論文 / 著書情報
Article / Book Information

論題(和文)	フレッシュアイズ 映像研究現場紹介 東京工業大学 篠田研究室
Title(English)	
著者(和文)	井上中順, 篠田浩一
Authors(English)	Nakamasa Inoue, Koichi Shinoda
出典(和文)	映像情報メディア学会誌, Vol. 63, No. 8, pp. 1116-1119
Citation(English)	, Vol. 63, No. 8, pp. 1116-1119
発行日 / Pub. date	2009,
権利情報 / Copyright	本著作物の著作権は映像情報メディア学会に帰属します。 Copyright (c) 2009 Institute of Image Information and Television Engineers.

研究室のホームページアドレス
↓
http://www.ks.cs.titech.ac.jp/

東京工業大学 篠田研究室

井上中順†

†東京工業大学 情報理工学研究科 計算工学専攻

"Shinoda Laboratory, Tokyo Institute of Technology" by Nakamasa Inoue (Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo)

1. まえがき

私たちが所属している篠田研究室は東京工業大学の大岡山キャンパスにあります．篠田研究室では，パターン認識とその実世界応用について研究をしており，特に動画像や音声を対象とした認識・理解の研究を行っています．また，音声に関する研究は古井研究室と共同で行っています．

研究室には留学生も多く在籍し，留学生は母国語の音声を研究するなど，各々が自分の得意な分野の研究に励んでいます．

本稿では，当研究室の研究内容について紹介させていただきます．また，本稿は2007年のメディア工学研究会の発表で当研究室のメンバが優秀研究発表賞を受賞したことから執筆の機会を与えていただくことになりました．

2. 東京工業大学

東京工業大学（東工大）は，前身の東京職工学校が1881年に設置されてから，128年の長い伝統を持つ理工系大学です．東工大には大岡山キャンパスとすずかけ台キャンパス，田町キャンパスがあります．また，大学院には，理工学，生命理工学，総合理工学，情報理工学，社会理工学，イノベーションマネジメントの六つの研究科があり，篠田研究室は情報理工学研究科の計算工学専攻に属しています．

計算工学専攻では，情報認識や自然言語処理からソフトウェア開発，計算機アーキテクチャの分野まで，多岐にわたって情報工学に関わる研究が行われています．

また，東工大の大きな特徴として挙げられるのが，スーパーコンピュータの“TSUBAME”です．TSUBAMEは「みんなのスパコン」をキャッチフレーズとしていて，東工大の研究室に所属している人なら誰でも研究に利用することができます．TSUBAMEは国内でも最大級のスーパーコンピュータで，2008年11月にはアップグレードが行われ，世界で初めてGPUがスーパーコンピュータに導入されました．その結果，ピーク性能で77.48TFLOPSを記録し，今後も更なる高性能化が目指されています．

このTSUBAMEを使えば，大量の動画像データを対象とした研究など，大規模な研究も高速に行うことができ，充実した実験環境が整っていると言えるでしょう．

3. 篠田研究室

篠田研究室（**図1**）は2003年に設立された比較的新しい研究室です．現在，研究室には篠田浩一准教授の下に，博士課程後期2名，修士課程10名，学部生3名が在籍しており，各人が自分の研究テーマを持ち，時に熱い議論を交わし合いながら研究を進めています．

研究内容はパターン認識技術を応用した，動画像と音声の認識・理解が主となっています（**図2**）．動画像と音声には，どちらも時系列データであるという共通点があるので，音声認識の技術を動画像認識に応用することもできます．さらに，テレビ番組など多くの映像資源では動画像と音声が一緒に与えられるので，将来的には両者を統合した研究が重要になると考えています．

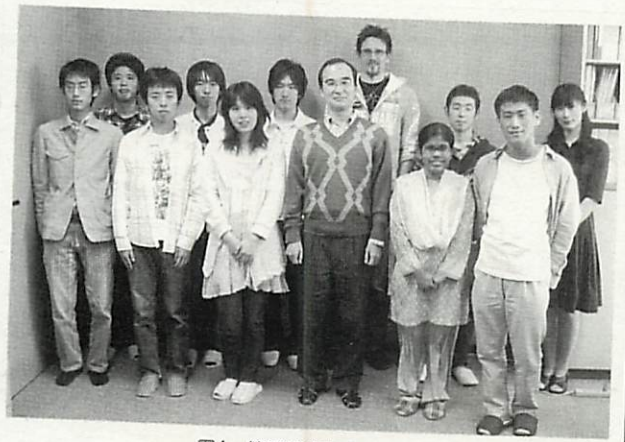

図1 篠田研究室のメンバ

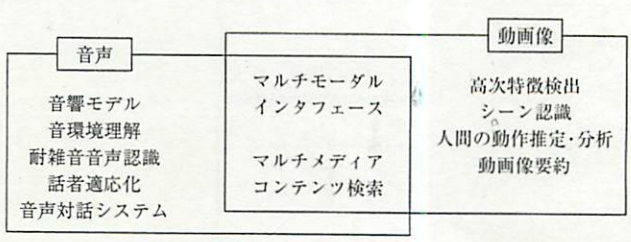

図2 篠田研究室の研究分野

映像情報メディア学会誌 Vol. 63, No. 8, pp. 1116～1119 (2009)

4. 研 究 紹 介

本章では，篠田研究室における動画像認識の研究と音声認識の研究から，最近のものを幾つか紹介します．

4.1 動画像に対する高次特徴検出

現在，テレビや録画機器のディジタル化とネットワークの高速化に伴い，世界中で映像資源の増加が著しいものとなっています．また，動画像の効率的な検索のために，動画像認識の技術が求められており，映像解析の研究がますます重要になってきています．ここでは，TRECVIDという国際的なワークショップで設けられたタスクの一つである，テレビ番組に対する高次特徴検出の研究を紹介します．

高次特徴とは，「飛行機」や「船」などと言った物体に加え，「町並み」や「夜」など，シーンを表す語を含めたものを指します（**図3**）．この研究の目的は，テレビ番組から，特定の高次特徴が映っているショットを検出することです．

本研究のシステムの概略を**図4**に示します．本手法では，まず，動画像のショットを代表する一つのフレーム画像（キーフレーム画像）を選び，そこからHarris-Affine, Hessian-Affineの二つ方法で局所領域の抽出を行います．次に，各局所領域からSIFT（Scale Invariant Feature Transform）特徴ベクトルを抽出します．さらに，学習データ全体のSIFT特徴ベクトルに対して，木構造クラスタリングを行うことで木構造辞書を作成し，それを基にSIFT特徴ベクトルを量子化することでVisual Wordという特徴量を作成します．

ここで，Visual Wordのみでも認識を行うことができますが，本手法では動画像の特性を生かすために，物体の「動き」に関する特徴も抽出します．具体的には，キーフレームの前後のフレームで，各局所領域に動きがあるかないかを判定し，Motion Wordと呼ばれる特徴量を作成します．最後に，Visual WordとMotion Wordを結合し，最大エントロピーモデルを用いて学習・認識を行います．さらに，現在では，領域情報を用いた認識方法や，キーフレーム以外の情報を有効的に活用する方法を検討しています．

図3 高次特徴の例

4.2 動画像要約の研究

ニュース番組やスポーツ番組のダイジェスト作成を始め，動画像の要約が必要な場面が数多くあります．本研究では，動画像の編集作業を支援することを目的とした，動画像要約を行っています．

研究の対象となる動画像は，テレビ局にある素材ビデオやホームビデオといった未編集の動画像です．未編集の動画像には，NGなどによる取り直し

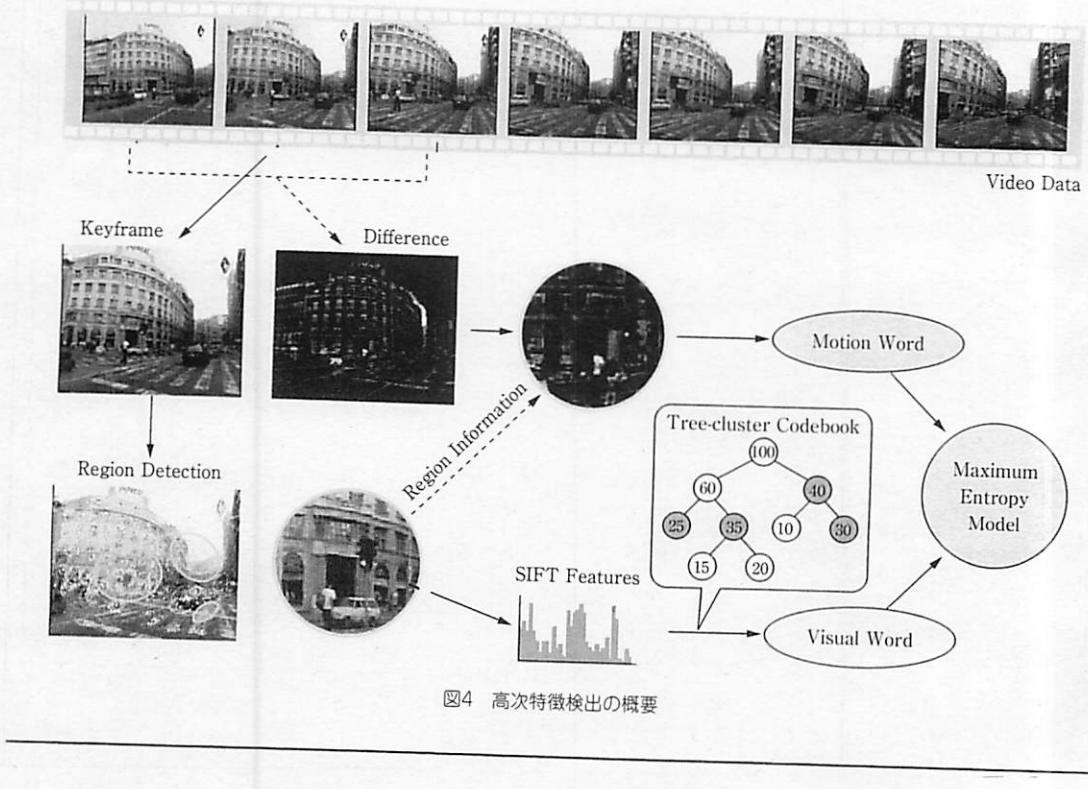

図4 高次特徴検出の概要

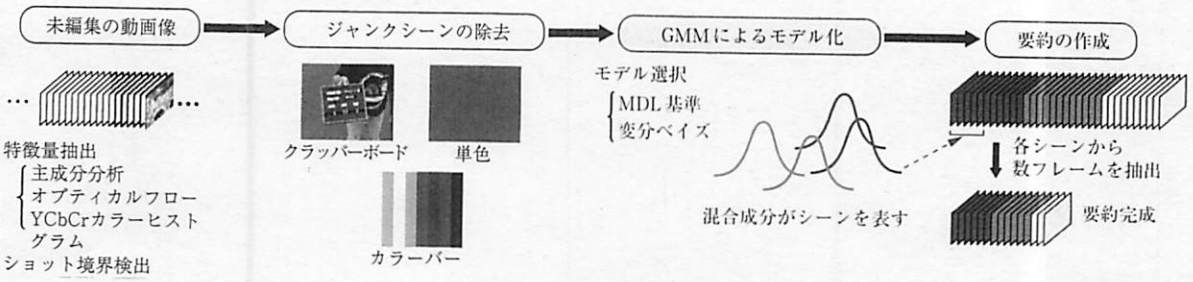

図5　動画像要約の流れ

が行われたリテイクシーンや，カラーバーなど意味のないジャンクシーンが含まれており，これらを除去することも必要となります．また，本研究では，動画像で「人が車に乗り込んだ」「車が走り去った」といった，人間にとって意味のある部分をシーンと呼び，動画像中のシーンに着目して要約を作成します．

システムの概略は**図5**のようで，ここでは動画像の大局的な特徴量を利用します．具体的には，主成分分析とYCbCrカラーヒストグラム，オプティカルフローから作成した特徴量を利用しています．また，初めにカメラの切替わり（ショット境界）を検出し，明らかなジャンクシーンも除去しておきます．

次に，動画像の各ショットにシーンが幾つ含まれているかを推定します．シーンの推定においては，ショットを混合ガウス分布（Gaussian Mixture Model; GMM）でモデル化し，GMMの各混合成分が一つのシーンを表現すると仮定します．そして，記述長最小（Minimum Description Length; MDL）基準または変分ベイズ法を用いて，モデル選択を行うことで，ショット中のシーンのクラスタリングを行うとともに，混合数（シーン数に一致）を推定します．

最後に，クラスタリングされた各シーンからそれを代表するフレームとその前後のフレームから要約を作成します．

4.3　野球動画の自動インデキシングの研究

野球の動画像に対して，見たいシーンの検索やハイライトの作成を行う際，「ホームラン」「ファール」「三振」などと言ったシーンのインデキシングが行われていると非常に便利です．また，インデキシングを人手で行うには多大なコストを要するため，自動的にインデキシングを行うことが求められています．ここでは，野球動画に対する自動インデキシングの研究を紹介します．

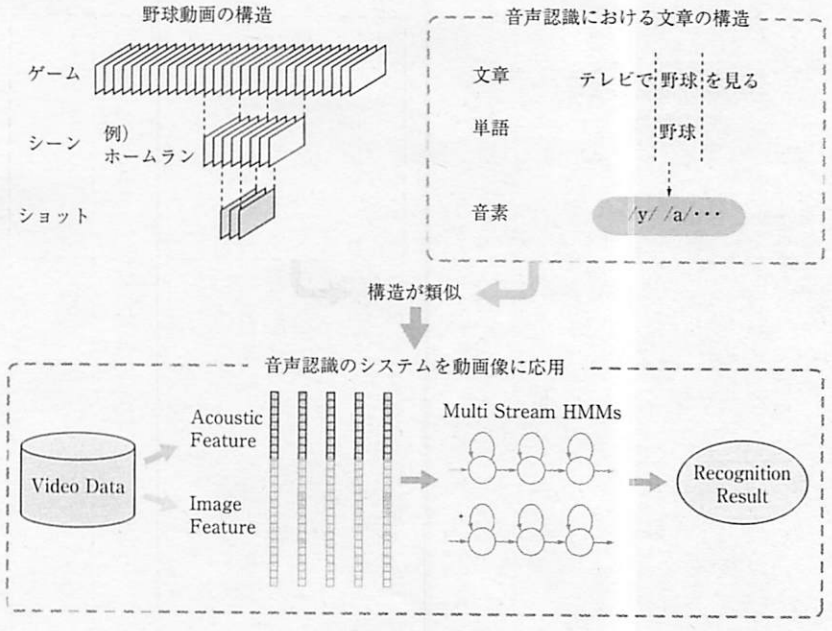

図6　野球動画と音声認識システム

この研究の大きな特徴は，音声認識の手法を動画像に応用していることです．音声と動画像の接点を見るために，まず，野球動画の構造を考えてみましょう．

図6に野球動画の構造を示します．ここで，野球の1ゲームは複数のシーンから成り，シーンは幾つかのショット（カメラの切替わりがない部分）から構成されています．一方，音声認識において，文章は複数の単語から成り，単語は幾つかの音素から構成されています．今，ゲームと文章，シーンと単語，ショットと音素をそれぞれ対応させて考えると，これらの構造が類似していることに気づきます．この類似性を基にして，音声認識の手法を動画像に応用することができるようになります．

具体的な手法としては，まず，野球動画の各フレーム画像から画像特徴量を抽出します．特徴量には，フレーム画像と差分画像に対する主成分分析特徴量，オプティカルフローによるカメラワーク特徴量を用いています．さらに，音のデータから，音響特徴量を抽出し，先の画像特徴量と結合します．

次に，結合して得られた特徴量を用いて，マルチストリームHMM（Hidden Markov Model）による学習を行います．ここで，マルチストリームHMMとは，ストリーム（ここでは音声と動画像）ご

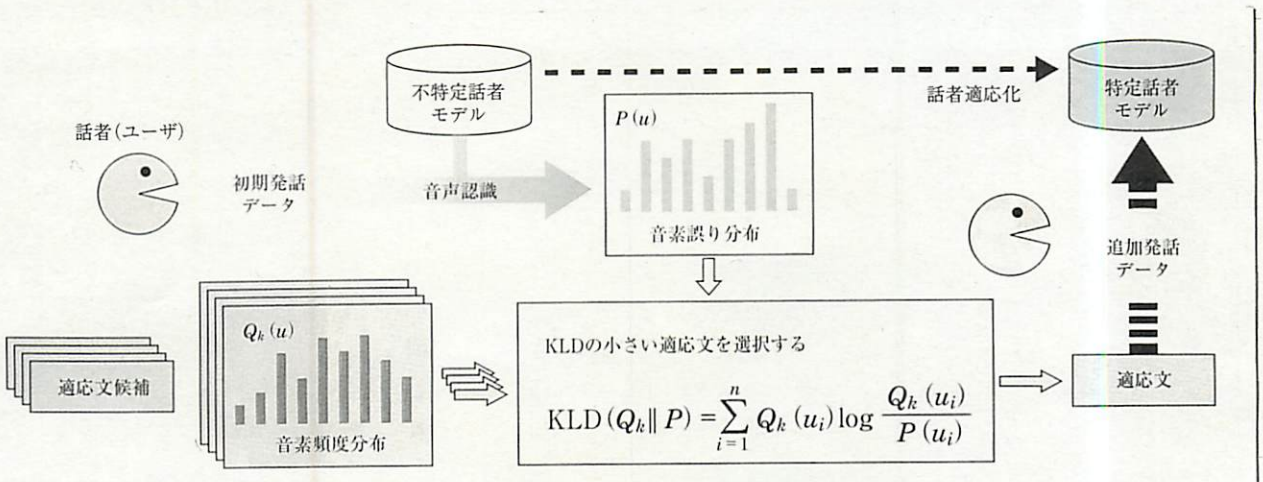

図7 能動的な適応文選択に基づく話者適応化

とに重み付けが可能なHMMのことです．この方法で得点シーンの認識を行った結果，F値が67.1%となりました．

さらに，音声認識における話者適応化の技術を応用した，ゲーム適応化やSVMを組合せたショット検出方法によって，F値が72.0%にまで向上しています．また，この研究では野球動画を対象としていましたが，今後はより一般的な動画像の構造を分析し，時系列の特徴を生かした動画像認識を行いたいと考えています．

4.4 音声認識における話者適応化の研究

篠田研究室では，音声に関する研究にも取組んでいます．ここでは，音声認識における話者適応化の研究について紹介します．

音声認識システムでは，不特定多数の話者から発話データを集め，そこから音声認識のモデル（不特定話者モデル）を学習するのが一般的です．しかし，音声は話者ごとに個性があり，特定のユーザが音声認識システムを利用する際には，その人の発話データのみから学習を行った方が，認識性能が高くなります．ただし，実際問題として，特定話者の発話データを大量に集めることは困難です．そこで，不特定話者モデルと特定話者の少量の発話データから，特定話者に適応したモデルを作る，話者適応化の研究が盛んに行われています．

特定話者の発話データは，ユーザに文章（適応文）を読み上げてもらうことで得られますが，どのような文章を読み上げてもらうかが問題となります．また，ユーザへの負担を軽減するためにも，適応の効果が大きい少量の適応文を選ぶべきです．さらに，話者毎に話し方や発声の仕方が異なるため，適切な適応文も話者毎に異なると考えられます．そこで，本研究では，話者の少量の発話から性能向上に役立つ情報を能動的に引き出して，適応文を選択することで，より効率的に話者適応を行う手法を提案しています．

提案手法（**図7**）では，最初に短い文章を読み上げてもらい，その認識結果から，どの音素の誤りが多かったかを表す音素誤り分布を求めます．ここで，次に読み上げてもらう文章は，認識誤りが多かった音素を多く含むものを選びます．より正確には，適応文の候補の音素頻度分布（音素の出現数の分布）から音素誤り分布へのカルバック・ライブラー情報量を用いて適応文を選択しています．これにより，ユーザが苦手とする音素を含むデータをより多く得ることができるため，話者適応の効果が大きくなると考えられます．最後に，選ばれた適応文を読み上げてもらい，MLLR法により話者適応を行います．

この手法により，従来よりもユーザが読み上げる文の量が少なくなり，効果的な話者適応が可能となりました．

5. むすび

本稿では，東京工業大学篠田研究室について紹介させていただきました．篠田研究室では，今日もより高精度な動画像・音声認識を目指して，研究を行っています．この研究紹介を通じて，動画像認識はもちろん，音声認識の研究にも興味を持っていただければ幸いです．

また，ホームページでも研究紹介をしている他，東工大の文化祭（工大祭）では，研究室公開も実施していますので，興味のある方は是非足を運んでみてください．

(2009年4月30日受付)

〔文 献〕

1) S. Hao, Y. Yoshizawa, K. Yamasaki, K. Shinoda, and S. Furui: "Tokyo Tech at TRECVID 2008", TRECVID 2008 workshop (2008-11).
2) K. Shinoda, K. Ishihara, S. Furui and T. Mochizuki: "Automatic Score Scene Detection for Baseball Video", LKR2008, pp.226-240 (2008-3).
3) 村上博子，篠田浩一，古井貞熙："能動的な適応文選択に基づく話者適応化"，日本音響学会2009年春季研究発表会講演論文集，pp.191-194 (Mar. 2009)

井上 中順（いのうえ なかまさ）
2009年，東京工業大学工学部情報工学科卒業．同年，同大学院情報理工学研究科計算工学専攻修士課程に進学．動画像認識，高次特徴検出に関する研究に従事．