Vol.2018-CVIM-210 No.31
2018/1/19

パーツ領域限定Dense TrajectoriesとLSTMを用いた行動認識

山田 花穂[1,a)]　金子 直史[1]　鷲見 和彦[1]

概要：本研究では，人体パーツ周辺に領域を限定した Dense Trajectories (DT) と Long Short Term Memory (LSTM) を用いた行動認識手法を提案する．従来の DT 特徴を用いた行動認識手法では，人物を含む画像全体から特徴を抽出しているため，身体の一部分だけの動きが特徴となる行動を捉えにくく，また特徴を識別する時間区間が一定長に区切られていることから，状態の遷移が特徴となる行動を識別できない問題があった．そこで人物の主要な身体パーツの周辺に検出領域を設定し，その領域から DT 特徴を抽出する．さらに時間方向の状態の遷移を特徴として識別できる LSTM を用いることにより，物を投げるや手を振るなど類似した細かい動きを，動作全体を通して正確に認識することを可能にした．

1. はじめに

防犯意識の高まりを背景に，公共施設や商業施設など様々な場所に防犯カメラの導入が進んでいる．その結果，警備員などが監視を担当するカメラの台数が増加して，犯罪等を見落とすリスクや長時間監視による負担が増大している．このような問題を解決するため，映像から自動的に人物の行動や状況を認識するシステムが望まれている．

人物の行動を自動的に認識する研究は従来から盛んに行われており，代表的な研究として Wang らの Dense Trajectories (DT) [1] や Simonyan らの Two-stream Convolutional Neural Network (CNN) [2] が挙げられる．Wang らは，入力動画に対して密な特徴点追跡を行うことで得られる軌跡を特徴として用いることにより，行動を認識する手法を提案した．また Simonyan らは，空間方向の CNN と時間方向の CNN を特徴として用いることにより，行動を認識する手法を提案した．しかし，これらの手法は画像全体から特徴を抽出しているため，身体の一部分だけの動きが特徴となる行動を捉えにくく，また特徴を識別する時間区間が一定長に区切られていることから，状態の遷移が特徴となる行動を識別できない問題がある．

そこで本研究では人物の主要な身体パーツの周辺に検出領域を設定し，その領域から Dense Trajectories を抽出する．さらに時間方向の状態の遷移を特徴として識別できる LSTM [3] を用いることにより，細かい動きが動作全体を通して正確に認識されることを可能にする．

[1] 青山学院大学
[a)] yamada.kaho@vss.it.aoyama.ac.jp

2. 関連研究

人物の行動を自動的に認識する手法には，人が設計した特徴 (Hand-crafted 特徴) を用いて行動を認識する手法と，深層学習により自動的に生成された特徴 (Deeply-learned 特徴) を用いて行動を認識する手法の 2 つに分けられる．

2.1 Hand-Crafted 特徴を用いた行動認識手法

Hand-crafted 特徴を用いる手法には Wang らの Dense Trajectories が挙げられる．この手法では，入力動画に対して密な特徴点追跡を行うことで得られる軌跡の形状特徴と，その周辺領域から Histograms of Oriented Gradients (HOG)，Histograms of Optical Flow (HOF)，Motion Boundary Histograms (MBH) [4] 特徴を抽出する．抽出された 4 つの特徴は Bag of Features (BoF) により組み合わされ，組み合わせた特徴を識別器 Support Vector Machine (SVM) を用いて学習することにより，行動の識別を行う．Dense Trajectories は Kanade-Lucas-Tomasi (KLT) tracker [5] や SIFT descriptors [6] によって得られた軌跡と比較して密な軌跡を抽出することができる．そのため身体全体の動きに関する特徴を，より正確に捉えることが可能となり，高精度な行動認識を実現した．

また Wang らは, Dense Trajectories を改良した Improved Dense Trajectories (IDT) を提案した [7] ．この手法は，Dense Trajectories を用いて軌跡を抽出した後，SURF と Random Sampling Consensus (RANSAC) [8] を用いてカメラの動きを推定することによって，カメラの動きによる余分な背景の軌跡を取り除く．また，抽出された 4 つの特

徴は Fisher Vector (FV) [9] を用いて組み合わされる．この手法によって，人物や物体の動きだけに着目した特徴を抽出することが可能となり，Dense Trajectories と比較して高精度な行動認識に成功した．

2.2 Deeply-Learned 特徴を用いた行動認識手法

Deeply-learned 特徴を用いる手法には Simonyan らの Two-stream CNN が挙げられる．この手法では，入力動画の RGB 画像と複数枚の Optical Flow 画像に対して個別の CNN を用いて特徴を抽出し，それぞれの CNN から得られた識別結果を統合することにより行動を認識する．RGB 画像から CNN 特徴を抽出することにより物体や背景のアピアランスの特徴を捉え，Optical Flow 画像から CNN 特徴を抽出することにより動きに関する特徴を捉えることが可能となった．

また，Sharma らは入力動画の RGB 画像に対して CNN を用いて特徴を抽出し，その特徴を LSTM によって学習させることで，行動を認識する手法を提案した [10]．LSTM は特徴を長期的に保持した状態で行動を識別することができるため，RGB 画像から CNN 特徴を抽出することで得られるフレームごとの物体や背景のアピアランス特徴を，動画全体を通して正確に捉えることが可能となった．

Deeply-Learned 特徴を用いた行動認識手法は近年盛んに研究されているが，高精度な認識を行うためには膨大な動画データを用いて特徴を学習させる必要があり，未だ十分な量の学習データセットは集められていない．

2.3 従来の行動認識手法の問題点

関連研究で挙げたこれらの手法は，画像全体から特徴を抽出しているため，身体の一部分だけの動きが特徴となる行動を認識しにくい問題が挙げられる．また，Dense Trajectories や CNN によって得られる特徴は，識別する時間区間が短く，一定長に区切られているため，状態の遷移が特徴となる行動を識別できない問題がある．

3. 提案手法

前節で述べた問題点に対して本研究では，人物の主要な身体パーツの周辺に検出領域を設定し，その領域から Dense Trajectories を抽出することにより，大域的な領域だけでなく局所的な領域からも動きに関する特徴を抽出できるようにする．さらに得られた特徴を LSTM によって識別することで，時間方向の状態の遷移を特徴として識別することを可能にする．この手法により，類似した細かい動きを，動作全体を通して正確に認識することを目指す．

提案手法の流れを図 1 に示す．提案手法は身体パーツ領域の設定と，特徴の抽出・識別の大きく 2 つの処理から成る．

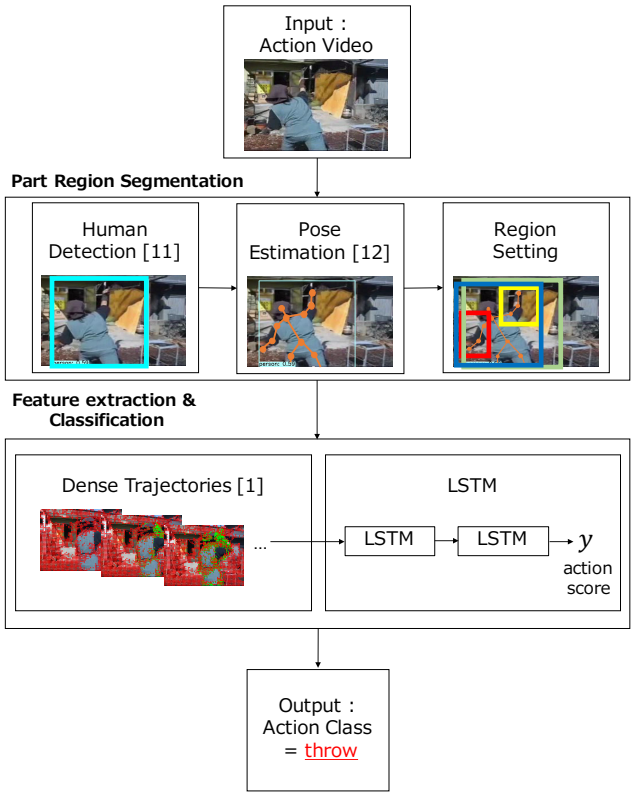


図 1: 提案手法の概要

3.1 身体パーツ領域の設定

1 つ目の処理では，入力された動画から右腕，左腕，上半身，全身の 4 つの身体パーツ領域を設定する．これらの領域の設定方法を図 2 に示す．まず動画の各フレームに対して，人物検出手法を用いて図 2 の赤枠のように全身領域を設定する．次に姿勢推定手法を用いて青枠のように関節位置情報を取得する．最後に姿勢推定結果から，右腕，左腕，上半身のパーツに該当する関節位置座標の最大・最小値を計算し，緑枠のように残りの 3 つのパーツ領域を設定する．ここで，全身領域の設定に姿勢推定でなく人物検出を用いる理由は，姿勢推定では下半身の関節位置の推定精度が低く，正確に全身領域を設定することが難しいためである．例えば，図 3 の画像に映る人物に対して，姿勢推定を用いると図 3(a) のように足の関節位置の推定に失敗するが，人物検出を用いると図 3(b) のように全身領域を正確に捉えることができる．

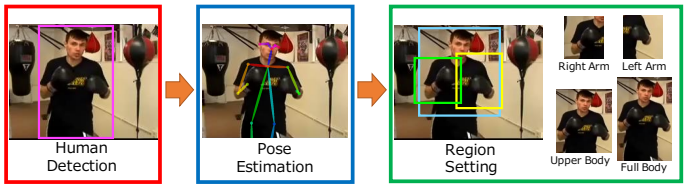


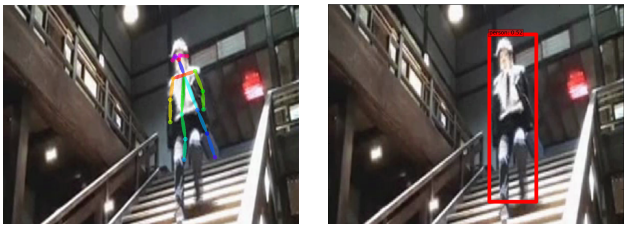
図 2: 身体パーツ領域の設定方法

人物の全身領域の設定には Liu らの物体検出手法 [11] を用いる．Liu らは物体の領域とクラスを単一のネットワー

クを用いて予測し，入力画像から物体を検出する手法を提案した．この手法は他の物体検出手法と比較して，低解像度の画像からも高精度に物体を検出することができる．本手法では，検出する物体クラスを人物のみに限定し，人物検出手法として用いる．行動クラス *climb stairs* の画像に人物検出を適用させた結果を図 3(b) に示す．図 3(b) から，検出された人物領域の範囲をそのまま全身領域として設定した場合，動きの特徴を捉えるには少し狭いことが分かるため，領域の幅と高さを 30%拡張し，この領域を全身領域として設定する．

残りの 3 つのパーツ領域を設定するために Cao らの姿勢推定手法 [12] を用いて，図 4(a) のように 18 個の関節位置を取得する．Cao らは入力画像から 2 つの CNN ネットワークを用いて，部位の位置をエンコードする Part Confidence Maps と，部位間の関連度をエンコードする Part Affinity Fields を計算し，それらを組み合わせることで画像内に映る複数人の姿勢をリアルタイムに推定する手法を提案した．この手法は，現在最も高精度に複数人の姿勢を推定することができる．また，本手法では 18 個の関節位置の取得後，各関節位置に対してカルマンフィルタを用いて追跡を行い，時刻 $t-1$ までのデータから時刻 t のときの位置を推定することで，動画全体を通して安定した推定を可能にする．

最後に，出力された関節位置座標を基に図 4(b) のように右腕，左腕，上半身の 3 つの領域を設定する．ただし，図 4(b) の領域では，実際のパーツ領域よりも範囲が狭いため，上半身の幅，右腕・左腕の幅と高さを 60%拡張し，上半身の高さは 80%拡張することにより，図 4(c) のような領域を最終的なパーツ領域として設定する．

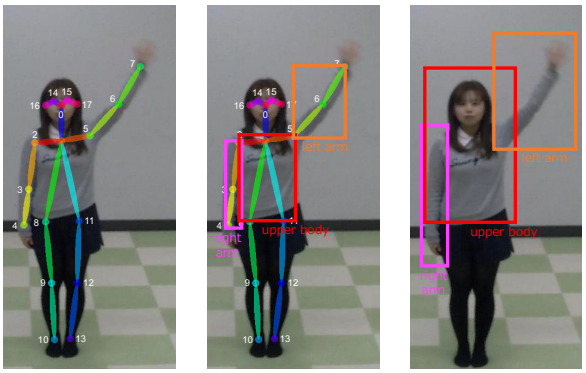


(a) 姿勢推定を用いた結果　(b) 人物検出を用いた結果

図 3: 姿勢推定・人物検出の結果による全身領域の違い

3.2 特徴の抽出と識別

2 つ目の処理では，設定した 4 つのパーツ領域から特徴を抽出し，行動を識別する．特徴抽出の流れを図 5 に示す．まず Wang らの Dense Trajectories を用いて 10 フレームごとに軌跡特徴を抽出する．これに伴い，パーツ領域は 10 フレーム分の関節位置座標を基に作成し，得られた領域から Dense Trajectories を抽出する．行動クラス *pour* の右腕の領域から Dense Trajectories を用いて抽出した軌跡を描画したものを図 6(b) に表す．赤色の点が特徴点，緑色の線が軌跡を表している．得られた特徴は FV によってエンコーディングを行い，その後各パーツ領域の特徴を連結する．これを時間ごとに LSTM に入力することによって，特徴の抽出と，行動の認識を行う．



(a) 姿勢推定結果　(b) パーツ領域　(c) 拡張パーツ領域

図 4: 3 つの身体パーツ領域の設定方法

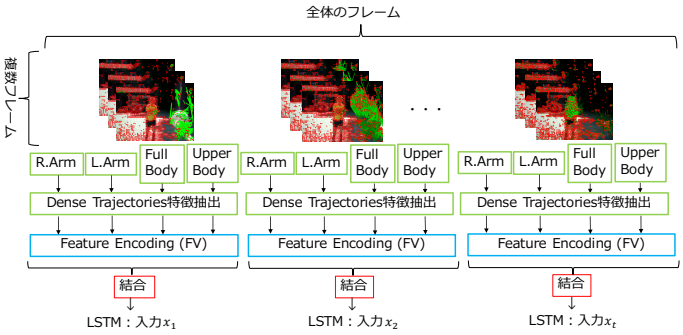
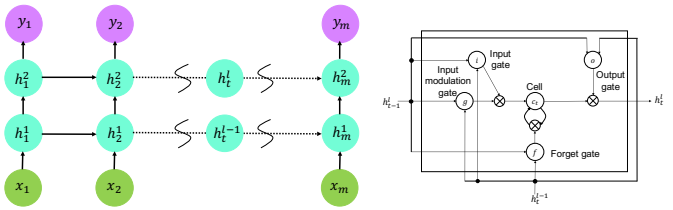


図 5: 特徴抽出の流れ

(a) 元画像　(b) 元画像の前後 10 フレームから得られる軌跡

図 6: 右腕から抽出される Dense Trajectories の軌跡

LSTM のネットワークは図 7 のように設計する．図 7(a) は全体のネットワークを表しており，2 層の LSTM ユニットから成る．図 7(a) の入力 x は Dense Trajectories によって抽出された特徴量であり，出力 y は行動クラスに属する確率を表し，h は LSTM ユニットを表す．最後の 10 フレームの特徴 x を入力した後に得られる出力 y を最終的な行動スコアとして行動を予測する．LSTM ユニットの内部構造は図 7(b) に示す．



(a) 全体のネットワーク　(b) LSTM ユニットの内部構造

図 7: LSTM のネットワーク構造

4. データセット

実験には JHMDB [13] と MPII Cooking Activities [14] の 2 つのデータセットを用いた．それぞれについて次節で詳しく説明する．

4.1 JHMDB

JHMDB [13] は HMDB [15] の一部であり，*brush hair*, *pick*, *golf*, *run*, *sit* など 21 クラスの人物行動を含むビデオデータセットである．ビデオは行動しているシーンのみに制限されており，1 つの行動につき 36 から 55 本の動画が含まれる合計 928 本で構成される．動画のフレームサイズは 320 × 240 ピクセルであり，その全てに人物の関節位置のアノテーションが付けられている．

4.2 MPII Cooking Activities (MPII Cook)

MPII Cooking Activities は 64 クラスの行動を含むビデオデータセットである．ビデオは全てキッチンで撮影されているため，背景はどの動画も同じである．人物行動は *cut dice*, *cut slices*, *cut stripes*, *wash hands*, *wash objects* など，動作が細かく類似した行動クラスが含まれており，合計 5609 本で構成される．動画のフレームサイズは 1624 × 1224 ピクセルであり，その全てに人物の関節位置のアノテーションが付けられている．

5. 実験

提案手法の有効性を示すために，前章で述べた 2 つのデータセットを用いて以下のような実験を行った．

5.1 パーツ領域限定特徴による行動認識精度の比較

はじめにパーツ領域限定を用いた Dense Trajectories によって行動認識精度がどの程度向上するのかを調査するため，提案手法，提案手法からパーツ領域限定を除いた手法，Ground-Truth (GT) の関節位置を用いてパーツ領域を設定した手法の 3 つの行動認識精度を JHMDB データセットを用いて比較した．この結果を表 1 に示す．表 1 から，パーツ領域限定を用いることにより行動認識精度が 8%程度向上することが分かった．これはパーツだけの特徴を抽出可能になったことにより，類似した細かい行動を認識できるようになったためである．認識できるようになった細かい行動についての詳細は次節に示す．

また，表 1 から姿勢の誤推定によって行動認識精度が 4%程度下がってしまうことも分かった．本研究で用いた姿勢推定手法は，他の手法と比較して画像内に映る人物のサイズが小さい場合 (図 8(a)) や，上半身のみしか映っていない場合 (図 8(b)) でも姿勢を正確に推定することができる．しかし，図 8(c) のように人物が複雑な姿勢をしている場合や，図 8(d) のように身体パーツと間違え得る紛らわしい物体が存在する場合等では正しく姿勢が推定できない．この場合，パーツ領域は間違った場所に設定され，Dense Trajectories を正しい領域から抽出できないため，他の行動クラスと誤認識されやすくなり，全体の行動認識精度も下がってしまった．

表 1: パーツ領域限定特徴による行動認識精度の比較

Method	JHMDB (%)
DT+FV+LSTM	61.4
パーツ領域限定+DT+FV+ LSTM	**69.8**
パーツ領域限定 (GT) +DT+FV+LSTM	**73.5**

(a) 成功例：*shoot ball*　(b) 成功例：*pour*

(c) 失敗例：*stand*　(d) 失敗例：*shoot bow*

図 8: 姿勢推定の成功例と失敗例 (JHMDB)

5.2 既存の行動認識手法との精度比較

次に既存の行動認識手法 (Dense Trajectories, Improved Dense Trajectories) と提案手法の精度を JHMDB と MPII Cook データセットの 2 つを用いて比較した．この結果を表 2 に示す．また JHMDB データセットを用いて Improved Dense Trajectories と提案手法の行動クラスごとの認識精度を比較した結果を図 9 に示す．表 2 から Dense Trajectories と行動認識精度を比較して JHMDB では 14%程度，MPII Cook では 12%程度上回ることができた．ま

た Improved Dense Trajectories と比較しても JHMDB で 4%程度，MPII Cook で 1%程度上回ることができた．これは前節でも述べたようにパーツ領域限定を用いることにより，パーツごとの細かい動きの特徴を捉えることが可能となったからである．実際に図 9 から *throw* や *shoot ball*，*clap* や *wave* など腕の動きが重要となる行動の精度は Improved Dense Trajectories と比較して大幅に向上していることが分かる．しかし，姿勢の誤推定によって精度が下がってしまった行動もあるため，全体の精度を大きく向上させることはできなかった．また，MPII Cook データセットはいくつかのサブアクションを含む行動クラスが多く含まれる．例えば *open close cupboard* は食器棚を開ける，食器棚を閉めるという 2 つのサブアクションで構成され，*take put in oven* はオーブンの扉を開ける，オーブンに食べ物を入れる，オーブンの扉を閉めるという 3 つのサブアクションで構成される．この場合，サブアクションの遷移特徴を捉えることが重要となり，LSTM を用いることによって，これら 2 つの行動であれば，Improved Dense Trajectories と比較して 10%以上行動認識精度を向上させることができた．

MPII Cook が JHMDB よりも精度向上が低い理由は，図 10(a)，図 10(c) のように身体パーツの隠蔽が多く，その場合でも隠蔽されているパーツの関節位置を予測して領域が設定されるため (図 10(b)，図 10(d))，他のパーツの特徴を隠蔽パーツの特徴として抽出してしまい，別の行動クラスとの誤認識が生じてしまうケースが多かったためであると考えられる．

さらに提案手法で誤認識する行動クラスの一例を図 11，図 12 に示す．*sit* と *pick* は図 11 のようにどちらもかがめる動きをしており，軌跡特徴が類似したものになる．この場合，動き特徴のみでは正しい行動クラスを見分けることは困難である．*sit* であれば椅子，*pick* であれば拾う物体が認識できれば 2 つを見分けることが容易になると考えられ，動き特徴に加えて物体や背景のアピアランス特徴も抽出することが重要になると考えられる．また *cut slices* と *cut dice* は図 12 のように全体を通して見た目にも動きにも大きな違いのない動画が多く存在した．この場合，短時間のわずかな動きでも見逃さずに認識できる方法が必要になると考えられる．

表 2: 提案手法と既存手法の行動認識精度比較

Method	JHMDB (%)	MPII Cook (%)
DT [1]	55.7	56.9
IDT [7]	65.8	67.6
Our method	**69.8**	**68.2**

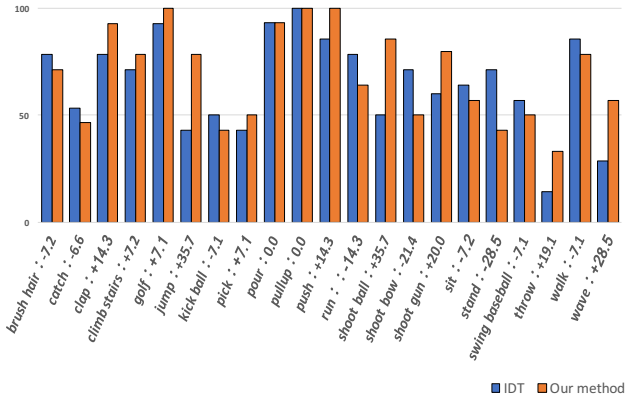


図 9: 各行動クラスの精度比較 (JHMDB)

(a) 元画像：*take put in spice holder*

(b) 姿勢推定結果：*take put in spice holder*

(c) 元画像：*take out from oven*

(d) 姿勢推定結果：*take out from oven*

図 10: 身体パーツが隠蔽された時の姿勢推定結果 (MPII Cook)

6. おわりに

本研究では，パーツ領域を限定した Dense Trajectories と LSTM を用いた行動認識手法を提案した．実験の結果，既存手法である Dense Trajectories，Improved Dense Trajectories と比較して，提案手法の方が行動認識精度を上回ることができた．この結果から，パーツ領域を設定し，設定した領域ごとに特徴を抽出することにより，類似した細かい行動でも認識できるようになったと言える．また LSTM の導入により時間方向の状態の遷移を特徴として識別することで，複数のサブアクションを含む行動を認識できるようになった．今後の課題は，動きの特徴だけでなく，物体や背景などのアピアランス特徴も抽出することで，より高精度な行動認識を行うことである．

(a) 行動クラス *sit* の 10 フレームごとの画像変化

(b) 行動クラス *pick* の 10 フレームごとの画像変化

図 11: 誤認識する行動クラス (JHMDB)

(a) 行動クラス *cut slices* の 10 フレームごとの画像変化

(b) 行動クラス *cut dice* の 10 フレームごとの画像変化

図 12: 誤認識する行動クラス (MPII Cook)

参考文献

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

[2] K.Simonyan and A.Zisserman: Two-stream convolutional networks for action recognition in videos, *NIPS*, pp. 568–576 (2014).

[3] S.Hochreiter and J.Schmidhuber: Long short-term memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).

[4] N.Dalal, B.Triggs and C.Schmid: Human Detection using Oriented Histograms of Flow and Appearance, *ECCV*, pp. 428–441 (2006).

[5] P.Matikainen, M.Hebert and R.Sukthankar: Trajectons: Action recognition through the motion analysis of tracked features, *ICCV Workshops*, pp. 514–521 (2009).

[6] J.Sun, X.Wu, S.Yan, L.-F.Cheong, T.-S.Chua and J.Li: Hierarchical spatio-temporal context modeling for action recognition, *CVPR*, pp. 2004–2011 (2009).

[7] H.Wang and C.Schmid: Action recognition with improved trajectories, *ICCV*, pp. 3551–3558 (2013).

[8] M. A.Fischler and R. C.Bolles: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM*, Vol. 24, No. 6, pp. 381–395 (1981).

[9] F.Perronnin, J.Sánchez and T.Mensink: Improving the fisher kernel for large-scale image classification, *ECCV*, pp. 143–156 (2010).

[10] S.Sharma, R.Kiros and R.Salakhutdinov: Action recognition using visual attention, *arXiv preprint arXiv:1511.04119* (2015).

[11] W.Liu, D.Anguelov, D.Erhan, C.Szegedy, S.Reed, C.-Y.Fu and A. C.Berg: Ssd: Single shot multibox detector, *ECCV*, pp. 21–37 (2016).

[12] Z.Cao, T.Simon, S.-E.Wei and Y.Sheikh: Realtime multi-person 2d pose estimation using part affinity fields, *arXiv preprint arXiv:1611.08050* (2016).

[13] H.Wang and C.Schmid: Towards understanding action recognition, *ICCV*, pp. 3192–3199.

[14] M.Rohrbach, S.Amin, M.Andriluka and B.Schiele: A database for fine grained activity detection of cooking activities, *CVPR*, pp. 1194–1201 (2012).

[15] H.Kuehne, H.Jhuang, E.Garrote, T.Poggio and T.Serre: HMDB: a large video database for human motion recognition, *ICCV*, pp. 2556–2563 (2011).