2G1-3

FV-HMM/MKL-SVM を用いた 局所スケルトン特徴の選択・統合による多クラス運動認識

FV-HMM/MKL-SVM using Local Skeleton Features for Multi-class Motion Recognition

郷津　優介 *1　高野　渉 *1　中村　仁彦 *1
Yusuke Goutsu　Wataru Takano　Yoshihiko Nakamura

*1東京大学大学院　情報理工学系研究科　知能機械情報学専攻
Department of Mechano-Informatics, Graduate School of Information Science and Technology, The University of Tokyo

Multi-class motion recognition is one of the most important problem to solve. We have proposed a framework which is available to classify multi-class motion and improves the accuracy of motion recognition. In this system, skeleton features, which consist of spatio-temporal data of position, speed and acceleration calculated by Inverse Kinematics, are derived from several combinations of local joints in human body and then are represented as motion features by Fisher Vector parameterized by Hidden Markov Model. The kernel of motion features are selected and integrated in response to motion target by learning parameters of Multiple Kernel Learning and Support Vector Machine in the same time. This approach makes it possible for robots to recognize various human motions of our daily life. The experiments demonstrates the availability of FV-HMM/SVM and this indicates the capability of FV-HMM/MKL-SVM.

1. はじめに

いわゆる人工知能を持った機械の出現により人間と機械の関係が大きな転換点を迎えている．例えば，プロセッサ速度の劇的な向上，ビッグデータの利用，NUI を実現したデバイスの登場などを要因として，マウスやキーボードのようにある程度の訓練を必要とする人間が機械に合わせる関係からジェスチャーなどを使って直感的に機械を操作できる機械が人間に合わせる関係へとシフトしている．このような関係において，機械が日常生活における人間の様々な行動やジェスチャー指令を理解して行動支援に繋げていくことは重要である．

行動認識で解決すべき課題の 1 つとして，多クラスの運動認識が挙げられる．例えば，100 クラス以上の運動認識は世界的にほとんど前例のない研究であり，それ故に重要な課題であると言える．ここで，運動とは単一モーダルなデータから構成されるものとし，これに対して行動は運動情報も含めた複数モーダルで得られるデータ源であると定義する．先の文脈においても機械が人間の多種多様な運動を識別することで行動支援が実現されていくと言える．また，筆者らの先行研究 [Goutsu 14] においても識別する運動のクラス数が増えると，それに比例して元々高次元であった特徴ベクトルがさらに高次元となり，計算速度やメモリ容量などの面で破綻するという問題があった．

本稿で提案する手法は，スケルトン情報から運動に関連する局所的なジョイント列を選択し，それらの位置に関する高次の微分情報で記述されたスケルトン特徴の時系列情報を隠れマルコフモデル (Hidden Markov Model:HMM) で学習し，HMM パラメータに基づくフィッシャーベクトル (Fisher Vector:FV) で表現することにより各ジョイント列に対応した運動特徴を作成する．作成された特徴をマルチカーネル学習 (Multiple Kernel Learning:MKL) により重み付け統合し，これにより最終的なクラス識別を行う（図 1 参照）．また，多クラスの運動認識に対応し且つ識別に有効な特徴量の設計を検討する．

連絡先: 郷津優介，東京大学大学院情報理工学系研究科，〒 113-8656 東京都文京区本郷 7-3-1，Tel: 03-5841-6381，Fax: 03-3818-0835，Email: goutsu@ynl.t.u-tokyo.ac.jp

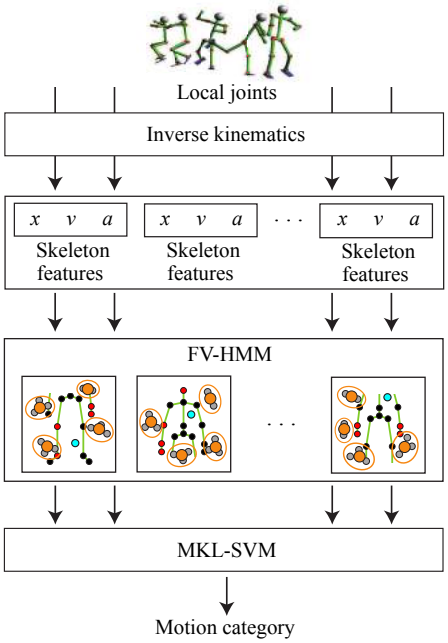

図 1: Overview of our proposed system(FV-HMM/MKL-SVM) for multi-class motion recognition.

2. 多クラス運動認識システム

現在，パターン認識の分野などで様々な行動認識の研究が行われている．Kinect から得られる深度画像を用いた人間の 3 次元姿勢推定の研究により，画像と奥行き情報に続く第 3 のモーダルとして人間のスケルトン情報まで取得できるようになった．これに伴い，色画像，シルエット画像 [Li 10]，深度画像 [Oreifej 13]，スケルトン情報 [Wang 12] [Zanfir 13] [Evangelidis 14]，時空間占有率 [Vieira 12] などのデータが行動認識の特徴量として利用されるようになり，先行研究を比較した場合にスケルトン特徴を用いた手法は識別率が高くなる傾向にあると言え，本稿

でもこのアプローチを採用している．また，行動認識過程におけるスケルトン特徴抽出後の処理には類似度計算をフレーム単位で行う方法 [Zanfir 13] とまとまった動作単位で行う方法 [Wang 12][Evangelidis 14] の 2 つのアプローチがある．後者では，行動データのセグメンテーションが必要になる場合があり，これに関して同じラベルを持つと推定される連続フレーム群にまとめる手法や時系列情報の変化点を検出する手法などがある．本稿では，スケルトン特徴の時系列情報で記述される人間の身体運動を HMM による離散的な運動記号として表現学習しているため，後者の立場を取っている．

2.1 局所スケルトン特徴

先述のように，スケルトン情報で構成された特徴を用いることで，識別率は比較的に高くなる傾向にある．ここでは，ジョイントの位置に基づいた情報をスケルトン特徴として用いる．また，スケルトン特徴に関して，スケルトン全体の大局的なジョイント列よりも行動と密接に関連した局所的なジョイント列を用いた方が良いこと [Wang 12][Evangelidis 14] や，ジョイントの位置情報の他に，姿勢は似ているが方向の異なる運動（立つと座る）に対して速度，速さと方向の異なる運動（円を描くと直線を描く）に対して加速度の情報も組み合わせて用いることで，識別に有効に働くことが知られている [Zanfir 13]．このことは，慣性などの物理的な制約条件を考慮することで，身体運動の時系列情報が以下のような二次のテイラー展開（二次関数）で近似できることにも表れている．

$$\boldsymbol{o}(t) \approx \boldsymbol{o}(t_0) + \delta\boldsymbol{o}(t_0)(t-t_0) + 1/2\delta^2\boldsymbol{o}(t_0)(t-t_0)^2 \quad (1)$$

ここで，$\boldsymbol{o}(t_0)$，$\delta\boldsymbol{o}(t_0)$，$\delta^2\boldsymbol{o}(t_0)$ は時刻 t_0 における位置，速度，加速度で構成されるスケルトン情報をそれぞれ表し，高次の微分情報を含むことで時刻 t_0 周辺の姿勢変化まで捉えていることを意味する．

本稿では，ジェスチャー認識のような上半身の動作だけの場合，局所的なジョイント列として表 1 に示すような 4 つのジョイントで構成される全 24 種類を考えている．ここで，表中のジョイント名は Kinect のマーカー点に付けられた名前と一致させており，L, C, R はそれぞれ Left, Center, Right を意味する．列の組み合わせはヒューリスティックに決めているが，この方法でも識別性能はそれほど落ちないことが [Evangelidis 14] で知られている．日常生活における動作の場合には，全身のジョイントを使った列の組み合わせを考える必要がある．また，ジョイント列のスケルトン特徴として，各ジョイントの位置，速度，加速度を縦に連結した特徴ベクトルを用いることにする．ここで，ジョイントの速度，加速度はジョイントの 3 次元位置から逆運動学計算 (Inverse Kinematics:IK) を行うことで求めている．

2.2 FV-HMM

人間の身体運動は時系列情報であり，信号の空間的ずれや時間方向の伸縮・移動などのゆらぎにロバストな HMM により学習し，離散的な運動記号を獲得する．これは隠れ状態の集合 $\boldsymbol{Q}$，状態遷移確率行列 $\boldsymbol{A}$，出力確率分布の集合 $\boldsymbol{B}$，初期状態確率の集合 $\boldsymbol{\pi}$ とした場合に，以下の 4 つのパラメータの集合 $\boldsymbol{\lambda}$ で表現される．

$$\boldsymbol{\lambda} = \{\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\} \quad (2)$$

運動記号 $\boldsymbol{\lambda}$ が時系列の運動情報 $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T\}$ を生成する確率を $P(\boldsymbol{O}|\boldsymbol{\lambda})$ とすると，この尤度が最大になるように EM アルゴリズムの一種である Baum-Welch アルゴリズムにより $\boldsymbol{\lambda}$ の最適化計算を行う．ここでの確率計算には Forward-Backward アルゴリズムを利用している．

表 1: 24 combinations of 4 joints

No.	J_1	J_2	J_3	J_4
1	ShoulderC	Head	ShoulderL	ElbowL
2	ShoulderC	Head	ElbowL	WristL
3	ShoulderC	Head	WristL	HandL
4	ShoulderC	Head	ShoulderR	ElbowR
5	ShoulderC	Head	ElbowR	WristR
6	ShoulderC	Head	WristR	HandR
7	ShoulderC	ShoulderL	ElbowL	WristL
8	ShoulderC	ShoulderL	WristL	HandL
9	ShoulderC	ShoulderR	ElbowR	WristR
10	Head	ShoulderL	ElbowL	WristL
11	Head	ShoulderL	WristL	HandL
12	Head	ElbowL	WristL	HandL
13	Head	ShoulderR	ElbowR	WristR
14	Head	ShoulderR	ElbowR	WristR
15	Head	ShoulderR	WristR	HandR
16	Head	ElbowR	WristR	HandR
17	ShoulderL	ElbowL	WristL	HandL
18	ShoulderL	ElbowL	ShoulderR	ElbowR
19	ShoulderL	WristL	ShoulderR	WristR
20	ShoulderL	HandL	ShoulderR	HandR
21	ElbowL	WristL	ElbowR	WristR
22	ElbowL	HandL	ElbowR	HandR
23	WristL	HandL	WristR	HandR
24	ShoulderR	ElbowR	WristR	HandR

このようにして訓練データ内の運動時系列ごとに HMM による学習を行い，複数の運動記号を獲得する．次に，運動記号間の距離として Kullback Leibler 情報量，距離構造として Ward 法を用いることにより，得られた運号記号群に対して階層構造クラスタリングを行う．これにより N_k 個の集合が得られ，それぞれの集合に対しても同様に運動記号を作成する．この代表的な運動記号に対して，各運動記号が表現する運動データに最も適合するように HMM パラメータ $\boldsymbol{\lambda}$ に関する対数尤度の勾配を以下のように計算する．

$$FS(\boldsymbol{O}, \boldsymbol{\lambda}) = \nabla_\lambda \log P(\boldsymbol{O}|\boldsymbol{\lambda}) \quad (3)$$

$$= \nabla_\lambda L(\boldsymbol{O}|\boldsymbol{\lambda}) \quad (4)$$

ここで，$FS(\boldsymbol{O}, \boldsymbol{\lambda})$ はフィッシャースコアを意味する．また，運動記号 $\boldsymbol{\lambda}$ は，初期状態確率 π_i，状態遷移確率 a_{ij}，出力確率（混合ガウス分布の場合，平均 μ_j と分散 σ_j）の 4 つのパラメータとなるため，以下のように定義される．

$$\nabla_\lambda L(\boldsymbol{O}|\boldsymbol{\lambda}) = \left[\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \pi_i}, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial a_{ij}}, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \mu_i}, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \sigma_i}\right]^T \quad (5)$$

各パラメータに関する具体的な勾配計算については [Goutsu 14] を参照されたい．この修正すべき方向を表現した値を構成要素とするベクトルをフィッシャーベクトルと呼び，クラスタリング後の代表的な運動記号からのフィッシャースコア $FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_k)$ を縦に結合した以下のような式で定義される．

$$FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}) = F_\lambda^{-1/2}[FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_1)^T, ..., FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_{N_K})^T]^T \quad (6)$$

ここで，F_λ はフィッシャー情報行列と呼ばれ，対数尤度の勾配の正規化を行っている．また，SVM の内積計算は FV-HMM に

よるフィッシャーカーネルとなり，以下の様な式で定義される．

$$FK(\boldsymbol{O}_i, \boldsymbol{O}_j) = \\ < FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}), FV_{HMM}(\boldsymbol{O}_j, \{\boldsymbol{\lambda}_k\}) > \quad (7)$$

2.3 MKL-SVM

2.1 節で説明した局所的なジョイント列によるスケルトン特徴に対して，2.2 節ではスケルトン特徴の時系列情報を FV-HMM で特徴表現することにより運動特徴を作成する．この様々な局所ジョイント列による運動特徴を対象に応じて選択的に利用することで識別率がさらに向上すると期待される．また，筆者らの従来手法では運動やジェスチャーのクラスごとに訓練データの運動記号群がまとまるように階層構造クラスタリングを適用しており，識別するクラス数の増加に比例して FV-HMM が高次元になっていく問題があった．しかし，これが局所ジョイント列ごとのまとまりに置き換わり，クラス数がさらに増加したとしても運動は局所ジョイント列による運動特徴の選択・統合で表現されるため，FV-HMM による次元の拡大を抑えることができる利点もある．ここでは，特徴選択と特徴統合に MKL を用いる．この手法では，複数の運動特徴のカーネルを線形結合することにより結合カーネルを作成し，それを SVM に適用することで特徴統合による運動認識を実現する．最適なカーネル（カーネルの重み付きで線形結合したカーネル）のサブカーネルに対する重みを β_j とすると，統合カーネルは以下の式で定義される．

$$FK_{combined}(\boldsymbol{O}_i, \boldsymbol{O}_j) = \sum_{k=1}^{K} \beta_k FK_k(\boldsymbol{O}_i, \boldsymbol{O}_j) \quad (8)$$

ここで，$\beta_j \leq 0$，$\sum_{k=1}^{K} \beta_k = 1$ とする．また，K はカーネル数のことであり，すなわち局所ジョイント列数を意味する．MKL は各サブカーネルをそれぞれの特徴と対応させることにより特徴選択や特徴統合を実現し，これにより最終的な運動ラベルを決定する．[Sonnenburg 06] では，単一カーネルの SVM 学習の反復により最適なカーネル重み β_j を SVM の学習パラメータと同時に求める方法を提案しており，本稿でも同様の手法を利用している．

3. 実験

FV-HMM/SVM のジャスチャー認識における有効性を検証する実験を行った．本実験には，ジャスチャー認識のコンペティション ChaLearn Looking at People Challenge 2014 で提供されたデータセットを利用した [*1]．20 クラスのジャスチャーに対して，人手でラベル付けされた 6830 個の訓練データを HMM や SVM の学習に使用し，ラベル有りの 3200 個の評価データをジャスチャー認識の性能評価に使用した．ここで，HMM で学習させるスケルトン特徴として，全身の中心座標系からみた上半身のみのジョイント点群に対する相対位置で構成される 33 次元のスケルトン特徴を用いた．図 2 に 2.1 節で説明した階層構造クラスタリングの結果を示す．ここで，$N_k = 22$ とした．いくつかのジェスチャーは，類似した特徴を持ったカテゴリごとにまとまって分類され，はっきりと分類できなかったジェスチャーに関しては，個人差による影響が原因であると考えられる．また，4 つの異なる識別手法 (A):HMM/1-NN, (B):HMM/350-NN, (C):Similarity-based-HMM/1-NN (Generative embedding), (D):FV-HMM/SVM (Generative kernel) による比較を行った．図 3 に各手法の概略図を示す．また，各手法はそれぞれ以下のようにカテゴリを選択する．

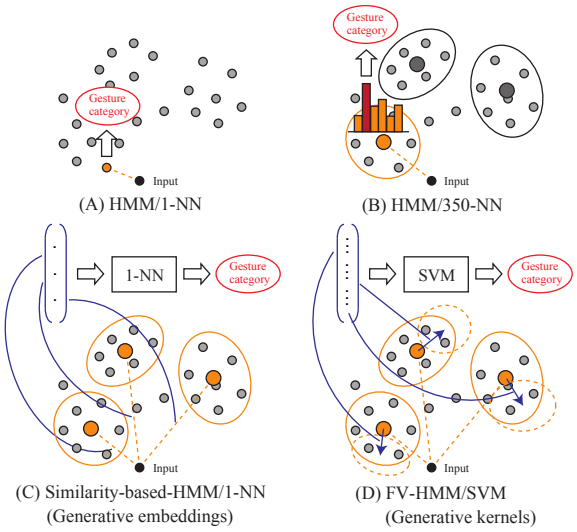

図 3: Four categorization methods for comparing. This figure shows the overviews of each method when given an input motion symbol.

・*HMM/1-NN*: 入力に対して一番近い運動記号が属するカテゴリを選択する．
・*HMM/350-NN*: 入力の最近傍クラスタ内で運動記号のカテゴリ投票を行い，一番高い投票率を得たカテゴリを選択する．
・*Similarity-based-HMM/1-NN*: 入力に対して各クラスタから得られる対数尤度を連結したベクトルを作成し，それに最も類似したベクトルの属するカテゴリを選択する．
・*FV-HMM/SVM*: FV-HMM により作成された運動特徴を SVM に入力し，識別されたカテゴリを選択する．

ここで，HMM ノードの接続方法に関して，全ての識別手法に Left-to-right 型を用いている．表 2 に識別手法の比較結果を示す．これより提案手法である (D) が最もカテゴリ平均識別率が高いと分かる．また，(A) と (D) の比較より生成的・識別的アプローチのハイブリッド手法が標準的な HMM を用いたアプローチよりも識別性能が良く，(C) と (D) の比較より Generative kernel アプローチが Generative embedding アプローチよりも識別性能が良いと言える．また，(A) の結果に関して，クラスタリングではっきりと分類できたカテゴリについては識別率が高くなる傾向にあった．

4. 結言

本稿では，多クラスの運動認識に対応し且つ識別に有効な特徴量の設計を検討するために，FV-HMM により表現された局所ジョイント列による運動特徴を MKL-SVM により運動の対象に応じて選択・統合することで識別率を向上させる手法を提案した．FV-HMM/SVM を用いた実験では，時系列情報の表現能力が高い HMM とクラス識別能力の高い SVM の双方の利点を活かしてお互いを結合することで，識別に有効に働くことを確認した．また，このことは FV-HMM/MKL-SVM に拡張することで運動認識の精度が向上することを示唆している．

*1 http://gesture.chalearn.org/

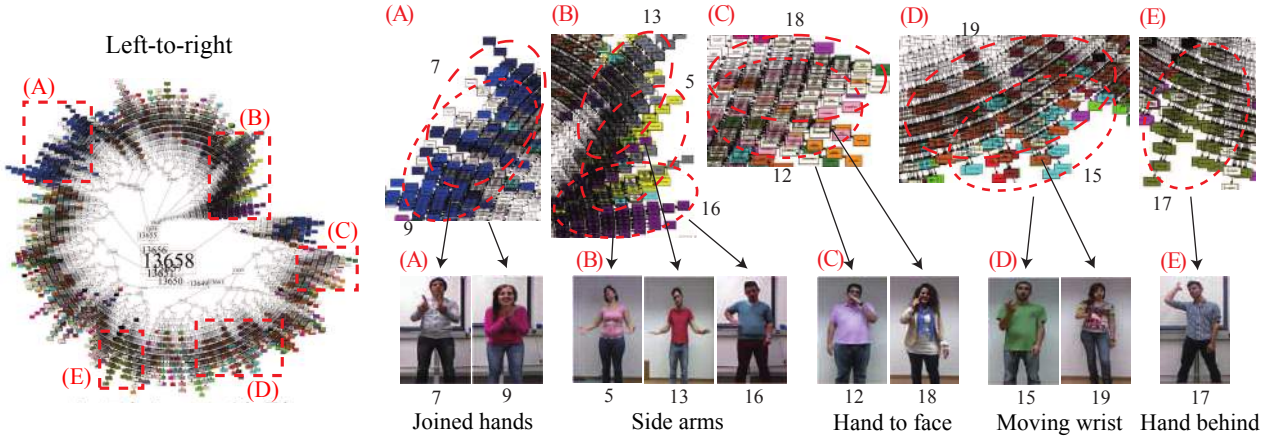

図 2: Result of hierarchically-structured clustering. The left figure shows overall views of the clustering in the left-to-right type. We represent scalable tree structures as circular shape. The upper right shows magnified views of remarkable area. The bottom right shows the gesture images. The number under each image corresponds to the number pointed out each ellipse.

表 2: Comparison result of correct recognition rate to all gesture categories when varying the categorization method. (A):HMM/1-NN, (B):HMM/350-N, (C):Similarity-based-HMM/1-NN, (D):FV-HMM/SVM (refer to Fig. 3).

	(A) LtoR	(B) LtoR	(C) LtoR	(D) LtoR		(A) LtoR	(B) LtoR	(C) LtoR	(D) LtoR		(A) LtoR	(B) LtoR	(C) LtoR	(D) LtoR
1	68.1	15.2	49.4	61.9	8	31.8	28.2	40.6	53.8	15	22.7	26.9	36.3	33.1
2	18.2	18.8	36.3	48.1	9	50.0	55.4	76.9	82.5	16	86.4	81.6	90.0	88.1
3	27.3	0.0	27.5	50.0	10	27.3	0.0	32.5	38.1	17	54.5	58.5	60.0	69.4
4	18.2	30.5	29.4	42.5	11	4.5	5.0	35.0	48.1	18	18.2	0.0	33.1	38.8
5	68.2	0.0	92.5	88.1	12	27.3	0.0	34.4	37.5	19	45.5	0.0	36.3	51.9
6	22.7	40.1	58.1	81.9	13	100	67.2	72.5	81.9	20	0.0	0.0	40.6	42.5
7	54.5	0.0	66.9	81.9	14	13.6	0.0	36.9	60.6	Avg	38.0	21.9	49.3	59.0

本研究は，平成 26 年度文部科学省科学研究費補助金若手 (A)「運動データベースからロボットの実世界運動制御への展開」（代表者：高野渉）の支援を受けて行った.

参考文献

[Goutsu 14] 郷津優介, 高野渉, 中村仁彦: Fisher Vector を用いた HMM と SVM のハイブリッド手法に基づくジャスチャー認識, 第 32 回日本ロボット学会学術講演会, 3B1-01, 2014

[Evangelidis 14] Evangelidis, G., Singh, G. and Horaud, R.: Skeletal quads: Human action recognition using joint quadruples, in *IEEE International Conference on Pattern Recognition (ICPR)*, pp.4513-4518, 2014

[Li 10] Li, W., Zhang, Z. and Liu, Z.: Action recognition based on a bag of 3d points, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.9-14, 2010

[Oreifej 13] Oreifej, O. and Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.716-723, 2013

[Sonnenburg 06] Sonnenburg, S., Rätsch, G., Schäfer, C. and Schölkopf, B.: Large scale multiple kernel learning, in *The Journal of Machine Learning Research*, Vol. 7, pp.1531-1565, JMLR. org, 2006

[Vieira 12] Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z. and Campos, M. F.: Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp.252-259, Springer, 2012

[Wang 12] Wang, J., Liu, Z., Wu, Y. and Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1290-1297, 2012

[Zanfir 13] Zanfir, M., Leordeanu, M. and Sminchisescu, C.: The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection, in *IEEE International Conference on Computer Vision (ICCV)*, pp.2752-2759, 2013