人間動作のみからの物体の分類

Do Hang Nga[1,a)]　Vu Gia Truong[1,b)]　柳井　啓司[1,c)]

1. はじめに

本研究は物体と人間の作用関係に基づいて動画における物体の分類を行う。物体の分類は物体のシェープやカラーなどといった従来の特徴だけではなく、その物体に関連した人間の動作の特徴によっても実現可能であるということを証明する。人間は物体を操作して動作を行う際、一般的には動くボディパーツは主に手であるという事実に基づいて、手の動きを人間と物体の作用関係の表現とする。手の動き特徴を抽出するためにまず手の検出を行う必要があるが、動画における手検出は非常に挑戦的なタスクだと考えられる。ここではフロー、上半身の位置、スキンカラーおよび手のシェープによる有効的な手の自動検出器を提案する。検出できた手の領域からトラジェクトリー特徴を抽出してその特徴を使って動画の分類を行う。トラジェクトリー特徴としてはデンストラジェクトリー [4] およびここで新しく提案される手トラックを利用する。

提案の手検出器の有効性を検証するには VideoPose2 データセット [*1] を利用する。また、物体の分類の実験には UCF-101 データセット [2] の楽器演奏のグループを利用する。実験の結果によると従来のデンストラジェクトリーに対して、手を検出して，手の動きと手領域から抽出したトラジェクトリーを用いると精度が向上できることがわかる。

2. 提案手法

2.1 提案手法の流れ

提案手法の大まかな流れは次のように纏められる。

(1) 与えられた動画に対し k フレームごとに手検出を行う。
(2) 検出結果を L 連続フレームで追跡する。
(3) 追跡結果から特徴を抽出する。
(4) 各特徴の Fisher ベクトルを作成し、統合したベクトルを動画の表現とする。

実験では $k = 3, L = 15$ とする。Fisher エンコーディングには [7] の手法を適用する。

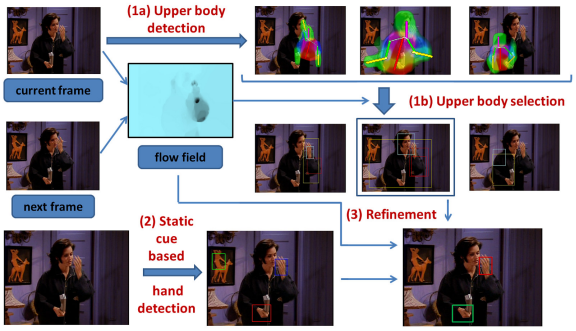

図 1 提案の手検出法の検出例。与えられた動画のフレームに対して、(1a) まず上半身検出器を利用して人間の上半身ポーズの候補を得る。(1b) 次にモーション情報を使って最も重要らしいポーズ（真ん中のポーズ）を選択する。(2)Mittal らの検出器を適用して手の候補を検出する。選択した上半身の顔領域のスキンカラーはこの手検出器に利用される．この検出の上位の結果は最後の行の中央の画像に表す (赤: 1 位、緑: 2 位, 青: 3 位)。(3) 最後にモーション情報および上半身の位置に基づいて得た手検出結果を再スコアする。

2.2 手検出

ここで動画における手の領域の自動推定手法を提案する。モーション情報および上半身検出結果を用いて Mittal らによって提案された静止画像における手検出器 [3] を改良する。フローを計算するには DeepFlow[5]、上半身を検出するには Calvin 検出器 [1] を利用する。提案の手検出法の流れは図 1 にて説明する。

検出結果の再スコアは次の式に従って行う。

$$s = s_m + w_f * s_f + w_u * s_u \tag{1}$$

ここで s_m は Mittal らの検出器によるスコアを示す。s_f はフローによるスコアを示し、検出結果領域のピクセルのフローの勾配の平均として計算される。このスコアは値が 0 から 1 までの間にあるように正規化される。また、s_u は上半身によるスコアを示し、検出結果領域の上半身のオーバーラップの割合として計算される。w_f と w_u は重みであり、実験結果によって w_f と w_u の最適の値はそれぞれ 0.7 と 0.2 であることがわかった。

2.3 手追跡

計算コストを減らすために k フレームごとに手検出を行った。その結果を連結するおよび信頼性が高い結果を探すために手追跡を行う。検出が行われたフレームに対して上位 h の検出結果を前方の L フレームでトラッキングする。結果の領域のフローの平均を使って結果を次のフレー

[1] 電気通信大学　〒 182–0021 東京都調布市調布ヶ丘 1–5–1
[a)] dohang@mm.cs.uec.ac.jp
[b)] vugia@mm.cs.uec.ac.jp
[c)] yanai@cs.uec.ac.jp
[*1] http://vision.grasp.upenn.edu/cgi-bin/index.php?n=VideoLearning.VideoPose2

表 1 手検出の結果。Our (+upper body)、Our (+flow)、Our (+flow+body) はそれぞれ Mittal らの検出器に選択した顔が使用された結果、モーション情報を再スコアに使用された結果、提案システムの全体が適用された結果を示す。

Method	Precision
Mittal ら [3]	41.7%
Our (+upper body)	42.6%
Our (+flow)	45.5%
Our (+flow+body)	46.3%

図 2 手検出結果例。 各フレームの上位 2 検出結果を表す（赤：1 位、緑：2 位)。上の 2 行は VideoPose2.0 の結果例である（一番目の行：ベースライン、2 番目の行：提案手法)。最後の行は UCF-101 の楽器演奏の検出結果例を表す。

ムに移す。使用するとするトラックは少なくても n フレームの上位の検索結果と 50%以上オーバーラップする。実験では $h = 2, L = 15, k = 3, n = 2$ とする。

2.4 特徴抽出

手トラックは手の中心点のトラジェクトリーのことである。ポイント P の長さ L のトラジェクトリーのシェープはそのポイントの変位ベクトルの $\Delta P_t = (P_{t+1} - P_t)$ シーケンス $S = (\Delta P_t, ..., \Delta P_{t+L-1})$ によって表される。ここで $P_t = (x_t, y_t)$ はフレーム t におけるポイント P の座標を示す。特徴のディスクリプタは次の式によって計算される。

$$S' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||} \quad (2)$$

上記の手トラック特徴に加えて手領域のデンストラジェクトリー [4] も抽出する。

3. 実験と結果

3.1 手検出の実験

VideoPose2.0 データセットを用いて提案の手検出器の有効性を検証する。このデータセットはムービーソースから収集された 14 ビデオショットを含む。全部で 2453 フレームがあり、3814 手がアノテーションされた。アノテーションは我々に行われた。検出結果 B_d のオーバーラップスコアは $O = \frac{area(B_g \bigcap B_d)}{area(B_g \bigcup B_d)}$ として定義される。ここで B_g はグラウンドトゥルースを示す。結果は表 1 にて表す。ベースラインは Mittal らの検出器とする。結果例は図 2 に示す。実験結果によるとモーションと人間検出の導入によって検出結果が改良された。

3.2 動画分類の実験

ここで UCF-101 の楽器演奏のグループのデータに実験を行った。このグループには 10 種類の楽器：cello、guitar、violin、daf（大型の枠太鼓)、dhol（両面太鼓)、piano、tabla、sitar（弦楽器の一種)、flute と drum の演奏のシーンを含む 1428 ビデオショットがある。実験の評価は ICCV2013 ワークショップ [*2] に従って行う。分類にはマルチクラス SVM[6] を学習する。ここでのベースラインは Wang らのデンストラジェクトリー [4] とする。実験結果は表 2 にて表す。実験結果から提案手法によってより有用な特徴が検出され、物体をよりよく分類できたことがわかる。

表 2 分類の結果。Our (Dense trajectories on hands)、Our (Hand tracks)、Combined feature はそれぞれ検出された手領域に限られて抽出されたデンストラジェクトリー、提案の手トラック特徴、統合特徴を利用した結果を示す。

Method	Precision
Wang et al. [4](Dense trajectories)	58.7%
Our (Dense trajectories on hands)	61.5%
Our (Hand tracks)	33.4%
Our (Combined feature)	64.5%

4. おわりに

人間の操作による物体の分類法を提案した。また動画における有効な手自動検出器を構成した。今回は我々の研究目的に着目した公開のデータは UCF-101 の楽器演奏のデータのみであるため楽器分類しか行わなかった。今後はデータを収集して他の種類の物体にも提案手法の有効性を検証する予定である。

参考文献

[1] Eichner, M. and Ferrari, V.: Better Appearance Models for Pictorial Structures, *British Machine Vision Conference* (2009).

[2] Khurram, S., Amir, R. and Mubarak, S.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, *CoRR*, vol. abs/1212.0402 (2012).

[3] Mittal, A., Zisserman, A. and Torr, P. H.: Hand detection using multiple proposals., *Proc. of British Machine Vision Conference*, pp. 1–11 (2011).

[4] Wang, H., Klaser, A., Schmid, C. and Liu, C.-L.: Dense Trajectories and Motion Boundary Descriptors for Action Recognition, *International Journal of Computer Vision*, Vol. 103, No. 1, pp. 60–79 (2013).

[5] Weinzaepfel, P., Revaud, J., Harchaoui, Z. and Schmid, C.: DeepFlow: Large displacement optical flow with deep matching, *Proc. of IEEE International Conference on Computer Vision* (2013).

[6] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables, *The Journal of Machine Learning Research*, Vol. 6, pp. 1453–1484 (2005).

[7] Perronnin, F., Sanchez, J. and Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification, *eccv*, Lecture Notes in Computer Science, Vol. 6314, pp. 143–156 (2010).

[*2] http://crcv.ucf.edu/ICCV13-Action-Workshop/