様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　９　日現在

機関番号：　１４５０１
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2016
課題番号：　２６２８００４０

研究課題名（和文）社会規模での大規模コーパス収集による映像検索エンジンの再構築

研究課題名（英文）Development of video retrieval engine by using a large-scale video corpus

研究代表者
上原　邦昭（Uehara, Kuniaki）

神戸大学・システム情報学研究科・教授

研究者番号：６０１６０２０６

交付決定額（研究期間全体）：（直接経費）　　13,700,000円

研究成果の概要（和文）：本研究では，大規模映像データから，様々な概念（物体，動作，シーンなど知覚可能な意味内容の総称）の認識結果を組み合わせて，高次のクエリ（例えば「屋外でギターを弾いている」）に適合する映像を検索する手法を開発した．近年，概念認識精度は大幅に向上したが，あらゆる概念を高精度に認識することは未だに困難である．そこで，Dempster-Shafer Theoryに基づいて，概念認識の不確実性を定量化し，不確実な認識結果からでも高精度な検索が行える確率的手法を開発した．さらに，人間の視覚的注意のモデルを導入して，映像中で注視されている概念を判別しながら，ユーザの意図に即した検索を行う手法を開発した．

研究成果の概要（英文）：We have developed a large-scale video retrieval method that can identify videos relevant to a semantically high-level query (e.g., “playing a guitar outdoors), by combing recognition results of concepts (abstracted names of meanings that a human can perceive from a video). Despite the recent advancement of concept detection, it is still difficult to accurately recognise various concepts. Thus, based on Dempster-Shafer Theory, we have quantised uncertainties in concept detection, and developed a probabilistic method that can perform accurate retrieval even using uncertain (erroneous) concept detection results. In addition, by adopting a visual attention model, we have developed a video retrieval method that reflects user’s intention by considering which concepts attract his/her attention.

研究分野：　人工知能、特に機械学習

キーワード：　機械学習　情報検索　映像データ　コーパス　映像検索

１．研究開始当初の背景

大量の Web 映像から，高次のクエリ（「屋外で人がギターを弾いている」，「日中，路上で人がポスターを持っている」など）に適合する映像を検索するためには，様々な概念の認識結果を組み合わせる必要がある．概念とは，物体，動作，シーンなど，映像から人が知覚する様々な意味内容の総称である．ここで，各概念の認識モデルは，機械学習の枠組みで，その概念の出現もしくは非出現を表すラベルが付与された映像集合（コーパス）を解析して構築される．つまり，概念が出現している映像に特有の視覚的特徴を捉えたモデルが構築される．

この枠組みの中で，概念の形状や向き，カメラ位置，照明条件といった変動要因に対して頑健な認識を行うためには，各変動要因に対応した映像を含んだ大規模コーパスが必要になってくる．言い換えると，概念認識モデルは，コーパス中の映像と見た目が類似した映像に対しては正確な認識が行えるが，そうでない映像に対しては挙動が不確定である．近年，大規模コーパスの整備により，概念認識精度は大幅に向上したが，インターネットは Open World であり，考えうる全ての映像をコーパスで吸収し切ることは実質不可能であるため，100%の認識精度は望めない．そのため，“ 不確実な（エラーを含む）” 概念認識結果に頼って検索を行うと，多数の誤検索や検索漏れが生じる．そこで，本研究では，概念認識モデルの不確実性を考慮しながら，映像がクエリに適合するか否か確率的に推論する映像検索エンジンを開発する．

また，近年，上記のような概念認識結果を用いた映像検索の有効性が多数報告されているが，概念認識と映像検索の本質的な違いが考慮されていない．具体的には，概念認識では，変動要因に頑健なモデルの構築に重点が置かれているが，このようなモデルは検索には適切ではない．例えば，「犬が映っている映像」を検索する際，「背景領域に小さく後ろ向きに映っている犬」が検索されても，ユーザにとって有用ではない．むしろ，興味の対象である領域（注目領域）に映る犬を検索すべきである．つまり，検索プロセスに人間の知覚機構を導入しなければ，ユーザにとって有意義な検索が行えない．このような着想の元，人間の視覚機構（顕著性に基づく視覚的注意）に基づいて映像から注目領域の抽出し，概念認識結果の意義を考慮した検索手法を開発する．

２．研究の目的

本研究では，Dempster-Shafer Theory (DST)という，無知量を表現するための確率論の一般型を用いて，概念認識の不確実性を表現する．従来の確率論では，概念が出現しているかどうか分からない無知状態は，2 つの確率変数 A(ppearance)，D(isappearance) に確率 0.5 を割り当てて表現することしかできなかった．これに対して，DST を用いれば，無知状態を表す{A,D}という「確率変数の部分集合」にも確率を割り当てて，物体認識の不確実性を詳細に表現できるようになる．

上記を踏まえて，以下の 3 つの研究目的と設定した．

(1) 不確実性の定量化：各概念の認識モデルに対して，無知状態に対する確率，すなわち不確実性を算出する．具体的には，認識モデルの評価値に対して，物体が出現している映像と出現していない映像の“ 密度比 ” を推定し，不確実性を定量化する．なお，この不確実性を用いれば，認識モデルが致命的に誤っている可能性のある映像を特定できる．そして，その真偽を，別で開発済みの映像アノテーションゲームを介してユーザに判定してもらえば，概念認識モデルの性能を効率的に向上させることが可能なコーパスを構築できる．

(2) 不確実性を考慮した映像検索：クエリに関連する概念を選択し，(1)で定量化した不確実性を導入した“ 最尤推定 ” を行って，不確実な概念認識の結果から，クエリに適合する映像を検索する確率的手法を開発する．

(3) 映像からの注目領域の検出：選択的注意という脳のメカニズムに基づいて，ユーザが注視している領域（注目領域）を検出する技術を Focus of Attention (FoA)と呼ぶ．ここで，ユーザが概念の出現を注釈付けするということは，注目領域に概念が出現していると仮定できる．これに基づいて，弱教師つき学習という手法を用いて，概念の出現が注釈された映像とそうでない映像から，注目領域を特定する．そして，注目領域内外の視覚的特徴を比較して，未知の映像中の注目領域を検出する FoA 手法を開発する．

３．研究の方法

上記の 3 つの研究テーマに関して，下記のように取り組んだ．

(1) 不確実性の定量化：図 1 を用いて，概念認識の不確実性を定量化する手法について概説する．図 1 では，Crowd という概念を対象として認識モデルを構築し，それを 1000 本の映像に適用して得られる評価値の分布を示している．横軸が評価値，＋と×印がそれぞれ，実際に Crowd が映っている映像とをそうでない映像を表している．ここで，見やすくするために，左側の縦軸に従って，各映像を ID ごとに縦方向に散らしてプロットしている．理想は，Crowd が映っている映像（＋印）と映っていない映像（×印）の評価値が 2 分割されることであるが，実際には，図 1 から分かるように，これらの映像の評価値が混在している．この「混在の度合い」が，概念認識の不確実性である．

DST を用いる上での最大の難点として，無知状態（本研究の場合，「いくつかの概念のうちの 1 つが出現している」という事象）を客観的にモデル化することが困難であると

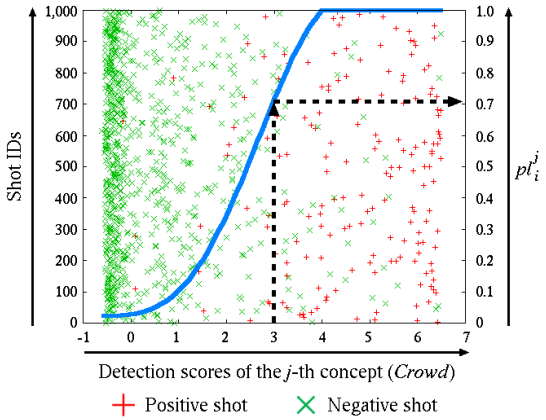

図 1：概念識の不確実性の定量化手法の概要（Crowd を対象）

いう点が挙げられる．最もシンプルには，無知状態に関する確率を求めるために，ユーザに無知状態であるかどうか注釈付けさせた映像を用いることが考えられるが，注釈付けが非常に主観的になってしまう．そこで，映像 x における概念 c の出現に関する不確実性を，c を含む無知状態の和集合で表される“出現可能性 $pl_c(x)$（概念が出現しているかも知れない確率）”としてモデル化する（無知状態を用いることと出現可能性を用いることは，数式的に同値であると証明できる）．

具体的には，x における c の認識モデルによる評価値を $S_c(x)$としたとき，$pl_c(x)$を以下のような“確率密度比”として定義する．

$$pl_c(x) = \frac{P_a(S_c(x))}{P_d(S_c(x))}$$

ここで，$P_a(S_c(x))$は，$S_c(x)$周辺に c が出現していると注釈付けされた映像が分布している確率密度，$P_d(S_c(x))$は，そうでない映像が分布している確率密度である．例えば，図 1 の青線は，上記のような確率密度比としてモデル化された Crowd の出現可能性（右側の縦軸が値）を表している．これによると，中間的な評価値 3.0 でも，周辺に Crowd が出現している映像が多数分布しているため，出現可能性としては 0.7 となっている．このように，認識モデルの評価値が小さくても，概念が出現している映像がその周辺に分布していれば，出現可能性が大きくなる．その結果，認識モデルにより誤って概念が出現していないと判定されてしまった映像をリカバリーでき，検索漏れを抑えることができる．

(2) 不確実性を考慮した映像検索：あるクエリに関して，N 個のサンプル映像が与えられたとき，未知の映像がクエリに適合するか否か判定する検索モデルを構築する．上記の出現可能性を用いると，検索漏れが抑えられる一方で，検索されるべきでない映像までも誤検索されてしまう．そこで，下記の線形結合モデルを用いて，映像 x に対して，各概念 c ごとに，出現可能性 pl_c とクエリに対する関連度 θ_c を掛け合わせて，x のクエリに対する適合性 $P(x)$を算出する．

$$P(x) = \sum_{c \in C} \theta_c pl_c(x)$$

ここで，θ_c は，最尤推定法を用いて，N 個のサンプル映像に対する適合性が最大化されるように推定される．最終的に，未知の映像のクエリに対する適合性を計算し，それらが高い順に映像をソートすれば検索が行えたことになる．

(3) 映像からの注目領域の検出：認知心理学の知見に基づいて，“ボトムアップ”と“トップダウン”という 2 つの処理を組み合わせて，映像中の注目領域を推定する FoA 手法を開発した．ボトムアップ処理は，外発的成分による網膜上の視覚刺激によって引き起こされ，色，エッジといった特徴量に関して周囲と異なる領域が注目領域と判定される．しかしながら，特徴量という低次の物理量だけからでは，人間の知覚に関する高次の注目領域を高精度に検出することは困難である．そこで，人間の意図，知識といった内発的成分によって引き起こされるトップダウン処理を用いて，ボトムアップ処理による注目領域を洗練させる．具体的には，文脈手がかりという，「空間的レイアウトが類似している映像では，同じような領域が注目領域と見なされやすい」という知見に基づいて，事前に注目領域がラベル付けされている映像との類似性から，対象映像の注目領域を修正する．

特に，開発手法では，弱教師付き学習を用いて，注目領域のラベル付けに係る人的コストを削減している．弱教師あり学習とは，大まかにラベル付けされたデータから，詳細なラベルを出力するためのモデルを構築する機械学習手法である．これを用いて，概念名のみが注釈付けされた映像から，トップダウン処理における注目領域（概念が映っている領域）を自動推定している．

(4) 深層学習を用いた概念認識モデルの高精度化：研究開始当初に設定した上記の 3 つのテーマに関して，初めの 2 年度内に一定の研究成果を得たため，最終年度は深層学習を用いた概念認識の高精度化に取り組んだ．なお，このテーマは，研究申請書の「多方面からの検討事項」に記載していたものである．

本研究では，畳み込みとプーリングという 2 種類の層を交互に積み重ねた畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）に焦点を当てた．これまでに，CNN を用いた高精度な概念認識手法が提案されているが，映像検索に必要な概念の全てがカバーされている訳ではない．言い換えると，多様なクエリに対応するためには，これまでは対象となっていなかった概念も認識できなければならない．そこで，転移学

習の枠組みで，学習済みの CNN を再利用して，検索に有用な概念を認識する手法を開発した．概要としては，学習済みの CNN の中間層の出力を入力とする小規模ニューラルネットワーク（microNN: micro Neural Network）を構築する．特に，できる限り多くの学習データを考慮するために，microNN を画像ドメイン，映像ドメイン 1（フレーム間の時間関係を考慮しない），映像ドメイン 2（長短記憶モデル（LSTM: Long-Short Term Memory）を用いて時間関係を考慮する）と段階的に転移させていく手法を開発した．

４．研究成果

まず，不確実性の定量化，及びそれを用いた映像検索の有効性を検証するために，米国標準技術局（NIST）主催の国際競争型ワークショップ TRECVID 2009 のサーチタスクで提供されている大規模ベンチマークデータを用いた．タスクとしては，24 個のクエリ（例えば，「4 階以上の高い建物が映っている」，「人，テーブル，コンピュータが映っている」など）のそれぞれに関して，291 本の学習用映像に含まれる 36,106 ショットを用いて検索モデルを構築し，619 本のテスト用映像に含まれる 97,150 ショットの中からクエリに適合するショットを検索する．検索結果の評価尺度として，クエリに関する適合度高い上位 1,000 ショット中に含まれる正解ショットの割合（精度）を用いる．この尺度を用いて，概念認識モデルの評価値を直接用いて検索を行った場合と，評価値を出現可能性に変換して検索を行った場合を比較すれば，不確実性を考慮することにより，どれだけの正解ショットがリカバリーできたか検証できる．結果として，24 個のクエリに関する平均の精度に関して，前者が 9.4%，後者が 11.2%となり，不確実性を考慮することの有効性を実証した．

図 2 に，開発した不確実性に基づく検索手法と TRECVID 2009 で開発された 88 手法との比較結果を示す．クエリごとに，四角とひし形がそれぞれ 88 手法による精度の中央値，最大値を表しており，開発手法の精度は三角形で表されている．特に，丸で囲まれた 3 つのクエリでは，開発手法による精度が 88 手法中のどれよりも高く，点線の丸で囲まれた 4 つのクエリでは，88 手法の精度と比較して，開発手法による精度が上位 5 位以内に含まれている．最終的に，24 個のクエリに関する平均精度では，開発手法は，88 手法中の 18 位にランク付けされる．ここで，比較手法は，概念認識モデルの評価値に加えて，テキスト解析や音声・視覚特徴を組み合わせた検索を行っている．これに対して，開発手法は，概念認識モデルの評価値のみからでも，上述の 7 つのクエリにおいてそれらを上回る検索性能を達しており，その有効性が分かる．

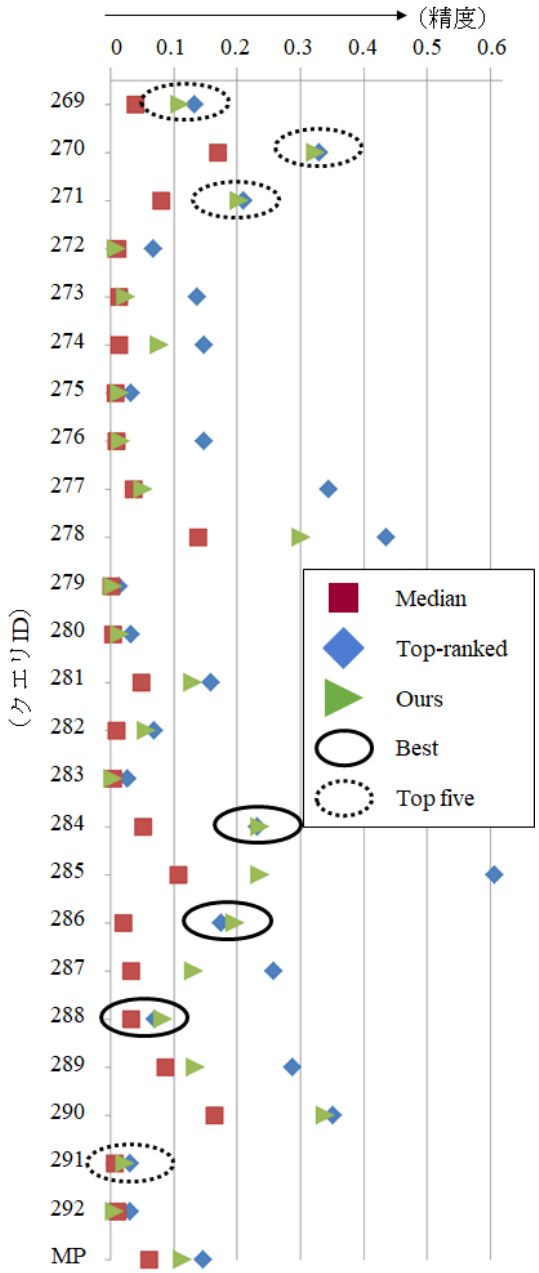

図 2：TRECVID 2009 のサーチタスクで開発された手法と開発手法との比較

次に，映像から注目領域を検出することの有効性を検証するために，TRECVID 2011 の意味インデキシングタスクで提供されたベンチマークデータを用いた．ここでは，図 3 に示す 23 個の概念のそれぞれに関して，11,485 本の学習用映像に含まれる 240,918 ショットを用いて認識モデルを構築し，その認識精度（正確には，平均精度（AP: Average Precision））を 8,215 本のテスト用映像に含まれる 125,880 ショットを用いて評価する．実験目的は，注目領域を考慮した場合としない場合を比較して，人間の知覚機構を導入することの有効性を検証することである．

図 3 の青棒は，注目領域を考慮せずに概念認識を行った場合の精度，赤棒は，注目領域中の視覚特徴を重み付けして他領域の特徴

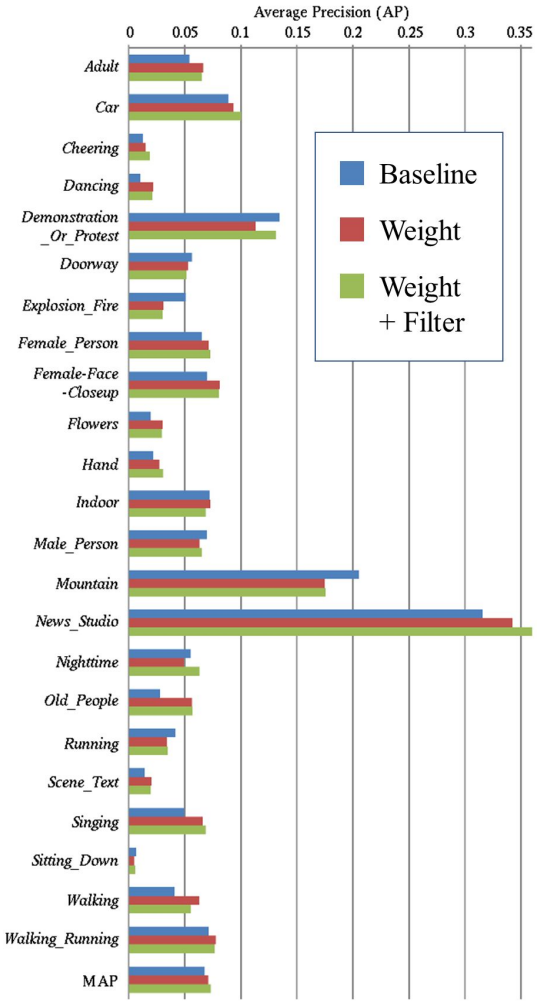

図 3：注目領域検出の有効性の検証結果

より重要視した場合の精度を表している．さらに，緑棒は，赤棒の重み付けに加えて，概念が注目領域に映ってそうにないショットをあらかじめフィルタリングした場合の精度を表している．図 3 の最下部は，23 個の概念に対する平均の精度を表しており，青，赤，緑の順に精度が向上しており，注目領域を考慮することの有効性が実証されている．

最後に，TRECVID 2016 のアドホックサーチタスクで提供されたベンチマークデータを用いて，深層学習を用いて高精度化した概念認識結果に基づく映像検索の有効性を検証する．ここでのタスクは，30 個のクエリ（例えば，「屋外でギターを弾いている」，「本棚を背景にして人がカメラに向かって話している」など）のそれぞれに関して，27,963 本の映像に含まれる 545,872 ショットを用いて検索モデルを構築し，4,593 本の映像に含まれる 335,944 ショットの中からクエリに適合するショットを検索することである．検索性能は，クエリに関する適合度の高い上位 1,000 ショットに対する精度（正確には，平均精度）を用いて評価される．

図 4 に，TRECVID 2016 のアドホックサーチタスク（Manually-assisted）で開発された全 22 手法の検索性能のランキングを示す．各棒は，1 つの検索手法の精度を表しており，黄色の棒で示す 3 手法が本研究で開発したものである（時間関係の考慮の有無等に関して 3 つの異なるバージョンがある）．特に，開発手法を含む図 4 の 22 手法は，NIST によって公式に評価されている．図 4 から，開発手法は全 22 手法中第 5 位，上位 4 手法は同一の研究チームによって開発されているため，チームとして全 8 チーム中第 2 位の検索精度を達成している．つまり，本研究で開発した検索手法は，世界トップクラスの検索性能を達成していると言える．

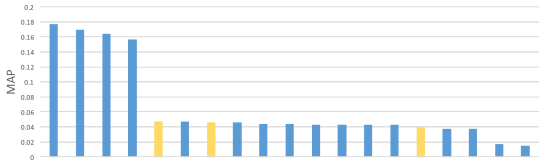

図 4：TRECVID 2016 アドホック検索タスク（Manually-assisted）で開発された検索手法と開発手法との比較

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① Kimiaki Shirahama, Tadashi Matsumura, Marcin Grzegorzek and Kuniaki Uehara: Semantic Indexing based on Focus of Attention Extended by Weakly Supervised Learning, International Journal on Advances in Software, 査読有, Vol. 8, No. 3-4, pp. 410-419, 2015
https://www.thinkmind.org/index.php?view=article&articleid=soft_v8_n34_2015_10

② Kimiaki Shirahama, Marcin Grzegorzek and Kuniaki Uehara: Weakly Supervised Detection of Video Events Using Hidden Conditional Random Fields, 査読有, International Journal of Multimedia Information Retrieval, Vol. 4, No. 1, pp. 17-32, 2015
DOI: 10.1007/s13735-014-0068-6

③ Kimiaki Shirahama, Yuta Matsuoka and Kuniaki Uehara: Hybrid Negative Example Selection Using Visual and Conceptual Features, Multimedia Tools and Applications, 査読有, Vol. 71, No. 3, pp. 967-989, 2014
DOI: 10.1007/s11042-011-0886-y

〔学会発表〕（計 7 件）

① 松本泰幸，篠崎隆志，白浜公章，上原邦昭，Curriculum Learning を用いたネ

ットワーク群による効率的な大規模動画像検索，情報処理学会コンピュータビジョンとイメージメディア（CVIM）研究会，2017.3.9～2017.3.10，国立情報学研究所（東京都）

② Yasuyuki Matsumoto, Takashi Shinozaki, Kimiaki Shirahama, Marcin Grzegorzek and Kuniaki Uehara, Kobe University, NICT and University of Siegen at TRECVID 2016 AVS Task, TREC Video Retrieval Evaluation (TRECVID) 2016 Workshop, 2016.11.14 ～ 2016.11.16, Gaithersburg (USA)

③ Zeyd Boukhers, Yicon Wang, Kimiaki Shirahama, Kuniaki Uehara and Marcin Grzegorzek, Convoy Detection in Crowded Surveillance Videos, The Seventh International Workshop on Human Behavior Understanding (HBU 2016), 2016.10.16, Amsterdam, (The Netherlands)

④ Kimiaki Shirahama, Takashi Shinozaki, Yasuyuki Matsumoto, Marcin Grzegorzek and Kuniaki Uehara, University of Siegen, Kobe University and NICT at TRECVID 2015 SIN and MED Tasks, TREC Video Retrieval Evaluation (TRECVID) 2015 Workshop, 2015.11.16 ～ 2015.11.18, Gaithersburg (US)

⑤ 松本泰幸，篠崎隆志，上原邦昭，Deep Learning の中間層学習表現を利用した動画像の意味解析（2C1-OS-06a-3），第 29 回人工知能学会全国大会（JSAI2015），2015.5.31，公立はこだて未来大学（北海道）

⑥ Kimiaki Shirahama, Tadashi Matsumura, Marcin Grzegorzek and Kuniaki Uehara, Empowering Semantic Indexing with Focus of Attention, The Seventh International Conferences on Advances in Multimedia (MMEDIA 2015), 2015.4.19～2015.4.24, Barcelona (Spain)

⑦ Kimiaki Shirahama, Marcin Grzegorzek and Kuniaki Uehara, Multimedia Event Detection Using Hidden Conditional Random Fields, The Forth ACM International Conference on Multimedia Retrieval (ICMR 2014), 2014.4.1～2014.4.4, Glasgow (Scotland)

〔図書〕（計 1 件）

① Kimiaki Shirahama, Kenji Kumabuchi, Marcin Grzegorzek and Kuniaki Uehara, Springer International Publishing, Multimedia Data Mining and Analytics: Disruptive Innovation (Aaron K. Baughman, Jiang Gao, Jia-Yu Pan and Valery Petrushin eds.), 2015, 454, (269-294 (Chapter 12): Video Retrieval Based on Uncertain Concept Detection Using Dempster-Shafer Theory)

〔その他〕

ホームページ等

① TRECVID 2016 アドホック検索の結果概要：http://www-nlpir.nist.gov/projects/tvpubs/tv16.slides/tv16.avs.slides.pdf
13 枚目のスライドに，我々の手法が，manually-assisted カテゴリにおいて、参加 8 チーム中第 2 位（全 22 手法中第 5 位）の検索精度を達成していることが示されている．

② TRECVID 2016 アドホック検索部門（manually-assisted カテゴリ）における開発手法に関する論文、口頭発表資料、ポスター資料は、下記の NIST のサイトで公開されている。
論文：http://www-nlpir.nist.gov/projects/tvpubs/tv16.papers/kobe_nict_siegen.pdf
発表資料：http://www-nlpir.nist.gov/projects/tvpubs/tv16.slides/tv16.avs.kobe_nict_siegen.slides.pdf
ポスター：http://www-nlpir.nist.gov/projects/tvpubs/tv16.slides/tv16.avs.kobe_nict_siegen.poster.pdf

③ MMEDIA 2015 で発表した FoA 手法に関する論文が，会議の優秀論文に選ばれたことが，下記のページに記載されている．
http://www.iaria.org/conferences2015/AwardsMMEDIA15.html

6．研究組織

(1)研究代表者
上原 邦昭 （UEHARA, Kuniaki）
神戸大学・大学院システム情報学研究科・教授
研究者番号：60160206

(2)研究分担者
松原 崇 （MATSUBARA, Takashi）
神戸大学・大学院システム情報学研究科・助教
研究者番号：70756197

(3)研究協力者
白浜 公章 （SHIRAHAMA, Kimiaki）
ドイツ・ジーゲン大学・パターン認識グループ・ポスドク
研究者番号：なし