画像・映像意味理解の現状と検索インタフェース

Recent Trends of Image and Video Semantic Analysis and Retrieval Interfaces

長谷山美紀

Abstract

画像・映像意味理解の研究動向とその検索への応用について紹介する．更に，画像及び映像が持つ固有の多義性とあいまい性から検索結果の可視化システムの必要性を議論し，その実現の試みについて紹介するとともに今後の展開について考える．

キーワード：画像・映像検索，意味理解，統計的機械学習，Bag of Keypoints，可視化インタフェース

1. はじめに

地上波ディジタル放送の開始，高速通信網の普及，更には，記録媒体の大容量化と低廉化も伴い，我々の周りは急速にディジタル化し，大量のディジタルデータが蓄積されている．そのデータから価値を見いだすことが重要とされ，様々な研究が行われている．特に，画像や映像はその娯楽性から，希望のコンテンツを視聴可能な検索サービスが存在し，活発に利用されている[(1)]．

既存検索サービスは，コンテンツに付与されたメタデータにより行われている．このメタデータとして，過去には，撮像時に得られる情報（撮像日時や場所等）や人手により付与されたキーワードが用いられてきたが，この数年間に，画像・映像解析手法が発展し，自動でメタデータを抽出する手法が実用に向けて大きく前進した．以前から研究されてきた機械学習，特に統計的機械学習によるメディア解析手法の高度化が，画像情報の認識と映像情報の意味理解の高精度化に大きく貢献している．

一方，このような手法の高度化と並行して，新たに検討すべき問題も顕在化した．我々は，望むコンテンツを明確に表現するクエリ（質問となるテキストや画像等）を想定できない場合があり，そのような場合には，検索という作業では望むコンテンツを取得できない[(2)]．本稿では，画像・映像の意味理解についての研究動向を紹介し，新たに現れた課題を解決する試みとして，画像の特徴に基づくデータ群の可視化手法を紹介し，本研究分野の今後の展開について考えてみたい．

長谷山美紀　正員　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
E-mail miki@its.hokudai.ac.jp
Miki HASEYAMA, Member (Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan).
電子情報通信学会誌　Vol.93 No.9 pp.764-769 2010年9月

2. 画像・映像の意味理解の研究動向と検索への応用

1.で述べたように，画像・映像検索は一般にコンテンツに付与されたメタデータにより行われる．大量の画像・映像データを検索対象とする時代の到来を予見して，画像・映像の意味を理解するための研究が始まった[(3),(4)]．これら手法の多くは，低レベル特徴に注目したものであり，示されたクエリに対して色やテクスチャが類似した画像の検索が可能であったが，真にユーザが望むコンテンツの検索のためには，高レベル特徴，つまり，映像の意味を表すメタデータを付与する必要がある．過去において，このような技術の実現は，困難であるとされていたが，ここ数年，コーパスベース(用語)のメディア解析手法[(5)]が提案され，映像情報の意味解析技術は実用に向けて大きく前進した．先行するコーパスベースの解析技術といえば，音声認識や自然言語処理のほか，文字認識等が挙げられる．映像解析においても，このアプローチを採用することで映像情報が持つ多様性やあいまいさに柔軟に対応できる手法が実現されており，特に，SVM（サポートベクトルマシン）やベイズ分類

等の統計的機械学習の貢献が大きい[6]．このような手法は，画像・映像検索サービスを提供する一部のサイトで，有害画像の検出や，著作権違反映像を検出するフィルタとして用いられている．また，画像・映像の意味理解に有効とされる多くの特徴量が提案され，近年では，テキスト検索における Bag of Words 法に対応する，Bag of Keypoints 法[7]が有力とされている．これも，統計的機械学習手法との高い親和性によるものと予想される．

ところで，機械学習には，学習のためのデータセットが必要である．どのようにしてデータセットを準備するかによって，適用可能な画像の種類が限定される場合やメタデータ付与の精度が低下する等の問題がある．また，正解データ（Ground Truth）をいかにして集めるかについても検討しなければならないが，現在は，研究者が人的労力をかけて多種多様な画像群を準備するのではなく，実現された手法の比較を可能とする共通のデータセットの提供が各所で行われている[8]．

一方，サーベイランス目的で取得された映像や，個人が撮像した画像や映像の増加から，更に複雑な問題の解法が望まれている．この種の映像では，識別したい事象によってはデータが少なく，共有できない場合があり，多様性を備えたデータセットの準備が不可能となるため，現在提案されている手法による解決が困難となる．このような場合には，その意味を理解するために，階層性を含む映像の構造解析を検討する等して画像が持つ多義性やあいまいさを高度に把握する必要がある．

3. 画像・映像検索結果の可視化とインタフェース

現状までの研究は，画像・映像が持つ，多様性や多義性，あいまい性から生じる問題を解決し，高度な意味理解を実現するために進められてきた．しかしながら，それが実現に向けて前進するのと並行して，新たに検討すべき問題も顕在化した．我々は，望むコンテンツを的確に表現するクエリを想定できない場合があり，常にクエリを意識して画像や映像を見ているわけではない．このような状況で，いかにして情報を提供すべきかを検討する試みが始まっている．画像・映像が持つあいまいさを受容し，更には人間側にも存在する多様性やあいまいさを許容しながら，それらを積極的に活用することで，個人が望む情報を獲得するための適応的な情報可視化インタフェースの実現である．

類似の問題がテキストデータにおいても存在し，その解決法の一つとして連想型検索（Associative Search）[9]が知られている．連想型検索は，ユーザが検索結果（この場合の検索結果とは，検索の過程で形成される探索領域と理解して頂きたい）の中から新たなデータを選択し，それをクエリとして繰り返し選択を行うことで，望むデータを取得する．この手法に基づけば，クエリを具体的に想定することが困難な場合においても，検索結果から新しいクエリを選択する繰返しの操作により徐々に検索精度を向上させ，最終的に目的の情報に到達することが可能となる．テキストデータについてはこの方法が実用化のフェーズに進んでいる[10]．画像及び映像においても，先の問題の解決のために，連想型検索が，有効な知見を与えるものと考える．そこで，以下に，画像及び映像の連想型検索実現のための試みとして筆者が行った，検索結果やデータベース全体を俯瞰する可視化インタフェースについて簡単に紹介する．

（1） 三次元連想型画像検索インタフェース Image Vortex と Image Cruiser への展開

Image Vortex は，従来の検索手法では困難であった，ユーザが明確なクエリを持ち合わせない場合の画像検索を実現するための試みである[注1]．蓄積された大量の画像の中からインタフェースを通して，ユーザが希望する画像を効果的に獲得するシステムの実現を目指している．提案システムでは，データベース中の画像から算出される特徴[11]を要素とするベクトルを低次元特徴空間へ射影し，得られた特徴ベクトルを用いて，画像間の差異を距離で定義する[11),(12]．なお，距離算出の際には，適宜画像特徴から得られた重みを併用している．更に，図1に示すように，得られた距離に基づいて三次元空間上に画像を配置し，これを操作することで，ユーザは画像データベースの全体を俯瞰し，効率良く希望の画像にたどり着くことが可能となる．三次元空間上の画像は，多次元尺度構成法を拡張した手法によって定義される距離尺度に基づき配置が決定される．Image Vortex の実用化に向けて実現された大規模データベース俯瞰型検索エンジン Image Cruiser[注2]を図2に示す．

用語解説

コーパスベース　実データの持つ多様性を十分に反映した正解データ付きの大規模コーパスを整備した上で，統計的手法や機械学習等を用いる解析アプローチ．

ライフログ　一般に人の行動（life）をディジタルデータとして記録（log）したもの．本稿では，行動に伴い，人が取得した情報も含めてライフログと呼ぶ．

セマンティックギャップ　マルチメディアデータ（画像，映像データ等）から抽出される特徴と人間が理解する意味との不一致．

（注1）　CEATEC JAPAN 2006（2006 年 10 月 3 日～10 月 7 日，幕張メッセ）「情報大航海ゾーン」にて「Cyber Space Navigator～次世代情報アクセス～」を出展．

（注2）　Image Cruiser は，経済産業省情報大航海プロジェクトにおけるサービス実証事業において開発が行われた．（http://imagecruiser.jp/light/demo.html）

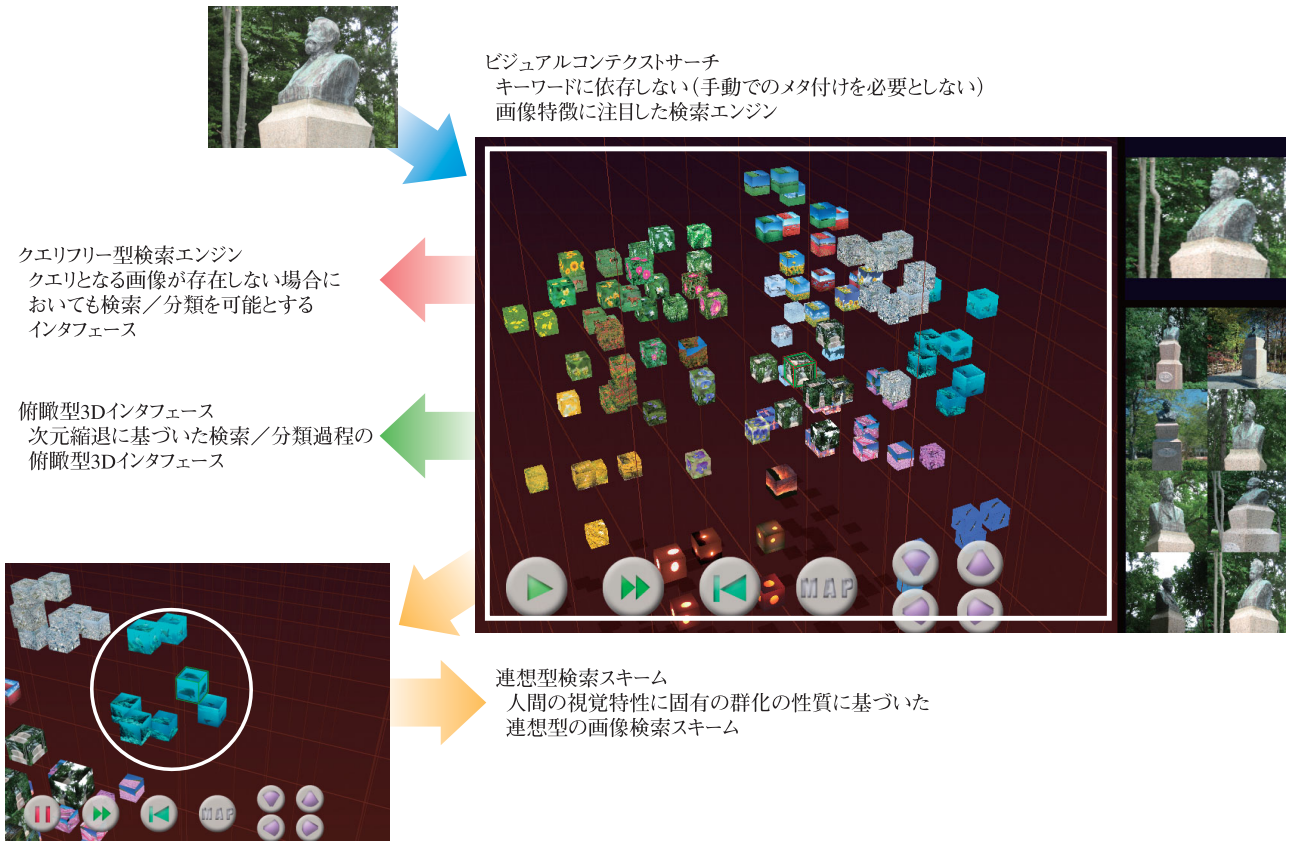

図1　Image Vortex のインタフェース　　キーワード等のクエリを必要とせず，データベース中に存在する画像を俯瞰して閲覧する 3D インタフェースを導入することで，効率良く希望の画像へ到達することを可能とする．また，類似する画像群間で隣接する配置を行うことで，人間の視覚特性に固有の群化の性質に基づく連想型の検索を実現している．

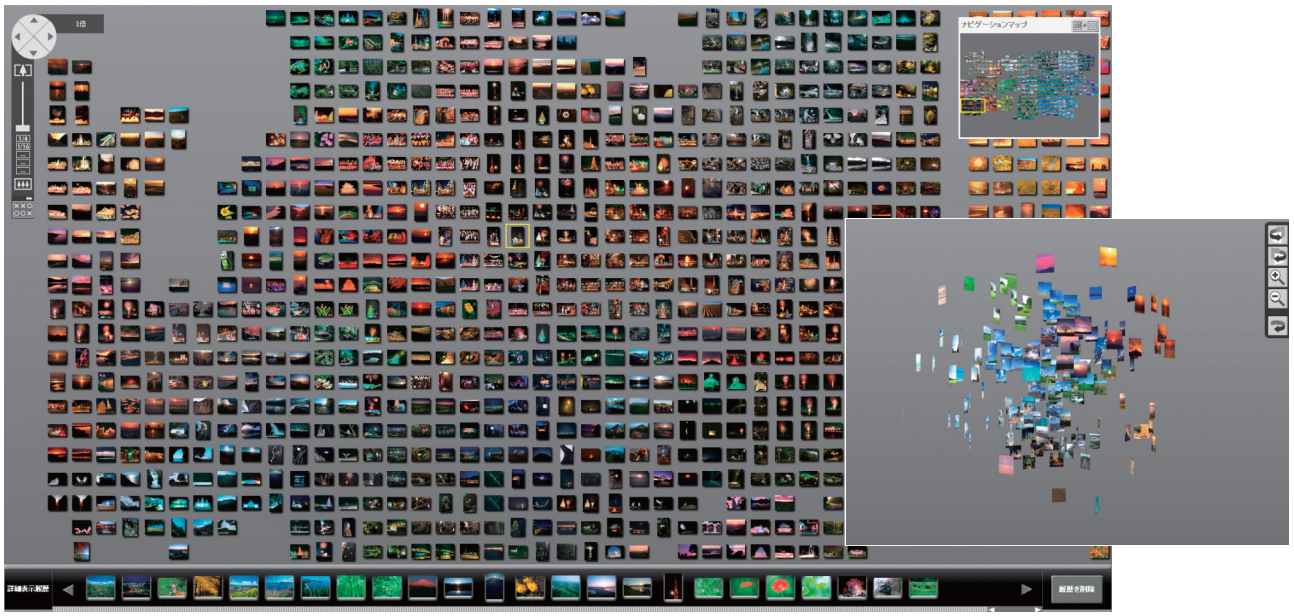

図2　Image Cruiser のインタフェース　　Image Cruiser では，Image Vortex で定義される距離尺度に基づき，高速な画像の配置を実現している．ユーザに親和性のあるインタフェースを実現することによって，膨大な量の画像を俯瞰し，容易かつ高速に望む画像へ到達することが可能となる．更に，本インタフェースでは，キーワード等の「望むものを的確に表現する」クエリを一切必要としない特徴を持つ．本図に示すデータベースは，36,000 枚の北海道の風景に関する画像を含んでいる．なお，現状のシステムは，100 万枚の画像に対応するサービスが可能である．

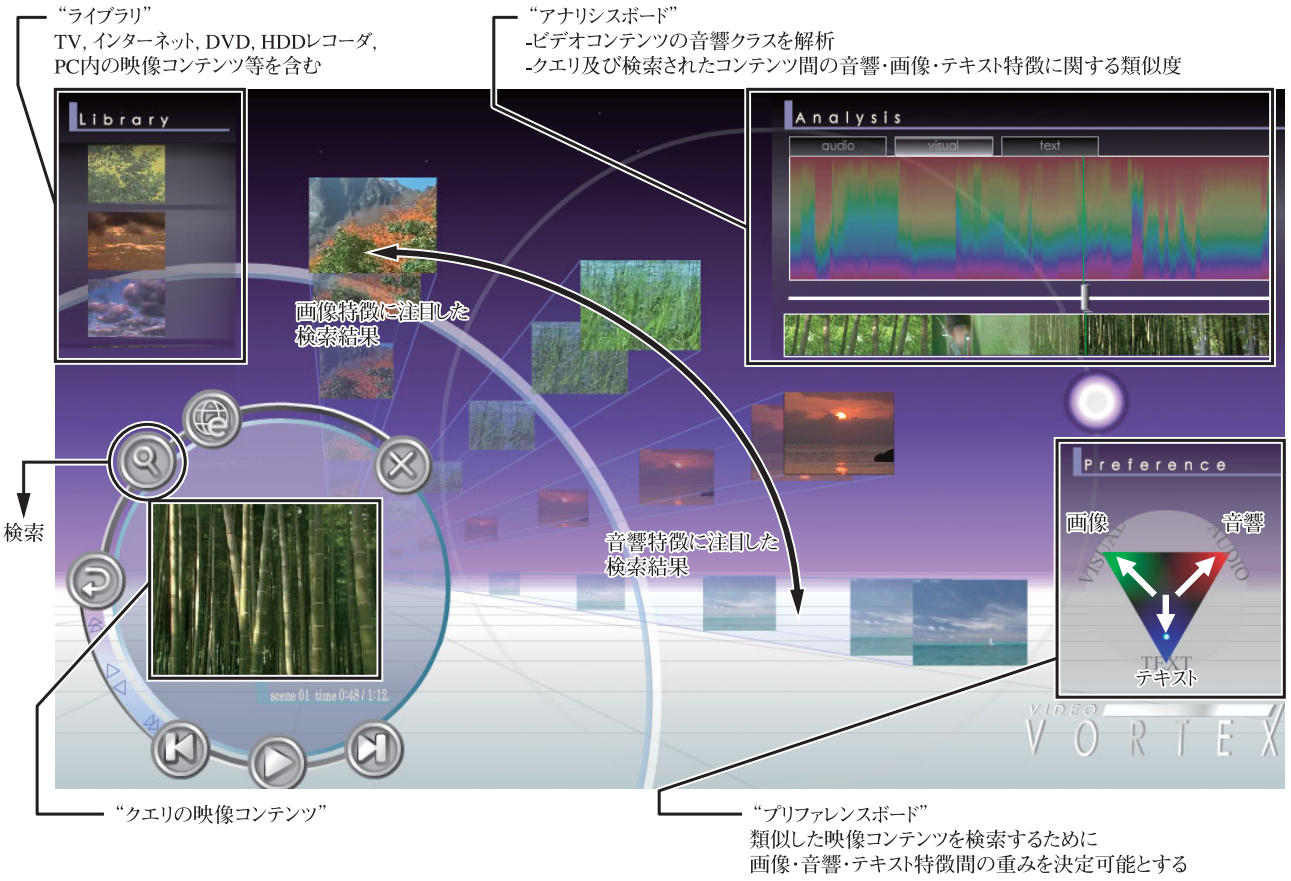

図 3　Video Vortex のインタフェース　　画像，音響，楽曲，及びテキスト特徴を統合したマルチモーダル処理を導入し，ユーザのし好をモデル化するインタフェースと組み合わせて用いることで，より容易に望む映像へ到達することを可能としている．

（2）　三次元連想型映像検索インタフェース Video Vortex

Image Vortex が，画像を対象とするのに対し，Video Vortex は，蓄積された大量の映像からユーザが希望する映像を効果的に獲得するための検索インタフェースの試みである(注3)．Video Vortex は，映像間の類似度に基づいて三次元の空間上に映像を配置することにより，空間の距離によって映像の類似性を理解することが可能なインタフェースとなっている．なお，本システムでは，個人の映像の好みをモデル化するために，視聴行動を表すデータを取得できる「プリファレンスボード」を準備しており，得られた操作履歴をフィードバックとして利用することで，個人の好みや視聴の状況に合わせた適応的な可視化を可能とする[13]．図 3 に，提案システムによる映像検索の様子を示す．

提案システムでは，画像・音響特徴を同時に用いるマルチモーダル処理[14]を導入したファジー制御規則の適用により映像をシーンごとに分割し，クエリに類似したシーンを提示する．このとき，シーン間の距離は，画像・音響特徴量[15]，楽曲特徴量[16]，及び音声信号や画像信号から算出されるテキスト特徴量（2. で紹介した手法で得られるメタデータ等を含む）に対して映像の時間方向の伸縮に対応可能な動的計画法を導入することで算出する．また，得られる距離に関し，ユーザはプリファレンスボードを通して“視覚”，“聴覚”，“テキスト”に自由に重みを設定しながら検索を繰り返すことで，希望の映像にたどり着くことが可能となる．この方法によって，映像データから得られる複数の異種特徴量の連携利用が可能となり，インタフェースによる効果的な検索結果の可視化によって，ユーザがクエリを持ち合わせていない場合でも，希望する映像の獲得が期待できる．

(注 3)　CEATEC JAPAN 2007（2007 年 10 月 2 日～10 月 6 日，幕張メッセ）「情報大航海プロジェクトブース」にて「Cyber Space Navigator～映像の次世代検索システム～」を出展．

4. 画像・映像に固有な検索インタフェース実現の試み

望むコンテンツを的確に表現するクエリを想定せずに人間が画像・映像を視聴し，望む情報を獲得する際には，異なるメディアの情報による想起が重要な役割を果たすことが予想される．例えば，静止画から動画像や音楽等，異なるメディアを横断し，望む情報にたどり着く検索である．

4. では，音響・動画像信号及び意味的特徴に基づいたメディア横断型の検索・分類，及びその可視化インタフェースの実現の試みについて紹介する(注4)．提案するインタフェースでは，データベース中に含まれる動画，静止画，音楽，更にはキーワードが付与されているコンテンツ及びユーザが視聴した履歴に基づき，画像特徴，音響・音楽特徴，及びテキスト特徴間での相関を求め，その結果から各コンテンツに対して，画像特徴，音響特徴，意味的特徴のすべての推定を可能とする．これらは，既知の特徴量から未知の特徴量を推定するために，カーネル PCA 及びカーネル CCA を利用している[17),(18)]．これにより，メディアが異なるコンテンツ間においても類似性の判定を行うことが可能となり，メディア横断型の検索及び分類が可能となる．また，実現された検索システムには，ユーザの画像・映像及び音楽の好みを，データセットと操作履歴からモデル化する手法が含まれている．望む情報を獲得するため，人がメディアを区別せずに，ほかのユーザの好みから自身の新たな好みのコンテンツに気付く知識創出の誘発を支援することを予見して実装を試みたものである．ライフログ(用語)の利用により更なる高度化と真の知識創出の実現に向けての前進が期待される．

5. ま と め

本稿では，画像・映像意味理解の研究動向について紹介し，その発展について，確率的情報処理の展開から分析した．更に，画像及び映像が持つ固有の多義性とあいまい性から検索結果の可視化システムの実現に関する試みを紹介し，今後の展開について検討した．

本文 3. 以降については，検討の必要性と筆者の試みの紹介にとどまっており，すぐに解決される容易な問題ではないことは十分に理解している．問題を解くためには，人間の認識のメカニズムの解明までも含み，以前より議論されているセマンティックギャップ(用語)の克服についても検討しなければならない．その前進のためには，多様な学問分野の融合は大きな鍵となり，ビジョンコンピューティングにおける情報処理も，更にそのステージを進めて，人間及び画像・映像の両者が持つ多様性とあいまい性をいかにして解き明かすかという大きな問題に対峙することになると予想する．

最後に，本文では「検索」の言葉を使用したが，これについて筆者の考えを申し添えたい．「検索」という言葉を使いながらも，本文の議論は，検索の定義を超えていると考える．筆者はかつて，探し出す意味で「探索」の言葉を使ったが，それだけでも表現できない．個人の思考とし好に合わせて，システム自身が検索だけでなく，探索空間を提示する推薦ともいえない，独自の融合形態を形成することで実現される，知識創出支援のようなものであろうか．このように考えると，大量に保持された画像や映像からの価値創出は，知識の享受と共有のメカニズムを実現することかもしれない．用語については，容赦頂き，読者自身の考えから適宜読み替えをお願いしたい．

(注 4) CEATEC JAPAN 2008（2008 年 9 月 30 日～10 月 4 日，幕張メッセ）「情報大航海プロジェクトブース」にて「Cyber Space Navigator～メディア横断型次世代検索～」を出展．

文 献

（1） 平成 21 年版情報通信白書，第 4 章第 1 節 1(4)，インターネットの利用目的，『平成 19 年末から最も利用が伸びたのはデジタルコンテンツ(音楽・音声，映像，ゲームソフト等)の入手・聴取であり，前年から 3.1 ポイント増となっている』ことが報告された．http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h21/index.html

（2） I. Campbell, “Applying ostensive functionalism in the place of descriptive proceduralism : the query is dead,” C.W. Johnston and M.D. Dunlop, eds., Proceedings of the Workshop on Information Retrieval and Human Computer Interaction, pp. 77-81, University of Glasgow, Sept. 1996.

（3） A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1349-1380, 2000.

（4） M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, “Query by image and video content : the QBIC system,” Computer, vol. 28, no. 9, pp. 23-32, 1995.

（5） 佐藤真一，“コーパスベース映像解析，”信学技報，PRMU 2006-75，pp. 111-120, Sept. 2006.

（6） その様子は，映像コーパスを用いた検索技術のための競争型ワークショップの結果により知ることができる．例えば，よく知られるワークショップとして以下を示す．TREC（The Text Retrieval Conference）Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/，その結果は，http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

（7） G. Csurka, C.R. Dance, L. Fan, and C. Bray, “Visual categorization with bags of keypoints,” Proc. Of European Conference on Computer Vision（ECCV）, pp. 1-22, 2004.

（8） 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（CVIM_19）, pp. 1-24, 2007.

（9） M. Kawamoto, Y. Kiyoki, N. Yoshida, S. Fujishima, and S. Aiso, “An implementation of a semantic associative search space for medical document databases,” Proceedings of the 2004 Symposium on Applications and the Internet-Workshops（SAINT 2004 Workshops）, pp. 488-493, 2004.

（10） 想--IMAGINE Book Search（http://imagine.bookmap.info）, ASSOCIE（http://www.nri.co.jp/renso/）, reflexa（http://labs.preferred.jp/reflexa/）

（11） 渡辺隆志，長谷山美紀，“エッジを考慮した類似画像分類の高精度化に関する考察，”信学技報，ITS2006-44, IE2006-229, pp. 7-10, Feb. 2007.

（12） R. Tokumoto and M. Haseyama, “Color distribution-based similar image clustering and its performance evaluation,” International Conference on Kansei Engineering and Emotion Research 2007（KEER2007）, no. C-25, 2007.

（13） S. Takahashi and M. Haseyama, “Realization of personalized video recommendation based on audio-visual features,” International Conference on Kansei Engineering and Emotion Research 2007（KEER2007）, no. I-1, 2007.

（14） N. Nitanda and M. Haseyama, “Audio-based shot classification for audiovisual indexing using PCA, MGD and fuzzy algorithm,” IEICE Trans. Fundamentals, vol. E90-A, no. 8, pp. 1542-1548, Aug. 2007.

（15） M. Yamamoto and M. Haseyama, “An accurate scene segmentation

method based on graph analysis using object matching and audio feature,” IEICE Trans. Fundamentals, vol. E92-A, no. 8, pp. 1883-1891, Aug. 2009.
(16) 今野聡司，二反田直己，長谷山美紀，“メロディーとリズムに着眼した音楽信号の類似度に関する一考察，”信学技報，ITS2006-65, IE2006-250, pp. 125-128, Feb. 2007.
(17) T. Ogawa and M. Haseyama, “POCS-based annotation method using kernel PCA for semantic image retrieval,” IEICE Trans. Fundamentals, vol. E91-A, no. 8, pp. 1915-1923, Aug. 2008.
(18) Y. Hatakeyama, T. Ogawa, S. Asamizu, and M. Haseyama, “A novel video retrieval method based on web community extraction using features of video materials,” IEICE Trans. Fundamentals, vol. E92-A, no. 8, pp. 1961-1969, Aug. 2009.

（平成 22 年 4 月 5 日受付　平成 22 年 4 月 23 日最終受付）

長谷山　美紀（正員）

1988 北大大学院工学研究科修士課程了．1989 北大・電子科学研究所・助手．1994 北大大学院工学研究科助教授．2005～2006 米国ワシントン大客員助教授．2006 北大大学院情報科学研究科教授，現在に至る．博士（工学）．画像及び映像処理とその意味的解析への応用の研究に従事．総務省情報通信審議会専門委員，経済産業省情報大航海プロジェクト研究会第 1 分科会次世代情報アクセスに関するビジョンと技術委員会委員，経済産業省情報大航海プロジェクト評議員/技術アドバイザー，日本放送協会（NHK）放送技術審議会委員．IEEE，映像情報メディア学会，日本音響学会各会員．

国際会議

2010 IEEE International Conference on Acoustics, Speech and Signal Processing（ICASSP2010）

主　催：IEEE Signal Processing Society
日　時：2010 年 3 月 14～19 日（6 日間）
会　場：Dallas Sheraton Hotel（米国，ダラス）
参加者：約 2,000 名
主要参加国：米国，中国，フランス，日本，カナダ，英国，ドイツ，オーストラリア，韓国ほか
セッション数及び論文数：153 セッション（内訳：Lecture：70，Poster：83），1,411 件（採択率約 48%）
デモンストレーション（Show & Tell）：32 件
Proceedings 発行所：IEEE
主たるトピックス：

ICASSP は，信号処理研究の最大規模の国際会議であり，信号処理，音声音響処理，画像処理，無線通信など幅広い分野をカバーする．中でも音声系の発表が多く，世界各国から著名な研究者が集う．本分野の研究動向を把握するには最適な会議である．

会期中は，オーラルセッション六つとポスターセッション七つがパラレルで行われた．今回はポスター会場が広大でデモンストレーションブースや coffee break の会場も兼ねており，更にレクチャーセッション会場も隣接しており参加者が一か所に集まったため，研究者間の意見交換や交流は例年に増して活発に行われていた．

報告者が主に聴講した音声・音響信号処理に関するセッションでは，従来から活発に研究が行われている音声強調，雑音抑圧が相変わらず目立っていたほか，個々は独立して動作するマイクを一つの集合として利用するアドホックマイクアレーのような新しい取組みに関する発表も見られた．また Compressive Sampling など信号処理の新しい原理についても盛んに議論が行われていた．

次回は，2011 年 5 月 22～27 日，チェコ共和国プラハにて開催予定である．

（執筆者　日岡裕輔　正員
日本電信電話株式会社 NTT サイバースペース研究所）