Evolutionary Joint Selection to Improve Human Action Recognition with RGB-D devices

Alexandros André Chaaraoui[a], José Ramón Padilla-López[a], Pau Climent-Pérez[b], Francisco Flórez-Revuelta[b,*]

[a] *Department of Computer Technology, University of Alicante, P.O. Box 99, E-03080, Alicante, Spain*
[b] *Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road, KT1 2EE, Kingston upon Thames, United Kingdom*

Abstract

Interest in RGB-D devices is increasing due to their low price and the wide range of possible applications that come along. These devices provide a marker-less body pose estimation by means of skeletal data consisting of 3D positions of body joints. These can be further used for pose, gesture or action recognition. In this work, an evolutionary algorithm is used to determine the optimal subset of skeleton joints, taking into account the topological structure of the skeleton, in order to improve the final success rate. The proposed method has been validated using a state-of-the-art RGB action recognition approach, and applying it to the *MSR-Action3D* dataset. Results show that the proposed algorithm is able to significantly improve the initial recognition rate and to yield similar or better success rates than the state-of-the-art methods.

Keywords: RGB-D devices, human action recognition, evolutionary computation, instance selection, feature subset selection

*Corresponding author

Email addresses: alexandros@dtic.ua.es (Alexandros André Chaaraoui), jpadilla@dtic.ua.es (José Ramón Padilla-López), P.Climent@kingston.ac.uk (Pau Climent-Pérez), F.Florez@kingston.ac.uk (Francisco Flórez-Revuelta)

1. Introduction

Recently, interest has grown on affordable devices as the *Microsoft Kinect* (Microsoft Corporation, 2013) or the *ASUS Xtion Pro* (ASUSTeK Computer Inc, 2013), which can capture depth quite reliably. These image sensors provide a depth image (D), besides the regular color image (RGB). The resulting RGB-D data can be used to obtain a marker-less body pose estimation. Specifically, a skeleton model consisting of a set of joints is generated. This characteristic data can be used in order to learn and classify human poses, actions or even activities of daily living (ADL). These depth sensors have become popular due to their low cost, high sample rate and capability of combining visual and depth information. Usage can be found both in research and commercial applications. Although they were initially designed for gaming purposes, other applications, where natural human-computer interaction (HCI) is required, are extensively employing these technologies (e.g. Seo & Lee (2013)). In particular, RGB-D devices are used in ambient assisted living for fall detection (Mastorakis & Makris, 2012), physical rehabilitation (Chang et al., 2011; Huang, 2011), medical image exploration in operating rooms (Gallo et al., 2011) and gait analysis (Stone & Skubic, 2011) among other applications.

Reliability and accuracy of RGB-D devices have been studied in several works (Obdrzalek et al., 2012; Alnowami et al., 2012), which show that the extraction of a skeleton from depth information is not straightforward. Among several difficulties, lack of precision and occlusions caused by body parts or other objects present in the scene stand out (Khoshelham & Elberink, 2012; Shotton et al., 2011).

Most of the existing works that employ RGB-devices for human action recognition use all the available joints obtained by the devices. However, some actions or gestures involve moving the whole body, whereas others are performed using only the arms or the hands. Therefore, it is interesting to determine which joints have a greater value to the success of the recognition method being used, and which ones can be discarded because they are not relevant for a specific application, since they introduce confusion or noise and reduce the recognition rate.

This paper proposes an evolutionary method for the selection of the subset of relevant joints that improve action recognition using RGB-D devices. A method based on a bag of key poses and dynamic time warping (DTW) is used as recognition algorithm in order to calculate the fitness of the different

solutions obtained in the evolution. This evolutionary feature subset selection method employs specific knowledge about the topological structure of the skeleton in order to obtain better solutions in less time.

The remainder of this paper is organised as follows: Section 2 reviews the state of the art on human action recognition with RGB-D devices and evolutionary feature subset selection. Section 3 presents the proposed evolutionary algorithm. Section 4 deals with the human action recognition method employed to evaluate the fitness of the individuals in the population. Section 5 presents the results obtained applying our proposed method to a well-known dataset. Finally, in Section 6 some discussion and conclusions are drawn.

2. Related work

This section reviews the most relevant state of the art on human action recognition with RGB-D devices and evolutionary feature subset selection related to this work.

2.1. Human action recognition with RGB-D devices

Experimental results show that humans are able to recognise different activities seeing only a few points of light attached to the joints of the human body (Moving Light Display, Johansson (1973); Polana & Nelson (1997)). Therefore, it seems that the position, orientation and motion of joints contain enough characteristic data in order to recognise activities using computers. Furthermore, good performance may be achieved using only the spatial distribution of the joints.

In the state-of-the-art works in the research field, it can be observed that an increasing number of applications for RGB-D-based human action recognition are being developed. The necessary datasets, so as to perform initial evaluations and compare the results, have been recorded and made publicly available. Both datasets designed for gesture or action recognition for natural user interfaces (NUI) or gaming (Li et al., 2010), and more complex activities involving interactions with objects (Wang et al., 2012b; Sung et al., 2011; Ni et al., 2011; Janoch et al., 2011) have been published.

There are several methods to extract a structured set of joints and their connections, i.e. the skeletal information, from depth maps (Shotton et al., 2011). These methods provide different kinds of skeleton models. The *Microsoft Kinect SDK* (Microsoft Corporation, 2013) provides a skeleton model

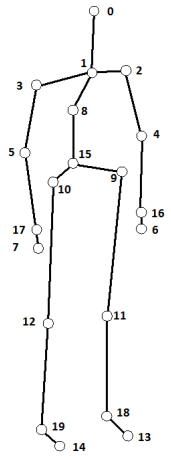


Figure 1: The 20 joints from a skeleton in the *MSR-Action3D* dataset.

with 20 joints (see Fig. 1), whereas the *OpenNI/NITE* (PrimeSense, Ltd., 2013) skeleton tracks a set of 15 joints.

The use of the different data provided by the RGB-D devices for human action recognition goes from employing only the depth data, or only the skeleton data extracted from the depth, to the fusion of both the depth and the skeleton data.

Li et al. (2010) use a simple but effective projection scheme to obtain a representation set of 3D points from the depth map. Dynamics of human motion are modelled based on a set of salient postures shared among the actions. These postures are described using a bag-of-points. Yang et al. (2012) propose a method to recognise human actions from sequences of depth maps. They project the depth maps onto three orthogonal planes and accumulate the whole sequence generating a depth motion map (DMM), similar to the motion history images (Bobick & Davis, 2001). Histograms of oriented gradients (HOG) are obtained for each DMM. The concatenation of the three HOG serves as input feature to a linear SVM classifier. Wang et al. (2012a) treat an action sequence as a 4D shape and propose random occupancy pattern features, which are extracted from randomly sampled 4D sub-volumes with different sizes and at different locations. These features are robust to noise and less sensitive to occlusions. An Elastic-Net regularization is employed to select a sparse subset of features that are the most discriminative for the classification. Finally a SVM classifier is trained for action classification.

Miranda et al. (2012) described each pose using a spherical angular rep-

resentation of the skeleton joints obtained with *Kinect*. Those descriptors serve to identify key poses through a multi-class classifier derived from support vector learning machines. A gesture is represented as a sequence of key poses and labelled on the fly through a decision forest, that naturally performs the gesture time warping and avoids the requirement for an initial or neutral pose. Xia et al. (2012) use histograms of 3D joint locations computed from the action depth sequences. These features are re-projected using LDA and then clustered into several posture visual words, which represent the prototypical poses of actions. The temporal evolutions of those visual words are modelled by discrete hidden Markov models. Azary & Savakis (2012) use sparse representations of spatio-temporal kinematic joint features and raw depth features which are invariant to scale and position. They create overcomplete dictionaries and classify input patterns using both L1-norm and L2-norm minimisation. Yang & Tian (2012) propose a new type of features, the EigenJoints. They employ 3D position differences of joints to characterise action information including posture, motion and offset features. After a normalisation process, PCA is applied to compute the EigenJoints. Then they employ a naïve-Bayes-nearest-neighbour classifier for multi-class action classification. Soh & Demiris (2012) propose an online echo state Gaussian process (OESGP), a novel Bayesian-based online method, to iteratively learn complex temporal dynamics and produce predictive distributions. They use a generative modelling approach whereby each action class is represented by a separate OESGP model. Inference is performed using a Bayes filter to iteratively update a probability distribution over the model classes. Fothergill et al. (2012) employ joint angles, joint angle velocities and xyz-velocities of joints as feature vector at each frame. Then gesture recognition is carried out using random forests.

Other methods fuse depth and skeletal data. Wang et al. (2012b) use the pairwise relative difference between joints' positions as features. The authors state that 3D joint positions are insufficient to fully model an action, especially when the action includes the interactions between the subject and other objects. Therefore, these interactions are characterised by local occupancy patterns (LOP) at each joint. This LOP feature computes the local occupancy information based on the 3D point cloud around a particular joint. A Fourier temporal pyramid is then used to obtain a more robust representation. Then an actionlet ensemble model is learnt to represent each action and to capture the intra-class variance.

2.2. Evolutionary feature subset selection

Feature vectors provide a set of characteristic data that represents the object to recognise. Along with the useful data, that may lead to a successful classification by means of machine learning algorithms, the set can include irrelevant or redundant information which could complicate the classification (Cantú-Paz, 2004). This unnecessary information can also be affected by noise, which hinders the classifiers to isolate the appropriate elements (Lanzi, 1997; Kira & Rendell, 1992).

Through feature selection, a subset of variables is chosen in order to optimise the classification and obtain higher success rates. In addition, a reduced feature subset also leads to a lower computational cost (Casado Yusta, 2009; Yang & Honavar, 1998).

When dealing specifically with skeletal data obtained with RGB-D devices, it can be seen that some joints are more important than others if pose or motion recognition is targeted (Raptis et al., 2011). There are several joints in the torso, as for instance the shoulders or the hips, which do not show an independent motion and rather move along with the whole body. Therefore, taking this knowledge into account, the characteristic value of the motion can be retained, and at the same time dimensionality reduction can be performed in order to improve the performance of the classification (Raptis et al., 2011).

Evolutionary feature subset selection has been used for decades (Siedlecki & Sklansky, 1989). The basic approach is to consider a binary vector where each gene represents the further consideration or not of a specific feature. Two main models are presented to implement this (Cantú-Paz, 2004; Casado Yusta, 2009): the *filter* model, and the *wrapper* model (John et al., 1994). The *filter* model performs *a priori* decisions in order to determine the relevance of the features based on their intrinsic properties (Liu & Motoda, 1998), but it ignores the learning algorithm underneath. On the contrary in the *wrapper* model, the feature selection algorithm encloses the learning algorithm. It uses it in order to perform evaluations on the possible feature subset selections and to find the optimal one (John et al., 1994). This approach presents a disadvantage over the former, since each feature subset evaluation may take a considerable amount of time (Lanzi, 1997; Wang & Huang, 2009). Nonetheless, wrapper-based approaches are usually preferred because the resulting feature subset selections show better results (Cantú-Paz, 2004).

In our case, an evolutionary algorithm is applied as a wrapper method. Therefore, the best feature subset selection is sought by iteratively evaluating the possible selections, and creating new selections by means of evolution of individuals.

3. Evolutionary algorithm for feature selection

This paper presents an evolutionary algorithm specifically designed for joint selection for skeletal data. The structure of the evolutionary algorithm follows the process presented in Algorithm 1. The algorithm has the following characteristics:

- the selection of the individuals to be affected by recombination and mutation operations is performed following a ranking method (Jong, 2006), according to which those individuals scoring higher in the fitness function have a larger probability of being selected;
- a specific crossover operator (see Section 3.2) has been developed which is aware of the topological structure of the skeletons;
- the standard mutation is used, i.e. each gene changes its value according to probability p_{mut}; and
- the fitness of each individual is obtained as the success rate using it as input to a recognition algorithm (see Section 4).

3.1. Individuals' representation

As it is usual in evolutionary feature subset selection, the chromosomes are encoded as binary vectors, each gene representing the use or not of that element during the recognition (Fig. 2), i.e. in the calculation of the fitness function.

3.2. Crossover

In a previous work (Climent-Pérez et al., 2013) we used a 1-point crossover operator. This operator was not aware of the skeleton topology, and therefore the recombination was performed considering at the same time different parts of the body, because the order of the joints in the chromosome is not representative of the tree topology where the HEAD joint is the root and hands and feet are the leaves (Fig. 3).

Algorithm 1 Evolutionary algorithm

Initialise the populations with N individuals generated randomly
Rank the population by fitness
repeat
  for number of new individuals to be created **do**
    ——— Generate a new individual ———
    Create one new individual i by crossover
    Mutate i
    ——— Calculate fitness ———
    Calculate $fitness(i)$ as the classification rate
  end for
  ——— Generate next generation's population ———
  Rank the population by fitness
  Select next generation's population with elitism
until $generations\_without\_changes > gen_{max}$

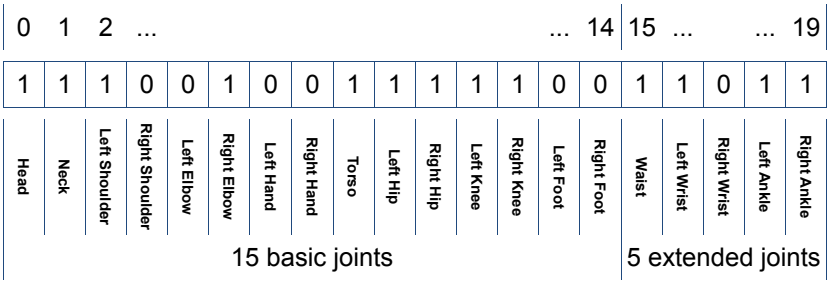


Figure 2: Chromosome. Joints 0 to 14 are employed by the skeleton models from both *Microsoft* and *OpenNI/NITE*. Joints 15 to 19 are only present in the 20-joint model from *Microsoft*.

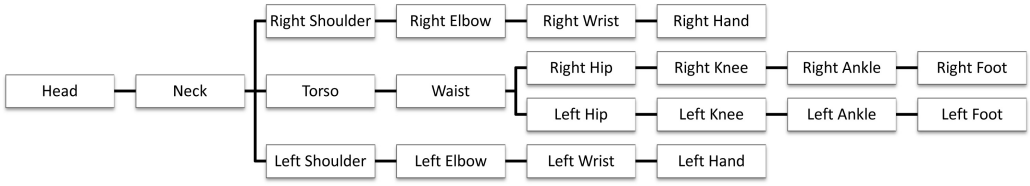


Figure 3: Skeleton topology.

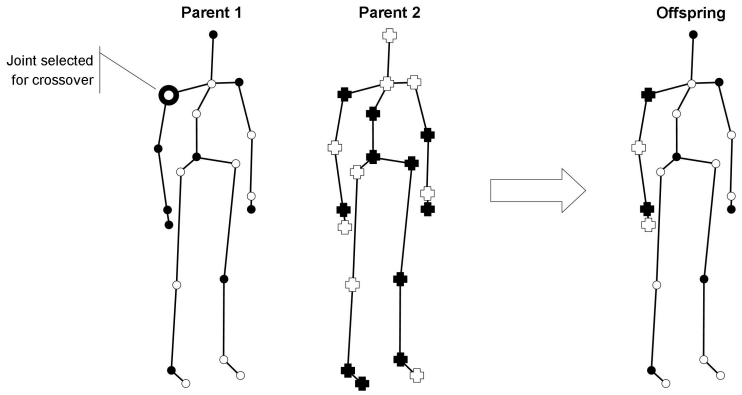


Figure 4: Crossover between skeletons. Gene values are represented by the colour of the joint (black=1, white=0).

For this reason, this paper proposes a crossover operator that is aware of the skeleton's tree topology. It works similar to the typical crossover in genetic programming, where a node in one parent is randomly selected, and all the branch below it is substituted by the same branch from a different parent (Fig. 4).

4. Human action recognition method

In order to evaluate a specific feature subset selection and to obtain the fitness value of the corresponding individual, a human action recognition method based on the one from Chaaraoui et al. (2012) is proposed. An overview of the method can be seen in Fig. 5. The method relies on key poses in order to model the most representative human poses. Then, sequence of key poses are obtained to capture the temporal relationship of action performances. Finally, action recognition is performed with DTW-based sequence matching. The adaptation of this recognition method to our particular skeletal representation obtained from depth images is detailed in this section.

4.1. Skeletal representation

Regarding the feature extraction process, human silhouettes are used in Chaaraoui et al. (2012). Translating this to the domain of skeletal data, a

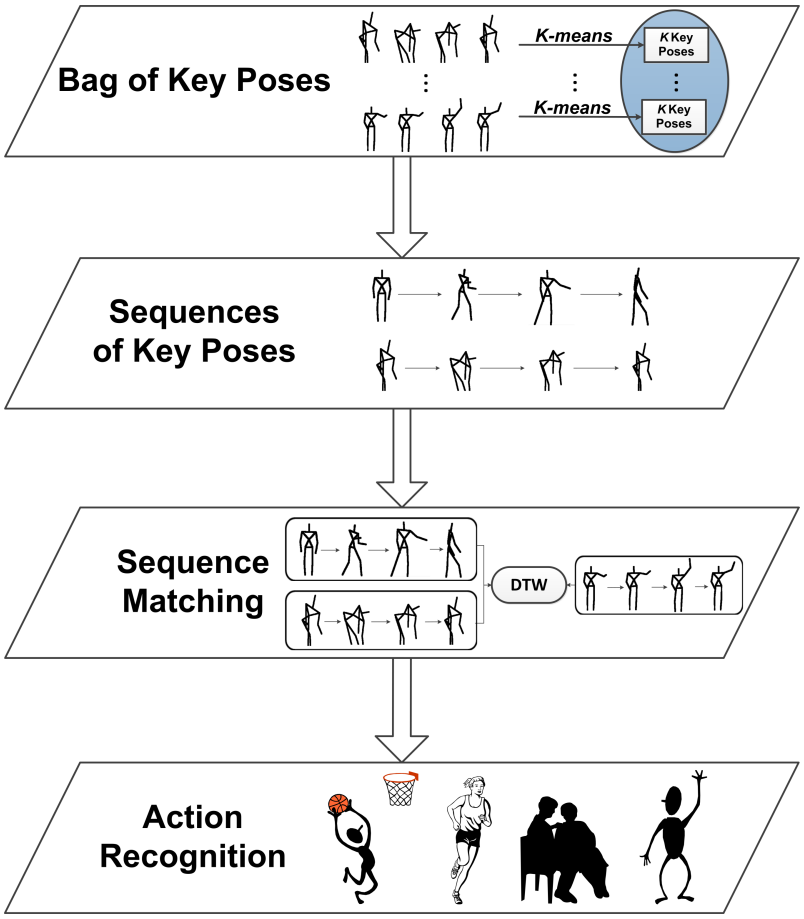


Figure 5: Overview of the employed human action recognition method.

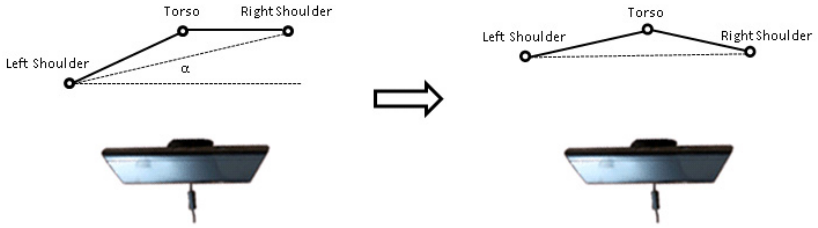


Figure 6: Rotation of the skeleton with respect to the Kinect.

body pose feature based on the data received from the RGB-D device needs to be provided.

In this sense, position, scale and rotation invariance is achieved by applying the proposed normalisation process detailed in Algorithm 2. As it can be seen, measures taken from the first instance are used in order to normalise the whole sequence. Then, the final feature vector is built up with the 3D coordinates of the joints, where only the joints that are selected during the evolutionary joint selection are used (Fig. 7).

Algorithm 2 Normalising algorithm

Determine the **normalising length** as the distance from the TORSO to the NECK in the first skeleton of the sequence

Determine the **y-axis rotation** α of the line connecting both shoulders with respect to the kinect (Fig. 6) in the first skeleton of the sequence

for all the skeletons in the sequence **do**
Set as (0,0,0) coordinate of the skeleton the average location of the TORSO, the LEFTSHOULDER, the RIGHTSHOULDER, the LEFTHIP and the RIGHTHIP.
for all the joints **do**
Translate according to the new reference centre
Normalise the coordinates according to the normalising length
end for
Rotate the skeleton α degrees about the vertical axis passing by the NECK (Fig. 6)
end for

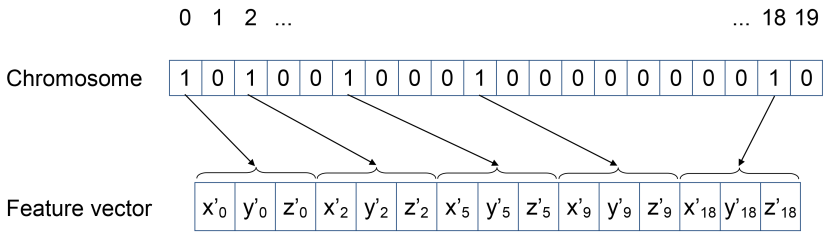


Figure 7: Construction of the feature vector. (x'_i, y'_i, z'_i) represent the 3D coordinates of joint i after the application of the normalisation algorithm.

4.2. Bag of key poses

Once the feature vectors are obtained they are employed in the learning algorithm. Key poses are generated in order to represent the most common and characteristic pose representations of each action class, and to reduce the scale of the problem. In this sense, K-means clustering is performed for the instances of each of the c action classes. The returned cluster centres are used as representative values and joined together. The resulting bag of key poses constitutes a dictionary of the most relevant poses for each action class (Fig. 8 shows an overview of the process).

4.3. Sequence matching

In this step, the temporal relationship between key poses is modelled. Since the evolution of the human pose over time can provide useful data so as to improve the recognition, a classification based on sequence matching is targeted. For this purpose, sequences of key poses are built. For each of the available training sequences, the individual skeletal representations are substituted with the *nearest neighbour* key pose out of the bag of key poses. The successive key poses make up a simplified sequence of key poses: $S = \{kp_1, kp_2, ..., kp_t\}$. At this point, not only the typical order and transitions between key poses are captured, but also noise and outlier values are filtered.

In the recognition stage, the same procedure is initially performed: 1) the skeletal representations are obtained for the RGB-D images of the video sequence to recognise, and 2) the sequence of key poses is built by finding the successive *nearest neighbour* key poses using the bag of key poses. Then, sequence matching can be performed. For this purpose, the DTW algorithm has been chosen, as it is able to successfully align sequences with consistent temporal order, but meaningful differences in speed. This is very

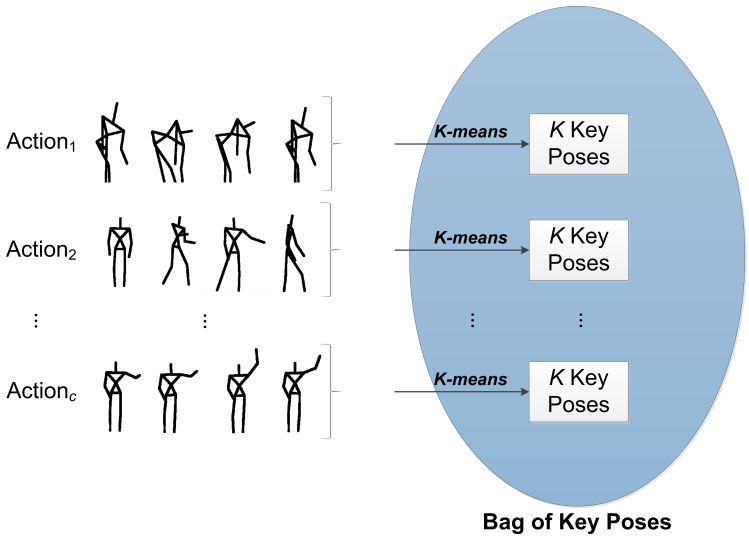


Figure 8: The bag of key poses is obtained by generating K key poses for each action class using K-means clustering, and merging the corresponding key poses together. Samples of the actions *bend*, *forward punch* and *high arm wave* are shown.

desirable in the comparison of action performances due to the different pace at which humans of unlike age and condition perform actions.

The DTW distance $d_{DTW}(S_{train}, S_{test})$ between two sequences of key poses $S_{train} = \{kp_1, kp_2, ..., kp_t\}$ and $S_{test} = \{kp'_1, kp'_2, ..., kp'_u\}$ is defined as:

$$d_{DTW}(S_{train}, S_{test}) = dtw\,(t, u) \quad , \tag{1}$$

$$dtw\,(i, j) = \min \left\{ \begin{array}{l} dtw\,(i-1, j)\,, \\ dtw\,(i, j-1)\,, \\ dtw\,(i-1, j-1) \end{array} \right\} + d(kp_i, kp'_j) \quad , \tag{2}$$

where $d(kp_i, kp'_j)$ is the Euclidean distance between two key poses. Hence, the label of the closest training sequence, i.e. the best match, will be returned as the final result of the classification. Fig. 9 shows an example for simplified one-dimensional sequence elements. It can be observed how the first and last elements are always matched, and the alignment in between is chosen based on the lowest distance. In this case, the final DTW distance $dtw(5, 6) = 9$.

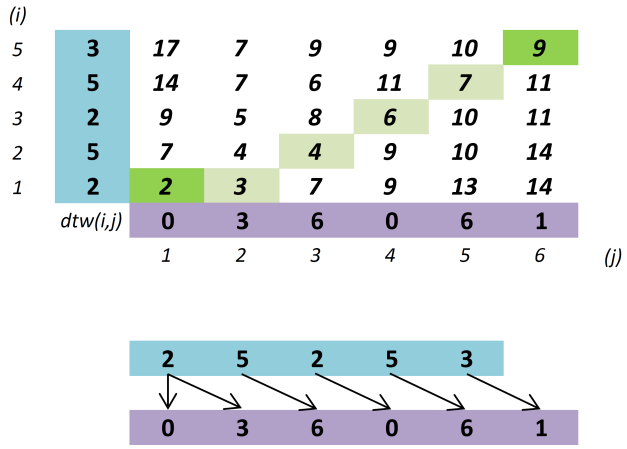


Figure 9: An example of the DTW algorithm for a simplified one-dimensional case (top), and the final alignment between elements (bottom) are shown. Note that the matrix elements indicate the accumulated distance between elements for the partial alignment with the lowest distance (following Eq. 2).

5. Experimentation

The proposed method has been evaluated with the Microsoft Action3D dataset (Li et al., 2010). This dataset contains 20 different actions (see Table 1), performed by 10 different subjects and with up to 3 different repetitions which makes a total of 567 sequences. However, 10 sequences are not used because the skeletons were either missing or wrong, as explained by the authors[1]. This dataset uses the 20-joint model described in Sec. 3.1.

Table 2 shows the three subsets of 8 gestures each, which have been commonly used in order to reduce the computational cost of the tests (Li et al., 2010). The AS1 and AS2 subsets group actions with similar movement, whereas AS3 groups more complex actions together.

Similarly to Li et al. who first used this dataset, we perform a cross-subject validation. Since most works do not state how they divided the subjects into two groups, i.e. the train and test data, we decided to employ the approach from Azary & Savakis (2012), which performs a 2-fold cross validation.

[1]MSR Action Recognition Datasets and Codes, http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm (last access: 24/05/2013)

Table 1: Actions in the *MSR-Action3D* dataset.

Label	Action name	Label	Action name	Label	Action name
a01	High arm wave	a08	Draw tick	a15	Side-kick
a02	Horizontal arm wave	a09	Draw circle	a16	Jogging
a03	Hammer	a10	Hand clap	a17	Tennis swing
a04	Hand catch	a11	Two-hand wave	a18	Tennis serve
a05	Forward punch	a12	Side-boxing	a19	Golf swing
a06	High throw	a13	Bend	a20	Pick-up and throw
a07	Draw cross	a14	Forward kick		

Table 2: Actions in each of the *MSR-Action3D* subsets.

AS1	AS2	AS3
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw cross	Side-kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two-hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pick-up and throw	Side-boxing	Pick-up and throw

Table 3: Classification rate for each subset.

	Dataset			
Method	**AS1**	**AS2**	**AS3**	**Average**
DMM-HOG (Yang et al., 2012)	**96.2**%	84.1%	94.6%	91.63%
EigenJoints (Yang & Tian, 2012)	74.5%	76.1%	96.4%	82.33%
OESGP (Soh & Demiris, 2012)	80.6%	74.9%	87.1%	80.87%
Sparse Repr. (L1-norm) (Azary & Savakis, 2012)	77.66%	73.17%	91.58%	80.80%
Sparse Repr. (L2-norm) (Azary & Savakis, 2012)	76.60%	75.61%	89.47%	80.56%
Keyposes and Decision Forests (Miranda et al., 2012)	93.5%	52%	95.4%	80.30%
Histograms of 3D Joints (Xia et al., 2012)	87.98%	85.48%	63.46%	78.97%
Bag of 3D Points (Li et al., 2010)	72.9%	71.9%	79.2%	74.67%
Bag of key poses and DTW (Chaaraoui et al., 2012)	78.90%	74.12%	89.21%	80.74%
Our method	91.59%	**90.83%**	**97.28%**	**93.23%**

The size of the population has been set to $N = 10$ individuals. Instead of having a static mutation probability, each time a random value for p_{mut} in the interval $[0, 0.2]$ is selected. This is done to try to avoid early convergence of the evolution. The termination condition is established as $gen_{max} = 250$ generations without changes in the fitness of the best individual of the population. All these parameters have been chosen experimentally.

Table 3 shows the best results we have obtained with our evolutionary algorithm (each test has been performed 10 times). In the subsets AS2 and AS3, the available recognition rates are outperformed reaching, to the best of our knowledge, the highest results so far. A promising result is achieved in AS1 where the result is better than the obtained by most of the methods. In average our approach also obtains the best results. As it was expected, and happens with most of the other algorithms, the results for AS1 and AS2 are worse than those for AS3, as gestures are more similar. Comparing with the results of the original (without the evolutionary optimisation) bag of key poses and DTW method, the optimisation improves the results considerably (16.08% for AS1, 22.54% for AS2, and 9.05% for AS3).

We have repeated similar tests using Leave-One-Actor-Out (LOAO) validation. In this cross validation test, actor-invariance is specifically tested by training with all but one actor, and testing the method with the unseen one. This is repeated for all actors, averaging the returned accuracy scores. Results are presented in Table 4.

Fig. 10 shows the confusion matrices for the recognition method without and with the application of the evolutionary optimisation applying the LOAO cross validation. In all the cases, the success rates match or improve the

Table 4: LOAO classification rate for each subset.

	Dataset			
Method	**AS1**	**AS2**	**AS3**	**Average**
Bag of key poses and DTW (Chaaraoui et al., 2012)	82.35%	77.88%	93.48%	84.57%
Our method	91.46%	91.78%	97.13%	93.46%

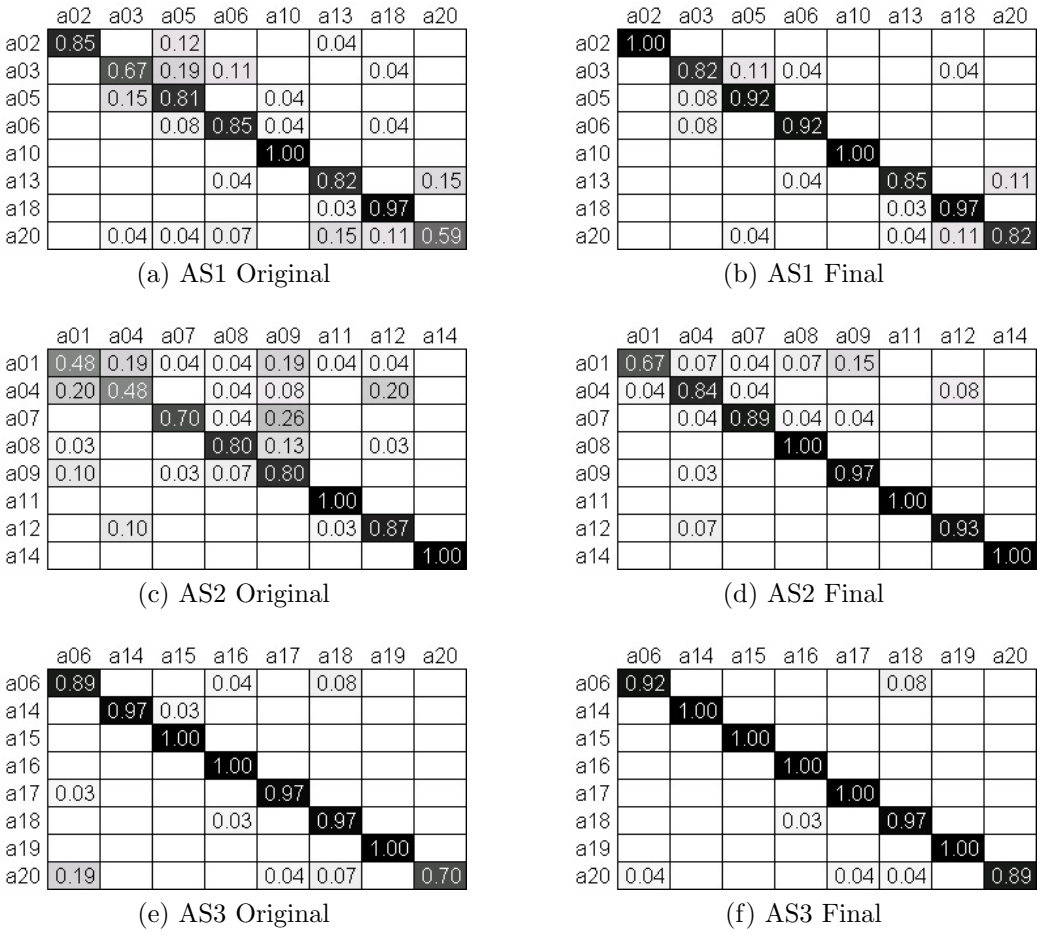
Table 5: Computational cost (in frames per second) for each subset.

	Dataset		
Method	**AS1**	**AS2**	**AS3**
Bag of key poses and DTW (Chaaraoui et al., 2012)	280.35	286.60	263.34
Our method	645.91	529.83	445.29
Improvement	**119.7%**	**89.0%**	**69.1%**

previous rates. In the case of the AS1 dataset, results are quite good. *Pick-up and throw (a20)* is a complex action composed of the sequences of *bend (a13)* and *high throw (a06)*, both actions also included in AS1. Besides, there are many *Pick-up and throw* sequences included in the dataset where the skeletons are wrongly calculated, but we did not remove them in order to get the results in the same conditions than the previous works. In AS2, the gestures are very similar as all are performed with the arms except *Forward kick (a14)*. Because of that, there is an important confusion between gestures that is considerably reduced after applying the evolutionary algorithm.

Fig. 11 shows the joints which have been selected for each of the datasets, as those are the ones that contribute the most to the recognition algorithm. Since AS1 includes mainly gestures performed with either one or both arms, these are basically the selected joints; in addition to the HEAD which is significant in the *bend* action. Similar conclusions can be drawn for AS2 and AS3, where joints from the legs or feet are selected, since actions as *forward kick* or *jogging* are included.

The reduction in the size of the feature vector has also an important effect in the computational cost of the recognition process. The Kinect device captures RGB and depth at a rate of 30 frames per second (fps). Table 5 shows the rates that the original bag-of-key-poses method and our evolutionary optimisation are able to reach. These rates are far superior to those 30 fps, therefore allowing real-time processing.

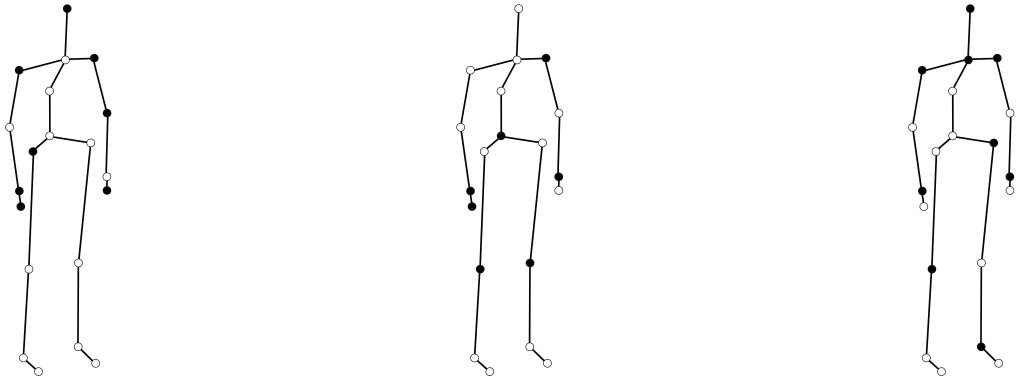


(a) AS1 Original (b) AS1 Final

(c) AS2 Original (d) AS2 Final

(e) AS3 Original (f) AS3 Final

Figure 10: Confusion matrices for each one of the datasets without (Original) and with (Final) the application of the evolutionary algorithm.



(a) AS1 (b) AS2 (c) AS3

Figure 11: Final feature subsets for each of the datasets (selected joints are shaded in black, ignored ones are left unshaded).

6. Discussion and future work

In this paper, we have proposed an evolutionary algorithm to improve action recognition using RGB-D cameras. Joint selection allows to improve the success rate obtaining better results than the state of the art. The reduction of the size of the feature vector, i.e. the active joints that have been chosen by the evolutionary algorithm, has an important impact on the processing rate, since the recognition is performed faster. That will allow the development of more complex recognition algorithms and their application in real-time. We have also presented a tree representation of the individual and a crossover operator that is aware of this structure in order to obtain better solutions faster. Once the relevant joints are selected, the feature vector used as input for the classifier is simple (composed by the 3D coordinates of the selected joints). Existing research employs more complex and larger feature vectors composed by distances between joints, quaternions, etc. These are alternatives that we need to consider in the future. In order to use the 3D coordinates of the joints, we propose an algorithm to normalise the sequences of skeletons to scale and rotation.

Two main difficulties have been encountered in our approach: the high computational cost of the wrapper approach and early convergence. A wrapper-based evolutionary feature subset selection approach requires the calculation of the fitness of an important number of solutions (individuals) until the

final solution is obtained. As a single fitness calculation involves a complete training and recognition process, the whole evolution could take a considerable time. Nevertheless, this calculation is offline. Once the best subset of features is obtained, this is used for recognition with a low computational cost, which allows real-time recognition. Early convergence happens when the evolutionary search is stacked in a local minimum and cannot achieve a good solution. In order to solve these problems, future work will be devoted to the design of variants of this proposal or more specialised evolutionary approaches. For instance, memetic algorithms (Moscato & Cotta, 2010; Ang et al., 2010), which combine evolutionary operators and local search procedures, would be a good option. Real-coded chromosomes would allow to assign weights to each of the joints. This will affect all the process, as this weighted joint selection must also be considered in the recognition method.

Besides, we have applied our evolutionary approach to an action recognition method that we developed for RGB images. This algorithm, when applied to skeletal data, obtains classification rates in the middle of the ranking of the state of the art. So, our optimisation approach obtains very good results using a recognition method that was not specifically designed for RGB-D devices. Therefore, the next step will be to apply it to the best and more recent methods in the state of the art (Wang et al., 2012a; Yang et al., 2012; Wang et al., 2012b) in order to study the level of optimisation achieved.

Acknowledgements.

This work has been partially supported by the European Commission under project "caring4U - A study on people activity in private spaces: towards a multisensor network that meets privacy requirements" (PIEF-GA-2010-274649) and by the Spanish Ministry of Science and Innovation under project "Sistema de visión para la monitorización de la actividad de la vida diaria en el hogar" (TIN2010-20510-C04-02). Alexandros Andre Chaaraoui and José Ramón Padilla-López acknowledge financial support by the Conselleria d'Educació, Formació i Ocupació of the Generalitat Valenciana (fellowships ACIF/2011/160 and ACIF/2012/064 respectively). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

This article was originally published in *Alexandros André Chaaraoui, José Ramón Padilla-López, Pau Climent-Pérez, Francisco Flórez-Revuelta, Evolutionary Joint Selection to Improve Human Action Recognition with*

RGB-D devices, Expert Systems with Applications, Available online 22 August 2013, ISSN 0957-4174, `http://dx.doi.org/10.1016/j.eswa.2013.08.009`.

References

Alnowami, M., Alnwaimi, B., Tahavori, F., Copland, M., & Wells, K. (2012). A quantitative assessment of using the Kinect for Xbox360 for respiratory surface motion tracking, . (pp. 83161T–83161T–10).

Ang, J., Tan, K., & Mamun, A. (2010). An evolutionary memetic algorithm for rule extraction. *Expert Systems with Applications*, *37*, 1302 – 1315.

ASUSTeK Computer Inc (2013). ASUS - Xtion PRO. http://www.asus.com/Multimedia/Xtion_PRO. Last access: 15/05/2013.

Azary, S., & Savakis, A. (2012). 3D Action Classification Using Sparse Spatio-temporal Feature Representations. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller, & M. Papka (Eds.), *Advances in Visual Computing* (pp. 166–175). Springer Berlin / Heidelberg volume 7432 of *Lecture Notes in Computer Science*.

Bobick, A., & Davis, J. (2001). The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence*, *23*, 257–267.

Cantú-Paz, E. (2004). Feature Subset Selection, Class Separability, and Genetic Algorithms. In K. Deb (Ed.), *Genetic and Evolutionary Computation GECCO 2004* (pp. 959–970). Springer Berlin / Heidelberg volume 3102 of *Lecture Notes in Computer Science*.

Casado Yusta, S. (2009). Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters*, *30*, 525–534.

Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2012). An Efficient Approach for Multi-view Human Action Recognition Based on Bag-of-Key-Poses. In A. A. Salah, J. Ruiz-del Solar, C. Meriçli, & P.-Y. Oudeyer (Eds.), *Human Behavior Understanding* (pp. 29–40). Springer Berlin Heidelberg volume 7559 of *Lecture Notes in Computer Science*.

Chang, Y.-J., Chen, S.-F., & Huang, J.-D. (2011). A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in Developmental Disabilities*, *32*, 2566 – 2570.

Climent-Pérez, P., Chaaraoui, A. A., Padilla-López, J. R., & Flórez-Revuelta, F. (2013). Optimal Joint Selection for Skeletal Data from RGB-D Devices Using a Genetic Algorithm. In I. Batyrshin, & M. Mendoza (Eds.), *Advances in Computational Intelligence* (pp. 163–174). Springer Berlin Heidelberg volume 7630 of *Lecture Notes in Computer Science*.

Fothergill, S., Mentis, H., Kohli, P., & Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* CHI '12 (pp. 1737–1746). New York, NY, USA: ACM.

Gallo, L., Placitelli, A., & Ciampi, M. (2011). Controller-free exploration of medical image data: Experiencing the Kinect. In *24th International Symposium on Computer-Based Medical Systems (CBMS), 2011* (pp. 1 –6).

Huang, J.-D. (2011). Kinerehab: a kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. In *The proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* ASSETS '11 (pp. 319–320). New York, NY, USA: ACM.

Janoch, A., Karayev, S., Jia, Y., Barron, J. T., Fritz, M., Saenko, K., & Darrell, T. (2011). A category-level 3-D object dataset: Putting the Kinect to work. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 1168–1174).

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Attention, Perception, & Psychophysics*, *14*, 201–211.

John, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning* (pp. 121–129). San Francisco, CA: Morgan Kaufmann.

Jong, K. A. D. (2006). *Evolutionary computation - a unified approach*. MIT Press.

Khoshelham, K., & Elberink, S. O. (2012). Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors*, *12*, 1437–1454.

Kira, K., & Rendell, L. A. (1992). The feature selection problem: traditional methods and a new algorithm. In *Proceedings of the tenth national conference on Artificial intelligence* AAAI'92 (pp. 129–134). AAAI Press.

Lanzi, P. (1997). Fast feature selection with genetic algorithms: a filter approach. In *IEEE International Conference on Evolutionary Computation, 1997* (pp. 537–540).

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3D points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 9 –14).

Liu, H., & Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers.

Mastorakis, G., & Makris, D. (2012). Fall detection system using Kinect's infrared sensor. *Journal of Real-Time Image Processing*, (pp. 1–12).

Microsoft Corporation (2013). Kinect for Windows. Voice, Movement & Gesture Recognition Technology. http://www.microsoft.com/en-us/kinectforwindows. Last access: 15/05/2013.

Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A. W., & Campos, M. F. M. (2012). Real-time gesture recognition from depth data through key poses learning and decision forests. In *Sibgrapi 2012 (XXV Conference on Graphics, Patterns and Images)*. Ouro Preto, MG: IEEE.

Moscato, P., & Cotta, C. (2010). A Modern Introduction to Memetic Algorithms. In M. Gendreau, & J.-Y. Potvin (Eds.), *Handbook of Metaheuristics* (pp. 141–183). Springer US volume 146 of *International Series in Operations Research & Management Science*.

Ni, B., Wang, G., & Moulin, P. (2011). RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 1147–1153).

Obdrzalek, S., Kurillo, G., Ofli, F., Bajcsy, R., Seto, E., Jimison, H., & Pavel, M. (2012). Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (pp. 1188 –1193).

Polana, R., & Nelson, A. (1997). Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision*, *23*, 261–282.

PrimeSense, Ltd. (2013). OpenNI. The standard framework for 3D sensing. http://www.openni.org. Last access: 15/05/2013.

Raptis, M., Kirovski, D., & Hoppe, H. (2011). Real-Time Classification of Dance Gestures from Skeleton Animation. In *Proceedings of the 10th Annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2011* (pp. 147–156).

Seo, D. W., & Lee, J. Y. (2013). Direct hand touchable interactions in augmented reality environments for natural and intuitive user experiences. *Expert Systems with Applications*, *40*, 3784 – 3793.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1297 –1304).

Siedlecki, W., & Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, *10*, 335 – 347.

Soh, H., & Demiris, Y. (2012). Iterative temporal learning and prediction with the sparse online echo state gaussian process. In *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Stone, E., & Skubic, M. (2011). Passive in-home measurement of stride-to-stride gait variability comparing vision and Kinect sensing. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (pp. 6491 –6494).

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2011). Human activity detection from RGBD images. In *AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*.

Wang, C.-M., & Huang, Y.-F. (2009). Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. *Expert Systems with Applications*, *36*, 5900 – 5908.

Wang, J., Liu, Z., Chorowski, J., Chen, Z., & Wu, Y. (2012a). Robust 3D Action Recognition with Random Occupancy Patterns. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision ECCV 2012* Lecture Notes in Computer Science (pp. 872–885). Springer Berlin Heidelberg.

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012b). Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*. Providence, Rhode Island.

Xia, L., Chen, C.-C., & Aggarwal, J. (2012). View invariant human action recognition using histograms of 3D joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 20 –27).

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. *Intelligent Systems and their Applications, IEEE*, *13*, 44–49.

Yang, X., & Tian, Y. (2012). EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. In *Second International Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR, 2012* (pp. 14–19). Providence, Rhode Island.

Yang, X., Zhang, C., & Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In N. Babaguchi, K. Aizawa, J. R. Smith, S. Satoh, T. Plagemann, X.-S. Hua, & R. Yan (Eds.), *ACM Multimedia* (pp. 1057–1060). ACM.