World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:9, 2018

Relevant LMA Features for Human Motion Recognition

Insaf Ajili, Malik Mallem, Jean-Yves Didier

Abstract**—**Motion recognition from videos is actually a very complex task due to the high variability of motions. This paper describes the challenges of human motion recognition, especially motion representation step with relevant features. Our descriptor vector is inspired from Laban Movement Analysis method. We propose discriminative features using the Random Forest algorithm in order to remove redundant features and make learning algorithms operate faster and more effectively. We validate our method on MSRC-12 and UTKinect datasets.

Keywords**—**Human motion recognition, Discriminative LMA features, random forest, features reduction.

I. Introduction

HUMAN motion recognition is a fundamental topic in the field of computer vision. One of the main challenges with motion recognition is that the same motion may be performed in different ways by different persons, and even by the same person. Although a significant amount of research has been focused on human motion representation, it remains still not enough. First motion features proposed in the computer vision literature were based on interest points [8], [11]. After, another type of features were studied based on depth informations provided from depth sensors. The depth cameras in general provide better quality 3D data than those estimated from monocular video sensors. This allows to focus on the analysis of the human motion through the identification of his joints [7], [16], [12], [10]. However, such proposed features did not take into account the semantic aspects of motion, for example to classify two actions with the same movement but performed with different intentions or rythms. In such case these features are not relevant to discriminate between these motions. Another key challenge in motion representation is to realize a good compromise between robustness performance and computational costs. The assumption that increasing the number of features can provide more informations about motion is not always valid in practice, because it can be time consuming and may lead to finding a less optimal solution. For feature reduction, several approaches were proposed and could be divided in two groups: methods based on statistical measures [5], [9], and methods based on learning algorithms [13], [15]. The first category consists in ranking features according to some statistical measures. It is fast and independ of any classifier, but it requires a threshold to select the top ranked features. Finally, in statistical approaches, some important features that are less informative on their own, but they are informative when combined with others can be discarded. The second category evaluates the importance of a random subset of features by training a model on it. A learning method is used to evaluate the importance of each combination of features on the classification performance. Despite the effectiveness of these methods, they have the constraint to be computationally more expensive compared to the statistics methods due to the repeated learning and cross validation steps.

M. Mallem and J. Didier are with IBISC, Univ Evry, Universit Paris-Saclay, 91025, Evry, France.

Insaf Ajili is with IBISC, Univ Evry, Universit Paris-Saclay, 91025, Evry, France (e-mail: insafajili@gmail.com).

In this paper, which extends our preliminary work presented in [1], [2], we address the problem of analyzing human motion from skeleton sequences captured by depth cameras. Particularly, our work focuses on representing human motions by keeping most salient and complementary features based on Random Forest algorithm. Our descriptor vector is inspired from Laban Movement Analysis method (LMA) to describe quantitative and qualitative representations of motions. The rest of the paper is structured as follows. Section 2 describes our proposed approach with motion recognition steps. Section 3 presents the experimental results on MSRC-12 and UTKinect datasets. Finally, conclusions and future work are stated in Section 4.

II. Proposed Approach

A. Data Acquisition

We use kinect sensor for data acquisition to extract 3D skeleton joints in real time. The first step is the normalization of all skeletons which consists in aligning all skeletons in the center of the kinect coordinate sytem with the base B at initial frame. Given a motion sequence $S = \{J_{j,t}\}$, $j \in 1, \dots, N, t \in 1, \dots, T$, $J_{j,t}$ corresponds to the coordinates of the joint j captured at frame t. We define a local coordinate system to the skeleton anchored to the hip center joint (J_c), represented by the base B', equipped with three unit vectors, the left hip joint vector $\vec{n_{lh}}$, the spine vector $\vec{n_s}$ and their cross product $\vec{n_c} = \vec{n_{lh}} \wedge \vec{n_s}$. For each sequence, we first apply translation to move the skeleton to the center of kinect, and after a rotation to align both coordinate systems (Fig. 1). The transformed joint yields to:

$$[J_{j,t}]_{B'} = R^{-1}_{B \leftarrow B'}([J_{j,t}]_B - [J_{c,1}]_B) \tag{1}$$

$$R_{B \leftarrow B'} = [\frac{\vec{n_{lh}}}{\|\vec{n_{lh}}\|} \ \frac{\vec{n_s}}{\|\vec{n_s}\|} \ \frac{\vec{n_c}}{\|\vec{n_c}\|}] \tag{2}$$

$$\vec{n_{lh}} = [J_{lhi,1}]_B - [J_{c,1}]_B \tag{3}$$

$$\vec{n_s} = [J_{s,1}]_B - [J_{c,1}]_B \tag{4}$$

Once we applied transformations to all sequences, our system is independent of the initial position and orientation of the subject in the scene. Then, we pass to the next step which consists in converting the skeleton joints data to a descriptor vector based on LMA method.

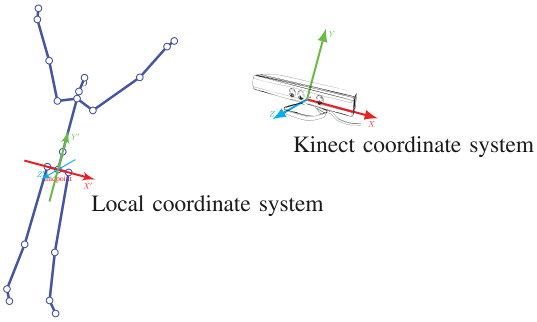


Fig. 1 Kinect and local coordinate systems

B. Motions Representation

LMA approach employs a multilayered description of movement, focusing on four components: Body, Space, Shape, and Effort. **Body** component has the responsibility of highlighting the body part which is moving, making the connection between the moving parts and taking in consideration the issues of locomotion and kinematics. For this category, we describe the organization and connection between the different joints (Fig. 2). We consider two parts, the upper and lower part. For the first one, the extension of the different joints is described by computing the following angles in left and right parts respectively: between hands and shoulders (θ_1^l, θ_1^r), between elbows and hips (θ_2^l, θ_2^r), between elbows and shoulders in the symmetrical part (θ_3^l, θ_3^r). We also calculate the distances between the two hands (d_{Hs}) as well as the distances between the shoulder center and both hands $(d_{shc,lh}, d_{shc,rh})$. It allows us to have a more idea on the way of the two hands. For the lower part of the body, the extension of the knees has been described with the angles between the feet and the hips (θ_3^l, θ_3^r). These two characteristics allow to characterize specific actions like crouch or hide gestures. We also characterize the opening of the legs with the angle computed between the two knees θ_{Ls}. Mean, standard deviation, and range of the body features are computed to quantify the Body component. We compute the length of trajectories (L) made by upper and lower body extremities (head, hands, and feet) to quantify the **Space** component.

$$L = \sum_{t=1}^{T-1} \| J_{j,t+1} - J_{j,t} \| \tag{5}$$

In **Shape** component, we describe the way the body changes shape during movement with three qualities. In the first *Shape flow* factor, we charachterize the change shape in a self-to-self relationship by computing the volume of the smallest convex envelope of the human body based on Quickhull algorithm [3], as shown in Fig. 2. The second factor is the *Directional movement*, we define the pathway of the movement of upper body extremities (hands and head) through space by computing their curvatures (C).

$$C = \sum_{t=2}^{T-1} \arccos\left(\frac{\overrightarrow{J_{j,t-1}J_{j,t}}}{\left\|\overrightarrow{J_{j,t-1}J_{j,t}}\right\|} \cdot \frac{\overrightarrow{J_{j,t}J_{j,t+1}}}{\left\|\overrightarrow{J_{j,t}J_{j,t+1}}\right\|}\right) \tag{6}$$

Finally, we quantify the *Carving* factor of the Shape component which describes the qualitative changes in the shape relating to spine joint pose at initial frame $J_{s,1}$, according to three planes: Horizontal (D_H), Frontal (D_F), and Sagittal (D_S), relating them to bipolar descriptors: spreading/enclosing, rising/sinking, and retreating/advancing, respectively (Fig. 2).

$$D_H = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} sqrt((J_{x_{j,t}} - J_{x_{s,1}})^2) \right) \tag{7}$$

$$D_F = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} sqrt((J_{y_{j,t}} - J_{y_{s,1}})^2) \right) \tag{8}$$

$$D_S = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} sqrt((J_{z_{j,t}} - J_{z_{s,1}})^2) \right) \tag{9}$$

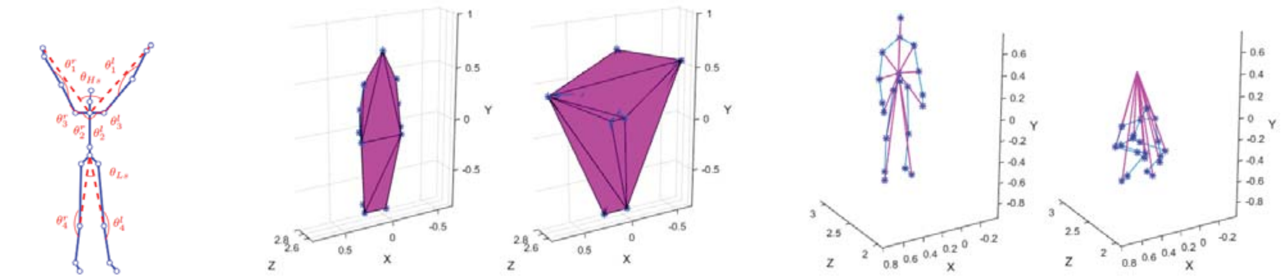
Effort component describes how the body concentrates its effort while performing a motion and characterizes expressive behaviors based on four factors: *Time*: Sudden/Sustained, *Weight*: Light/Strong, *Flow*: Bound/Free, and *Space*: Direct/Indirect (Fig. 3). In Effort component we focus on the upper body part (head, hands, and spine), since it was the most expressive part during human motion. Joints velocities and accelerations are computed for quantifying *Time* and *Weight* factors, respectively. Three measures of variability (Mean, standard deviation, and range) are used for both features. We quantify the *Flow* factor by computing the yaw and pitch range of joints motion. For Free motion we will obtain a higher range compared to Bound motion. To describe the direction of the movement in space for *Space* factor, we compute the Straightness index (S) of joints motion as the ratio of the distance between the first and last frame (D) to the sum of the displacements between two successives frames (L).

$$S = \frac{D}{L} \tag{10}$$

C. Motions Recognition

For motions training and classification, we apply the Random Forest approach (RF) [4]. This method consists of an ensemble of decision trees, each tree is grown using by a different bootstrap sample from the training data. Let the feature vector be $v = \{f_i\}, i = 1, \ldots, d$, where d is the number of features for each sample. At each node, best split is chosen from a random sample of p features from d. Consider a node k comprising S_k samples, splitted into left and right child nodes with subsamples of S_{kl} and S_{kr}, respectively, the tree is then grown by selecting the splitting condition that maximizes the purity of the resulting tree. Gini index $I(S_k)$ is

World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:9, 2018



(a) Body features (b) Shape flow features (c) Carving features

Fig. 2 Some LMA features

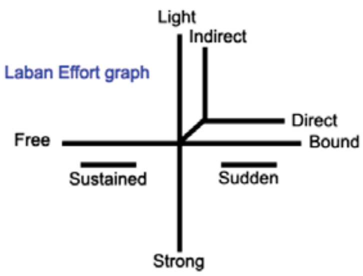


Fig. 3 Effort factors

used to select the feature at each internal node k. The amount of homogeneity gain achieved by the splitting node k in feature f can be evaluated in the following equation:

$$G(f, S_k) = I(S_k) - \sum_{i \in l,r} \left(\frac{|S_{ki}|\, I(S_{ki})}{|S_k|} \right) \tag{11}$$

where $I(S_k) = 1 - \sum_{j=1}^{l} (\frac{S_k^j}{S_k})^2$, l is the number of classes in node k, S_k^j denotes the number of learning samples which belong to class j at node k. Therefore, after several selections for f, the one producing the lowest value of Gini index is picked as the split criterion for the node. In the testing step, each test sample is simultaneously pushed through all trees, starting from the root, and assigning the data to the right or left child recursively until a leaf node is reached. Finally, the forest chooses the classification having the majority of votes from each of the decision trees made.

D. Features Reduction

Feature reduction step consists in keeping the smallest subset of most relevant features for motions representation to achieve a good compromise between accuracy and runtime in the classification process. The measure of relevant feature is returned by RF method. During training phase, each tree is grown using a different bootstrap sample from the original training data, leaving 1/3 as OOB (Out Of Bag) to estimate the prediction error of OOB. The importance of the feature f_i is measured as the difference between OOB error rate of each tree before and after permuting f_i.

$$I^t(f_i) = \frac{\sum_{j \in O^t} I(y_j \neq y_{jt})}{|O^t|} - \frac{\sum_{j \in O^t} I(y_j \neq y_{jt}^i)}{|O^t|} \tag{12}$$

$$I(f_i) = \frac{\sum_{t=1}^{T} I^t(f_i)}{T} \tag{13}$$

O^t corresponds to OOB samples for a tree t, y_j is true class label of the j^{th} training sample. y_{jt} and y_{jt}^i are the predicted classes for the j^{th} sample by tree t before and after permuting the feature f_i, respectively. Finally, a high decrease in accuracy is an indication of the feature importance. We start with the whole set of features, we compute and record the OOB error rate. After we sort the features in descending order of importance, and we remove the feature of small importance f_{min}. Moreover, we apply the Tukey's test ($\alpha = 0.05$) to simultaneously remove features that do not give a significant difference of the OOB error rate result.

Algorithm 1: Feature reduction process

Input : $v_0 = \{f_i\}, i = 1, \ldots, p$ $\triangleright$ v_0 is the whole feature set.
Output: $v^* = \{f_j\}, j = 1, \ldots, p^*$ $\triangleright$ v^* subset of most relevant features.

```
k = 0
while |v_k| ≥ 1 do
    Compute and record OOB error rate: E_k(v_k)
    for i = 1 to p do
        Compute I(f_i)        ▷ Importance of each feature in v_k.
    end
    Sort {f_i} in descending order according to values of I(f_i)
    f_min = argmin_i {I(f_i)}
    Apply Tukey's test and select set of features {f_t} that does
    not lead to a significant changement of E_k
    R = f_min ⋃ {f_t}
    v_{k+1} = v_k \ R
    k = k + 1
end
v* = argmin_k {E_k(v_k))} ▷ v* is the optimal feature subset with
minimal OOB error.
```

III. Experimental Results

To evaluate the performance of our method, we use two public action datasets, MSRC-12 [6] and UTKinect [14], we report two measures: the mean of fscores and the OOB error rate. For the first measure, we adopt the 5-fold cross validation to optimize the RF parameters, and compute the averaged results. We employ the commonly used F-score as the performance measure.

1) Evaluation on MSRC-12 Dataset: MSRC-12: is a dataset composed of 594 sequences, containing the performances of 12 gestures by 30 people. In total, there are 6244 gesture instances. The gesture classes are divided into two groups: metaphoric gestures, and iconic gestures. The motion files contain 3D coordinates of 20 joints captured at a sample rate of 30Hz.

We converted the raw data into a descriptor vectors based on our LMA qualities. Our descriptor vector composed of 85 features was fed into a learning algorithm, RF. Most important parameters of RF, the number of trees (n_{trees}) and the number of features to consider when splitting a node ($max_{features}$) were adjusted. We varied n_{trees} starting from 10 until 200 trees, and we tested three values of $max_{features}$(85, $log_2(85)$, and $\sqrt{85}$). Best recognition rate of 94.89% was achieved when setting $n_{trees} = 100$ and $max_{features} = log_2(85)$, and almost the same recognition result of 94.88% was obtained for $n_{trees} = 100$ and $max_{features} = \sqrt{85}$. We also confirmed the RF parameters values by computing OOB error rate while varying n_{trees} and $max_{features}$. As we can see in Fig. 4, the two curves of $max_{features} = log_2(85)$, and $\sqrt{85}$ are very close with a very little superiority result of $max_{features} = log_2(85)$. We identified the minimum value of n_{trees} where OOB error stabilize (around 0.008), we found $n_{trees} = 100$, which confirms the result obtained with the recognition rate measure. Table I illustrates the recognition

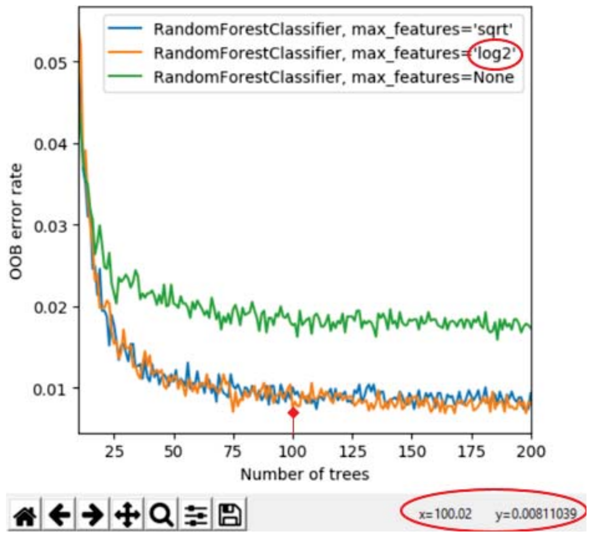


Fig. 4 OOB error rates in terms of RF parameters (n_{trees} and $max_{features}$)

results of our method compared to the state of the art methods on MSRC-12 dataset.

Our method outperforms the state of the art methods, and is very close to the result obtained in [12], which confirms the robustness of our descriptor in characterizing both iconic and metaphoric gestures with an accuracy rates of 99% and 93%, respectively. After evaluating our descriptor vector on gestures recognition in MRC12-dataset, we applied our feature

TABLE I
RECOGNITION RATES OF OUR METHOD COMPARED TO THE STATE OF THE ART METHODS ON MSRC-12 DATASET

Methods	**Recognition rates (%)**
Hussein et al. [7]	91.70
Zhou et al. [16]	90.22
Wang et al. [12]	94.86
Lehrmann et al. [10]	90.90
Our method	**94.89**

TABLE II
COMPARISON OF (MEAN-FSCORES, OOB ERROR VALUES, NUMBER OF FEATURES) BETWEEN BEFORE AND AFTER FEATURES REDUCTION STEP IN MSRC-12 DATASET

	$Mean\ F - score$	$errOOB$	N
Before reduction	0.94 (+/-0.02)	0.008	85
After reduction	0.94 (+/-0.02)	0.004	67

reduction algorithm (Algorithm 1) in order to keep only most discriminant features according to this dataset. To obtain more stable results and better estimations for the expected OOB error rate, we repeated this procedure 30 times and average the results. In Table II, we make a comparison between results obtained before and after feature reduction process, in terms of $Mean\ F - score$, OOB error rate ($errOOB$), and number of features (N). We notice that the number of relevant features is decreased about 20% achieving the same F-score results and decreasing the OOB error rate. We obtained a low OOB error rate of 0.004 with a number of relevant features $N = 67 < 85$ (see Fig. 5).

2) Evaluation on UTKinect Dataset: We also evaluated our descriptor with UTkinect dataset which is composed of 10 subjects performing 10 different activities in varied views namely walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands. A total number of 199 sequences are available. Each action is repeated twice by the actor. Sequences are captured using one Kinect in indoor settings and their length ranges from 5 to 120 frames. This is a challenging dataset due to variations in the view point and high intra-class variations where each actor performs actions in different views. We applied both methodes, feature

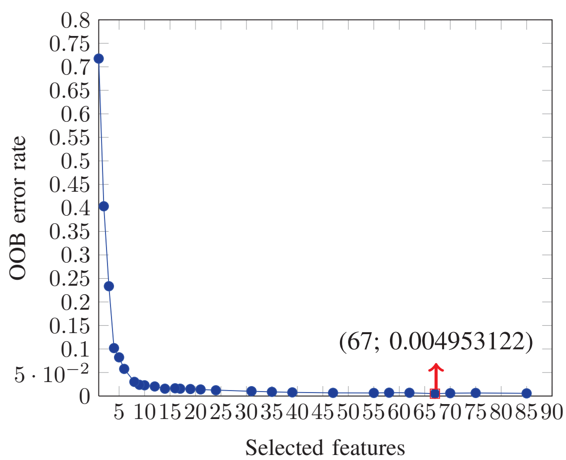


Fig. 5 Optimal feature subset with minimal OOB error using Tukey's test ($\alpha = 0.05$)

TABLE III
COMPARISON OF (MEAN-FSCORES, OOB ERROR VALUES, NUMBER OF FEATURES) BETWEEN BEFORE AND AFTER FEATURES REDUCTION STEP IN UTKINECT DATASE

	$Mean\ F-score$	$errOOB$	N
Before reduction	0.96 (+/-0.02)	0.0075	85
After reduction	0.96 (+/-0.01)	0.006	36

extraction with LMA approach and after RDF method for actions recognition. We measured the recognition rate using same validation method as MSRC-12 dataset, the 5-fold cross validation technique. We obtained as result the mean of f-scores $0.96(+/-0.01)/$. With the same RDF parameters setting in MSRC-12 dataset, we applied features reduction step. Table III summarizes recognition results ($Mean\ F-score$, OOB error rate ($errOOB$), and number of features (N)) before and after applying features reduction step. With 36 features we obtained same mean f-score and a lower OOB error value. So we can say that our method managed to reduce features while keeping most relevant features and same recognition results.

IV. CONCLUSION

In this paper, we presented an efficient method for extracting most relevant motion descriptors for human motion recognition. Our descriptor was inspired from LMA technique to combine both quantitative and qualitative characteristics of motion. Furthermore, an effective feature reduction algorithm was applied to keep only the most informatives features which had a great impact on the computational latency while maintaining or even improving the reported results. Based on these results, we plan to recognize expressive motions and study the importance of each LMA features to characterize human emotions.

REFERENCES

[1] I. Ajili, M. Mallem, and J. Y. Didier. Gesture recognition for humanoid robot teleoperation. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1115–1120, Aug 2017.
[2] I. Ajili, M. Mallem, and J.-Y. Didier. Robust human action recognition system using laban movement analysis. *Procedia Computer Science*, 112(Supplement C):554 – 563, 2017. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.
[3] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4):469–483, Dec. 1996.
[4] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.
[5] A. B.Surendiran1. Feature selection using stepwise anova discriminant analysis for mammogram mass classification. *International Journal on Signal & Image Processing*, 2(1):4, January 2011.
[6] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1737–1746, New York, NY, USA, 2012. ACM.
[7] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2466–2472. AAAI Press, 2013.
[8] I. Laptev and T. Lindeberg. Space-time interest points. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 432–439. IEEE, 2003.
[9] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 9(4):1106–1119, July 2012.
[10] A. M. Lehrmann, P. V. Gehler, and S. Nowozin. Efficient nonlinear markov models for human motion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, June 2014.
[11] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.*, 103(1):60–79, 2013.
[12] P. Wang, Z. Li, Y. Hou, and W. Li. Action recognition based on joint trajectory maps using convolutional neural networks. *CoRR*, abs/1611.02447, 2016.
[13] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 647–653, Cambridge, MA, USA, 2000. MIT Press.
[14] L. Xia, C. C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27, June 2012.
[15] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama. High-dimensional feature selection by feature-wise non-linear lasso. *ArXiv e-prints*, Feb. 2012.
[16] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang. Discriminative key pose extraction using extended lc-ksvd for action recognition. In *2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, Nov 2014.