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　　１　日現在

機関番号：　１４４０１
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３００９０

研究課題名（和文）フォークソノミーによる大規模タグ付き映像集合を利用した動詞的概念の視覚モデル化

研究課題名（英文）Visual Concept Modeling of Verbs Based on a Large Scale Set of Tagged Videos Provided by Folksonomy

研究代表者

中村　和晃（Kazuaki, Nakamura）

大阪大学・工学研究科　・助教

研究者番号：１０５８４０４７

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：YouTube等の映像共有サービスに投稿されるタグ付き映像を利用して動詞的概念（動詞により表される概念）の視覚モデルを構築する技術について検討した．一般に，映像に付与されているタグが映像中のどのシーン（区間）を表現したものであるかは自明ではない．そこで本研究では，同一のタグが付与された複数の映像群に共通して現れる区間（共通区間）を当該タグに対応する区間として抽出する手法，および抽出した区間を利用して当該タグが表す概念の視覚モデルを構築する手法を開発した．また，共通区間の抽出に際しては，区間同士の類似度を定める必要があるが，これをタグ付き画像の集合に基づいて定量化する手法も併せて開発した．

研究成果の概要（英文）：This research project investigates a technology for constructing visual models of concepts represented by verbs using tagged videos which are stored in web-based video sharing services such as YouTube. In general, a video consists of several segments, and each of them has a different semantic content. Therefore, we cannot obtain the complete correspondence between tags and segments. To cope with this problem, this project proposes a method to extract a set of segments whose semantic content commonly appears in videos that have the same tag (called “common segments” in this report), and construct a visual model of the tag using the extracted segments. In the process of common segment extraction, it is quite important to measure the similarity between two segments. This project also proposes a method for calculating the similarity based on a large scale set of tagged images.

研究分野：　視覚メディア処理

キーワード：　視覚メディア処理　視覚概念学習　動作認識　映像処理　画像処理

１．研究開始当初の背景

スマートフォン等のカメラ搭載型端末や監視カメラの普及，並びに YouTube に代表される映像共有サイトの発展に伴って，全世界的に映像コンテンツの量が増加しつつある．これらの映像にはカメラの設置者や撮影者の意図を超えて有用な情報が含まれており，その有効活用が望まれているが，膨大な量の映像を人間の解析者が実際に視聴することは現実的とは言えないため，映像の内容を自動的に認識・理解する技術が強く要請されている．

映像認識・理解技術の実現を困難なものとしている要因の一つにセマンティックギャップがある．これは，映像の意味内容と表現量（視覚特徴量）が直接的に結びつかない，という問題として知られる．この問題に対し，近年，Web 上の映像共有サイトに投稿されるタグ付き映像が有望な情報源として期待を集めている．タグ付き映像に付与されるテキストタグは映像の意味内容を表しており，このタグと視覚特徴量の関係を統計的に解析することによりタグが表す概念の視覚モデルが構築できれば，映像認識・理解技術の発展に大きく資するものと予想される．しかし，タグ付き映像では一般に，各タグが映像中のどのシーン（区間）に対応するのかが不明である（これを本研究では「タグ・シーン対応の不完全性」と呼ぶ）という問題がある．この問題のため，上述のような統計的解析は実際には困難なものとなっている．

２．研究の目的

本研究では，タグ付き映像に付与されるタグの中でも特に動詞（動名詞）に着目し，それにより表現される動詞的概念の視覚モデルをタグ付き映像の集合から構築することを目指す．また，その際に問題となる「タグ・シーン対応の不完全性」について，その対処法を考究する．

３．研究の方法

「タグ・シーン対応の不完全性」への対処法として，同一の動詞タグが付与された複数の映像群から，各々に共通する区間（共通区間）を抽出することを考える．ここで，共通区間抽出の際の「共通性」は，単純な見た目の共通性ではなく，意味内容の共通性として定量化する必要がある．この際にもやはりセマンティックギャップの存在が大きな問題となる．そこで本研究では，「タグ・シーン対応の不完全性」に類する問題が存在しないタグ付き画像にまず着目し，それに対する統計的解析を通して，名詞タグにより表される名詞的概念の視覚モデルを構築する．次に，構築したモデルに基づいて映像中の各区間を名詞的概念の集合として表現し，その集合

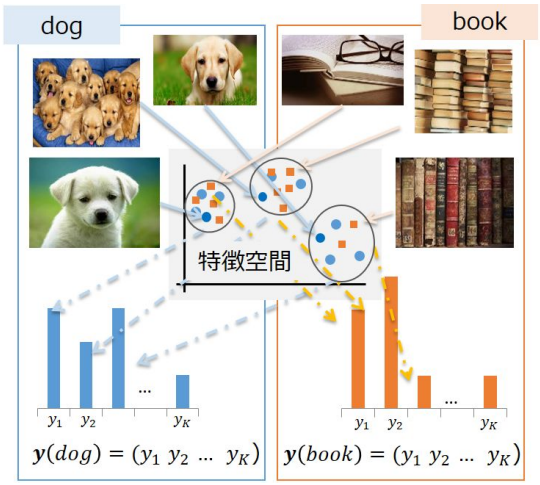

図１　名詞的概念の視覚モデル化

の類似度として区間の共通性を定量化する手法を開発する．その後，複数の映像群から共通区間を抽出する手法を確立する．また，以上の内容と並行して，タグ・シーン対応のとれた映像群から動詞的概念の視覚モデルを構築する手法を開発し，それを用いた人間動作解析を試みる．

まとめると，次の４つの課題を通して研究目的の達成を図る．

(1) タグ付き画像集合からの名詞的概念の視覚モデル化
(2) 名詞的概念の視覚モデルに基づく映像区間の共通性尺度の設計
(3) タグ付き映像集合からの共通区間抽出
(4) タグ・シーン対応のとれた映像群からの動詞的概念の視覚モデル化

４．研究成果

(1) タグ付き画像集合からの名詞的概念の視覚モデル化

同一の名詞タグが付与された大量の画像に基づいて，そのタグが表す名詞的概念を視覚モデルとして表現する手法を開発した[学会発表③④⑦⑧⑪]．本手法の基本的な考え方は次の通りである．

まず，一枚一枚の画像から一つずつ単一の視覚特徴ベクトルを抽出する．この際の視覚特徴ベクトルとしては，画像中の色ヒストグラムや画像の局所的な形状を表現する SIFT (Scale Invariant Feature Transform) など様々なものを検討した[学会発表⑪]が，結論としては，予め学習された Convolutional Neural Network の中間出力を特徴として用いたものが最も高い性能を示すことが分かった[学会発表③④]．次に，抽出した視覚特徴ベクトルをクラスタリングし，特徴空間を複数の部分領域に分割する．最後に，ある名詞タグが付与された画像の視覚特徴ベクトルが特徴空間中どの領域に属するかを求め，そのヒストグラムとして当該タグの表す名詞

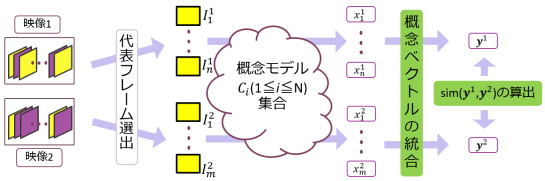

図 2　名詞的概念による映像の表現と類似度評価

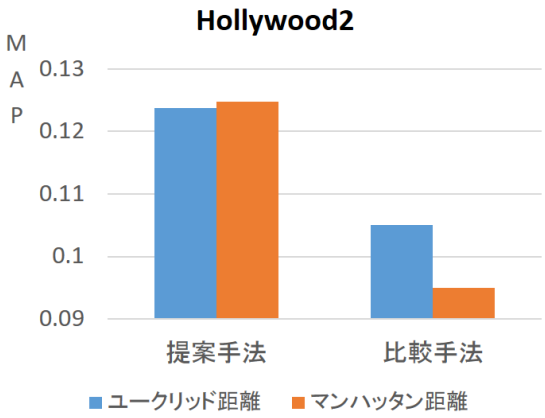

図 3　HollyWood2 に対する検索精度の比較

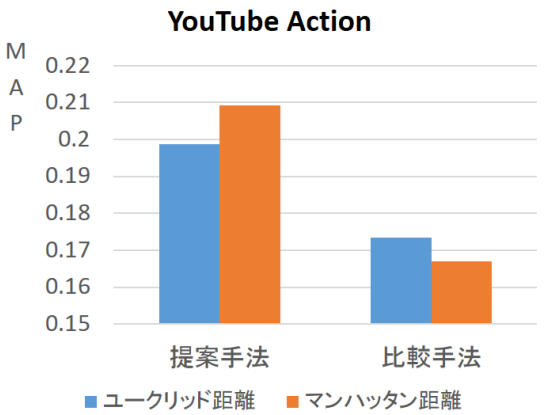

図 4　YouTube Action に対する検索精度の比較

的概念の視覚モデルを構築する．本手法の概要を図 1 に示す．

上記の方法により構築された視覚モデルを実験的に評価した結果，視覚化することの難しい抽象的な概念ほどヒストグラムが一様に近くなり，また，類似した意味内容を持つ概念同士ではヒストグラムも類似する傾向が見られるなど，モデルとしての妥当性を示す結果が得られた[学会発表③④⑦]．

(2) 名詞的概念の視覚モデルに基づく
映像区間の共通性尺度の設計

前項で述べた名詞的概念の視覚モデルは，その構築に用いた視覚特徴ベクトルと同種の特徴量を未知の画像からも抽出し，これと視覚モデルとの合致度を評価することにより，その画像中に当該の名詞的概念がどの程度含まれているかを推定するのに役立てることができる．この性質を利用し，映像区間中の代表フレームにどのような名詞的概念がどの程度含まれているかを推定し，その結果を複数の代表フレームについて統合することにより，当該映像を名詞的概念の集合として表現する手法を開発した．さらに，名詞的概念の集合を一種のベクトルとみなし，その間の距離をユークリッド距離やマンハッタン距離として算出することにより，名詞的概念に基づいて複数映像間の類似度を定量化する手法を開発した[学会発表⑨]．本手法の概要を図 2 に示す．

本手法の有効性を確認するために，定量化した類似度に基づいて映像検索を行った場合の検索精度を調べる実験を行った．比較対象として，名詞的概念を利用せず，代表フレームから抽出される視覚特徴ベクトル自体の類似度を映像の類似度とみなす手法を用意し，提案手法との精度差を検証した．図 3 および図 4 に，検索対象のデータセットとして Hollywood2 Human Actions and Scenes Dataset (Hollywood2) および YouTube Action Dataset (YouTube Action) を用い，また，ベクトル間の距離尺度としてユークリッド距離とマンハッタン距離を採用した場合の実験結果を示す．これらの結果から，名詞的概念を用いることにより映像検索の精度が 20～40%程度向上することが分かる．このことから，映像を名詞的概念の集合として表現し，その集合の類似度として複数映像間の共通性を定量化する本手法の有効性が確かめられたと言える．

(3) タグ付き映像集合からの共通区間抽出

前項で開発した共通性尺度に基づいて，同一のタグが付与された複数の映像群から共通区間を抽出する手法を開発した[学会発表②]．

研究開始当初は，映像中の各区間を表現する手段として名詞的概念の集合のみを考えていた．しかし，名詞的概念は区間中の各フレームから抽出されるものであり，本来的には動きに関する情報を含んでいないため，映像を表現する手段として必ずしも十分ではないことが後の検討結果から判明した．この点に関して，単一の人物画像のみからその人物の身体各位の運動方向を推定する研究も試みた[学会発表①⑤]が，その有効性は限定的であったことからも，各フレームから独立に抽出される特徴に基づいて動き情報を表現することは困難であることが示唆される．

以上の点を踏まえ，本研究では，映像を人物領域（前景）とそれ以外の領域（背景）に分けることを考える．動詞的概念は基本的に人間の動作を表す概念であるため，動き情報としては特に人間の運動に着目すべきと考えられる．そこで，前景からは動きに関する特徴量を，背景からは前項の方法により名詞的概念の集合を抽出し，それらの組合せによ

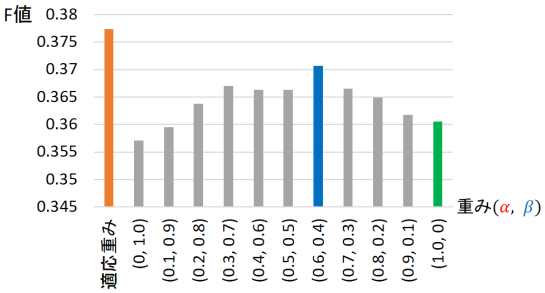

図 5　共通区間抽出に関する実験結果

り映像中の各区間を表現することを提案した．

共通区間の抽出には吸収マルコフ連鎖と呼ばれる外れ値除去手法を採用した．同一のタグが付与された複数の映像が存在するとき，それらの映像には当該タグに対応する区間が共通に存在する．一方，当該タグと無関係の区間は一部の映像にしか存在しない．従って，非共通区間は共通区間に比べ数が少なくなり，区間集合全体の中で外れ値としてふるまうことになる．これを除去することにより，残った区間を共通区間として抽出することが可能になる．ここで，外れ値は，他のどの区間とも類似していない区間として解釈されるため，これを除去するためには，区間同士の類似度を定義する必要がある．本研究では，この類似度を，動き特徴量の類似度および前項で開発した名詞的概念集合の類似度の和として定義する．具体的には，2 つの映像区間に対し，各々の動き特徴を x1, x2，名詞的概念集合を y1, y2 としたとき，両者の類似度を

$$\alpha\ \mathrm{sim}(x1, x2) + \beta\ \mathrm{sim}(y1, y2)$$

として定義する．ここで sim はベクトル間の類似度を定義する関数であり，ユークリッド距離の逆数などが考えられる．また，α, β はどちらの情報をより重視するかを定める重みパラメータである．

本手法の有効性を調べるため，THUMOS 2014 Dataset に含まれる 20 種類 700 本程度の動作映像を対象に，共通区間の抽出を試みた．抽出精度を F 値により評価した結果を図 5 に示す．図 5 において，(α, β)=(0, 1) は動き特徴量のみを用いた場合の抽出精度，(α, β)=(1, 0) は名詞的概念のみを用いた場合の抽出精度に相当する．この結果から，動き特徴量と名詞的概念集合の各々をそれぞれ単独で用いるよりも，両者を組み合わせた提案手法の方が良好な抽出精度が得られていることが分かる．また，重み (α, β) を固定値ではなく動作の性質に応じて適応的に与えた場合には，さらに良好な抽出精度が得られることが分かった．これらの結果から，提案する共通区間抽出手法の有効性が確かめられたと言える．

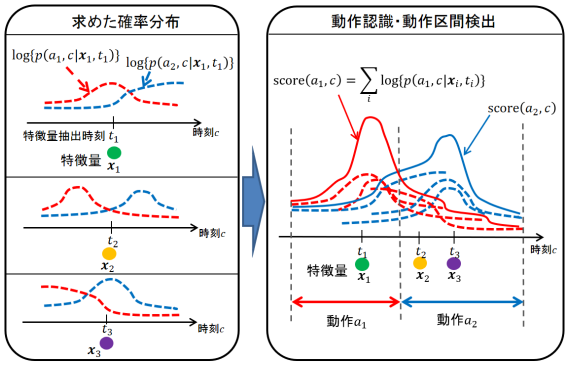

図 6　動詞的概念の視覚モデル化とその応用

(4) タグ・シーン対応のとれた映像群からの動詞的概念の視覚モデル化

前項までの手法により，タグ・シーン対応のとれた映像群を得ることがある程度可能になると期待される．そこで次に，そのような映像群が得られているという前提の下で，実際に動詞的概念を視覚モデル化する手法を開発した[学会発表⑥⑩]．

タグ・シーン対応が取れている映像群からの動詞的概念の視覚モデル化は，動作解析・認識の分野で一般的に行われている処理である．この分野では，構築された視覚モデルの応用例として，入力映像全体が単一の動作に対応しているという前提の下でその動作の種類をモデルに基づいて認識することを想定しているものが多い．しかし，一般の映像において，その全体が単一の動作に対応している例は極めて少ない．これを踏まえ，本研究では，まず入力映像中で行われている動作の種類を認識するだけでなく，その動作が行われている区間を検出することを想定し，その用途に適した視覚モデルを構築することを目指した．提案した手法の概要は次の通りである．

まず，同一の動詞タグ a に対応する多数の映像（区間）に基づいて，「時刻 t に特徴量 x が観測された場合に動作 a が時刻 c において行われている確率」p(a,c|x,t) を統計的に求める．この確率分布が a により表される動詞的概念の視覚モデルとなる．このモデルを動作認識および動作区間検出に適用する際には，対象となる映像の各フレーム i から特徴量 x(i)を抽出したのち，p(a,c|x(i),i) を動作 a および時刻 c に対するスコアとみなして投票する．これを任意の i に対して実行し投票スコアの総和を求め，その値が一定値以上となる区間を動作種別とともに検出することにより，動作認識と動作区間検出を同時に実現する．以上の枠組みの模式図を図 6 に示す．

以上の提案手法の有効性を確かめるために実験を行った．この実験では，標準的な動作映像データセットである KTH Dataset に含まれる 6 種類の動作映像を任意の順序でつなぐことにより疑似的に一本の長時間映像を

ground truth \ spotting result	boxing	hand-clapping	hand-waving	jogging	running	walking
boxing	65.5	13.3	6.0	6.8	0.0	8.3
handclapping	20.4	50.4	11.7	7.4	0.5	9.6
handwaving	2.7	6.5	81.6	3.3	0.0	6.0
jogging	0.7	1.0	3.3	59.8	0.0	35.2
running	0.0	0.4	1.1	83.1	4.0	11.4
walking	0.8	0.7	3.9	8.9	1.0	84.8

図 7　提案手法による動作認識・検出結果

ground truth \ spotting result	boxing	hand-clapping	hand-waving	jogging	running	walking
boxing	7.2	21.6	48.1	4.2	14.6	4.4
handclapping	8.1	18.1	32.8	21.7	9.0	1.0
handwaving	7.7	16.3	28.6	23.7	11.0	12.6
jogging	3.7	18.3	23.3	25.9	16.3	12.5
running	6.5	14.8	27.6	19.6	20.9	10.6
walking	5.4	3.2	29.1	30.4	26.5	5.4

図 8　比較手法による動作認識・検出結果

作成し，その映像から 6 種類の動作をそれぞれ認識するとともにその区間を検出することを試みた．比較対象として，一般的な動作認識・検出法である Dynamic Time Warping に基づく手法を用意し，提案手法との精度差を検証した．この結果を図 7，8 に示す．

図 7，8 の結果から，提案手法が比較手法に比べ高い精度で動作認識・検出を実現できていることが分かり，提案する動詞的概念の視覚モデル化手法の有効性が確かめられたと言える．なお，6 種類の動作のうち jogging，running，walking の 3 動作については，提案手法でも認識誤りが高い割合で発生しているが，この理由としては，これらの 3 動作が互いに類似した動作であり，その区別が本来困難なものであることが挙げられる．

５．主な発表論文等

〔学会発表〕（計 11 件）

① 櫻井皓介，中村和晃，新田直子，馬場口登：“隣接領域間での整合性を考慮した静止画からの運動フロー場推定，”電子情報通信学会 2017 年総合大会，D-12-22，p.78，名城大学，2017 年 3 月．

② 河村圭将，中村和晃，新田直子，馬場口登：“前景の運動と背景の外観を併用した複数映像からの共通動作区間の検出，” 電子情報通信学会技術研究報告，PRMU2016-221，pp.149-154，名城大学．2017 年 3 月．

③ 長澤優佑，中村和晃，新田直子，馬場口登：“ジャンク画像を考慮したマルチモーダルトピックモデルによる概念間距離の算出，” 電子情報通信学会技術研究報告，PRMU2016-213，pp.105-110，名城大学，2017 年 3 月．

④ Y. Nagasawa，K. Nakamura，N. Nitta，and N. Babaguchi: “Effect of Junk Images on Inter-concept Distance Measurement: Positive or Negative?，” Proc. of 23rd Int’l Conf. on Multimedia Modeling，pp.173-184，Reykjavik，Iceland，January 2017.（査読有）

⑤ C. Bender-Saebelkampf，K. Nakamura，N. Nitta，and N. Babaguchi: “Motion Prediction from Still Images Considering the Ambiguity of the Relationship Between Appearance and Motion Features," Proc. of 19th Meeting on Image Recognition and Understanding，PS-2-73，Hamamatsu，August 2016.

⑥ 原啓太，中村和晃，馬場口登：“動作の階層構造を考慮した投票法に基づく人間動作の時間的スポッティング，”電子情報通信学会技術研究報告，PRMU2015-168，pp.7-12，産業技術総合研究所，2016 年 3 月．

⑦ 三輪智宏，中村和晃，馬場口登：“類似性評価の観点の多様性を考慮した概念間類似度の算出，” 電子情報通信学会 2016 年総合大会，D-12-30，p.99，九州大学，2016 年 3 月．

⑧ 長澤優佑，中村和晃，馬場口登：“画像信頼度の反復推定に基づく概念間類似度の算出，”電子情報通信学会技術研究報告，PRMU2015-101，pp.7-12，信州大学，2015 年 12 月．

⑨ K. Kawamura，K. Nakamura，and N. Babaguchi: “Similarity Measurement of Human Action Videos Using Image-Based Models of Diverse Concepts，” Proc. of 18th Meeting on Image Recognition and Understanding，SS-2-2，Osaka，July 2015.

⑩ K. Hara，K. Nakamura，and N. Babaguchi: “Temporal Spotting of Human Actions from Videos Containing Actor's Unintentional Motions，” Proc. of IEEE Int’l Conf. on Multimedia and Expo，Torino，Italy，July 2015.（査読有）

⑪ K. Nakamura and N. Babaguchi: “Inter-Concept Distance Measurement with Adaptively Weighted Multiple Visual Features，” Proc. of 1st Int’l Workshop on Feature and Similarity Learning for Computer Vision，Singapore，November 2014.（査読有）

６．研究組織

(1)研究代表者
中村 和晃（NAKAMURA，Kazuaki）
大阪大学・大学院工学研究科・助教
研究者番号： 10584047