科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１６　日現在

機関番号：　３２６９０
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００１８８

研究課題名（和文）選択的注意を組み込んだ増強型確率的学習に基づく人と物体のインタラクションの理解

研究課題名（英文）Attention-guided object and action recognition based on probabilistic learning and feature boosting for understanding human-object interaction

研究代表者
渥美　雅保（ATSUMI, Masayasu）

創価大学・工学部・准教授

研究者番号：００１９２９８０

交付決定額（研究期間全体）：（直接経費）　　4,100,000円、（間接経費）　　1,230,000円

研究成果の概要（和文）：人と物体のインタラクションの理解に向けて，選択的注意に基づく物体の学習・認識と物体への働きかけ動作の学習・認識・推論のための確率的手法を提案した．前者に関して，選択的注意にガイドされた動的マルコフ確率場での物体セグメンテーション，及び確率潜在コンポーネント木の学習とブースティングによる特徴選択に基づく物体認識の手法を構築した．後者に関して，物体に働きかけるアクションの視覚的動き特徴を表すクラスとその解釈を与える意味素とからなる確率意味ネットワークを確率潜在コンポーネント解析に基づき学習し，それを用いて物体指向アクションを認識・推論する手法を構築した．そして，それらの有効性を実験により確かめた．

研究成果の概要（英文）：This research proposed probabilistic methods of attention-guided object recognition and object-oriented action recognition for understanding human-object interaction. In the proposed methods, attention-guided object recognition is performed in the context of co-occurring objects by using a classification tree which is learned based on the probabilistic latent component tree analysis and feature boosting. Also object-oriented action recognition is performed in the mutual contexts of objects and actions by using a probabilistic semantic network of visual motion classes and their semantic tags which is learned based on the incremental probabilistic latent component analysis. It was shown that the proposed method achieved high recognition accuracy through experiments using image data sets of plural object categories and also a set of video clips of object-oriented actions captured by a Kinect sensor mounted on a robot.

研究分野：　知能情報学

科研費の分科・細目：　情報学、知能情報学

キーワード：　注意　確率的学習　ブースティング　物体認識　アクション認識　コンテキスト　意味ネットワーク　確率的推論

１．研究開始当初の背景

日常生活空間において人を支援するためには，人が何をしているのかを理解することが必要である．人の動作には物体を用いるための物体への働きかけ動作が多くみられる．このとき，動作を理解するためには，動きとその動きが働きかける対象である物体を認識することが必要であり，動作と物体はペアで実世界における意味を形成する．本研究では，このような観点から捉えられる動作を「物体指向動作」と呼ぶ．ところで，物体はそれを含む多くの情景内物体からなるコンテキストの中に置かれ，同じく動作も一連の動作からなるコンテキストの中で行われることが普通である．物体の認識がそのコンテキストにより促進されることはよく知られた知見であるが，１つ１つの動作の認識も一連の動作からなるコンテキストにより促進されると考えられる．本研究では，前者の１つ１つの動作を「アクション」，後者の一連の動作を「アクティビティ」と呼び，アクティビティがアクションのコンテキストを与えてアクションの認識を促進すると仮定する．これら観点のもとで，本研究において，物体指向動作の理解に向けて当初設定した研究課題は次の２つである．

(1)雑然とした情景の中での物体の認識に関して，筆者が従来の研究で行ってきた注意に基づくセグメンテーションと知覚体制化，及び確率潜在コンポーネント解析に基づく物体認識の研究を発展させて，選択的注意により定められるコンテキストのもとで物体の学習と認識を行うとともに，物体の特徴選択にブースティング(増強)を導入して認識性能を高める手法を構築する．

(2)物体への働きかけ動作に関して，確率潜在コンポーネント解析に基づく手法を物体指向動作の学習と認識に拡張し，また，物体と動作に与えられる言語ラベルを相互に確率的に関連付けることにより，物体とそれに働きかけるアクションを相互コンテキストとして用いるとともに，一連の動作からなるアクティビティをアクションのコンテキストとして用いて，物体指向動作を視覚的・言語的に認識・推論する手法を構築する．

２．研究の目的

雑然とした情景の中で物体への働きかけ動作を理解することが可能なビジョンシステムを実現するために，選択的注意を組み込んだ増強型確率潜在コンポーネント木に基づく物体の学習と認識の手法，及び物体に働きかけるアクションとアクティビティの動き特徴とその格表現に基づく言語ラベルを確率的に相互関連付けて意味ネットワークに学習することにより，アクティビティをコンテキストとしてアクションを認識・推論する手法を構築する．そして，画像データセット，及び映像データセットを用いた実験により提案手法の有効性を評価して，ロボットが人を自律的に支援するために人の物体指向動作を理解する基盤技術を確立する．

３．研究の方法

(1)選択的注意を組み込んだ増強型確率潜在コンポーネント木に基づく物体の学習と認識の手法を構築する．本手法の特徴は次の３つである．第一に，選択的注意のもとで動的に形成されるマルコフ確率場でセグメンテーションを行い，共起セグメント集合を物体コンテキストとして学習する点，第二に，インクリメンタル確率潜在コンポーネント解析(Incremental Probabilistic Latent Component Analysis，I-PLCA)に基づいて分類木としての確率潜在コンポーネント木を学習し，また，ブースティングにより分類木上で特徴選択を行って分類性能を向上させている点，第三に，共起制約のもとで情景内の前景物体とその後景物体を認識する点である．そして，Caltech-256 画像データセットとMSRC ラベル付き画像 DB v2 を用いた実験により本手法の有効性を評価する．

(2)物体指向アクションとアクティビティの動き特徴とその格表現に基づく言語ラベルを確率意味ネットワークに学習し，アクティビティをコンテキストとしてアクションを認識・推論する手法を構築する．本手法の特徴は次の３つである．第一に，アクション，及びアクティビティの視覚的な動き特徴を表すクラス集合をインクリメンタル確率潜在コンポーネント解析(I-PLCA) により学習する点，第二に，アクション，及びアクティビティの視覚的な動き特徴を表すクラスとそれらの言語的意味を与える格３つ組の意味素の間の関連を確率的な意味ネットワークに獲得して，視覚レベルの動作認識と言語レベルの動作推論を融合している点，第三に，アクションとアクティビティの共起関係を求めて，それを用いてアクティビティをコンテキストとしたアクションの認識を実現している点である．そして，ロボットに搭載した Kinect センサーでキャプチャした物体指向動作のビデオクリップデータセットを作成し，それを用いた実験により本手法の有効性を評価する．

４．研究成果

(1)選択的注意を組み込んだ増強型確率潜在コンポーネント木に基づく物体の学習と認識手法の概要は次のとおりである．

注意にガイドされたセグメント集合の画像からの抽出は，画像の顕著性マップから選ばれた複数の高顕著度の前注意点の周りに動的に形成されるマルコフ確率場でのセグメンテーションと，それらセグメントに対して計算される注意度が大きいセグメントの選択とグルーピングによりなされる．図１に，人物を前景とする画像を例として，その流れを示す．このとき，あるカテゴリの物体を前景とし他のカテゴリの物体を後景とする画

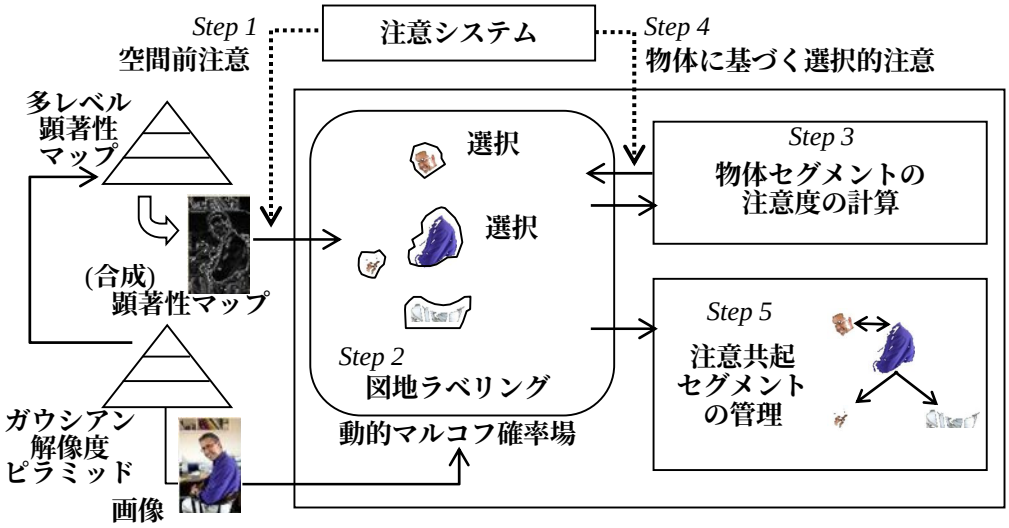

図１　注意にガイドされたセグメント集合の抽出

像を，その前景物体の情景カテゴリ画像と呼ぶ．物体セグメントをその局所キー特徴のBoF(Bag of Features)により表現する．

各情景カテゴリ画像集合から抽出された物体セグメントのBoF集合に対して，I-PLCAにより物体クラス集合を求める．また，同時に，それら物体クラス集合を用いて，前景物体カテゴリのコンテキストの特徴を計算する．次に，全情景カテゴリから求められる物体クラスの全集合に対して，確率潜在コンポーネント木をそれらクラスを葉に持つ分類木として生成する．ここで，葉ノードの物体クラスのカテゴリラベルは，各物体クラスで最大確率を持つ物体セグメントに順次与えられる物体カテゴリラベルを用いて半教師付きでラベル付けされる．そして，確率潜在コンポーネント木の各ブランチノードにおいて，与えられた物体セグメントのカテゴリが２分岐のどちらにあるかの判定に用いるキー特徴の部分集合をブースティングに基づき信頼度付きで選択する．

各情景カテゴリにおける物体カテゴリの共起関係は，その情景カテゴリでの物体カテゴリの出現確率とすべての情景カテゴリでの物体カテゴリの出現確率を用いて，前景物体カテゴリと後景物体カテゴリの間で一種の自己相互情報量を計算することに基づき計算される．

情景内物体の認識では，与えられた情景画像に対して，それに含まれる物体のカテゴリを，確率潜在コンポーネント木の信頼度付きブーストキー特徴を用いた探索と物体カテゴリ間の共起に基づき求める．また，それら物体カテゴリの中から前景となる物体カテゴリを，前景物体カテゴリのコンテキスト特徴を用いて推定する．

(2)選択的注意を組み込んだ増強型確率潜在コンポーネント木に基づく物体の学習と認識の実験結果の概要は次のとおりである．

①注意にガイドされて選択されたセグメント集合からのI-PLCAによる情景内物体クラス解析と前景物体カテゴリのコンテキストの特徴づけの実験を，Caltech-256画像データセットの20個のカテゴリの画像を用いて行った．局所特徴には128次元のSIFT特徴を用い，キー特徴集合のサイズは438であった．情景カテゴリ内の物体クラス数は平均7.55個で，前景物体カテゴリのコンテキストの特徴量間の距離の算出より，前景物体カテゴリがそのコンテキストによりうまく特徴づけられることを確かめた．

②確率潜在コンポーネント木を用いた認識におけるブースト特徴と共起の効果の評価を，MSRCラベル付き画像DB v2に含まれる429枚の画像から16個の情景カテゴリを構成して，５分割交差検定により行った．局所特徴には，疎関心点での128次元のグレイSIFT(Interest Point Grey SIFT,IPGS)と密格子点での384次元の反対色SIFT(Dense Opponent Color SIFT,DOCS)を用い，IPGSとDOCS特徴に対するキー特徴集合のサイズはそれぞれ719と720であった．IPGSまたはDOCS特徴のもとで16個の情景カテゴリから生成された物体クラスの総数の平均は97.6個(情景カテゴリあたり6.1個)で，これらクラスを葉に持つ分類木の深さの平均は11.93であった．また，前景物体カテゴリと平均2.03個の後景物体カテゴリとの間に強い共起がみられた．表１に，特徴選択有・無，共起利用有・無に対する物体カテゴリの分類正確度を示す．特徴選択，及び共起利用により分類正確度が高くなることが確認された．特に，前景物体カテゴリの分類正確度は，DOCS特徴では，特徴選択あり・共起ありの場合0.988,特徴選択あり・共起なしの場合0.979,IPGS特徴では，特徴選択ありで共起ありの場合もなしの場合も0.996と高い性能を示した．また，特徴選択及び共起により，少ない探索数で高い分類正確度が得られることが確かめられた．一般に，物体認識性能は，学

習・認識手法のみでなく，特徴量や学習データセットに依存する．本手法が同じ特徴量と学習データセットを用いた既存手法と比較して高い性能を示すことが確かめられた．

表１　物体カテゴリの分類正確度

認識方法	特徴選択	有	無	有	無
	共起利用	有	有	無	無
特徴記述子	DOCS	0.760	0.740	0.742	0.728
	IPGS	0.681	0.674	0.655	0.649

(3)物体指向アクション・アクティビティの確率意味ネットワークの学習，及び認識・推論手法の概要は次のとおりである．

人の動作を身体スケルトンのジョイント点の３次元座標の時系列としてキャプチャする．本研究では，両手による物体指向動作を扱うため，肩中心に対する両手の相対３次元座標の時系列を利用する．これら相対３次元座標は，Kinect センサーを用いて得られるスケルトンのジョイント座標から計算することが可能である．

両手の相対３次元座標の時系列から，両手の動き特徴量を次の手順により求める．まず，両手の相対３次元座標をある間隔で量子化し，量子化された相対位置とその変位の時系列を計算する．次に，それら時系列に対して，アクション，及びその系列であるアクティビティを，それらの開始フレームと終了フレーム，及び格３つ組＜対象意味素(target synset)[名詞]，格(case)，動作意味素(motion synset)[動詞]＞のアノテーションを付与することにより抽出する．そして，各アクション，及びアクティビティの動き特徴量を，それらの相対位置と変位の時系列に対して，肩中心を原点として身体周りの３次元空間をある大きさで分割したブロックごとの変位のヒストグラムの連結ヒストグラムとして求める．

アクションの確率意味ネットワークの学習では，アクションの格３つ組付きヒストグラムの集合を入力として，動き特徴を表すクラス集合とそれらと格３つ組の意味素との確率ネットワークを求める．まず，アクションのヒストグラムの集合から，I-PLCA によりアクションクラス集合を求める．次に，アクションクラスと意味素の結合確率の計算に基づいて，アクションクラスと意味素のネットワークを生成する．アクティビティの確率意味ネットワークの学習でも，同様に，アクティビティの格３つ組付きヒストグラムの集合を入力として，動き特徴を表すクラス集合とそれらと格３つ組の意味素との確率ネットワークが求められる．また，アクションとアクティビティの共起関係をアクションとアクティビティの格３つ組の確率から自己相互情報量を計算することにより求める．このアクション・アクティビティの共起関係づけられた確率意味ネットワークを本論では ACTNET と呼ぶ．図２に ACTNET の構成を示す．

アクション及びアクティビティの認識と推論では，アクションのヒストグラムの系列入力に対して，それらアクションの格３つ組，及びアクティビティの格３つ組を求める．ま

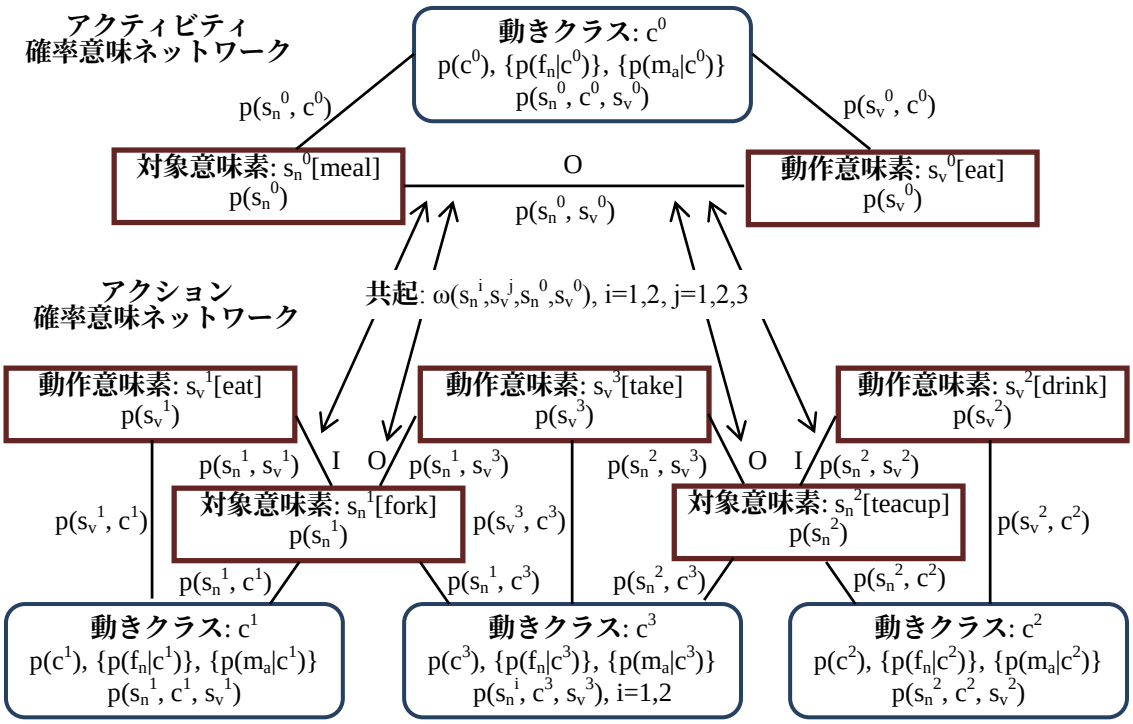

図２　ACTNET の構成

ず，各アクションのヒストグラムに対して，アクションクラスを確信度付きで求める．同時に，アクション系列のヒストグラムの和に対して，アクティビティクラスを確信度付きで求める．次に，アクションとアクティビティの格３つ組を，それらの求められたクラスに基づいて，確率意味ネットワーク上での確率推論とアクションとアクティビティの共起関係を用いて求める．

(4)物体指向アクション・アクティビティの確率意味ネットワークの学習，及び認識・推論の実験結果の概要は次のとおりである．

物体指向動作の ACTNET への学習，及び ACTNET を用いた認識と推論の評価を，Kinect センサーを用いてキャプチャしたビデオクリップデータセットを２つ作成して行った．１つは提案手法の動作評価用の小さいデータセットで，格３つ組＜食事，を，食べる＞と＜イラスト，を，描く＞によりラベル付けられた２つのアクティビティを含む．アクティビティ＜食事，を，食べる＞には，３つの物体とそれらに対する９個のアクション，アクティビティ＜イラスト，を，描く＞には２つの物体とそれらに対する７個のアクションが含まれる．アクションの総数は 16 個である．図３にアクションのスナップショットを示す．もう１つは提案手法の性能評価用の大きいデータセットで，４個のアクティビティ，10 個の物体，27 個のアクションを含む．動きのヒストグラム化における身体周りのブロック分けは，身体の近傍の前方と側方をそれぞれ１辺 30cm の９ブロック，その外側の前方と側方をそれぞれ大きく９ブロックと８ブロック，後方を１つのブロックとする．これよりブロック数は 36 となり，動きヒストグラムの次元は 972 次元である．

図３　アクションの例

①小さいデータセットを用いた実験１において学習された ACTNET の構成を表２に示す．また，図２はこの ACTNET の一部である．ACTNET の解析により，アクティビティ，アクション，及びそれらの間の共起が適切に学習されていることが確かめられた．表３に ACTNET による認識・推論の評価結果を示す．アクティビティとの共起を利用せずに独立にアクションの認識・推論を行った場合の正解率は 75%であったのに対して，アクティビティとの共起を利用した場合は，アクションの正解率が 93.8%に上昇した．また，物体が何かの追加情報が与えられたときのアクションの正解率は 93.8%であった．アクションの次善解までの正解率は，共起なしの場合 93.8%，共起ありの場合 100%，物体が何かの追加情報が与えられた場合 100%であった．

表２　実験１の ACTNET の構成

	アクティビティ	アクション
クラス数	2	16
対象意味素数	2	5
動作意味素数	2	10
意味素ペア数	2	16

表３　実験１の認識・推論結果

アクティビティ正解率	100%
アクション正解率(共起なし)	75.0%
アクション正解率(共起あり)	81.3%

②大きいデータセットを用いた実験２では，４つのアクティビティの各々に対して４つのビデオクリップを用意して４分割交差検定により性能評価を行った．実験２において学習された ACTNET の構成を表４に示す．表５に ACTNET による認識・推論の評価結果を示す．アクティビティとの共起を利用せずに独立にアクションの認識・推論を行った場合の正解率は 53.3%であったのに対して，アクティビティとの共起を利用した場合は，アクションの正解率が 62.5%に上昇した．また，物体が何かの追加情報が与えられたときのアクションの正解率は，共起なしの場合 75.8%，共起ありの場合 83.4%であった．アクションの次善解までの正解率は，共起なしの場合 59.2%，共起ありの場合 76.7%で，物体が何かの追加情報が与えられた場合はそれぞれ 85.8%と 96.7%であった．

表４　実験２の ACTNET の構成

	アクティビティ	アクション
クラス数	12	53.5
対象意味素数	4	10
動作意味素数	4	16
意味素ペア数	4	27

表５　実験２の認識・推論結果

アクティビティ正解率	93.8%
アクション正解率(共起なし)	53.3%
アクション正解率(共起あり)	62.5%

③これら２つの実験により，学習された ACTNET によりアクションとアクティビティの認識が可能で，特に，アクション認識のあいまいさが追加情報を用いた推論により解消されること，コンテキストを与えるアクティビティとの共起によりアクションの認識性能をあげられることが示された．本実験では独自のデータセットを作成して利用した

ため，認識性能を既存手法と単純に比較することは難しいが，同様の問題を扱う既存研究と比較して十分に高い性能を達成していることを確かめた．
(5)本研究の成果である選択的注意のもとでの物体指向動作の学習と認識手法は，日常生活空間においてコンピュータやロボットが人を自律的に支援するために必要な知能情報処理の基盤技術を提供するものである．今後，本研究の成果を，人との対話やロボットの行動プランニングに結び付けていくことが課題である．

５．主な発表論文等

〔雑誌論文〕（計６件）

① Masayasu Atsumi, Learning Probabilistic Semantic Network of Object-oriented Action and Activity, Lecture Notes in Artificial Intelligence, 査読有，掲載決定，2014.
② Masayasu Atsumi, Object Categorization in Context based on Probabilistic Learning of Classification Tree with Boosted Features and Co-occurrence Structure, Lecture Notes in Computer Science: Advances in Visual Computing, 査読有，Vol.8033，pp.416-426，2013，DOI: 10.1007/978-3-642-41914-0_41.
③ Masayasu Atsumi, Attention-Guided Organized Perception and Learning of Object Categories Based on Probabilistic Latent Variable Models, Journal of Intelligent Learning Systems and Applications, 査読有，Vol.5，No.2，pp.123-133，2013，DOI: 10.4236/jilsa.2013.52014.
④ Masayasu Atsumi, Learning Visual Categories based on Probabilistic Latent Component Models with Semi-supervised Labeling, GSTF International Journal on Computing, 査読有，Vol.2，No.1，pp.88-93，2012，DOI:10.5176_2010-2283_2.1.133.
⑤ Masayasu Atsumi, Visual Categorization based on Learning Contextual Probabilistic Latent Component Tree, Lecture Notes in Computer Science: Artificial Neural Networks and Machine Learning，査読有，Vol.7552，pp.419-426，2012，DOI:10.1007/978-3-642-33269-2_53.
⑥ Masayasu Atsumi, Object and Scene Recognition based on Learning Probabilistic Latent Component Tree with Boosted Features, International Journal of Machine Learning and Computing，査読有，Vol.2，No.6，pp.762-766，2012，DOI:10.7763/IJMLC.2012.V2.232.

〔学会発表〕（計９件）

① Masayasu Atsumi, Learning Probabilistic Semantic Network of Object-oriented Action and Activity, The 16th International Conference on Artificial Intelligence: Methodology, Systems, Applications, Sept.11-13, 2014, Varna, Bulgaria.
② 渥美雅保，物体指向動作の心象と表象の確率的カテゴリゼーション，2014 年度人工知能学会第 28 回全国大会，2014 年 5 月 12-15 日，松山市，日本．
③ Masayasu Atsumi, Object Categorization in Context based on Probabilistic Learning of Classification Tree with Boosted Features and Co-occurrence Structure, 9th International Symposium on Visual Computing, July 29-31, 2013, Rethymnon, Crete, Greece.
④ 渥美雅保，注意と共起に基づくシーンの学習と認識，2013 年度人工知能学会第 27 回全国大会，2013 年 06 月 4-7 日，富山市，日本．
⑤ Masayasu Atsumi, Object and Scene Recognition based on Learning Probabilistic Latent Component Tree with Boosted Features, 2012 International Conference on Information and Intelligent Computing, Dec.8-9, 2012, Chengdu, China.
⑥ Masayasu Atsumi, Visual Categorization based on Learning Contextual Probabilistic Latent Component Tree, 22nd International Conference on Artificial Neural Networks, Sept. 11-14, 2012, Lausanne, Switzerland.
⑦ 渥美雅保，前景・後景コンテキストからのラベル付き物体カテゴリ木学習に基づく共起物体認識，2012 年度人工知能学会第 26 回全国大会，2012 年 6 月 12-15 日，山口市，日本．
⑧ Masayasu Atsumi, Visual Category Learning based on Probabilistic Latent Component Models with Semi-supervised Labeling, 2nd Annual International Conference on Advanced Topics in Artificial Intelligence, Nov.24-25, 2011, Singapore.
⑨ 渥美雅保，確率潜在コンポーネント木による物体カテゴリ構成の学習，2011 年度人工知能学会第 25 回全国大会，2011 年 6 月 1-3 日，盛岡市，日本．

６．研究組織

(1)研究代表者

渥美　雅保（ATSUMI, Masayasu）
創価大学・工学部・准教授
研究者番号：００１９２９８０