The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

長時間動作の文脈と身体動作の相互認識

Bidirectional recognition between motion context on long-term observation and human motion

小椋 忠志 *1　稲邑 哲也 *1*2
Tadashi Ogura　Tetsunari Inamura

*1総合研究大学院大学
SOKENDAI(The Graduate University for Advanced Studies)

*2国立情報学研究所
National Institute of Informatics

This paper describes a motion recognition method to reduce recognition error, which has two-layered structure; motion recognition is affected by context estimation in the first layer, and context estimation is affected by motion recognition in the second layer. We introduce an algorithm to integrate the motion recognition by conventional HMM and motion label production by the topic model in the first layer. We also introduce particle filter to estimate and update the context based on the result of motion recognition in the second layer. A set of particles present a probabilistic distribution of motion topics, and motion recognition and particle update procedures are performed on each particle. In an evaluation experiment, we used a sequential motion which is a sequential connection of 33 motion primitives as a long-term observation target. The results showed that the proposed method reduced recognition errors and tracked motion context by topic probability compared with conventional methods.

1. はじめに

長時間人の動作を観測した場合，個々の動作はランダムに発生しているのではなく，とある文脈のもとで実行されている．その個々の動作において，ここでは身体の動作を文章化できる最小の単位を動作プリミティブと呼ぶことにする．具体的には，ホコリをはたく，床を掃く，バイバイする，と言った意味のある時間長で区切った動作のことである．上記の 3 つの動作を例にとると，それらの動作は，掃除という文脈のもとで実行されている動作プリミティブであると言える．また，観測者が掃除をしているという文脈を知ることが出来れば，上記の 3 つの認識結果のうち，バイバイするは誤認識で，似た動作である「窓を拭く」という掃除の文脈の動作プリミティブであったと，認識を改めることが出来る．この例は，ホコリをはたく，床を掃くという動作プリミティブの観測情報から，対象が掃除をしているという文脈を認識し，その文脈情報を基に動作プリミティブの認識を改めるという，相互的な認識が行われている．本研究では，文脈認識と動作認識が相互に影響しあう関係性を利用して長時間の動作に対する文脈推定とそれに基づく動作プリミティブの誤認識低減を目指す．

文脈を用いた動作認識に着眼点を置いた研究をいくつか紹介する．[Sminchisescu 06] は CRF に基づいて時系列データの前後関係性をモデル化し動作の認識を改善している．[Zhang 08] は Shape Context と呼ばれる画像の特徴表現を動作の文脈ととらえ，動作認識手法を提案している．しかしながら，これらの研究は認識のために文脈という表現を利用しているものの，本研究の目指す相互に影響しあう関係性を用いていない．

関連研究との違いを明確にするため，長時間観察における一連の動作を説明するような概念となる文脈をここではトピックと呼ぶことにする．このトピックという表現は，Blei らが提案したトピックモデル [Blei 03] の考え方に基づいている．すなわち，単語認識や文章理解をするために文脈に注目する考え方を，動作認識に応用する．トピックモデルを動作認識に応用する例では，[Wang 09]，[Tavenard 13]，[Ogura 16] が挙げられる．しかし，これらの手法もトピックと動作認識の 2 層が相互に影響しあう関係性を扱っていない．

本研究では，トピックの推定と動作プリミティブの認識の 2 つのプロセスが相互の関係性をもつ仕組みを実現することを目的とする．本稿では，相互認識の計算モデルを提案するとともに，長時間の時系列データに対して認識を行い，文脈情報を用いない手法と比較して，動作の誤認識の低減に効果があることを示す．

連絡先: 小椋忠志，総合研究大学院大学，東京都千代田区一ツ橋 2-1-2，t-ogura@nii.ac.jp

2. 提案手法

図 1および Algorithm 1 に提案する認識手法の処理手順を示す．今どんなトピックであるかという確率を離散の粒の集合で表現し，その 1 粒をパーティクルフィルタの概念と同様にパーティクルと呼ぶ．本手法では動作の文脈を表現するトピックの分布をパーティクルの集合として表現する．k 番目のパーティクル k が所属するトピック q_k は次のようなトピックのインデックスを持つ．

$$j \in \{1, 2, \cdots, J\} \tag{1}$$

ここで J はトピックの総数である．また，パーティクルの集合 $\boldsymbol{Q}$ は次のようになる．

$$\boldsymbol{Q} = \{1, 2, \cdots, K\} \tag{2}$$

ここで，K はパーティクルの個数を表す．このパーティクル集合を用いて現在注目している時刻におけるトピックが j である確率 $P(j)$ を次のように求める．

$$P(j) = \frac{n(\boldsymbol{R})}{K} \tag{3}$$

$$\boldsymbol{R} \in \{k; q_k = j\} \tag{4}$$

ここでの $n(X)$ は X の個数を数える関数で，$\boldsymbol{R}$ は $\boldsymbol{Q}$ のうち q_k が j となる部分集合である．今後の処理はパーティクルごとに動作認識とそのパーティクルの更新が行われる．

各動作ラベル m_i ごとに HMM によってパラメータ λ_i を求めておく．また，トピック j ごとの m_i の出現確率 $P(m_i|j)$ は

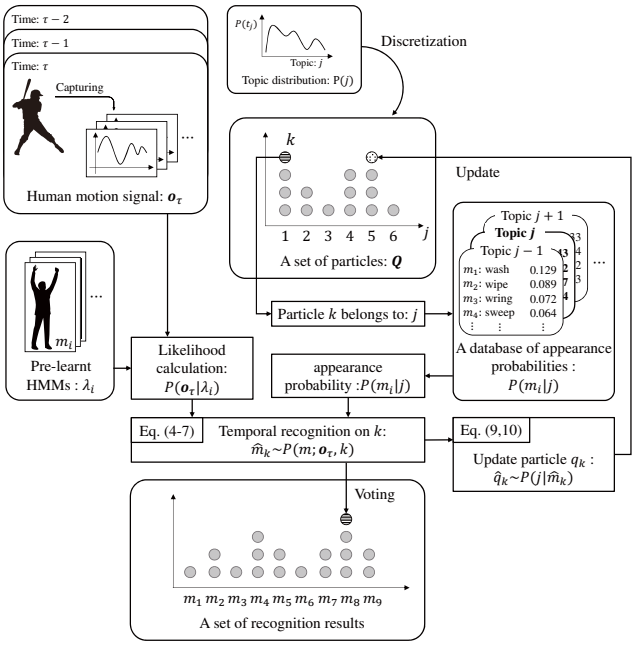

図 1: パーティクルによるトピック情報を用いた動作認識とトピック分布の更新手順

Algorithm 1 Recognition and topic update with particles

for $\tau = 1$ to End of data **do**
  for $k = 1$ to K **do**
    $j = q_k$
    $C = 2\,|\max \log L(o_\tau|\lambda_i)|$ ▷ Eq. (6)
    Calculate α using Eq. (7) ▷ Eq. (7)
    Calculate $S(o_\tau, i, k)$ using Eq. (5) ▷ Eq. (5)
    $P(m; o_\tau, k) = S(o_\tau, i, k) / \sum_i S(o_\tau, i, k)$ ▷ Eq. (9)
    $\hat{m}_k \sim P(m; o_\tau, k)$ ▷ Eq. (8)
    $P(\hat{m}_k) = \sum_{j=1}^{J} P(\hat{m}_k|j)$ ▷ Eq. (12)
    $P(j|\hat{m}_k) = P(\hat{m}_k|j)P(j)/P(\hat{m}_k)$ ▷ Eq. (11)
    $\hat{q}_k \sim P(j|\hat{m}_k)$ ▷ Eq. (10)
  end for
  Update $\boldsymbol{Q}$ using $\{\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_K\}$
  $i_{result} = \arg\max_i n((\hat{m}_k = m_i); \{\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_K\})$
end for

あらかじめ用意することとする [Ogura 16]．次に，時刻 τ で観測された身体動作信号 o_τ に対する尤度 $L(o_\tau|\lambda_i)$ を求める．HMM による認識尤度 $L(o_\tau|\lambda_i)$ および文脈に基づく動作出現確率 $P(m_i|j)$ を用いて，認識スコア $S(o_\tau, i, k)$ を次式のように計算する．

$$\log S(o_\tau, i, k) = \alpha\,(\log L(o_\tau|\lambda_i) - C) + \log P(m_i|j) \quad (5)$$

ここで定数 C は式 (9) で計算する際に相殺されるため任意の値を用いても構わないが，本手法では，$\log L(o_\tau|\lambda_i)$ がすべての場合において負の値になるよう，定数 C を次のように設定する．

$$C = 2\,|\max \log L(o_\tau|\lambda_i)| \quad (6)$$

式 (5) における α は，HMM による認識尤度とトピックごとの動作出現確率の影響度を調整する役割を持ち，本稿では試験的に次のように設定した．

$$\alpha = \frac{\max(\log P(m_i|j) - C)}{\max \log L(o_\tau|\lambda_i)} \quad (7)$$

これは，HMM による認識尤度とトピックごとの動作出現確率の最大値を基準として，両者を統合する際の重みが同一になることを狙ったものである．

続いてパーティクル k における動作認識結果 $\hat{m}_k$ を次のようにサンプリングする．

$$\hat{m}_k \sim P(m; o_\tau, k) \quad (8)$$

ここで，認識結果が m_i となる確率 $P(m; o_\tau, k)$ は次のように求める．

$$P(m; o_\tau, k) = \frac{S(o_\tau, i, k)}{\sum_i S(o_\tau, i, k)} \quad (9)$$

動作認識結果 $\hat{m}_k$ は，パーティクル k と同様に K 個出力され，認識結果の分布を得る．

最後にパーティクルの更新を次のように行う．

$$\hat{q}_k \sim P(j|\hat{m}_k) \quad (10)$$

ここで，動作 $\hat{m}_k$ に基づいて，それが属するトピック j が選ばれる確率 $P(j|\hat{m}_k)$ は次のように求める．

$$P(j|\hat{m}_k) = \frac{P(\hat{m}_k|j)P(j)}{P(\hat{m}_k)} \quad (11)$$

上記の式における $P(\hat{m}_k)$ は $P(\hat{m}_k|j)$ を j について次のように周辺化して求める．

$$P(\hat{m}_k) = \sum_{j=1}^{J} P(\hat{m}_k|j) \quad (12)$$

これらの処理をすべてのパーティクルにおいて実施し，認識結果の出力とパーティクル集合 $\boldsymbol{Q}$ の更新を行う．また，更新されたパーティクル集合 $\boldsymbol{Q}$ を用いて，次の時刻の認識および更新を実行する．

このループ構造を用いることで，トピックの推定とその時間的変化を捉え，動作の認識へ役立てる仕組みを実現する．

3. 検証実験

提案する認識手法により，トピックを推定し認識性能が高まることを確認し，提案手法の有効性を検証する実験を行う．この実験では，時系列に観察される動作を対象として，長時間観察の動作認識におけるトピックの有用性と，トピックの変化に対する本文脈認識法の追従性に焦点を当てる．表 1に実験に使用した動作のラベルとその動作の所属するトピックを示す．また，図 2に実際にキャプチャした動作の様子を示す．対象の動作の中では図 2(a) と図 2(b) の様によく似た身体動作を含んでいる．身体動作信号は，頭・両肩・両手首の各 3 次元と，肘・全指の各 1 次元を対象とした，合計 27 次元の関節角で，モーションキャプチャデバイスを用いて収集される．各動作ラベルにおいて 22 サンプルを収集し，21 サンプルを学習データとして用いる．各動作は 60[Hz] でキャプチャされ，約 4 秒間の長さとする．Left-to-right モデルの HMM を用い，状態数は 16，GMM の混合数は 5 で学習を行った．また認識対象となるテストデータは，収集した 22 サンプルのうち学習データに用いなかった 1 サンプルを，動作ラベル m_1 から m_{33} まで順番に時系列に繋げたものを用いる．

表 2に本実験で使用するトピックごとの動作出現確率 $P(m_i|j)$ の一部を示す．これらは実験者の主観に基づいて手動で作成し

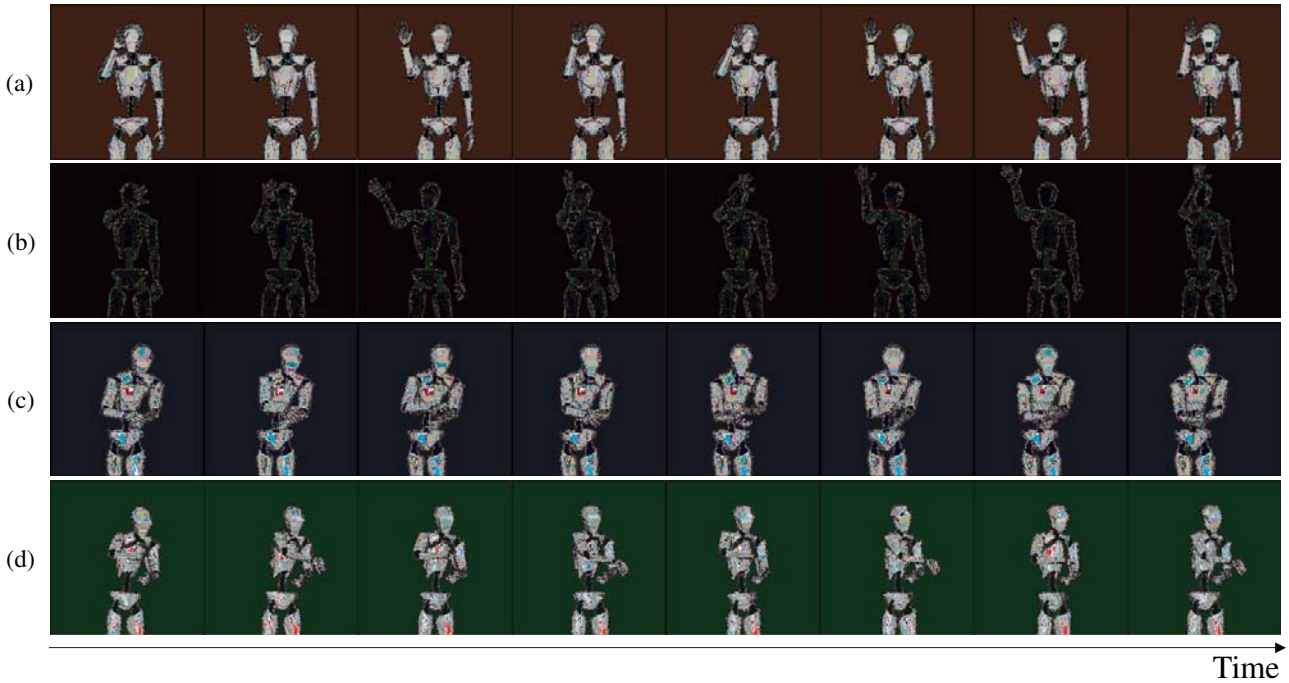

図 2: 対象の動作の様子とよく似た動作の例．(a) m_6: wiping window．(b) m_{33}: waving hand．(c) m_8: washing．(d) m_{13}: frying with pan．(a) と (b) の動作，(c) と (d) の動作はそれぞれよく似ており，誤認識を引き起こしやすい．

表 1: 対象の動作と所属するトピック

m_i: Label name	j: Topic
m_1: dusting off	1: cleaning
m_2: dustpan	
m_3: sweeping floor	
m_4: vacuuming	
m_5: washing	
m_6: wiping window	
m_7: wiping table	
m_8: wringing cloth	
m_9: cutting	2: cooking
m_{10}: dishing foods	
m_{11}: breaking egg	
m_{12}: opening fridge	
m_{13}: frying with pan	
m_{14}: igniting	
m_{15}: mixing	
m_{16}: pouring oil	
m_{17}: replacing	
m_{18}: seasoning	
m_{19}: tearing	
m_{20}: check temperature	
m_{21}: air hockey	3: game center
m_{22}: inserting coins	
m_{23}: driving	
m_{24}: playing drum	
m_{25}: gun shooting	
m_{26}: playing slot	
m_{27}: playing taiko	
m_{28}: victory pose	
m_{29}: bowing	4: greeting
m_{30}: beckoning	
m_{31}: exchanging card	
m_{32}: shaking hands	
m_{33}: waving hand	

た．ここでは，表 1で示した分類と同様にトピックの数 J は 4 で，それぞれのトピック上で動作出現確率の高いもの記載した．またこの実験ではパーティクルの数 K は 400 とし，トピック分布の初期分布として一様分布を用いた．

比較対象として次式のようにパーティクル，トピックを用いずに HMM のみを用いた手法と比較する．

$$i_{result} = \arg\max_i L(\boldsymbol{o}_\tau|\lambda_i) \tag{13}$$

表 2: トピック j ごとの動作出現確率 $P(m_i|j)$ の例

j	m_i:$P(m_i\|j)$	j	m_i:$P(m_i\|j)$
1	m_6: 0.112	3	m_{28}: 0.111
	m_8: 0.110		m_{24}: 0.111
	m_7: 0.109		m_{22}: 0.111
	m_4: 0.107		m_{23}: 0.109
	m_5: 0.106		m_{26}: 0.109
	m_2: 0.105		m_{21}: 0.108
	m_3: 0.104		m_{25}: 0.107
	m_1: 0.104		m_{27}: 0.106
	m_{16}: 0.008		m_2: 0.007
	m_{13}: 0.007		m_3: 0.006
	⋮		⋮
2	m_{17}: 0.076	4	m_{30}: 0.157
	m_{10}: 0.075		m_{32}: 0.154
	m_{12}: 0.075		m_{33}: 0.153
	m_{19}: 0.074		m_{31}: 0.153
	m_{18}: 0.074		m_{29}: 0.150
	m_{15}: 0.074		m_2: 0.013
	m_{11}: 0.074		m_8: 0.013
	m_{20}: 0.073		m_3: 0.011
	m_9: 0.073		m_7: 0.011
	m_{13}: 0.073		m_6: 0.010
	⋮		⋮

ここで i_{result} は認識した動作ラベルのインデックスを指す．図 3にこの手法による認識結果を示す．縦軸は i_{result}，横軸はテストデータに対して区切られた区間のインデックスを示す．青いプロットは認識成功を表し，赤いプロットは認識誤りを表す．この際の認識率は 51.5[%] であった．動作の認識誤りを確認すると，例えば m_6 の窓を拭くという動作 (図 2(a)) と m_{33} のバイバイするという動作 (図 2(b)) は非常によく似た動作であり，それぞれ誤認識が生じていることがわかる．図 4に提案手法を用いた認識結果を示す．認識誤りの数が減り，認識率は 72.7[%] に向上した．その際のトピックの分布の推移を図 5に示す．横軸は同様で，縦軸は式 (3) で計算されるトピックの確率である．また，判例の数字はトピックのインデックスを示している．確率の推移を追うと，おおよそ正しいトピックを追従

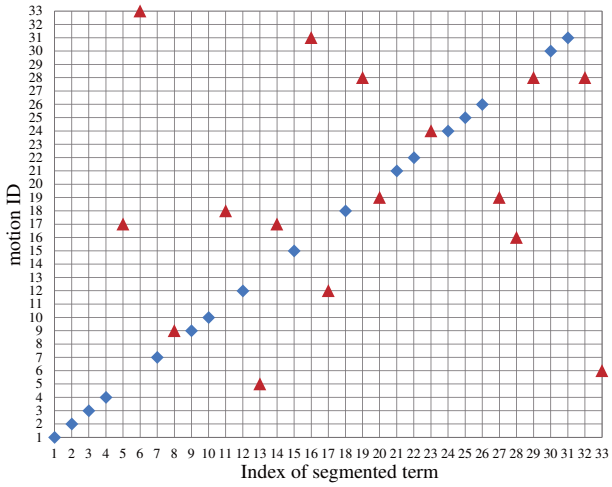

図 3: 提案手法を用いない手法による認識結果

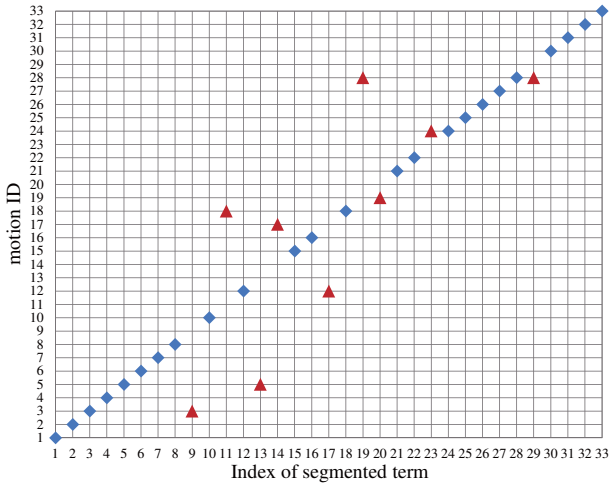

図 4: 提案手法による認識結果

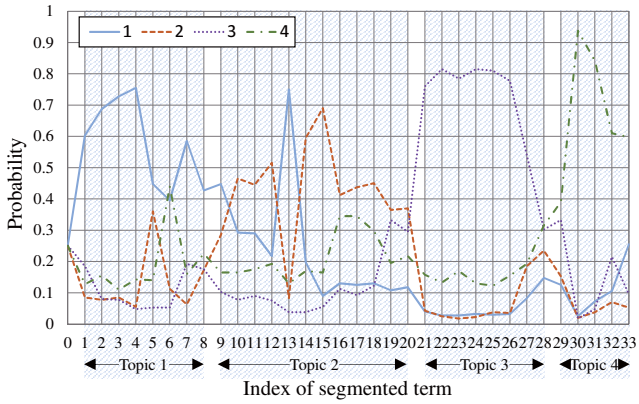

図 5: トピックの分布の推移

できていることが確認できる．また，トピックの情報を基に図 4における m_6 および m_{33} の誤認識を防ぎ，正しい認識を実現していることがわかる．一方で図 5の 13 番目のトピックの確率が突然変動していることが確認できる．図 4の該当箇所を確認すると m_{13} と認識するべきところ m_5 と誤認識している．実際の動作は図 2(c) と図 2(d) に示すように似ている．この誤認識の要因の 1 つは，HMM による認識尤度 $L(\boldsymbol{o}_\tau|\lambda_i)$ において，$L(\boldsymbol{o}_\tau|\lambda_5)$ と $L(\boldsymbol{o}_\tau|\lambda_{13})$ が大きく離れていたため，式 (5) を計算する際にトピックの持つ情報の影響力が少なかったこと挙げられる．これを解決する手段として，HMM の認識尤度に対して，トピックが与える影響力が大きくなるように α の値を調整することが挙げられる

4. おわりに

本稿は，長時間にわたる一連の動作に対して文脈と動作の認識が相互に影響しあう関係性を持つ認識手法を提案し，その有効性を検証した．トピックの分布はパーティクルの集合で表現し，パーティクルごとに動作の認識とパーティクル更新を行う手法を提案した．実験では，4 つのトピックに所属する 33 種の動作を時系列に順番に実施した長時間動作を仮定したテストデータを対象とし，トピック情報を用いない従来手法と比較して，認識率を向上させた．また，長時間観察においてトピックの変化を追従できることを確認した．

今後は，式 (5) における α の値をより認識性能が向上するように調整し，トピック認識に対する影響度の検討を行う．また，道具などの環境とのインタラクションや身体部位ごとの動作情報をトピック情報へ影響させることで，本手法の特性を生かした認識を実現する．

謝辞

本研究は，JST，CREST（グラント番号 JPMJCR15E3）の支援を受けたものである．

参考文献

[Sminchisescu 06] Sminchisescu, C., Kanaujia, A., and Metaxas, D.: Conditional models for contextual human motion recognition, Computer Vision and Image Understanding, Vol. 104, Iss. 2, pp. 210–220, (2006).

[Zhang 08] Zhang, Z., Hu, Y., Chan, S., and Chia, L.-T.: Motion Context: A New Representation for Human Action Recognition, In Proceedings of the European Conference on Computer Vision (ECCV), pp. 817-829, (2008).

[Blei 03] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp. 993–1022, (2003).

[Wang 09] Wang, Y., and Mori, G.: Human action recognition by semilatent topic models, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 31 No. 10, pp. 1762–1774, (2009).

[Tavenard 13] Tavenard, R., Emonet, R., Odobez, J.-M.: Time-sensitive topic models for action recognition in videos, In Proceedings of Conference on Image Processing (ICIP), pp. 2988–2992, (2013).

[Ogura 16] Ogura, T., Sakato, T., and Inamura, T.: Human motion recognition based on topic model, In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3533–3534, (2016).