犬視点映像からの動作認識

○岩下友美　高嶺朝理 (九州大学)　Michael Ryoo(ジェット推進研究所)　倉爪亮 (九州大学)

1.　はじめに

一人称視点映像とは，人の頭部や胸部に固定されたカメラから得られる映像である．被験者が見ている空間が被験者の視点で撮影されるために，人の日常動作の分析に適しており，これまでに人の行動支援や行動意図の理解などの研究が進められてきた．例えば，Kitani ら [1] は一人称視点映像を用い，自転車に乗るなどの自己動作の認識を行っており，Fathi ら [2] は料理など自己と物体とが関連する動作の認識を行った．また Ryoo ら [3] は個人と個人とが関連する動作の認識を行っている．

本研究では，これまでの一人称視点映像と異なり，人に代わり動物の視線に着目し，動物にカメラを装着して得られる動物視点の映像を用いた動作認識手法を提案する．特に本研究では犬にカメラを装着し，例えば人とのボール遊びや餌やりなど，これまでにまだ注目されていない，犬，物体，人間の 3 つが関連する状況での犬の動作認識を行うことを目的とする．本予稿の流れは以下の通りである．まず第 2 章では犬視点映像データベースの構築について述べ，第 3 章では犬視点映像を用いた動作認証手法について述べる．第 4 章では犬視点映像データベースに対して提案手法を適用して実験を行い，本手法の有効性を示す．第 5 章はまとめである．

2.　犬視点映像データベースの構築

本章では，犬視点映像データベースの構築について述べる．まず図 1 に示すように GoPro カメラ (GoPro, Hero3) を犬の背中に設置した．犬は 5 種類の動作（(a) ボール取り，(b) 食事，(c) 水飲み，(d) 人になでてもらう，(e) 引っ張り合い）を行った．図 2 にそれぞれの動作の撮影画像の例を示す．ここで，ボール取りは人間がボールを投げ，それを犬が追いかけて口で加えるという動作であり，人間，犬，物体の 3 つが関連する状況での動作となる．また食事と引っ張り合いでは，それぞれの動作は人間がおやつを手に持ち犬がそれを食べるという動作，またペットボトルを人間と犬が同時に引っ張り合うという動作であり，ボール取りと同様に人間，犬，物体の 3 つが関連した状況での動作となる．水飲みは犬と物体とが関連する動作，人になでてもらうは人と犬とが関連する動作となる．画像解像度は 384 × 288，フレームレートは 47Hz であり，また動作はそれぞれ 5 回行った．

図 1 カメラ設置

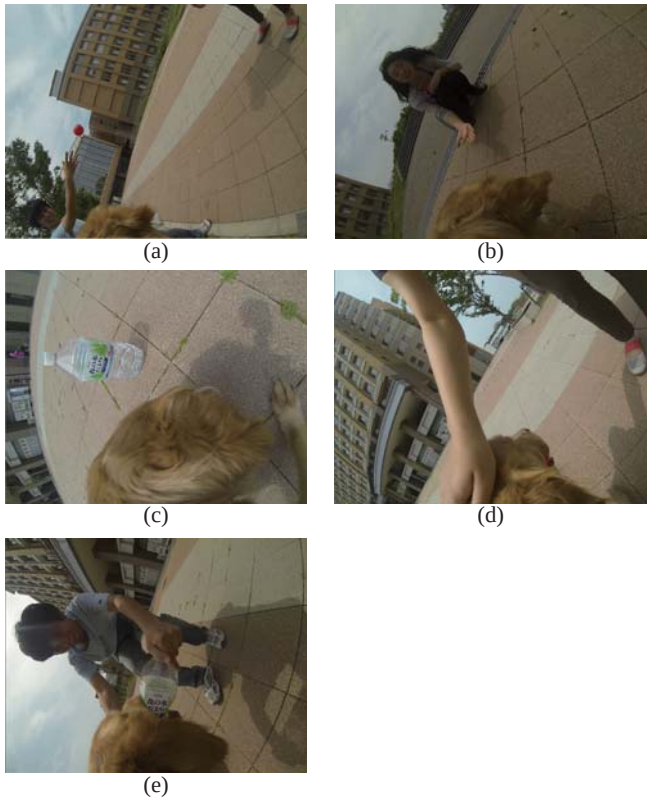

図 2 撮影画像の例: (a) ボール取り，(b) 食事，(c) 水飲み，(d) 人になでてもらう，(e) 引っ張り合い

3.　犬視点画像を用いた動作認識

本章では動作認識のための特徴抽出手法，および認識手法について述べる．本手法では，まず時空間画像から局所特徴を抽出するために cuboid[4] および STIP[5] を用いた．局所特徴量の記述では，cuboid では Dollar らが用いた特徴と同様の勾配情報を用いており，また STIP では Laptev が用いた HOG, HOF, HOG/HOF のいずれかを用いた．

次に *visual words* を用いて，それぞれの動作シーケンスの特徴量を抽出する．*visual words* では，まず学習用データから抽出された局所特徴量に対して k-means 法を適用して，局所特徴量をクラスタリングする．ここで cuboid の場合は k=100，STIP の場合は k=600 と

した．STIP により抽出された局所特徴量のクラスタリング結果の一例を図 3 に示す．次に各動作シーケンスにおいて，クラスタリング結果を用いてヒストグラムを求める．テスト用データの動作シーケンスの特徴量の抽出では，まず抽出された局所特徴量が属するクラスタを求め，次にヒストグラムを構築する．テスト用データの動作シーケンス動作識別は，Support Vector Machine を用いて行う．

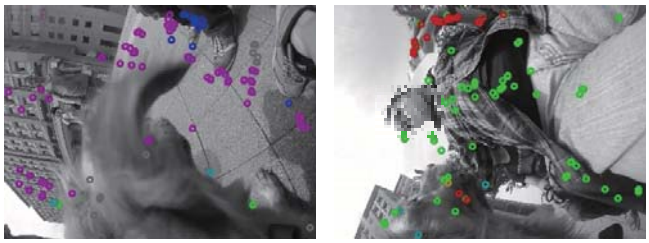

図 3 STIP により抽出された特徴点．異なる色は異なるクラスタに属することを意味する．

4. 実験

本章では犬視点映像データベースを用いた，動作識別実験について述べる．実験で用いる映像データベースは，2 章で述べたように 5 種類の動作 ((a) ボール取り，(b) 食事，(c) 水飲み，(d) 人になでてもらう，(e) 引っ張り合い)，それぞれ 5 回の試行の犬視点映像から構成される．

表 1〜4 に 4 種類の実験 ((特徴 1) cuboid により特徴を抽出し，勾配情報に基づき特徴を記述した場合，(特徴 2) STIP により特徴を抽出し，HOG に基づき特徴を記述した場合，(特徴 3) STIP により特徴を抽出し，HOF に基づき特徴を記述した場合，(特徴 4) STIP により特徴を抽出し，HOG/HOF に基づき特徴を記述した場合) により得られた混同行列を示す．これらの結果より，水飲みは他の動作と比較して，容易に識別できるが、一方、食事は他の動作と誤識別されやすいことがわかる。

表 1 (特徴 1) cuboid により特徴を抽出し，勾配情報に基づき特徴を記述した場合 [%]

	(a)	(b)	(c)	(d)	(e)
(a) ボール	60	20	0	20	0
(b) 食事	40	40	20	0	20
(c) 水飲み	0	0	100	0	0
(d) なでる	20	0	0	80	0
(e) 引張る	0	0	0	20	80

5. まとめ

本予稿では，これまでの一人称視点映像と異なり，人に代わり動物の視線に着目し，動物にカメラを装着して得られる動物視点の映像を用いた動作認識手法を提案した．また犬に GoPro カメラを装着して，犬視点の画像を撮影し，犬視点映像データベースを構築した．実験ではデータベースに対して提案手法を適用して，提案手法の有効性を示した．

表 2 (特徴 2) STIP により特徴を抽出し，HOG に基づき特徴を記述した場合 [%]

	(a)	(b)	(c)	(d)	(e)
(a) ボール	40	0	0	60	0
(b) 食事	0	20	60	20	0
(c) 水飲み	0	0	100	0	0
(d) なでる	0	0	0	100	0
(e) 引張る	0	0	0	20	80

表 3 (特徴 3) STIP により特徴を抽出し，HOF に基づき特徴を記述した場合 [%]

	(a)	(b)	(c)	(d)	(e)
(a) ボール	0	20	20	60	0
(b) 食事	60	20	0	0	20
(c) 水飲み	0	0	80	20	0
(d) なでる	40	0	0	60	0
(e) 引張る	20	0	20	20	40

表 4 (特徴 4) STIP により特徴を抽出し，HOG/HOF に基づき特徴を記述した場合 [%]

	(a)	(b)	(c)	(d)	(e)
(a) ボール	20	0	0	80	0
(b) 食事	0	0	20	60	20
(c) 水飲み	20	0	80	0	0
(d) なでる	0	0	0	100	0
(e) 引張る	0	0	0	0	100

参考文献

[1] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto: “Fast unsupervised ego-action learning for first-person sports videos”, In CVPR, 2011.

[2] A. Fathi, A. Farhadi, and J. M. Rehg: “Understanding egocentric activities”, In ICCV, 2011.

[3] M. Ryoo and L. Matthies: “First-Person Activity Recognition: What Are They Doing to Me?”, In CVPR, 2013.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie: “Behavior recognition via sparse spatio-temporal features”, In IEEE Workshop on VS-PETS, 2005.

[5] I. Laptev: “On Space-Time Interest Points”, Int. J. of Computer Vision, Vol.64, No.2-3, pp.107-123, 2005.