, v_2^T, \ldots, v_{N_T}^T\}$ と表し，その重み関数を定義する．ただし，ユーザの自由記述によって付与される性質上，タグには多くの誤りが含まれる．そのため，提案手法では，タグのフィルタリングを行い，その結果から映像内容との適合性の高いタグを抽出する処理を導入する．以降，タグのフィルタリングとタグ重み関数についてそれぞれ説明する．

5.2.2.1　タグフィルタリング

まず，自然言語処理において一般的に行われるタグのフィルタリングを施す．具体的に提案手法では，3 文字未満のタグ，時間を表すタグ，英語以外の言語で付与されたタグ，そして，ストップワードリスト [1] に該当するタグを除去する．次に，The Stanford Natural Language Processing (NLP) Group が開発している英語テキスト用の自然言語開発ツール [64] を用いた品詞タグ付けを基に形容詞タグ，そして，映像のメタ情報から得られた投稿ユーザ名に該当するタグを検出し除去する．最後に，次式に示すように，タグの出現頻度の低いタグ w を閾値 Th_w を用いて

$$|R_w| < Th_w \tag{5.2}$$

除去する．ただし，R_w はタグ w が付与された映像，$|R_w|$ はその総数をそれぞれ表す．また，式 (5.7) の算出の際には，大文字・小文字の区別はせず，自然言語処理ライブラリ NLTK [65] を用いてタグ w に語幹処理を施す．以降では，上記のタグフィルタリングが施されているものとして説明を進める．

[タグリストの拡張]

提案手法では各映像のタグの再割り当てを行う．そのための前処理としてまず，記述性及び適合性の高いタグの検出を容易にするためタグリストを拡張する．具体的にまず，映像 v のメタ情報からタイトルと説明文を単語分割しタグリストに

[1] [Online]. Available: http://bit.ly/8vBrVF

加え，初期タグリスト I_v を得る．次に，収集した映像に存在する全タグを T と表したとき，タグ w_1 ($w_1 \in I_v$) とタグ w_2 ($w_2 \in T$) の共起頻度 $o(w_1, w_2)$ を次式により算出する．

$$o(w_1, w_2) = \frac{|R_{w_1} \cap R_{w_2}|}{|R_{w_1}|} \tag{5.3}$$

ただし，式 (5.8) では個々のタグの出現頻度による影響を考慮するための正規化を施している．閾値 Th_o を用いて，$o(w_1, w_2) > Th_o$ を満たす w_2 を抽出する．続いて，初期タグリスト I_v と条件を満たすタグ候補を統合することで拡張されたタグリスト C_v を得る．

次に，得られたタグリスト C_v のランク付けを行うことで，適合性の高いタグを抽出する．提案手法では，文献 [66] を基にタグのランク付けを施す．文献 [66] では，類似映像に付与されるタグの内容も同じく類似するという仮定を設け，映像 v の k 近傍の映像とそれに付与されたタグへの投票に基づき映像 v におけるタグ w ($w \in C_v$) の適合度

$$tagRelevance(w, v, k) = n_w[N(v, k)] - Prior(w, k) \tag{5.4}$$

を算出する．ただし，$n_w[\cdot]$ はタグ w を含む映像のカウンタ，$N(v, k)$ は映像 v の k 近傍に存在する映像をそれぞれ表す．なお，提案手法では映像 v の k 近傍を，映像グラフ G^M を用いた最短経路距離に基づき定義することで，関連映像の収集を容易とする．そして，$Prior(w, k)$ は

$$Prior(w, k) = k\frac{|R_w|}{|\Phi|} \tag{5.5}$$

で算出される．ただし，Φ はタグが付与された全映像，$|\Phi|$ はその総数をそれぞれ表す．*tagRelevance* の算出のための処理をアルゴリズム 1 に示す．

ただし，式 (5.9) は負の値を取り得る．この場合提案手法では，アルゴリズム 1 に示すように，最小値を 1 に設定する．

Algorithm 1 Learning tag relevance by neighbor voting

Input: A user-tagged video v.

Output: $tagRelevance(w, v, k)$,i.e., the tag relevance value of each tag w in video v.

1: Find k nearest visual neighbors of v from the collection.
2: **for** tag w in tags of v **do**
3: $\quad tagRelevance(w, v, k) = 0$
4: **end for**
5: **for** video u in the neighbor set of v ($\in N(v, k)$) **do**
6: $\quad$ **for** tag w in ($R_v \cap R_u$) **do**
7: $\quad\quad tagRelevance(w, v, k) = tagRelevance(w, v, k) + 1$
8: $\quad$ **end for**
9: **end for**
10: $tagRelevance(w, v, k) = tagRelevance(w, v, k) - Prior(w, k)$
11: $tagRelevance(w, v, k) = \max(tagRlevance(w, v, k), 1)$

以上により得られた $tagRelavence$ を基に，提案手法ではタグ w の映像 v へのスコア $score(w, v)$ を次式で算出する．

$$\begin{aligned} score(w, v, k) =& tagRelevance(w, v, k) \cdot \frac{\epsilon_r}{\epsilon_r + (rank_w - 1)} \\ & \cdot \frac{\epsilon_d}{\epsilon_d + abs(\epsilon_d - \log(|R_w|))} \end{aligned} \tag{5.6}$$

ただし，ϵ_r および ϵ_d はパラメータであり，$rank_w$ は式 (5.9) で得られた値の順位を与える．なお，式 (5.11) の第 2 項目はランキング順位を保持するための項，そして，第 3 項目は出現頻度の高い一般的なタグの値を減衰させるための項として

それぞれ定義される [67]．最後に，式 (5.11) で得られた $score(w, v, k)$ の値が上位のタグを映像 v に与えるタグと決定する．以上により，提案手法では従来手法と比較して，記述性の低い汎用なタグを除去し，かつ適合性の高いタグを新たに映像に付与することが可能となる．

5.2.2.2　重み関数の算出

提案手法では，タグ v_i^T，v_j^T 間の関係性を評価するため Normalized Google Distance [68] に基づいて，タグ間距離 $\mathrm{NGD}(v_i^T, v_j^T)$ を次式により算出する．

$$\mathrm{NGD}(v_i^T, v_j^T) = \frac{\max\{\log g(v_i^T), \log g(v_j^T)\} - \log g(v_i^T, v_j^T)}{\log N_g - \min\{\log g(v_i^T), \log g(v_j^T)\}} \tag{5.7}$$

ただし，N_g は Google が保有する Web ページの総数である．また，$g(v_i^T)$，$g(v_j^T)$，そして，$g(v_i^T, v_j^T)$ はタグ v_i^T, v_j^T，そして，$v_i^T + v_j^T$ を Google に与えたときのヒット件数をそれぞれ表す．このとき，タグ v_i^T，v_j^T 間の重み関数 $w^T(i, j)$ は次式に示すように $\mathrm{NGD}(v_i^T, v_j^T)$ を用いて算出可能となる．

$$w^T(i, j) = \exp\left\{-\frac{\mathrm{NGD}(v_i^T, v_j^T)}{2\sigma_{TAG}^2}\right\} \tag{5.8}$$

ただし，σ_{TAG} はパラメータを表し，σ_l と同様にして得られる．

5.2.3　異種グラフの構築と重みの伝播

最後に，前節で得られた二つのサブグラフ G^k を用いて異種グラフを構築する．さらに，提案手法では，タググラフの重みを映像グラフに伝播させるため，文献 [69] で提案されている重み伝播手法を異種グラフに適用する．今，G^{MT} を二つのサブグラフ G^k を統合した異種グラフと表す．ただし，$V^{MT} = \{V^M \cup V^T\}$ は

ノード集合，E^{MT} は二つのサブグラフを相互に接続するエッジの集合を表し，提案手法では，映像 v_i^M に付与されているタグに対してエッジを張る．このとき，異種グラフ G^{MT} 上の映像 v_i^M とタグ v_j^T 間の重み $\pi^{MT}(i,j)$ を次式で算出する．

$$\pi^{MT}(i,j) = \begin{cases} \exp\left\{-\frac{score(v_j^T, v_i^M)}{2\sigma_{MT}^2}\right\} & \text{if } v_j^T \in S^T(v_i^M) \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

ただし，$S^T(v_i^M)$ はノード v_i^M と接続する G^T 内のノード集合を表す．また，$\pi^{MT}(i,j)$ を (i,j) 成分の要素として持つ重み行列を $\mathbf{\Pi}^{MT}$ と表したとき，提案手法では，次式に示すように重み伝播を施す．

$$\begin{cases} \mathbf{W}_{(t)}^T = \theta_T \mathbf{\Pi}^{MT\prime} \mathbf{W}_{(t-1)}^M \mathbf{\Pi}^{MT} + (1-\theta_T)\mathbf{W}_{(0)}^T \\ \mathbf{W}_{(t)}^M = \theta_M \mathbf{\Pi}^{MT} \mathbf{W}_{(t-1)}^T \mathbf{\Pi}^{MT\prime} + (1-\theta_M)\mathbf{W}_{(0)}^M \end{cases} \tag{5.10}$$

ただし，$\mathbf{\Pi}^{MT\prime}$ は重み行列 $\mathbf{\Pi}^{MT}$ の転置，下付き文字 (t) は t 回目の反復，(0) は初期状態を表す．また，θ_M, θ_T $(0 \leq \theta_M, \theta_T \leq 1)$ は，各サブグラフの重みの重要度を決定するパラメータを表す．提案手法では，次式に示す収束条件を満たすまで式 (5.10) の反復を行う．

$$(\mathbf{W}_{(C+1)}^k - \mathbf{W}_{(C)}^k) \simeq 0, \quad k = \{M, T\} \tag{5.11}$$

ただし，C は収束条件を満たす反復回数を表す．以上のように，提案手法では，タググラフの重みを映像グラフに伝播させることを可能とする．さらに，$\mathbf{W}_{(C)}^M$ を映像グラフの重み行列として更新して適用することで，映像特徴とタグ情報の類似性の両者を考慮した Web 映像検索を可能とする．次節では，映像グラフ G^M の重み行列 として $\mathbf{W}_{(C)}^M$ を適用した Web 映像検索法を詳述する．

5.3　グラフ構造に基づく検索結果のリランキング法

本節では，グラフ構造に基づく検索結果のリランキング法を提案する．今，説明のため映像グラフ G^M に含まれるノード集合 V^M から抽出される映像特徴を $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_M}\}$ と与える．ただし，$\mathbf{x}_i \in \mathbb{R}^d$ は映像 v_i の d 次元特徴ベクトルとして定義される．このとき，最適なランキングスコア $\mathbf{r}^*$ は，次式に示す事後確率

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} f(\mathbf{r}|\bar{\mathbf{r}}, \mathcal{X}) \tag{5.12}$$

を最大化するランキングスコア $\mathbf{r} = [r_1, r_2, \ldots, r_{N_M}]$ を推定することで得られる．ただし，$\bar{\mathbf{r}} = [\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_{N_M}]$ は初期ランキングスコアを表し，テキスト検索結果の順位を基に算出される．なお，r_i および $\bar{r}_i$（$0 \leq r_i, \bar{r}_i \leq 1$）は映像 v_i のクエリとの適合度を表すスコアとして定義される．本文では，式 (5.12) に従って，初期ランキングスコア $\bar{\mathbf{r}}$ の順列を変える処理をリランキングと定義している．次に，式 (5.12) は，ベイズの定理を基に，次式に示すように制約付き事前確率および尤度の積で表される．

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} p(\mathbf{r}|\mathcal{X}) \times p(\bar{\mathbf{r}}|\mathbf{r}) \tag{5.13}$$

式 (5.13) に示す通り，リランキングを行うためには，二つの項（条件付き事前確率 $p(\mathbf{r}|\mathcal{X})$ および尤度 $p(\bar{\mathbf{r}}|\mathbf{r})$）を定義する必要があり，5.3.1 および 5.3.2 では，それぞれについて詳述する．さらに，5.3.3 では，得られた項を基に式 (5.13) の解 $\mathbf{r}^*$ を推定する．

5.3.1　グラフ構造を考慮したリランキングスコアの推定

提案手法では，グラフの近傍に配置されるノード間の特徴は類似すると仮定し，サンプル $\mathbf{x}_i$ 毎に局所モデル $o_i(\cdot)$ を構築する．$o_i(\cdot)$ は，ノード毎に最適な平

滑化を実現するスコアを近傍ノード集合 $\mathcal{N}(\mathbf{x}_i) = \{\mathbf{x}_h^{(i)}, r_h^{(i)}\}_{h=1}^{n_i}$ を用いて学習することで得られる．また，$\mathbf{x}_h^{(i)}$ は，ノード $\mathbf{x}_i$ における h 近傍ノード，そして，n_i は近傍ノード数を表す．このとき，文献 [70] を利用して各ノードが属するコミュニティ検出することで，近傍ノード集合 $\mathcal{N}(\mathbf{x}_i)$ を同一コミュニティのみに限定することが可能となる．なお，提案手法では，サンプル $\mathbf{x}_i$ を起点とした最短経路距離に基づき近傍ノード集合 $\mathcal{N}(\mathbf{x}_i)$ を取得する．

以上により定義される局所モデル $o_i(\cdot)$ を用いて，式 (5.1) の $\psi_i(\mathbf{r}, \mathbf{X})$ は，ランキングスコアとの誤差を推定するため

$$\psi_i(\mathbf{r}, \mathcal{X}) = \{r_i - o_i(\mathbf{x}_i)\}^2 \tag{5.14}$$

で表される．式 (5.14) を用いることで正則化項 $Reg(\mathbf{r}, \mathbf{X})$ は次式のように書き換えられる．

$$Reg(\mathbf{r}, \mathcal{X}) = \sum_i \{r_i - o_i(\mathbf{x}_i)\}^2 \tag{5.15}$$

局所モデル $o_i(\cdot)$ は，近傍サンプル集合 $\mathcal{N}(\mathbf{x}_i)$ とそのスコア $\mathbf{r}^{(i)} = [r_h^{(i)}]'$ の依存性をモデル化するように表現される．

$$o_i(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x}) \tag{5.16}$$

本文では，次式に示すコスト関数を解くことで，式 (5.16) のモデルを得ることが可能である．

$$\min_{\mathbf{w}} \lambda \parallel \mathbf{w} \parallel^2 + \sum_{h=1}^{n_i} \{\mathbf{w}'\phi(\mathbf{x}_h^{(i)}) - r_h^{(i)}\}^2 \tag{5.17}$$

ただし，λ は係数，そして，$\phi(\mathbf{x}_h^{(i)}) \in \mathbb{R}^{d'}$ $(d' \gg d)$ は特徴ベクトルを高次元の特徴空間へ写像することで得られる．このとき，式 (5.17) の解は

$$\mathbf{w} = \Phi_i(\Phi'\Phi_i + \lambda\mathbf{I})^{-1}\mathbf{r}^{(i)} \tag{5.18}$$

である．ただし，$\Phi_i = [\mathbf{x}_h^{(i)}]'$ で定義される行列，$\mathbf{e}$ は全ての要素が 1 のベクトルと定義される．さらに，式 (5.18) を式 (5.16) に代入することで局所モデル $o_i(\cdot)$ を得ることを可能とする．

$$\begin{aligned} o_i(\mathbf{x}_i) &= \mathbf{k}'(\lambda\mathbf{I} + \mathbf{K})^{-1}\mathbf{r}^{(i)} \\ &= \beta_i'\mathbf{r}^{(i)} \end{aligned} \tag{5.19}$$

ただし，$\beta_i' = \mathbf{k}'(\lambda\mathbf{I} + \mathbf{K})^{-1}$, $\mathbf{k}$ は $k_h = \phi(\mathbf{x}_i)'\phi(\mathbf{x}_h^{(i)}) = k(\mathbf{x}_i, \mathbf{x}_h^{(i)})$ で定義されるベクトル，$\mathbf{K}$ は $k_{mn} = \phi(\mathbf{x}_m^{(i)})'\phi(\mathbf{x}_n^{(i)}) = k(\mathbf{x}_m^{(i)}, \mathbf{x}_n^{(i)})$ を要素とする行列，そして，$k(\cdot,\cdot)$ はカーネル関数としてそれぞれ定義される．したがって，式 (5.19) を式 (5.15) に代入することで次式が成り立つ．

$$\begin{aligned} Reg(\mathbf{r}, \mathcal{X}) &= \sum_i (r_i - \beta_i'\mathbf{r}^{(i)})^2 \\ &= \mathbf{r}'\mathbf{R}\mathbf{r} \end{aligned} \tag{5.20}$$

ただし，$\mathbf{R} = (\mathbf{I} - \mathbf{B})'(\mathbf{I} - \mathbf{B})$，$\mathbf{B} = [b_{ij}]_{N_M \times N_M}$ は b_{ij} を要素とする行列，b_{ij} は $\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)$ である場合，β_i の対応する要素, それ以外の場合は $b_{ij} = 0$ と定義される

5.3.2　尤度算出

式 (5.13) の第二項は，初期ランキングスコアと推定されるランキングスコアとの間の尤度である．これは，初期ランキングスコアの情報を保持するための項と

して以下に示すように定義される．

$$p(\bar{\mathbf{r}}|\mathbf{r}) = \frac{1}{Z_2}\exp\{-\epsilon \times Dist(\mathbf{r}, \bar{\mathbf{r}})\} \tag{5.21}$$

ただし，$Z_2 = \sum_{\mathbf{r}} \exp\{-\epsilon \times Dist(\mathbf{r}, \bar{\mathbf{r}})\}$ は正規化項，ϵ はスケーリングパラメータ，そして，$Dist(\mathbf{r}, \bar{\mathbf{r}})$ は，二つのスコアの間の非類似度を表し，次式で定義される．

$$Dist(\mathbf{r}, \bar{\mathbf{r}}) = \frac{1}{2}\sum_{i,j \in S_{\bar{\mathbf{r}}}} \left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j}\right)^2 \tag{5.22}$$

ただし，$S_{\bar{\mathbf{r}}}$ は全てのサンプルペア集合 $(\mathbf{x}_i, \mathbf{x}_j)$ に対応するスコアが $\bar{r}_i > \bar{r}_j$ を満たすペア (i, j) の集合として定義される．

5.3.3　リランキングスコアの最適化

式 (5.22) および式 (5.20) を用いると，式 (5.13) は，次式で定義されるエネルギー関数 $E(\mathbf{r})$ の最小化問題に帰着する．

$$E(\mathbf{r}) = \mathbf{r}^T\mathbf{R}\mathbf{r} + \rho\sum_{i,j \in S_{\bar{\mathbf{r}}}} \left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j}\right)^2 \tag{5.23}$$

ただし，ρ は二つの項を制御するトレードオフパラメータである．このとき，最適なランキングスコア $\mathbf{r}^*$ は

$$\mathbf{r}^* = \arg\min_{\mathbf{r}} \mathbf{r}^T(\mathbf{R} + \rho\mathbf{L}_\mathbf{A})\mathbf{r} - \tilde{\rho}^T\mathbf{r} \tag{5.24}$$

の解を求めることで得られる．ただし，$\tilde{\rho} = 2\rho(\mathbf{A}\mathbf{e})$，そして，$\alpha_{ij} = 1/(\bar{r}_i - \bar{r}_j)$ と与えたとき，非対称行列 $\mathbf{A} = [\alpha_{ij}]_{N_M \times N_M}$ と定義される．また，$\mathbf{L}_\mathbf{A}$ は，ノード

表 5.1: 10 Event Queries used in the experiments.

Query
FIFA WorldCup 2014 highlights
Sochi 2014 Winter Olympics opening ceremony
NBA Finals 2014 highlights
New York Fashion Week 2014
Speech at Apec China 2014
2014 Hong Kong protests
2014 Israel Gaza Conflict air strikes
Malaysia Airlines Flight 17 crash moment
Flood in Indonesia 2014
Calbuco Volcano Eruption in Chile

間の重みが $|\alpha_{ij}|$ であるグラフ G_A で定義される graph Laplacian である．最後に，式 (5.24) を $\mathbf{r}$ について微分した結果が 0 となる解を求めることで次式が得られる．

$$\mathbf{r}^* = \frac{1}{2}(\mathbf{R} + \rho \mathbf{L_A})^{-1} \tilde{\rho} \tag{5.25}$$

以上により，グラフの局所的構造を考慮したリランキングが実現され，得られたランキングスコア $\mathbf{r}^*$ の上位の映像を結果として与える検索が可能となる．

5.4　実験

本節では，提案手法の有効性を確認するため，YouTube から収集した Web 映像に対し提案手法を適用した実験を行う．

5.4.1　実験条件

提案手法を適用した実験を行うためまず，次の手順により Web 映像とタグを収集する．

Step I: Web 映像およびタグの収集 Wikipedia で分類されているイベントカテゴリ “Events by Topic”[2] を参考に YouTube に表 5.1 に示すクエリを与えて検索を行った．これにより，2014 年 2 月〜2015 年 3 月の期間にアップロードされた検索結果の上位 500 件のタグを含む Web 映像を取得した．さらに，Google が提供している YouTube DATA API を利用した関連映像検索から映像のリンクを定義し，映像グラフ G^M を構築した．

Step II: タグのフィルタリング 収集されたタグをフィルタリングし，同一映像に付与されているタグ間にエッジを張ることでタググラフ G^T を構築した．

また，式 (5.7) の重み α_l ($F \in \{HSV, SIFT, MFCC\}$) はそれぞれ 0.4，0.4，0.2 とし，式 (5.10) のパラメータ θ_M，θ_T をそれぞれ 0.2，0.2，そして，式 (5.25) のトレードオフパラメータ ρ を 0.5 と設定した．さらに，本実験では，初期ランキングスコア $\bar{\mathbf{r}}$ を YouTube の検索順位に対し Rank 法（$\bar{r}_i = N - \hat{r}_i$）[71] を適用し算出した．ただし，$N$ は検索総数，そして，$\hat{r}_i$ は映像 v_i の検索順位とそれぞれ定義される．

5.4.2　評価方法

評価方法としてまず，情報検索精度を評価する際に用いられる Mean Average Precision (MAP) を用いた．ただし，AP は全ての適合映像が出現するまでの Precision の平均，MAP は全キーワードにおける平均である．続いて，Web 検索の精度を評価するために用いられる Normalized Discounted Cumulative Gain (NDCG)

[2] http://en.wikipedia.org/wiki/Category:Events_by_topic

を算出した．NDCG では，次式により検索上位 d 件の検索精度を算出する．

$$\mathrm{NDCG@}d = Z_d \sum_{j=1}^{d} \frac{2^{t^j} - 1}{\log(1+j)} \tag{5.26}$$

ただし，t^j は j 番目の検索結果の適合度，そして，Z_d は正規化項であり，正解となる上位 d 件の検索結果の精度が 1 になるよう設定する．なお，適合度 t^j を 3 段階[3] 定義し，手動で各映像に与えた．本実験において，Okapi BM25 [72] をベースラインとして比較を行う．Okapi BM25 は，クエリ $Q = \{q_i\}_i^{N_q}$ と映像 v に含まれる文章 D の関連度 $score_{BM25}(D, Q)$ を次式により算出する．

$$score_{BM25}(D, Q) = \sum_{i=1}^{N_q} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k+1)}{f(q_i, D) + k \cdot (1 - \beta + \beta \cdot \frac{|D|}{\mathrm{avgdl}})} \tag{5.27}$$

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{5.28}$$

ただし，$f(q_i, D)$ は文章 D に含まれる単語 q_i の出現頻度，|D| は文章 D の長さ，avgdl は文章の長さの平均値，そして，$n(q_i)$ はクエリ q_i の出現頻度である．また，k および β はパラメータであり，それぞれ 1.2 および 0.75 と設定した．なお，本実験では，映像 v を含む Web ページ内のテキストから文章 D を取得した．

5.4.3　精度評価

提案手法で導入する正則化項の有効性を確認するため，表 5.2 に，従来の正則化項との MAP による比較を行う．表 5.2 では，4 つの正則化項：コミュニティ検出機能を含む局所学習正則化項（Ours），局所学習正則化項（Local），正規化 Laplacian 正則化項（N-Lap），そして Laplacian 正則化項（Lap）の結果を示す．また，比較のため，初期ランキングスコアのみによる検索結果（Initial），Okapi BM25 [72] による結果（BM25），そして，VisualRank [12] による結果をそれぞ

[3] 本実験では，“0:Irrelevant”, “1:Fair”, そして，“2:Relevant”を与えた．

表 5.2: MAP Comparison of each Regularization.

Method	MAP	
	VG	MVG
Ours	0.236	**0.239**
Local [13]	0.223	**0.227**
N-Lap [73]	0.206	**0.215**
Lap	0.202	**0.211**
Initial	0.196	
BM25 [72]	0.186	
VisualRank [12]	0.201	

れ示す．ただし，表 5.2 では，異種グラフによる重み伝播を施した重み行列を用いたリランキング結果を MVG，そして，重み伝播を施さない場合の結果を VG と表す．表 5.2 より，Ours が比較手法の結果を上回っており，提案手法が高精度なリランキングを実現していることが確認できる．さらに，全ての正則化項において MVG が VG の結果を上回っていることから，提案手法で利用する異種グラフを介した重み伝播処理の有効性が確認できる．したがって，以降の実験では重み伝播を施した重み行列を用いたリランキングによる結果を示す．

次に，本文で新たに導入した正則化項を学習する際の最大近傍数 K による精度の変化を，MAP により比較する．図 5.1 では，2 つの正則項：Ours，Local の最大近傍数 K を 1 から 150 まで変化させた場合の結果を示す．図 5.1 より，Ours が Local の結果を上回っており，提案手法で定義したコミュニティ検出機能を含む局所学習正則化項は，類似映像が存在する可能性が高い領域毎に局所モデルの学習が可能となり，より効果的なリランキングを実現していることが確認できる．

最後に，図 5.2 に検索上位件数を変化させたときの NDCG による精度の比較

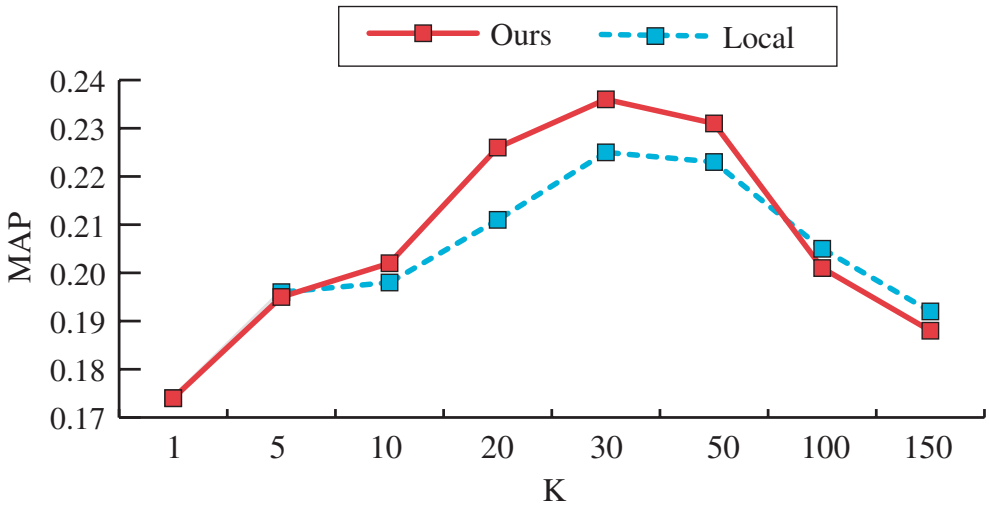


図 5.1: MAP Comparison of nearest-neighbor number for local learning regularization.

を示す．図 5.2 では，全てのキーワードにおける NDCG の平均を MNDCG と表す．図 5.2 に示す結果より，提案手法が比較手法を上回っており，NDCG を評価指標として用いた場合でも精度の向上が確認できる．一方で，検索上位件数を増やしたとき，比較手法が提案手法を上回っていることがわかる．これは，コミュニティ検出法の導入により適合映像同士が異なるコミュニティに属してしまったため，スコアの平滑化が上手く働かなかったことが原因であると考えられる．しかしながら，上位の検索精度の向上が確認できるため，提案手法は高精度なリランキングを実現しているといえる．

5.5　まとめ

本章では，映像特徴とタグ情報の類似性に注目したリランキング手法を提案した．提案手法ではまず，映像のリンク関係を基に，映像とタグそれぞれをノードとするサブグラフを統合した異種グラフを構築した．このとき，異種グラフの

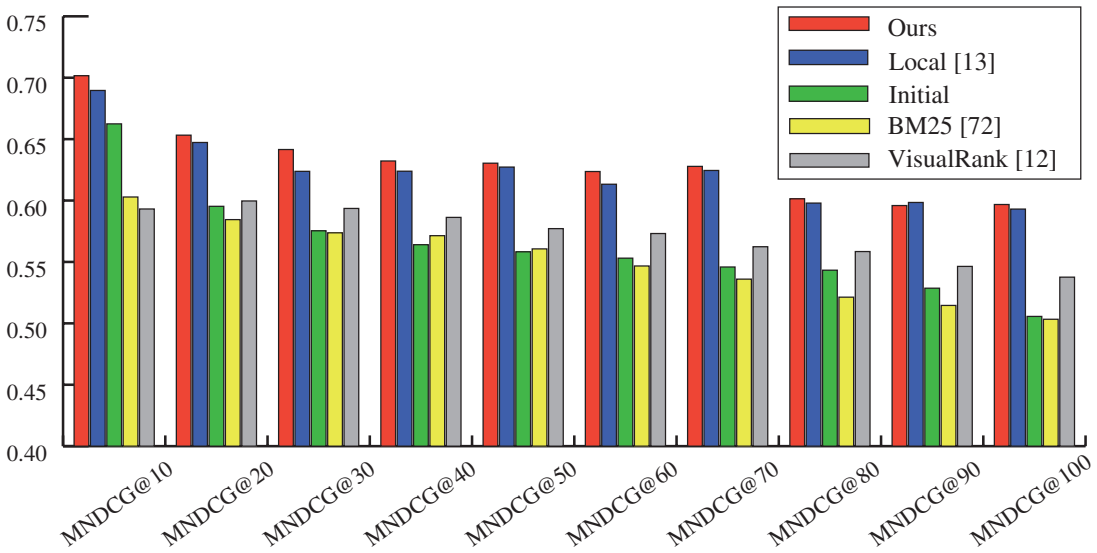


図 5.2: MNDCG Comparison of the Video Search Performance.

エッジには，映像特徴とタグ情報の類似性に基づき重み付けを施し，異種グラフのエッジを介して，タググラフの重みを映像グラフに伝播する処理を導入した．次に，提案手法では，キーワード検索の順位を基に算出するスコアを，グラフ近傍に位置する類似映像間で平滑化するため，ノード毎に局所モデルを構築した．したがって，提案手法では，推定されるスコアと局所モデルとの誤差を表す正則化項を加えたエネルギー関数の最小化問題を解くことで，近傍の類似映像間のスコアを平滑化可能とし，グラフの整合性を保つリランキングを可能とした．実験では，リランキングの高精度化を確認し，映像特徴とタグ情報の類似性を用いることの有効性を示した．

第6章　結論

6.1　本論文の成果

本論文では，映像検索の高精度化における課題である，メタデータ付与の自動化のための映像意味解析技術を映像のシーン分割と撮像物体識別に大別しそれぞれの高精度化手法を提案した．さらに，ユーザが与えるキーワードとメタデータに基づいて行われる検索に対して，望む映像を上位に配置するような検索結果を提示するためのリランキングを導入する手法を提案した．

まず，映像をシーンに分割する手法では，シーン境界毎に特徴の変化が大きいジャンルの映像では，画像，音響という特徴を利用するだけでは，映像が有する繰り返し構造の検出は困難となる問題が存在した．次に，撮像物体を識別する手法では，背景の多様性や同時に撮像されている他の物体の影響から，撮像物体の識別を行う上で雑音となる特徴が含まれ，識別精度が低下する問題が存在した．最後に，検索結果のリランキング手法では，キーワード検索で得られた映像間の類似性に基づいてその実現を可能とする手法が提案されてきたが，動画共有サービスから提供される映像には，映像の特徴以外にも映像間の関係を把握する上で有益な情報が含まれている．しかしながら，これらを活用してリランキングの精度を改善する手法が提案されていなかった．

そこで，本論文ではまず，映像の複数特徴から得られるシーン境界の判定結果を協調的に用いるシーン分割手法の提案を行った．提案手法では，各ジャンルの中でも映像構造が明確に存在する歌謡番組に焦点を当て，例えば,「司会区間」の次に「歌唱区間」が続き，それらが番組内で繰り返し現れるという歌謡番組固有の構造を利用した．この構造に基づいてショットの階層クラスタリングを行い，

その結果から取得される有向グラフのカット辺を探索する．なお，ショットの階層クラスタリングを行う際には，歌謡番組に注目した画像，音響特徴を導入した．画像特徴については，シーンが遷移するときに，同時に各シーンの中心となる出演者も変化することを考慮し，新たに出演者の変化を検出する特徴を顔認識法の結果から定義した．各々の特徴から得られる結果を統合する協調処理により歌謡番組固有の繰り返し構造を検出することで，シーン分割の高精度化を実現した．

次に，撮像対象の自動抽出処理を導入した物体識別手法の提案を行った．具体的に，SVM を用いた識別超平面を算出し，特徴空間における特徴ベクトルと超平面との距離に基づいて，撮像物体の存在する領域を内包する矩形を自動検出する処理を提案した．提案手法では，背景領域に存在する特徴を除去することで，撮像物体の領域に限定した特徴抽出を可能とし，新たな識別超平面を算出した．これにより，背景や他の物体による悪影響を抑制した物体識別を実現した．

最後に，映像特徴とタグ情報の類似性に注目したリランキング手法の提案を行った．提案手法では，まず映像のリンク関係を基に，映像とタグそれぞれの類似性を表現可能なグラフを構築し，そのグラフ上において，映像とタグ，各々の類似性を伝播させる．さらに，キーワード検索による順位が，類似性の高い映像ノード間で近くなるように目的関数を与え，その最小化問題を解くことで，グラフの整合性を保つようなリランキングが可能となり，望む映像が上位に配置される映像検索が実現した．

以下，各章で論じた内容について要約する．

- 第 2 章では，映像検索の高度化を目的とした，各種の映像解析手法について説明した．次に，映像のシーン分割，撮像物体識別，そして，検索結果のリランキングに関する従来研究についてそれぞれ説明し，現状の課題と本論文の解決策について明確にした．

- 第 3 章では，映像の複数特徴の利用による映像の構造に注目したシーン分割手法の提案を行った．まず，歌謡番組固有の構造を明確にするとともに，その構造を検出するのに適した画像，音響特徴を説明した．次に，特徴毎に階層クラスタリングを行い，繰り返し構造を検出し，その結果を統合す

ることで，シーン境界判定を行う手法を提案した．実際に放映された歌謡番組を用いた実験から，複数特徴を協調的に用いることで，歌謡番組固有の構造を高精度に検出可能であることを示した．さらに，その構造に従うことでシーン分割の精度向上が可能となることを示した．

- 第 4 章では，撮像対象の自動抽出処理を導入した物体識別手法の提案を行った．まず，対象とする識別器である SVM とその識別超平面の性質について説明した．次に，SVM から得られる識別超平面の特性を利用した物体領域の自動選定法を提案し，提案法を用いた特徴ベクトルと識別器の更新法をそれぞれ説明した．評価用画像データセット Caltech 101/256 を用いた実験では，SVM から得られる識別超平面の特性を利用して撮像物体の領域を自動で選定可能であることを示した．さらに，選定した領域のみから抽出した特徴に基づいて更新した識別器を用いて識別精度の向上が可能となることを示した．

- 第 5 章では，映像特徴とタグ情報の類似性を考慮したグラフ構造に基づくリランキング手法の提案を行った．まず，映像のリンク関係を基に，映像とタグそれぞれの類似性を表現可能なグラフの構築方法を示し，そのグラフ上において，映像とタグ，各々の類似性を伝播させる手法を説明した．次に，キーワード検索による順位が，類似性の高い映像ノード間で近くなるように目的関数を導出した．最後に，目的関数の最小化方法を説明した．実際に映像共有サービスから収集した映像を用いた実験では，映像特徴とタグ情報の類似性の両者の類似性を用いることの有効性を示した．

以上のように，映像のシーン分割と撮像物体の識別それぞれが抱える課題を解決し，映像意味解析技術の精度向上を実現したことで，メタデータ付与の自動化へ貢献するものと考えられる．さらに，望む映像を上位に配置するような検索結果を提示するため，映像特徴とタグ情報の両者を用いたリランキング手法を導入した．これによって，本研究は，ユーザが望む映像の効率的な検索と閲覧の実現に貢献するものと考えられる．

6.2　今後の課題

本論文では，映像検索の高精度化を実現するために必要である，複数特徴を利用した映像意味解析手法とリランキング手法について提案した．具体的に，複数特徴の利用による構造に注目した映像のシーン分割，および撮像対象の自動抽出手法を導入した物体識別手法の提案を行った．さらに，映像特徴とタグ情報の両者の類似性に着目した検索結果のリランキング手法の提案を行った．本研究の今後の課題を以下に示す．まず，第 3 章の映像のシーン分割手法に関しては，歌謡番組の編集構造に注目してシーン分割の高精度化を実現した．一方，映像のジャンルに汎用なシーン分割手法について課題を残している．この実現のためには，映像のジャンル毎に適応的に，使用する特徴を選択可能にする枠組みの提案が必要である．次に，第 4 章の撮像物体の識別手法に関しては，1 枚の画像に撮像される物体を高精度に識別する手法を示した．一方，映像の撮像内容をより高精度に解析するため，画像特徴に加えて，自動選定した物体領域の動き特徴などを考慮することが必要である．最後に，第 5 章の検索結果のリランキング手法に関しては，実際の映像共有サービスから得られたデータセットに対して実験を行い有効性を示した．しかしながら，大規模なデータセットに対しての適用については未検討であり，処理の高速化が必要である．また，各々の手法を統合した検索システムの開発についての論述には至っていない．これらを今後の課題として挙げ，映像検索技術のさらなる高精度化について取り組んでいく所存である．

謝辞

本研究は，著者が北海道大学大学院情報科学研究科メディアネットワーク専攻に在学中の，2012 年 4 月より現在にいたる約 4 年にわたって行ったものである．

本研究を進めるにあたり終始あたたかいご指導と激励を賜りました長谷山美紀教授に深く感謝致します．長谷山教授の貴重なご指導がなくては，本論文の完成はありませんでした．今こうして論文をまとめるにあたり，そして長谷山教授のご指導の下での博士課程の研究生活を終えるにあたり，心より感謝の意を表すとともに，これまでの貴重な体験を今後の研究生活に生かし，努力を重ねていきたいと思います．

本論文をまとめるにあたりご助言をいただき，副査をお引き受けいただいた北海道大学大学院情報科学研究科，坂本雄児教授ならびに山本強教授に深く感謝致します．

小川貴弘助教におきましては，日頃から研究の本質を考えるようご指導とご鞭撻を賜りました．研究の進め方，考え方，まとめ方など研究の基礎を丁寧にご教授していただきました．学生生活においても，日々の様子を気にかけていただき，どんなに忙しいときでも真摯に対応していただき，いくつものご助言をいただけたこと，心より感謝致します．研究補助員の北市佳穂里さんにおきましては，様々な手続きの至る所でお世話になりました．様々な手続きのおかげで今日まで研究生活に専念し，多くの経験を得ることができました．また，お忙しいときにも気遣いの言葉をかけていただけたこと，心より感謝致します．また，御協力いただいたメディアダイナミクス研究室諸先輩並びに学生の皆様に感謝致します．

最後に，これまで自分の思う道を進むことに対し，温かく見守りそして辛抱強く支援していただいた両親と兄に深い感謝の意を表し謝辞とさせていただきます．

参考文献

[1] M. Yeung, B.-L. Yeo, and B. Liu. Segmentation of video by clustering and graph analysis. *Comput. Vis. Image Und.*, Vol. 71, No. 1, pp. 94–109, 1998.

[2] 帆足啓一郎, 菅野勝, 内藤正樹, 松本一則, 菅谷史昭. 汎用的特徴量に基づく動画像話題分割手法. 信学論 (D), Vol. 89, No. 10, pp. 2305–2314, 2006.

[3] W. Hu, N. Xie., L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 41, No. 6, pp. 797–819, 2011.

[4] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: a literature survey. *ACM Comput. Surv.*, Vol. 46, No. 3, pp. 38:1–38:38, 2014.

[5] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *Int. J. Multimed. Info. Retr.*, Vol. 2, No. 2, pp. 73–101, 2013.

[6] L. Fei-Fei. and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 524–531, 2005.

[7] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, Vol. 2, pp. 121–167, 1998.

[8] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *Proc. ACM Int. Conf. Multimedia*, pp. 35–44, 2006.

[9] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Proc. Int. Conf. Content-Based Image and Video Retrieval*, pp. 238–247, 2003.

[10] R. Yan and A. Hauptmann. Co-retrieval: A boosted reranking approach for video retrieval. *IEE Proc. Vision, Image and Signal Processing*, Vol. 152, No. 6, pp. 888–895, 2005.

[11] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 297–300, 2008.

[12] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 30, No. 11, pp. 1877–1890, 2008.

[13] X. Tian, Y. Yang, J. Wang, X. Wu, and X. S. Hua. Bayesian visual reranking. *IEEE Trans. Multimedia*, Vol. 13, No. 4, pp. 639–652, 2011.

[14] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A formal study of shot boundary detection. *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 17, No. 2, pp. 168–186, 2007.

[15] 長坂晃朗, 田中譲. カラービデオ映像における自動索引付け法と物体探索法. 情処学論, Vol. 33, No. 4, pp. 543–550, 1992.

[16] 鈴木賢一郎, 中嶋正臣, 坂野鋭, 三部靖夫, 大塚作一. 動き方向ヒストグラム特徴を用いた映像データからのカット点検出法. 信学論 (D-II), Vol. 86, No. 4, pp. 468–478, 2003.

[17] H. Sundaram and S.-F. Chang. Video scene segmentation using video and audio features. In *Proc. IEEE Int. Conf. Multimedia Expo*, pp. 1145–1148, 2000.

[18] A. Hanjalic, R.L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 9, No. 4, pp. 580–588, 1999.

[19] Y. Zhai and M. Shah. Video scene segmentation using markov chain monte carlo. *IEEE Trans. Multimedia*, Vol. 8, No. 4, pp. 686–697, 2006.

[20] N. Goela, K. Wilson, Feng Niu, A. Divakaran, and I. Otsuka. An svm framework for genre-independent scene change detection. In *Proc. IEEE Int. Conf. Multimedia Expo*, pp. 532–535, 2007.

[21] J. Wang, L. Duan, Q. Liu, H. Lu, and J.S. Jin. A multimodal scheme for program segmentation and representation in broadcast video streams. *IEEE Trans. Multimedia*, Vol. 10, No. 3, pp. 393–408, 2008.

[22] 柳井啓司. 一般物体認識の現状と今後. 情処学論, Vol. 48, No. SIG 16, pp. 1–24, 2007.

[23] P. Duygulu, K. Barnard, and D.F.N. Freitas. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. Eur. Conf. Comput. Vis.*, pp. 97–112, 2002.

[24] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr benchmark: A new evaluation resource for visual information systems. In *Proc. Int. Conf. Language Resources and Evaluation*, pp. 13–23, 2006.

[25] L. Fei-Fei., R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Understand.*, Vol. 106, No. 1, pp. 59–70, 2007.

[26] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. In *Dept. Comput. Sci., California Inst. Technology, Tech. Rep.7694*, pp. 1–20, 2007.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[28] G.A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[29] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Proc. ACM MM WS on Multimedia Information Retrieval*, pp. 321–330, 2006.

[30] S. Zanetti, L. Zelnik-Manor, and P. Perona. A walk through the web's video clips. In *Proc. Int. Conf. Computer Vision and Pattern Recognition Workshops*, pp. 1–8, 2008.

[31] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, Vol. 60, No. 2, pp. 91–110, 2004.

[32] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 1, pp. 886–893, 2005.

[33] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, Vol. 110, No. 3, pp. 346–359, 2008.

[34] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[35] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 26–33, 2005.

[36] N. Otsu and T. Kurita. A new scheme for practical flexible and intelligent vision systems. In *Proc. IAPR Workshop on Computer Vision*, pp. 431–435, 1988.

[37] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Computer Vision*, Vol. 42, No. 3, pp. 145–175, 2001.

[38] J. Yang, K. Yu., Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1794–1801, 2009.

[39] J. Wang, J. Yang, K. Yu, F. Lv, and T. Huang. Locality-constrained linear coding for image classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3360–3367, 2010.

[40] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 2, pp. 2169–2178, 2006.

[41] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, 2008.

[42] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. In *Proc. Eur. Conf. Comput. Vis.*, pp. 696–709, 2008.

[43] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. learning mid-level features for recognition. In *Proc. Eur. Conf. Comput. Vis.*, pp. 2559–2566, 2008.

[44] S. Gao, I. Tsang, L. Chia, and P. Zhao. Local features are not lonely laplacian sparse coding for image classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3555–3561, 2010.

[45] J. Sanchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 1665–1672, 2011.

[46] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 2nd edition, 1998.

[47] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[48] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 107–117, 1998.

[49] 坂野鋭, 武川直樹, 中村太一. 核非線形相互部分空間法による物体認識. 信学論 (D-II), Vol. 84, pp. 1549–1556, 2001.

[50] 前田賢一, 渡辺貞一. 局所的構造を導入したパターン・マッチング法. 信学論 (D), Vol. J68-D, No. 3, pp. 345–352, 1985.

[51] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, Vol. 10, No. 5, pp. 1299–1319, 1998.

[52] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 511–518, 2001.

[53] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Analysis and Machine Inteligence*, Vol. 25, No. 5, pp. 564–577, 2003.

[54] N. Nitanda and M. Haseyama. Audio-based shot classification for audiovisual indexing using pca, mgd and fuzzy algorithm. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, Vol. E90-A, pp. 1542–1548, 2007.

[55] Open Computer Vision Library (OpenCV).[Online]. Available:http://opencvlibrary.sourceforge.net.

[56] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, Vol. 2, pp. 1–27, 2011.

[57] J.C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *In Advances In Large Margin Classifiers*, pp. 61–74, 1999.

[58] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, Vol. 54, pp. 45–66, 2004.

[59] H. Zhang, A.C. Berg, M. Maire, and J. Malik. SVM-KNN: discriminative nearest neighbor classification for visual category recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 2, pp. 2126–2136, 2006.

[60] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proc. ACM Int. Conf. Image and video retriecal*, pp. 401–408, 2007.

[61] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classication with sets of image features. In *Proc. Int. Conf. Comput. Vis.*, pp. 1458–1465, 2005.

[62] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, Vol. 73, pp. 213–238, 2007.

[63] E. Apostolidis and V. Mezaris. Fast shot segmentation combining global and local visual descriptors. In *Proc. Int. Conf. Acoust., Speech., Signal Process.*, pp. 6583–6587, 2014.

[64] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[65] NLTK 3.0 documentation. [Online]. Available: http://www.nltk.org.

[66] X. Li, C.G.M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Trans. Multimedia*, Vol. 11, No. 7, pp. 1310–1322, 2009.

[67] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. Int. Conf. World Wide Web*, pp. 327–336, 2008.

[68] R.L. Cilibrasi and P.M.B. Vitanyi. The google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, Vol. 19, No. 3, pp. 370–383, 2007.

[69] T. Yao, C. W. Ngo, and T. Mei. Circular reranking for visual search. *IEEE Trans. Image Processing*, Vol. 22, No. 4, pp. 1644–1655, 2013.

[70] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, p. P10008, 2008.

[71] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[72] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, Vol. 3, No. 4, pp. 333–389, 2009.

[73] M. Wu and B. Schölkopf. Transductive classification via local learning regularization. In *Proc. Int. Conf. Artificial Intelligence and Statistics*, pp. 628–635, 2007.

著者の研究業績

論 文

(A) 学会誌

[**A-1**] S. Yoshida, H. Okada, T. Ogawa, and M. Haseyama,“A Method for Improving SVM-Based Image Classification Performance Based on a Target Object Detection Scheme,” ITE Transactions on Media Technology and Applications, vol.1, no. 3, pp.237-243, July, 2013.

[**A-2**] 吉田 壮, 小川 貴弘, 長谷山 美紀,“歌謡番組における映像の構造に注目したシーン分割の高精度化,” 電子情報通信学会論文誌 (D), vol. J97-D, no. 7, pp. 1177-1188, 2014 年 7 月.

(B) 技術報告

[**B-1**] 吉田 壮，小川 貴弘，長谷山美紀，“歌謡番組における映像の構造に注目したシーン分割手法 〜顔認識のシーン分割精度に与える効果の検証〜”，映像情報メディア学会技術報告, vol. 37, no. 8, pp. 101-106, 2013 年 2 月.

[**B-2**] 吉田 壮，小川 貴弘，長谷山美紀，“Hidden Conditional Random Fields による映像の構造解析に基づくシーン分割の高精度化に関する検討”，映像情報メディア学会技術報告, vol. 38, no. 7, pp. 285-290, 2014 年 2 月.

[**B-3**] 吉田 壮，小川 貴弘，長谷山美紀，“映像特徴とタグ情報を考慮したネットワーク構造解析に基づく Web 映像検索に関する検討”，第 29 回信号処理シンポジウム, pp. 521-526, 2014 年 2 月.

[**B-4**] 吉田 壮，小川 貴弘，長谷山美紀，“映像特徴とタグ情報を考慮したネットワーク構造解析に基づく Web 映像検索に関する検討 ～ネットワークの局所的構造を考慮した正則化項の導入による高精度化～”，映像情報メディア学会技術報告, vol. 39, no. 7, pp. 77-82, 2015 年 2 月.

[**B-5**] 吉田 壮，小川 貴弘，長谷山美紀，“グラフ構造に基づくリランキングを利用した Web 映像検索に関する検討 ～初期検索結果の誤りを考慮した最適化に基づく高精度化～”，第 30 回信号処理シンポジウム, pp. 521-526 (2015)

(C) 国際学会

[**C-1**] S. Yoshida, T. Ogawa, and M. Haseyama,“Accurate Graph-Based Scene Segmentation for Videos of Music Programs,” Proceedings of the 27th International Technical Conference on Circuits/Systems, Computers and Communications, Sapporo, Japan, July, P-T1-09, 2012.

[**C-2**] S. Yoshida, T. Ogawa, and M. Haseyama,“Accurate Scene Segmentation Method Based on Video Structures in Music Programs,” Proceedings of International Workshop on Advanced Image Technology, Nagoya, Japan, January, pp. 621 -625, 2013.

[**C-3**] S. Yoshida, T. Ogawa, and M. Haseyama, “Effectiveness Verification of Audio-Based Preprocessing on a Scene Segmentation Method in Music Programs,” Proceedings of the 28th International Technical Conference on Circuits/Systems, Computers and Communications, Yeosu, Korea, July, pp. 846-849, 2013.

[**C-4**] S. Yoshida, T. Ogawa, and M. Haseyama, “Heterogeneous Graph-Based Video Search Reranking Using Web Knowledge via Social Media Network,” Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, Brisbane, Australia, October, pp. 871-874, 2015.

(D) 講演発表

[D-1] 吉田 壮，小川 貴弘，長谷山美紀，“相互部分空間法を用いた音楽番組のインデキシング高精度化に関する検討”，電気・情報関係学会北海道支部連合大会講演論文集, p. 144, 2011 年 10 月.

[D-2] 吉田 壮，小川 貴弘，長谷山美紀，“物体認識における識別器の高精度化に関する検討 〜SVM を用いた物体領域の自動選定手法の導入〜”，電気・情報関係学会北海道支部連合大会講演論文集, p. 122, 2012 年 10 月.