科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１８　日現在

機関番号：　１２６１２
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５００４４

研究課題名（和文）Ｗｅｂマルチメディアマイニングによる動詞概念と名詞概念およびその関係の自動学習

研究課題名（英文）Web Multimedia Mining for Learning Verb and Noun Concepts and Their Relations

研究代表者
柳井　啓司（Yanai, Keiji）

電気通信大学・情報理工学（系）研究科・准教授

研究者番号：２０３０１１７９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円、（間接経費）　　870,000円

研究成果の概要（和文）：本研究では，3年間の研究によって，次の成果が得られた．(1) キーワードを入れるだけで，YouTubeなどのWeb上の動画共有サイトから動画を収集し，その中から自動的にキーワードに相応しい動画ショットを選択するシステムの実現．(2) Web上の静止画像を利用し，(1)の精度を向上させる方法．(3) (1)の精度を向上させるために，キーワードと動画ショットの関係をさらに効率的に利用するためのVisualTextualRank法の提案．(4) 人間動作の手の動きに注目し，それを時空間特徴量として表現することによって，人間の動作から物体を推定，およびインタラクションを分析する手法の提案．

研究成果の概要（英文）：After the three-year research activities on this research grant, we have obtained the following results: (1) A new method to collect video shots corresponding to the given keywords from the Web video sharing site such as YouTube. (2) A method to improve (1) by using results of Web image search engines additionally. (3) A novel method, VisualTextualRank, for efficient use of relations between words and video shots. (4) A method to extract new spatio-temporal features by paying attention to human hands motion, which helps effectively recognize and analyze interactions between objects and human hands.

研究分野：　総合領域

科研費の分科・細目：　メディア情報学・データベース

キーワード：　動作認識　ビデオ認識　Webマルチメディアマイニング　Web動画　画像認識　一般物体認識

１．研究開始当初の背景

近年, Web 上には一般のユーザによってアップロードされた大量の画像や動画像が存在し，その多くには検索のための手がかりとなるように「タグ」とよばれるキーワードや説明文が付加されている．静止画像については，タグを利用してその内容を推定したり，画像認識のための物体認識モデルの自動学習を行ったりする研究が盛んに行われている．一方，限定された条件下で撮影された実験動画に対しては研究が行われているものの，タグ付きの Web 動画像からの動作認識モデルの自動学習はほとんど研究がなく，さらに物体認識と組み合わせた物体・動作の同時自動学習についてはまったく存在していない．

そうした状況に対して，我々は Web 動画認識に関する以下の研究において，新しい時空間特徴量抽出の手法，および動画の全フレームから動き特徴と静止画特徴を抽出しそれらを１つのベクトルにする Bag-of-Frames 表現を提案し，これら３種類の特徴を Multiple Kernel Learning で学習データに応じて最適な重みで統合して利用することによって，公開されている Web 動画分類ベンチマークセットにおいて世界最高の分類性能を達成している．

> Akitsugu Noguchi and Keiji Yanai: A SURF-based Spatio-Temporal Feature for Feature-fusion-based Action Recognition, Proc. of ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation, (2010/09).

さらに，我々は【科学研究費若手研究(A)・平成 20 年度～平成 22 年度・「1000 クラスに対応した大規模一般画像認識システムの実現」】において大量の Web 画像を用いた大規模一般物体認識のシステムをクラスタ計算機上に実装した経験があり，このシステムを元にファイルサーバを拡張することによって大量の Web 動画の分析に対応が可能である．

２．研究の目的

本研究では，動画の視覚特徴から抽出可能な「動詞」の概念と，動作対象もしくは動作主体の「名詞」概念，およびその関係を学習する確率モデルを提案し，それを Web 上の動画共有サイトに存在する大量のタグ付きの動画像に適用することによって，動詞概念と名詞概念の間の関係を分析し，認識に利用することを目的とする．例えば，「ラーメン」を「食べる」と「ハンバーガー」を「食べる」，「人間」が「歩く」と「ライオン」が「歩く」は「動詞」概念としては同じであるが，実際の動作としては異なるため，両者の関係を考慮することにより高精度な動作・物体認識が実現できることが期待される．こうした研究は，動詞概念のみの学習からさらに深い領域に踏み込んだ従来にない新しい研究であると言える．

３．研究の方法

本研究では，まず第一段階として初年度に動作と物体の同時生成確率モデルの研究を行ない，比較的小規模なデータセットで有効性を確認する．また，動作と動作対象もしくは動作主体が含まれる動画像を YouTube などの動画共有サイトより効率的に収集する方法についても検討する．具体的には，主に，動作と動作主体・動作対象の動詞概念，名詞概念の組み合わせを大量のタグ付き Web 動画から抽出する手法について研究する．

動詞に対応する Web 動画像の収集については，まず，動詞概念に対応する単語を 100 個程度準備する．次に， YouTube などの動画共有サイトからテキスト検索によって動詞に対応する動画のテキストタグ情報のみを大量に収集し，出現する頻度の高い名詞を 30 程度抽出する．次に，その名詞と元の動詞を組合せ，"eat+ramen"や"shoot+football"などをクエリでテキスト検索を行い動画を大量に収集する．ただし，動画共有サイトの動画に付与されているテキストタグは実際にその内容と対応していない場合があるので，実際にクエリと深く関連すると思われる動画像を見つけることが重要になってくる．本研究では，タグの共起性を考慮した手法でまずテキストタグのみで動画をフィルタリングして，ノイズ動画を除去する．最後に，動画のシーンが大きく変化した箇所を簡単な手法で検出し，ショットと呼ばれる数秒～数十秒の短い動画像集合に分割し，それぞれのショットから時空間特徴量，動き特徴，静的視覚特徴量を抽出する．なお，特徴量抽出時にはカメラモーションのフローを検出し補正処理を行い，複数の動作主体が検出された場合はそれぞれ別々の時空特徴，動き特徴量として抽出する．

次に抽出した各ショットの静的な画像特徴と動き特徴，時空間特徴を組み合わせて利用し，PLSA(Probabilistic Latent Semantic Analysis)もしくは LDA(Latent Dirichlet Allocation)を拡張した新しく提案する確率生成モデルを用いて，動作と動作主体，動作対象の自動分類を同時に実現する．実験では，動作毎に分けてモデル学習を行う場合と，様々な動作をすべてまとめて学習する２通りを実験する予定である．後者の場合，学習データ量が多くなるが，異なる動詞であっても，動作が同じである場合には同じカテゴリとして検出されるので，名詞概念と動詞概念の関係分析の結果としては新しい結果が得られる可能性がある．また，同時に，教師なしの画像のランキング手法である VisualRank 手法を改良し，動詞に対応する代

表的な名詞，もしくは各名詞に対応する代表的な動詞を抽出することも実験する予定である．

次に２年目以降は，大規模な Web 動画収集を行い，大規模実験を実施する予定である．年度の結果を踏まえて大規模な Web 動画収集を行う．さらに，前年度研究した手法を大規模 Web 動画に対して適用して，動詞概念と名詞概念の関係について，分析を行う．

Web 動画収集に当たっては，具体的には 1000 以上の動詞について，Web 動画を収集する予定である．1000 種類を人手で選ぶのは困難であるので，1000 の動詞についても，Web 動画に付いているタグの頻度ランキングなどから自動的に選択する予定である．

分析にあたっては，計算機クラスタを活用し，提案手法をスケーラビリティを考慮し並列分散処理可能な形でインプリメントを行って，大量の Web 動画像の処理を実施する．

最終年度の３年目では引き続き大規模実験を行い，シーンコンテキストや動画に含まれる音声情報の複合的な利用についても検討を予定している．前年度までは既に我々が提案している時空間特徴量，Bag-of-Frames による動き特徴と静的視覚特徴を統合して用いていたが，ここではさらにシーン認識を用いてその結果を統合して利用する予定である．シーン認識には bag-of-features と spatial pyramid およびサポートベクターマシンを併用した標準的な教師あり学習を利用し，学習データには 397 シーンからなる SUN Database を利用する予定である．シーンコンテキストも利用することで，動作の動詞概念，動作主体・動作対象の名詞概念，シーンのコンテキストの 3 種類の概念を複合した分析が可能となり，Web 動画からのマイニングとしてさらに深い領域に踏み込んだこれまでにない新しい研究が実施できる．

４．研究成果

本研究では，3 年間の研究によって，次の成果が得られた．

(1) キーワードを入れるだけで，YouTube などの Web 上の動画共有サイトから動画を収集し，その中から自動的にキーワードに相応しい動画ショットを選択するシステムの実現．実際に 100 種類以上のキーワードに関して動画ショットの収集事件を実施し，評価を行った．この成果は，主に雑誌論文[1]と学会発表[16]にまとめられている．

(2) Web 上の静止画像を利用し，(1)の精度を向上させる方法．主に人間の動作が写っている画像を Web から収集し，(1)の精度向上に利用した．この成果は，主に雑誌論文[1]と学会発表[13]にまとめられている．

(3) (1)の精度を向上させるために，キーワードと画像の関係をさらに効率的に利用するための VisualTextualRank 法の提案．(1)ではタグによる動画の選別と，画像特徴による選別が別々のステップになっていたが，本研究ではそれを同時に行う手法を提案し，精度向上を実現した．この成果は主に学会発表[5]の内容となっている．

(4) 人間動作の手の動きに注目し，それを時空間特徴量として表現することによって，人間の動作から物体を推定，およびインタラクションを分析する手法の提案．実験では 10 種類の楽器の分類を手の動きの特徴のみから実現した．どれも同士としては「楽器を弾く」という動作であるが，手の細かい動きを抽出することによって，バイオリンとビオラ，ギターなど似た種類の楽器でも手の動きだけから分類することが出来た．現在，研究成果を投稿中である．

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５件）

[1] Yoshiyuki Kawano and Keiji Yanai, FoodCam: A Real-time Food Recognition System on a Smartphone, Multimedia Tools and Applications (2014). (http://dx.doi.org/10.1007/s11042-014-2000-8) (in press) (査読有)

[2] Do Hang Nga and Keiji Yanai: Automatic Extraction of Relevant Video Shots of Specific Actions Exploiting Web Data, Computer Vision and Image Understanding, Vol. 118, pp. 2-15 (2014/01). (http://dx.doi.org/10.1016/j.cviu.2013.03.009) (査読有)

[3] 松田裕司，柳井啓司: 複数品目が含まれる食事画像の認識における共起関係の利用，電子情報通信学会論文誌，vol. J96-D, no.8, pp.1724-1730 (2013/08). (査読有)

[4] 松田裕司，甫足創，柳井啓司: 候補領域推定に基づく複数品目食事画像認識，電子情報通信学会論文誌，Vol.J95-D, No.8, pp.1554-1564 (2012/08). (査読有)

[5] 秋間雄太，川久保秀敏，柳井啓司: Folksonomy を用いた画像特徴とタグ共起に基づく画像オントロジーの自動構築，電子情報通信学会論文誌 Vol.J94-D, No.8, pp.1248-1259 (2011/08). (査読有)

〔学会発表〕（計１６件）

[1] Do Hang Nga and Keiji Yanai: A

Dense SURF and Triangulation based Spatio-Temporal Feature for Action Recognition, Proc. of Multimedia Modeling Conference (MMM), Dublin, Ireland (2014/01).

[2] Do Hang Nga and Keiji Yanai: A Spatio-Temporal Feature based on Triangulation of Dense SURF, Proc. of ICCV Workshop on Action Recognition with a Large Number of Classes, Sydney, Australia (2013/12).

[3] Yoshiyuki Kawano and Keiji Yanai: Rapid Mobile Food Recognition using Fisher Vector, Proc. of Asian Conference on Pattern Recognition, Okinawa, Japan (ACPR) (2013/11).

[4] Masaya Okamoto and Keiji Yanai: Summarization of Egocentric Moving Videos for Generating Walking Route Guidance, Proc. of Pacific-rim Symposium on Image and Vision Technology, Guanajuato, Mexico (PSIVT) (2013/10).

[5] Do Hang Nga and Keiji Yanai: Large-scale Web Video Shot Ranking Based on Visual Features and Tag Co-occurrence, Proc. of ACM Multimedia, Barcelona, Spain (2013/10).

[6] Takamu Kaneko and Keiji Yanai: Visual Event Mining from Geo-Tweet Photos, Proc. of ICME Workshop on Social Multimedia Research , San Jone, CA, USA (2013/07).

[7] Yoshiyuki Kawano and Keiji Yanai: Real-time Mobile Food Recognition System, Proc. of CVPR Workshop on Mobile Vision, Portland, OR, USA (2013/06).

[8] Yuya Kohaya and Keiji Yanai: Visual Analysis of Tag Co-occurrence on Nouns and Adjectives, Proc. of Multimedia Modelling Conference (MMM), Huangshan, China (2013/01).

[9] Yuji Matsuda and Keiji Yanai: Multiple-Food Recognition Considering Co-occurrence employing Manifold Ranking, Proc. of IAPR International Conference on Pattern Recognition (ICPR) , Tsukuba, Japan (2012/11).

[10] Takuma Maruyama and Keiji Yanai: Real-time Mobile Recipe Suggestion System using Food Ingredient Recognition, ACM MM WS on Interactive Multimedia on Mobile and Portable Devices (IMMPD) Nara, Japan (2012/11).

[11] Yusuke Nakaji and Keiji Yanai: Visualization of Real-World Events with Tweet Photos, Proc. of IEEE ICME Workshop on Social Media Computing, Melbourn, Australia (2012/07).

[12] Yuji Matsuda and Keiji Yanai: Recognition of Multiple Food Images by Detecting Candidate Regions, Proc. of IEEE International Conference on Multimedia and Expo (ICME), Melbourn, Australia (2012/07).

[13] Do Hang Nga and Keiji Yanai: Automatic Collection of Web Video Shots Corresponding to Specific Actions using Web Images, Proc. of IEEE CVPR Workshop on Large-Scale Video Search and Mining, Province, USA (2012/06)

[14] Keiji Yanai: World Seer: A Realtime Geo-Tweet Photo Mapping System, Proc. of ACM International Conference on Multimedia Retrieval (ICMR), Demo Paper, Hong Kong (2012/06).

[15] Kohya Okuyama and Keiji Yanai: A Travel Planning System Based on Travel Trajectories Extracted from a Large Number of Geotagged Photos on the Web, Proc. of Pacific-Rim Conference on Multimedia, Sydney, Australia (PCM) (2011/12).

[16] Do Hang Nga and Keiji Yanai: Automatic Construction of an Action Video Shot Database using Web Videos, Proc. of IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain (2011/11).

〔図書〕（計１件）

[1] Keiji Yanai, Hidetoshi Kawakubo and Kobus Barnard: Entropy-based Analysis of Visual Concepts, In "Multimedia Information Extraction", Wiley-IEEE CS Press, pp.63-80 (2012/05).

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等
http://mm.cs.uec.ac.jp/webvideo/

６．研究組織

(1)研究代表者

柳井　啓司　（YANAI, Keiji）
電気通信大学・大学院情報理工学研究科・

准教授
研究者番号：２０３０１１７９

(2)研究分担者
なし

(3)連携研究者
なし