Generating Local Temporal Poses from Gestures with Aligned Cluster Analysis for Human Action Recognition

Mike Edwards

Xianghua Xie

Swansea University
Swansea, UK

Abstract

The use of pose estimation for human action recognition has seen a resurgence in previous years, due in part to the natural representation of the activity as a sequence of key poses and gestures. The use of sequence alignment techniques has aided the process of comparing between sequences of differing temporal rates, with aligned cluster analysis segmenting an observation into lower level action primitives. We suggest that the representation of a given action class via its lower level gestures can help to identify the higher-level action class label. We therefore present a method for the generation of key poses via the initial segmentation of an action class into gestures that are similar across numerous observations. We treat all training observations as a single observation in which there are repetitions of the same action class. By applying segmentation, we then identify common gestures across the class, which are used to generate the key poses we optimize via evolutionary programming. Global recognition rates of 97.4% are achieved using a subset of the MSR Action3D dataset. We then expand the method to recognize interaction events between two individuals using the SBU Kinect Interaction dataset, achieving recognition rates of 83.9% and over 96.4% when observing the first 6 classes.

1 Introduction

Human Action Recognition (HAR) is a field concerned with the detection and identification of different human behavior classes observed within a scene. As such, it is a topic that provides benefit to numerous problem domains; including surveillance, human-computer interaction and medical diagnostics [1]. Events are often categorized based on complexity in gesture, action, interaction, or group activity, yet an activity can potentially be a mixture of lower-level gesture types, *e.g.*, 'walk' contains gestures including 'lift leg', 'swing leg forward', and 'lower leg' [21]. Current appearance and pose-based methods have lead to the accurate recognition of simplistic actions and gestures, such as 'waving', 'running', and 'jumping' [4, 9, 22]. There has been recent renewed interest in the use of pose-based HAR, partly due to the availability of commercial depth sensor systems which are able to track body joint locations with reasonable accuracy [16, 27].

There are several key issues to consider in HAR analysis. Often individuals may perform the same actions with both the intra- and inter-subject level variation in their spatial or temporal execution [24]. Therefore it is necessary to develop methodologies that are able to deal with the impact of spatio-temporal variation on an intra- and inter-subject level, whilst maintaining partitioning information at the inter-class level. Recent approaches have made use of sequence alignment to allow temporal comparison between actions, key pose representation to study the underlying gesture composition of an action, and segmentation to identify gestures within a sequence [11, 30].

Previous study on the use of pose estimation has promoted the use of a Bag of Key Poses (BoKP) model, in which representative key spatial poses form a bag of words, which can be compounded to describe higher-level actions [3, 9, 11, 13, 20]. To achieve this, k key poses are generated by clustering similar frames from a whole sequence. Transforming a sequence into a key pose representation reduces the impact minor frame-to-frame spatial variations, provided that sensible key poses are generated [30], as representative poses are produced for each class and stored within one bag [3, 9, 13, 20].

To align actions that are linear sequence of poses, which may vary in temporal execution, sequence alignment techniques such as Dynamic Time Warping (DTW) [18, 23], Dynamic Manifold Warping (DMW) [14, 15], and Canonical Time Warping (CTW) [29], have been employed to reduce the impact of temporal variations, [5, 6, 11, 22]. These methods have however been criticized in situations where the temporal execution rate may provide some key information between two classes, *e.g.* 'run' and 'walk' [7], or there are repeated cyclic gestures within the action [25]. In some cases, an action can be defined by its accumulated composition of primitive poses, forming a bag of words representation [11]. In both of these situations we believe it is beneficial to first segment the observation to identify any repeated primitive gestures. This will identify cyclical or compound gestures that form a higher level action. To segment an observation, [30, 31] utilize DTW to group varying length segments into k clusters by dynamic programming via Aligned Cluster Analysis (ACA) and Hierarchical Aligned Cluster Analysis (HACA).

Finding an optimal set of classification parameters is non-trivial, and optimization requires the selection of informative training samples and features to reduce the impact of outliers in the action space. Evolutionary programming methods have provided optimum selection of training instances [6], and also informative features for a given observed action class [8]. This online learning has an attractive application for HAR, learning new action classes without having to retrain a classifier in an offline fashion.

To efficiently recognize interactions between two people, whilst providing a method of key pose generation that reflects the composition of higher level actions in terms of their shared gesture dictionary, we present a means of using sequence alignment to obtain sub-action gesture segmentation across all training observations. The ability of ACA to cluster similar segments of frames from a sequence, combined with the benefit of recognizing repeated sub-action segments via temporally flexible DTW, presents a method of segmenting similar sub-actions between multiple observations of an action class. These segmented gestures are then represented as key poses in an evolving bag representation, thus identifying key poses of a local temporal region which is repeated across training instances. By moving towards the recognition of more complex scenarios we hope to eventually lead towards recognition of higher-level, complex interactions between individuals; such as the context specific interactions discussed by [11, 12].

The rest of the paper is organized as follows: In Section 2 we describe our method of using ACA to identify cross-subject gestures before extracting key poses for each gesture

cluster, and the evaluation techniques undertaken. In Sections 3 and 4 we draw conclusions on the predictive abilities of the proposed method, evaluating performance with two publicly available datasets.

2 Methods

We propose the use of sequence alignment and segmentation methodology to identify cross-subject gestures to generate a key pose representation for action and interaction recognition. By identifying descriptive poses within each gesture we are able to more accurately represent sub-action primitives which compound to form a given gesture. We propose that understanding these gestures may in turn benefit the learning of higher level actions. We utilize ACA to generate segments for a given action, using these gesture clusters to identify key poses. The key pose space allows reduction of spatial variation within the observations, providing more accurate sequence alignments. The sequences of key poses are then used to generate a nearest neighbour classifier for predicting labels of newly observed sequences. In order to identify suitable pose generation parameters we utilize evolutionary programming to select informative training observations and features from the input data.

2.1 Segmentation of Gestures

The observation of an action is often the compounding of numerous poses into a sub-action gesture, with multiple gestures then forming the given class. Therefore a set of observed sequences for an action, X_a, where A is all possible actions, often contains a set of sub-action gestures which best describe X_a. These gestures are a temporally ordered sequence of key poses, the frame-by-frame pose of the human body. With the intention to recognize similar primitive gestures across all observations of X_a, we represent all training instances of the given action a as a single continuous sequence. Unlike previous methods, which use k-means as a method of determining the k key poses, we make use of the ACA methodology presented by [30, 31] to first cluster similar action primitives into k' gesture clusters common across all training instances of a given action. A minimum and maximum segment length is selected and all possible segment sizes within that range are iteratively clustered using a dynamic time alignment kernel. Using DTW, possible segments are aligned to current members of each cluster and allocated to the most similar cluster, with each iteration minimizing within-cluster variance, segmenting out similar repeated gestures across subjects. Once we obtain the ACA segmentation, we find the k cluster centroid poses by k-means clustering over each frame of a gesture cluster.

Usual key pose generation draws representative poses from all frames of observed sequences, which may cause motion with a gesture to be lost in key pose representation. In comparison, by identifying key poses within sub-action gestures we are formulating a key pose representation that reflects the gestures that in turn compound to form an action. This produces key poses from gestures that are observable across numerous subjects, providing informative poses that compose each gesture.

2.2 Sequence Alignment and Prediction

The identified gesture poses are used to generate a DTW nearest neighbour classifier to provide label predictions for new observations of a given action class. For each of the training

samples we produce a key pose representation, reducing noisy spatial variation present between frames, by picking the nearest pose to the current frame. This is repeated for each of the testing samples. Using DTW as a nearest neighbour distance metric we are able to reduce the effects of slight temporal fluctuation in execution rates common in HAR. A test sample is predicted to share the label of the training sample with the shortest warping distance.

2.3 Parameter Optimization

To optimize our gesture-based key poses we expand upon the evolutionary programming explored in [6], allowing parameters for our models to be identified over each generation. We first construct a population $P_{1:n}$ containing n individuals, with each individual represented via a genomic sequence, $p_i = [g_1, ..., g_L]$; where each gene vector, g_l, represents a parameter of the model. The training instance vector, $i_{1:n} \in \{0,1\}$ for n training samples, is a binary indicator of whether a given sample in the training set is used to generate the key poses. The vector is updated by random initialization as training instances are added to the system. This vector aims to optimize the selection of informative training samples. The parameter vector, $p_{1:a} \in \{\mathbb{N}\{1, ..., K\}\}$, where K is an upper limit constraint on the possible number of key poses with which to populate the bag. Smaller k results in coarser approximation of action class a. Should the system learn a new action, then A is increased by 1 and representative key poses are learnt for the new class. The feature selection vector, $f_{1:s \times m} \in \{0,1\}$ for m possible joints and s subjects, is a binary indicator denoting if a given feature is used to generate key poses. By treating the individual subjects in a scene separately, we are able to optimize which joints are informative to the overall class; this is beneficial when an interaction class has the same label, but the two subjects react differently across instances. In each generation all individuals are ranked on poses they produce, maximizing the correct predicted class labels obtained by sequence alignment classification outlined in Section 2.2.

Standard Evolutionary Algorithm (EA) operators are used for *reproduction*, *recombination*, *mutation*, and *ranking* of the population within each generation. *Recombination* for i and p occurs as outlined within common practice of EA, via single point swapping [2]. Recombination of f occurs via a domain specific crossover method, in which a joint and its dependent branch is substituted with the second parent. Recombination helps to defer convergence onto a homogeneous population by introducing variation of genes between parents and offspring. Once a new offspring is generated, *mutation* provides variation within the population gene-pool, attempting to avoid optimizing towards a local minimum by widening the search space. Each gene within vectors $i_{1:n}$ and $f_{1:s \times m}$ are subject to a binary flip based on their respective mutation rates; whilst genes in vector p are either sampled from a random distribution over all possible values, or from a Gaussian distribution over a localized range, each with equal chance. For observations of human interaction we must modify the operators to handle two individuals, thus the recombination operator has a domain specific crossover method that accounts for the semantics that describe each of the two observed individuals.

2.4 Implementation and Evaluation

The method presented requires a number of initialization parameters to be selected before they are optimized using evolutionary programming. For all experiments, the mutation probability was dynamically selected via a random distribution from between 0.0 and 0.1 for instance vector i, and between 0.0 and 0.2 for vectors f and p. In our study, the Gaussian standard deviation for mutation of genes within parameter vector p was empirically set to

Table 1: Action class sets used for evaluation, with generation in which they are introduced to the population. The first two actions are introduced simultaneously for initialization.

Generation	AS2 [19]	SBU [28]
0	High arm wave	Approaching
0	Hand catch	Departing
50	Draw x	Pushing
100	Draw tick	Kicking
150	Draw circle	Punching
200	Two hand wave	Exchanging
250	Forward kick	Hugging
300	Side boxing	Handshake

$\sigma = 4$, producing a small localized mutation search space when using Gaussian gene mutation. For our evolutionary optimization, for comparability to [6], we selected an initial population size of $n = 10$, with 10 offspring created at each generation. For initial seeding of the population, genome vectors are randomly initialized. In both single action and interaction experiments we limited the number of generations per action to 50, increasing this to 100 made little difference in the overall accuracy of the populations; however evolutionary optimization can be repeated indefinitely to allow for the time restriction to be relaxed on the optimization. For k-means clustering, we limited the maximum value of k to 40 key poses; as it appeared to provide both a decrease in complexity, and marginal increase in accuracy on the 75 poses used by [6]. This is only an upper limit on the number of key poses generated, and increase in accuracy observed may be an artefact from initialization, however we noticed no noticeable hindrance to the system.

2.4.1 Single Action

For single person HAR, the proposed method was evaluated on the 20 tracked joints of the MSR Action3D dataset [19]. This dataset is a commonly used standard for single person HAR methods, and is comprised of 3 subsets containing various action classes; containing 20 action classes performed by 10 subjects, repeated up to 3 times. We evaluate our method on the AS2 subset, which is viewed as the most complex of the 3; utilizing the train/test split outlined by [19], producing a *leave-one-actor-out* cross-subject validation analysis. The AS2 set contains the 8 action classes listed in Table 1. Joint coordinates are utilized as features, with each gene of the feature vector f representing a joint marker, $f_{1:20}$. This subset was also used to evaluate the methods in [6], which we have also implemented here for cross-comparison. For ACA sequence segmentation on the MSR dataset, we empirically initialized the segmentation method to group the total-instance sequence into $k' = 5$ sub-action gesture clusters, with a segment length limitation of between $nMin = 1$ and $nMax = 10$ frames.

2.4.2 Two Person Interactions

To evaluate the method for the purpose of interaction recognition we utilize the SBU Kinect Interaction dataset [28]; a dataset consisting of 8 interaction classes, Table 1. 21 pairs of subjects performed actions up to 3 times, and 15 joints were tracked via Kinect. Following the 5-fold cross validation split outlined in [28], with 4-5 interaction pairs per fold. Joint

Table 2: Global recognition rate (%) across all action classes at final generation of evolutionary optimization.

Dataset	[6]	k-means	Proposed gesture key poses
MSR Action3D	88.6	96.3	**97.4**
SBU Kinect Interaction	-	83.3	**83.9**

coordinates features are utilized, with the pairwise interaction encoded as a 30 dimensional temporal sequence, $f_{1:15}$ representing person A and $f_{16:30}$ person B. To indicate this in population genetics, the recombination of a given feature vector $f_{1:30}$ occurred via a modified domain specific method; if the cross point fell between $f_{1:15}$, any dependent joints along the branch would be taken from the person A on the second parent, while cross points falling on genes in the range $f_{1:15}$ were selected from person B. By this method we have chosen to maintain domain specific recombination whilst applying it to handle the two individuals observed in the scene. To obtain ACA segmentation of the SBU dataset, we initialized the segmentation cluster value $k' = 5$ sub-action clusters, with gesture length between $nMin = 1$ and $nMax = 4$ frames as there are a lot of cyclic motions within the SBU class which have very short repetition rates.

3 Results

We present the findings of our proposed method within Table 2 and Figures 1 and 3, utilizing ACA segmentation to generate the key poses used in recognizing observed action input sequences. The results shown are the averaging of cross-fold validation as detailed in Section 2, over 3 replicate runs.

3.1 Single Person Action

The proposed method of obtaining key poses from segmented gestures achieved a global accuracy that improves upon the comparable k-means method outlined by [6]. The prior segmentation of class training samples is able to extract informative gestures from the action class, with subsequent clustering of within-gesture poses identifying poses that are able to more comprehensively describe the action classes observed. As expected, introduction of a new action class does have negative effect on recognition rates of the currently optimized population. This initially results in decreased accuracy due to random initialisation of the genetic representation of the new action class. However, evolutionary optimization returns the population to an acceptable level, as seen by the increase in accuracy in following generations, Figure 1.

Prior clustering of the action class into cross-subject gestures works well to produce key poses for sequence alignment based classification. The evolutionary method compliments this by adapting to the introduction of new action classes, optimizing towards the most informative set of parameters for the model. The MSR Action3D dataset is a common dataset within the pose estimation community, and recognition rates of 97.4% on perceivably its most complex subset are an indicator of the benefit to using gesture segmentation in the identification of key poses. From Figure 2 we can see that common error lies in partitioning between the classes '*high arm wave*' and '*side boxing*', where each individual in the population was unable to classify one of the testing samples. There was also a smaller level of
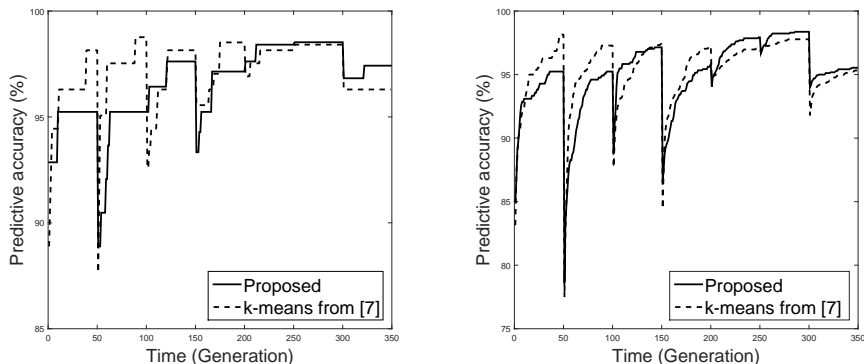
Figure 1: L-R: Maximum and Average predictive accuracy of population when classifying single person actions on the MSR AS2 dataset. A new class is introduced every 50 generations.

confusion in predicting between '*hand catch*' and '*side boxing*' classes. Surprisingly there was little confusion in the recognition of the classes '*draw x*', '*draw tick*' and '*draw circle*', those which we would presume to contain the most subtle action gestures. The large fall in classifier accuracy at generation 50 is coupled with the introduction of the '*draw x*' class, with the subsequent '*draw . . .*' classes providing a similar fall in population accuracy at generations 100 and 150. This is reasonably acceptable due to the complexity of the classes, and the small number of frames that their gestures are comprised of; however, within a few generations the population has optimized the parameters and returned previous accuracy levels. In the case of introducing '*two hand wave*' and '*forward kick*' there is an increase in population accuracy upon learning the new classes, this suggests that the training samples have then provided some benefit to partitioning the previously learnt actions, boosting the recognition of these classes.

3.2 Two Person Interaction Recognition

Similar improvement over the use of standard key pose generation can be seen from the interaction recognition evaluation. Figure 3 shows that for the majority of the action classes observed the predictive accuracy is in excess of 95% when the bag of key poses has been generated using ACA. In both methods used, the recognition rate between the '*approaching*' and '*departing*' classes reached 100% within a small number of generations, if not immediately; this is believed to be due to the simple, almost polar opposite sequence of poses that are generated during the creation of the bag. Despite this issue being discussed in [17, 21, 28], we decided to keep these classes as part of recognition testing due to the need for adaptation with later introductions of unobserved classes. During the adaptation to new interaction classes, we observe a decrease in recognition accuracy as expected; however the drop in accuracy is not as noticeable as with the single action recognition. This may be due to the more simplistic classes provided by the SBU dataset, or due to the higher dimensional embedding of features. There was some difficulty for both methods to return to their previous level of accuracy once a new action class was introduced; although a small increase occurs

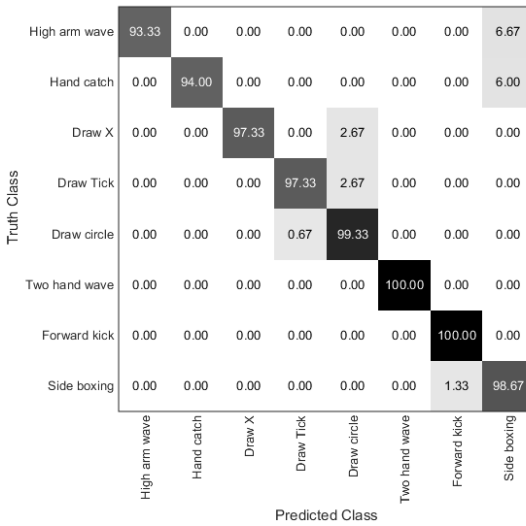| Truth Class \ Predicted Class	High arm wave	Hand catch	Draw X	Draw Tick	Draw circle	Two hand wave	Forward kick	Side boxing
High arm wave	93.33	0.00	0.00	0.00	0.00	0.00	0.00	6.67
Hand catch	0.00	94.00	0.00	0.00	0.00	0.00	0.00	6.00
Draw X	0.00	0.00	97.33	0.00	2.67	0.00	0.00	0.00
Draw Tick	0.00	0.00	0.00	97.33	2.67	0.00	0.00	0.00
Draw circle	0.00	0.00	0.00	0.67	99.33	0.00	0.00	0.00
Two hand wave	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
Forward kick	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
Side boxing	0.00	0.00	0.00	0.00	0.00	0.00	1.33	98.67

Figure 2: Confusion matrix of single person action class recognition. Values shown are predictive rates for the final generation of optimisation.

within the allotted 50 generation time frame, the final prediction accuracy does not reach the standard it achieved before the introduction of the new class, as can be observed with the MSR action recognition. This could be due to the rate of mutation or the generation length being cut short. Despite this drop in accuracy we are still able to generate strong recognition accuracy on multiple complex pairwise interactions by first segmenting the action class into lower level gestures.

4 Conclusion

This study has shown that key pose generation benefits from the initial segmentation of lower-level gestures from all observed training instances. This identifies key temporal sub-actions across instances of an action class, before then using these segments for the generation of the key poses. The use of aligned cluster analysis has allowed us to extract common gesture sequences from across all training observations by sequence alignment with Dynamic Time Warping. This segmentation has then in turn been utilized to create a bag of key poses that is able to accurately recognize action classes on both a single person, and two person interaction level. Although this method generates significantly more key poses for the bag, it is these informative poses that are able to assist in classifying new observations in the scene by describing gestures that are repeatedly observed across numerous instances. The use of ACA segmentation to generate key pose representations has benefits in the recognition of pairwise interactions between two individuals, providing an increase in the correct prediction through use of key poses. Although the initial accuracy of the classifier is variable, the evolutionary optimization of the tuning parameters is able to increase the predictive accuracy over time.
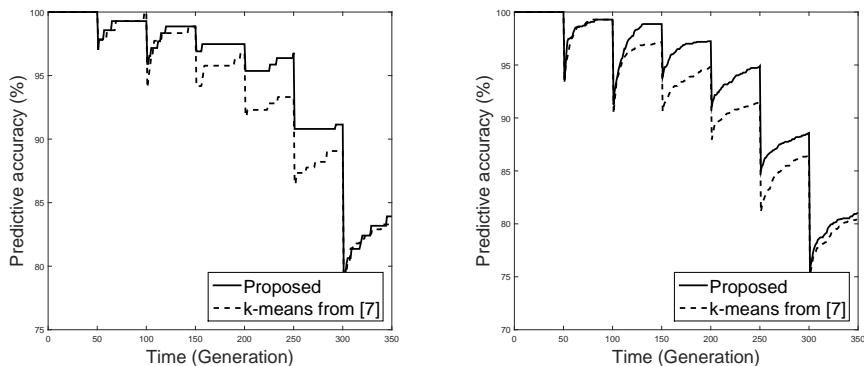
Figure 3: L-R: Maximum and Average predictive accuracy of population when classifying two person pairwise interactions. A new class is introduced every 50 generations.

The understanding of the underlying gestures are key to recognizing actions, as has been demonstrated by the use of key poses, sequences of key poses, bag of key pose, and sequence alignment techniques that have come to fruition over recent years. Further understanding of how an action execution can be comprised of gestures that are global across both subjects and observations will help to identify which portions of an event are beneficial to the partitioning of the action space.

In terms of performance; the number of key poses that this method creates is large, creating k key poses for each of the segmented gesture clusters. Therefore a reduction in the number of key poses that represent each segment cluster may be beneficial to the overall accuracy and speed of the system. Just as with the selection of training parameters, the use of evolutionary programming may guide the selection of an optimum ACA segmentation. The observed accuracies are acceptable for the HAR domain, and especially when considering the recognition of interactions between two individuals, in which level of variation in execution can vary on a large scale and the class labelling is broadly generalized.

In coming studies we intend to expand the complexity of the interaction problem further; incorporating the use of action classes that consist of multiple low level gestures accumulated over a period of time, such as conversational interactions. We will also look into the non-sequential key pose representations, improving predictive accuracy for classes in which there is little constraint on the linear ordering of poses.

References

[1] JK Aggarwal and MS Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3), 2011.

[2] Thomas Bäck. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York, New York, USA, 1996.

[3] Sermetcan Baysal, Mehmet Can Kurt, and Pinar Duygulu. Recognizing Human Actions Using Key Poses. In *Inernational Conference on Pattern Recognition*, pages 1727–1730, 2010. doi: 10.1109/ICPR.2010.427.

[4] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. 29(12):2247–53, 2007. ISSN 0162-8828. doi: 10.1109/TPAMI. 2007.70711.

[5] William Brendel and Sinisa Todorovic. Activities as time series of human postures. In *European Conference on Computer Vision*, number 1, pages 1–14, 2010.

[6] Alexandros Andre Chaaraoui and Francisco Flórez-revuelta. Adaptive human action recognition with an evolving bag of key poses. *Autonomous Mental Development*, 6 (2):139–152, 2014.

[7] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15):1799–1807, 2013. ISSN 01678655. doi: 10.1016/j.patrec. 2013.01.021.

[8] Alexandros Andre Chaaraoui, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert Systems with Applications*, 41(3):786–794, 2014. ISSN 09574174. doi: 10.1016/j.eswa.2013.08.009.

[9] Shahzad Cheema, Abdalrahman Eweiwi, Christian Thurau, Christian Bauckhage, Fraunhofer Iais, and Sankt Augustin. Action Recognition by Learning Discriminative Key Poses. In *Computer Vision Workshops*, pages 1302–1309, 2011.

[10] Yan Chen, Qiang Wu, and Xiangjian He. Using dynamic programming to match human behavior sequences. *Control, Automation, Robotics and Vision*, pages 17–20, 2008.

[11] Jingjing Deng, Xianghua Xie, and Ben Daubney. A bag of words approach to subject specific 3D human pose interaction classification with random decision forests. *Graphical Models*, 76(3):162–171, 2014. doi: 10.1016/j.gmod.2013.10.006.

[12] Jingjing Deng, Xianghua Xie, and Shangming Zhou. Conversational Interaction Recognition based on Bodily and Facial Movement. In *Int. Conf. on Image Analysis and Rec.*, 2014.

[13] Abdalrahman Eweiwi, Shahzad Cheema, Christian Thurau, and Christian Bauckhage. Temporal key poses for human action recognition. In *International Conference on Computer Vision Workshops*, pages 1310–1317, 2011. doi: 10.1109/ICCVW.2011. 6130403.

[14] Dian Gong and Gerard Medioni. Dynamic Manifold Warping for view invariant action recognition. In *International Conference on Computer Vision*, number 3, pages 571–578, 2011. doi: 10.1109/ICCV.2011.6126290.

[15] Dian Gong, Gérard Medioni, and Xuemei Zhao. Structured Time Series Analysis for Human Action Segmentation and Recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1414–1427, 2014.

[16] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Transactions on Cybern.*, 43(5):1318–1334, 2013. ISSN 2168-2275. doi: 10.1109/TCYB.2013.2265378.

[17] T. Hu, X. Zhu, and K. Guo, W.and Su. Efficient interaction recognition through positive action representation. *Mathematical Problems in Engineering*, 2013:1–11, 2013. ISSN 1024-123X. doi: 10.1155/2013/795360.

[18] Hong Li and Michael Greenspan. Multi-scale gesture recognition from time-varying contours. volume 1, pages 236–243, 2005. ISBN 0-7695-2334-X. doi: 10.1109/ICCV. 2005.156.

[19] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action Recognition Based on A Bag of 3D Points. In *International Workshop on CVPR for Human Communicative Behavior Analysis*, 2010.

[20] Fengjun Lv and Ramakant Nevatia. Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383131.

[21] Sangho Park and J. K. Aggarwal. A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia Systems*, 10(2):164–179, 2004. doi: 10.1007/s00530-004-0148-1.

[22] Michalis Raptis, Darko Kirovski, and Hugues Hoppe. Real-time classification of dance gestures from skeleton animation. In *SIGGRAPH/Eurographics Symposium on Computer Animation*, volume 1, page 147, 2011. doi: 10.1145/2019406.2019426.

[23] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech, and Signal Process.*, 26(1):43–49, 1978. ISSN 0096-3518. doi: 10.1109/TASSP.1978.1163055.

[24] Ashok Veeraraghavan, Anuj Srivastava, Amit K Roy-Chowdhury, and Rama Chellappa. Rate-invariant recognition of humans and their activities. *Transactions on Image Processing*, 18(6):1326–39, 2009. doi: 10.1109/TIP.2009.2017143.

[25] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. pages 1290–1297, 2012. ISBN 978-1-4673-1228-8. doi: 10.1109/CVPR.2012.6247813.

[26] Daniel Weinland, E Boyer, and R Ronfard. Action recognition from arbitrary views using 3D exemplars. pages 1–7, 2007. ISBN 9781424416318.

[27] Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Does human action recognition benefit from pose estimation? pages 67.1–67.11, 2011. ISBN 1-901725-43-X. doi: 10.5244/C.25.67.

[28] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, Dimitris Samaras, and Stony Brook. Two-person interaction detection using body-pose features and multiple instance learning. pages 28–35, 2012. ISBN 9781467316125.

[29] Feng Zhou and Fernando De la Torre. Canonical Time Warping for Alignment of Human Behavior. In *Advances in Neural Information Processing Systems Conference*, 2009.

[30] Feng Zhou, Fernando De Torre, and Jessica K Hodgins. Aligned cluster analysis for temporal segmentation of human motion. Technical report, Carnegie Mellon University, 2008.

[31] Feng Zhou, Fernando De la Torre, and Jessica K Hodgins. Hierarchical Aligned Cluster Analysis for Temporal Clustering of Human Motion. *Transactions on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2013.