Human Motion Recognition from 3D Pose Information
Trisarea: A New Pose-based Feature

M. Vinagre[1], J. Aranda[1,2] and A. Casals[1,2]
[1]*Robotics Group, Institute for Bioengineering of Catalonia, Baldiri Reixac 10-12, 08028 Barcelona, Spain*
[2] *Universitat Politcnia de Catalunya, BarcelonaTech, Jordi Girona 1-3, 08034 Barcelona, Spain*

Keywords: Pose-Based Features, Human Motion Recognition, Human System Interface.

Abstract: The use of pose-based features has demonstrated to be a promising approach for human motion recognition. Encouraged by the results achieved, a new relational pose-based feature, Trisarea, based on geometric relationship between human joints, is proposed and analysed. This feature is defined as the area of the triangle formed by connecting three joints. The paper shows how the variation of a selected set of Trisarea features over time constitutes a descriptor of human motion. It also demonstrates how this motion descriptor based on Trisarea features can provide useful information in terms of human motion for its application to action recognition tasks.

1 INTRODUCTION

In recent years, the study of computational methods that allow identifying and understanding human motion has been a field of interest in research and industry. The interest in this topic is motivated by its potential application in a large variety of systems, such as surveillance, patient monitoring, robotics, games, intelligent user interfaces, and in general, those activities that involve some kind of interaction between users and systems. This research is commonly known as gesture, action or activity recognition, and several surveys have been published related to this topic in the last years. In (Poppe, 2010) a detailed overview of current advances in the field is provided. The work in (Aggarwal and Ryoo, 2011) presents recognition methodologies developed for simple human actions, as well as, for more complex high-level activities. The study presents an approach-based taxonomy that compares the advantages and limitations of each approach.

Despite the efforts of a large number of researchers, motion recognition still remains an unsolved problem due to the variability of input data in intra-classes and similarity in inter-classes. In the real world a given motion can be performed by subjects anthropomorphically different and, a given subject can perform a determined motion with different absolute parameters of velocity or trajectory, affecting its appearance.In general, human motion recognition approaches exploit appearance and/or human body parts information by defining suitable features of an image sequence. From appearance features, some methods recognize motions as a sequence of local low-level features in images (Gorelick et al., 2007; Matikainen et al., 2010). Other methods use body parts information, as human pose estimation, to extract posture and body motion features (Ellis et al., 2013; Gu et al., 2010). A recent work (Yao et al., 2011) discusses about both approximations for action/motion recognition in home-monitoring scenarios, depicting that pose-based features outperform low-level appearance features, even when data are heavily corrupted by noise. The study also suggests that a combined approach of both techniques can be beneficial for motion recognition.

Different pose-based features extracted from positions of human joints have been used, which can be mainly classified into two methodologies. The first relies on obtaining features from joint parameters as orientation, position, velocity or acceleration. Many previous works use this kind of features to represent human' poses and motion (Gu et al., 2010; Xia et al., 2012). However, the problem of extracting a reliable similarity measure between the same type of motions or poses from individual properties of joints is still unresolved. In the second methodology this problem is reduced obtaining features from correspondences between joints, so called relational

Vinagre M., Aranda J. and Casals A..
Human Motion Recognition from 3D Pose Information - Trisarea: A New Pose-based Feature.
DOI: 10.5220/0004482800740082
In *Proceedings of the 10th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2013), pages 74-82
ISBN: 978-989-8565-71-6

pose-based features. Usually, they are geometric correspondences between joints, called relational geometric features, as the Euclidean distance between two joints or the distance between a joint to a plane spanned by other joints. (Yun et al., 2012).

The good prospects of relational geometric features as motion and pose representation (Chen et al., 2009) and the short number of such featured proposed up to now, has motivated this work. Thus, this research attempts to contribute with a new relational geometric feature and its use for motion recognition. We propose a relational geometric feature called Trisarea, which describes the geometric correspondence between joints by means of the area of the triangle that they define. We demonstrate how the variation of the Trisarea feature in a motion sequence retains useful information for its application in human motion recognition.

The rest of the paper is organized as follows. Section II gives a short review of the recent advances in pose-based features for their use in motion recognition. Section III presents this new pose-based feature called Trisarea. A motion representation as Trisarea evolution and a motion descriptor based on Trisarea features variation over time is presented in section IV. In section V, the performance of a motion recognition integrating our motion descriptor is presented. Experimental results are given in section VI and conclusion and future extensions in section VII.

2 RELATED WORK

Human motion recognition from pose-based features requires an inherent procedure for extracting human pose. Vision-based pose estimation faces the difficult problem of estimating kinematic parameters of a body model, either from static frame or a frame sequence. However, despite this complex initial processing, this approach has several advantages over motion recognition from appearance based features since it is invariant to the point of view and to appearance variations produced by environment conditions. It is also less sensitive to noise from intra-class variances in contrast to recognition from appearance based features.

Previous promising methods for interactive human pose estimation and tracking are those that use a volumetric model of the body as (Luo et al., 2010; Matthias Straka and Bischof, 2011) or utilize depth information extracted from structured light sensors, as the newly Microsoft Kinect camera or the Asus Wavi Xtion as (Shotton et al., 2011; Schwarz et al., 2010).

Motivated by the current progress in real-time pose estimation, some recent works in human motion recognition have been performed based on such information. The work in (Raptis et al., 2011) presents a real-time dance gestures classification system from pose representation. It uses a cascaded correlation-based classifier for multivariate time-series data and distance metric based on dynamic time-warping. This approach has an average accuracy of 96.9% for approximately 4 second motion recordings. In (Miranda et al., 2012) a method is introduced for real-time gesture recognition from a noisy skeleton stream extracted from Kinect depth sensors. They identify key poses through a multi-class classifier derived from support vector learning machines and gestures are labelled on-the-fly from key pose sequences through a decision forest tree.

These and other works as (Gu et al., 2010; Uddin et al., 2011; Xia et al., 2012; Sung et al., 2012) recognize human motion from direct measures of joint parameters of the human body as angles, instantaneous position, orientation, velocity, acceleration, etc. Such approaches have as inconvenient that different repetitions of a same action must be numerically similar, and, due to the irregularity in the periodicity of human actions and intra-person motion variability this assumption is not always true.

Other methods are more flexible which they use relational geometric features describing correspondences between joints in a single pose or a short sequence of poses. In (Müller et al., 2009) different relational geometric features are introduced, which have been used for single human action recognition (Yao et al., 2011; Wang et al., 2012) and for two-people interaction activities detection (Yun et al., 2012), with good results. In (Chen et al., 2009) different of these type of features are proposed, as:

- Distance feature. It is defined as the Euclidean distance between all pairs of joints of a human pose, at time t.
- Rotation feature. It is the rotational angle of the line spanned by two joints with respect to the reference pose.
- Plane feature.It computes the correspondence between the plane spanned by some joints with respect to a single joint, as the distance from this joint to the referred plane.
- Normal plane feature. Similar to plane feature, but here the plane is defined by its normal spanned by two joints and a joint belonging to the plane.

- Angle feature. It is the angle between two lines spanned by two pairs of different joints.

3 TRISAREA FEATURE

In our work, the human body or pose is interpreted as a connected graph which nodes are the joints themselves and edges represent body parts. Given a pose from the above representation, a feature extraction is performed in order to extract some joint correspondences that characterize this pose. In this way, a new feature called Trisarea is applied. Trisarea represents a geometric relation between joints given by the area of the triangle they define.

Mathematically, let p_1, p_2, p_3 be the coordinates of joints j_1, j_2, j_3 in a Euclidean space $\mathbb{R}^3$. Given a pose P, the Trisarea feature between j_1, j_2, j_3 joints is defined by:

$$\triangle(j_1, j_2, j_3, P) = \frac{1}{2} \cdot \|\overrightarrow{p_1p_2} \times \overrightarrow{p_1p_3}\| \quad (1)$$

Figure 1 Shows an example of Triarea features in a given pose where there are three geometric relations between eight joints.

4 MOTION DESCRIPTOR

In the previous section, Trisarea has been presented as a feature for an individual pose description. In this section, a descriptor of human motion is built from the evolution of different Trisarea features over time. The existence of irrelevant Trisarea features was realized by observation, thus, an automatic method to filter and select the most important components is shown. A single vector representation of these selected features evolution over time and its application as a motion descriptor are presented.

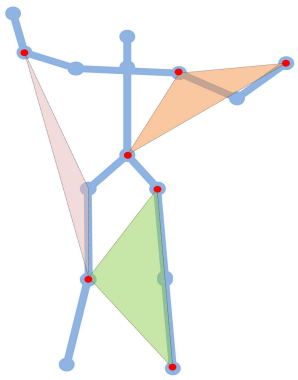

Figure 1: Example of Trisarea features.

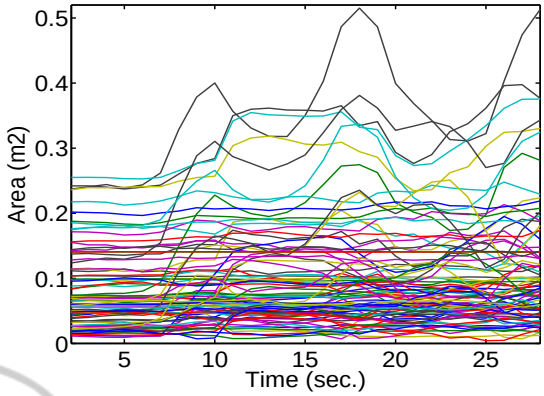

Figure 2: Representation of a *'arm wave'* motion with the evolution of Trisarea features.

4.1 Motion Representation as Trisarea Evolution

In order to represent the motion of human body, the evolution of Trisarea features from the pose sequence of motion is calculated. The number of Trisarea features in a motion representation depends on the number of joints in the pose representation. Being J the number of joints of a pose representation, the number of Trisarea features F is:

$$|F| = \frac{J!}{J \cdot (J-3)!} \quad (2)$$

Thus, given a motion m with a sequence of poses P_{SEQ} the motion representation $Mrep(m)$ is a matrix of $F \times \|P_{SEQ}\|$ defined as:

$$Mrep(m) = \begin{bmatrix} \triangle_1(j_i, j_j, j_k, p_0) & \cdots & \triangle_F(j_r, j_p, j_q, p_0) \\ \vdots & \ddots & \vdots \\ \triangle_1(j_i, j_j, j_k, p_C) & \cdots & \triangle_F(j_r, j_p, j_q, p_C) \end{bmatrix} \quad (3)$$

Figure 2 shows the motion representation of a *arm wave* motion. In this example, the pose representation contains $J = 15$ joints as shown in figure 1. So, the number of Trisarea features by applying the equation 2 is $|F| = 455$.

In this motion representation, Trisarea features contribute to encode useful information about motion. However, many of these features are irrelevant and can be obviated without loss of discrimination performance.

The selection of relevant features is not immediately intuitive, but few of them are clearly irrelevant, as those define invariant areas, formed by mutually constrained joints (i.e. torso, right shoulder, left shoulder).

We perform an unsupervised feature selection procedure in order to filter irrelevant Trisarea features; those that do not contribute to recognize certain motion from a motion set. This process is described below.

4.2 Dimensionality Reduction

The selection process is a common preprocessing filter step used for classification and pattern recognition applications. In this process, we want to determine relevant triangles to reduce computational cost and avoid undesired noise.

Hence, we have used a computationally feasible unsupervised component selection methodology called principal feature analysis (PFA) (Lu et al., 2007) to find the salient components of the initial feature vector. This method exploits the information that can be inferred from the principal component coefficients to obtain the optimal subset of joints relationships.

This feature selection methodology differs from common feature extraction methods as principal component analysis (PCA), independent component analysis (ICA) and Fisher Linear Discriminate Analysis (LDA). These methods apply a mapping from the original feature space to a lower dimensional feature space, having as disadvantage that all components of the original feature are needed in the projection to the lower dimensional space, so they must be always calculated. Instead, in PFA only a subset of relevant components in the original feature is selected, thus lessening computation time. In this case there is no mapping process and it is possible to work directly in a reduced feature space. Detailed information of this method can be read in (Lu et al., 2007).

PFA method is applied on a large set of Trisarea features extracted from randomized poses of all the available sampled motions to be recognized. Here, the variation to be retain is set in order to retain most important features. As a result, we obtain a reduced motion representation $Mrep_{filt}$ with a set of Trisarea features F_{filt} which the number of features is less than the original motion representation($|F| >> |F_{filt}|$).

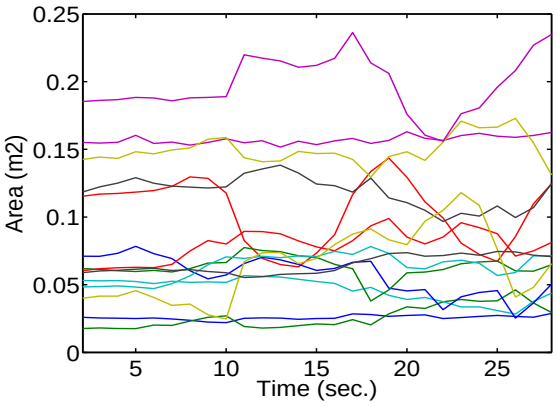

Figure 3: Filtered motion representation of *'arm wave'* example.

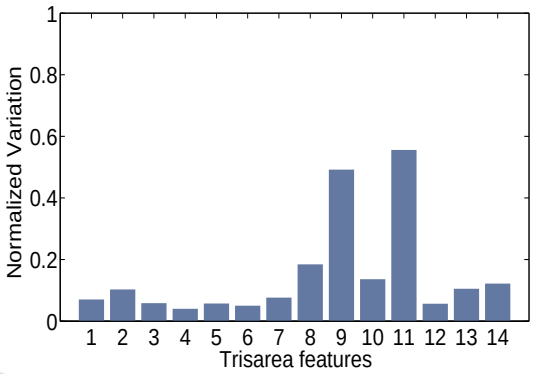

Figure 4: Motion Descriptor of a *'arm wave'* example.

As an example, figure 3 shows the remaining features ($C_{filt} = 14$) as a result of the PFA process that retains the 95% of data input from the original set of features($C = 455$) of the arm wave motion representation shown in figure 2.

4.3 Trisarea Variation as Motion Descriptor

The analysis of temporal trends in the variation of Trisarea features is useful for motion description. Our motion description uses a descriptive statistic parameter called Pearson's variation coefficient. This statistic parameter allows us to perform a temporal description and summarization of a univariate time series δ_i of Trisarea feature $\triangle_i$ over time T, calculated as:

$$D_{var}(\delta_i) = \frac{\sigma(\delta_i)}{\mu(\delta_i)} \quad (4)$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ perform the standard deviation and the mean of a univariate time series over time. Finally, the equation 4 is applied to every univariate time series on the set of filtered Trisarea features(F_{filt}). As a result, a single vector with a dimension $1 \times |F_{filt}|$ is calculated:

$$\varphi = \langle D_{var}(\delta_1), \cdots, D_{var}(\delta_{|F_{filt}|}) \rangle \quad (5)$$

As an example, figure 4 shows the instance of the motion 'wave arm' example shown in section 4.2. This motion is instantiated with the 14 Trisarea features selected by the FPA pre-process.

5 Human Motion Recognition

The core of a motion recognition process is exploring input data in order to identify it. In this work, the meaningful features about motion are represented in the motion descriptor explained in section 4.3. Since the motion descriptor is defined as a feature vector, classification methods using Machine Learning techniques can be used. The performance of the majority of these techniques depends on the feature space and the quality of data used in learning. In fact, our motion representation is not useful with techniques which deal with the temporal order of data patterns because this data has been reduced.

In order to evaluate the contribution of Trisarea features as a feature space in a recognition process, two classification techniques were performed. The first classification technique was a Nearest Centroid classifier. This classification is a supervised neighbors-based learning method that obtains a prototype class by the centroid of its training instances. Let T_{set}^{ζ} be the set of training instances of a certain class ζ used in the learning phase, the motion prototype or centroid of a class is calculated as:

$$Prot_{\zeta} = Mean(T_{set}^{\zeta}) \tag{6}$$

In order to determine the class of an unknown input motion, the minimal similarity against all motion prototypes using an Euclidean distance measurement is calculated. Let φ an unknown motion instance, which must be classified, and let K be the set of motion types (*classes*). The resulting motion class ζ of φ is given by:

$$\zeta = argmin_{\zeta}(\|\varphi - Prot_{\zeta}\|)_{\forall\zeta\in K} \tag{7}$$

The second classification method used was the Naive Bayes classifier. This classification is a supervised learning method based on applying Bayes' theorem with the *assumption* of independence between every pair of Trisarea features. It requires a small amount of training data to estimate necessary parameters. In order to classify an unknown motion instance φ from K types of motion, the maximum a posteriori (MAP) decision rule is applied. The estimation of the motion class ζ of φ is calculated as:

$$\zeta = \text{argmax}_{\zeta}(P(\zeta|\varphi)_{\forall\zeta\in K} \tag{8}$$

In this work, the likelihood function of the features given for each class was modelled as Gaussian mixtures.

6 EXPERIMENTATION

In this section, a test of the motion recognition approaches explained in section 5 is presented. For this test, a public dataset is used in order to compare the obtained results with other recognition methods in the literature. Finally, the obtained results are discussed.

6.1 Experimentation Setup

Dataset. We selected a public MSR Action 3D dataset (Li et al., 2010) which supplies the sequences of depth maps captured by a depth camera similar to a Kinect device, with a frame rate of 15 fps and down-sampled resolution of 320x240. This dataset contains 20 different actions that cover various movement of arms, legs, torso and their combinations without human-object interactions: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hands wave, side -boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing and pick up & throw. Each action was performed by 9-10 subjects two or three times. The subjects were advised to use their right arm or leg whenever an action is to be performed by a single limb. Altogether, 567 actions sequences in total were used, those provided by the dataset.

Pose model. The pose estimation was extracted from the original depth maps. After that, pose estimation results were inspected manually to filter possible pose estimation errors.

Our human pose model was a pose representation of 15 joints, as shown in figure 5. As some approaches explained in section 2, we normalized the position of joints relative to the torso position to make the pose

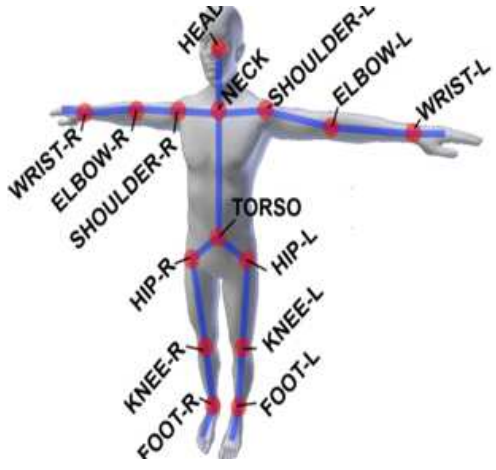

Figure 5: Pose model.

description independent to pose changes with respect to the world space and camera changes(e.g. point of view changes) and tolerant to anthropomorphic differences.

6.2 Results

In order to contrast our analysis, the dataset actions were divided into three subsets, suggested in (Li et al., 2010). This division has been followed in different works in order to obtain a public benchmarking. Concretely, every of subsets contain 8 actions, as shown in table 1. The action sets AS1 and AS2 were intended to group actions with similar movements, while the action set AS3 was intended to group complex actions together.

Table 1: Data subsets from MSR Action 3D dataset.

Subset AS1	Subset AS2	Subset AS3
Horiz. arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	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
Pickup&throw	Side boxing	Pickup&throw

For each experiment/subset, motions were represented as the Trisarea evolution representation explained in section 4.1. After that, the filtering step called Feature Principal analysis(*FPA*) and explained in section 4.2 was applied. For the *FPA*, a random set of poses from available motions to be recognized was calculated.Concretely, this set was built with a set of 1500 randomized poses spread evenly from actions.

In the *FPA* process, the number of features to be filtered was calculated taking into account the ratio of the amount of original input information to be retained (*variation retained ratio*). The behaviour of the original input information retained against the number of filtered features is shown in figure 6. The inflection points of each experiment were around 90%. So, we chose 90% as variation reatined ratio to preserve a good information-feature ratio. The dimensionality of the motion descriptors was 18, 16 and 21 for AS1, AS2 and AS3 respectively.

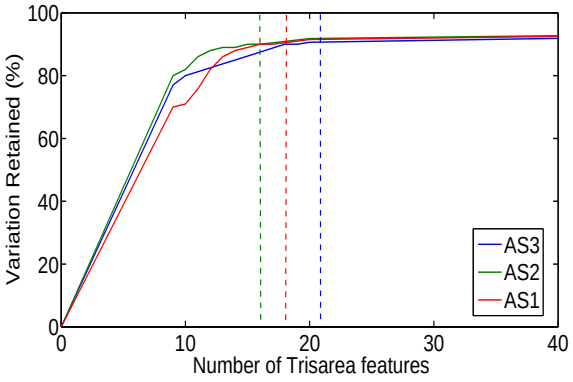

Figure 6: Input data variation retained against Trisarea feature dimensionality.

With the results of the filtering step, the remaining Trisarea features were used to deal motion data into a feature space. The data transformation was performed with applying equation 5 over data of subsets. The new data representation of actions in a single feature vector allow us to perform a motion/action recognition process based on the two classification methods introduced in section 5. Since these classification methods have a training step in a supervised learning way, 3/4 parts of action instances were selected as training data. The rest of data were used for the classification test.

The accuracy of both classification approaches for each experiment/subset are shown in table 2.

Table 2: Results of the classification performance.

	AS1	AS2	AS3
Nearest Centroid	75.4%	73.5%	79.6%
Naive Bayes	88.6%	86.3%	94.0%

We can observe a better accuracy of the Nave Bayes (NB) classifier than the Nearest Centroid (NC) technique. One of the reason is that the NC classifier performed was a non-generative model and it does not take into account the variability of the distances to the centroid within a class. Nevertheless, the Naive Bayes classifier is a generative model where classes are modelled by probability distributions which are generated by training data. This probabilistic framework accommodates asymmetric misclassification and class priors.

In order to analyze the NC classifier performance, similarity between test samples and learned motion prototypes has been calculated. Figure 7 shows the normalized mean distance between test samples belonging to the same class against the motion prototypes. Low values was expected along the diagonal indicating high similarity of the samples with the prototype of their class. Higher out of diagonal values indicate low similarity to other prototypes so lower probability of misclassification.

The dispersion of the test samples of the same class were not significant (*with stdev around 0.2*). This fact depicts good results of the Trisarea feature space representation of motion data. Figure 7 shows

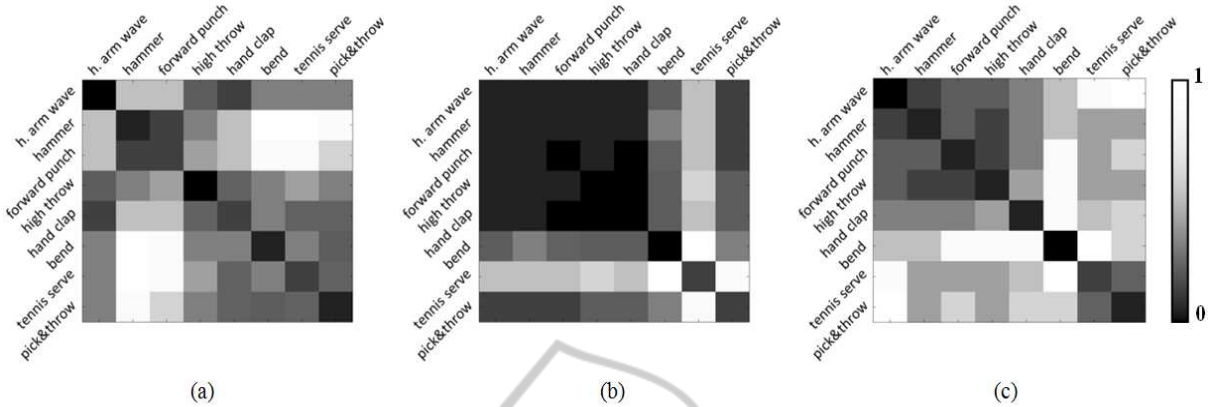

Figure 7: Normalized sum of distances between all tested instances belonging to a class against the class prototypes in (a) AS1, (b) AS2 and (c) AS3 action sets.

that some actions like *high arm wave* and *hammer*, similar motions present similar distance between their test samples and their prototypes. Indeed, in figure 7.b we can observe that half of the motions are very similar and it remarks the poor classification accuracy in AS2. These results show that NC classifier decreases its accuracy when similar motions exist.

In the case of high amount of training data and non-similar real movements, the NC classifier with our motion instances can be applied. This approach constitutes a fast recognition task with moderate recognition accuracy. For better accuracy results, the nave Bayes classifier can be applied. The result of NB classifier provides an average of 89.6% while NC classifier reaches an average of 76

In general, both classifiers results are good and they depict that our Trisarea feature seems to be useful feature for human motion recognition. Relevant features retain sufficient information to describe and discriminate different human motions.

Since NB shows a better performance than NC classifier, NB classifier was selected as our motion recognition method and a specific comparison against others approaches in the literature is presented in the next section.

6.3 Comparison

We compared our approach with three different state of the art methods (Li et al., 2010; Vieira et al., 2012; Miranda et al., 2012). These approaches uses different recognition strategies and they use the same dataset as benchmark. So, we can perform a testbed in order to compare and evaluate our motion recognition performance. In (Li et al., 2010) a method to recognize human actions from sequences of depth maps is presented. They obtain a projection of representative sampled 3D points to characterize a set of salient postures which are used as nodes to describe an action graph. In (Vieira et al., 2012) depth maps images are used too. They present Space-Time Occupancy Patterns (*STOP*) which depth information sequence is represented in a 4D space-time grid and an action graph based system is used to learn a statistical model for each action class. On the other hand, (Miranda et al., 2012) presents a real-time action gesture recognition from a pose stream with an angular representation. They capture key poses through a multi-class classifier and a gesture/motion is labelled from a key pose sequence through a decision forest algorithm

For the testbed comparison, the partition of dataset were performed in the same way that the others approaches.This partition consisted in performing a training set with half of the samples and the rest of samples for the test part. Accuracies of our approach and others with datasets AS1, AS2 and AS3 are shown in Table 3. In general, the

Table 3: Comparison of recognition accuracies (%).

	Li et al.	Vieira et al.	Miranda et al.	Our
AS1	72.9%	84.7%	93.5%	76.2%
AS2	71.9%	81.3%	52.0%	72.3%
AS3	79.2%	88.4%	95.4%	81.0%
Avg.	74.7%	84.8%	80.3%	76.5%

results in table 3 depict how the dynamics of the Trisarea feature is able to give useful information in terms of human motion recognition. The results of the dimensionality reduction of motion descriptors show a reduction from 455 possible Trisarea features to a maximum of 21, retaining around 90% of the original input information. This fact confirms

that only few Trisarea features are relevant to recognition processes. Statistics on relevant features was performed and this analysis shows that relevant triangles are, in most cases, formed by a swing of non-adjacent joints with unconstrained movements in 3D space (i.e. elbows, knees or wrist) with respect to a joint with a constrained position from torso reference (i.e. shoulder, hip or neck).

The comparison with other recognition approaches denotes that our approach performs reasonably well. For this testbed, our results outperform (Li et al., 2010) approach. On the other hand, (Vieira et al., 2012; Miranda et al., 2012) have better results because they have a more accurate temporal information about motion. However, our approach have the most compact representation and surely is faster than previous pose-based methods as (Miranda et al., 2012).

As conclusion, the results show the usefulness of Trisarea features extraction to perform a motion feature space that permits us the use of classical and powerful Machine Learning methods to perform human motion recognition. Although we have obtained promising results with the Pearsons coefficient of variation, this measure is ambiguous for similar movements and we have to explore other options to encode the dynamic behavior of Trisarea features in motion sequences.

7 CONCLUSIONS

A new relative pose feature called Trisarea has been presented and its use for human motion recognition has been proposed. Specifically, a motion descriptor based on the evolution of Trisarea features has been performed. A Principal Feature Analysis has provided good results selecting the most relevant Trisarea features with the objective of reducing the dimension of the obtained feature vector for posterior recognition steps.

With the filtered motion descriptors, a single instance of a motion has been proposed applying Pearsons relative coefficient of variation for each Trisarea feature over time. These motion instances have been generated over three different datasets in order to verify recognition results.

As a result, the experiments have demonstrated the usefulness of Trisarea features for human motion recognition tasks. A comparison with other approaches in the same scenario has revealed that our recognition results have not been far from other methodologies results. In addition, the presented approach has got a good accuracy/speed ratio because it has less preprocessing calculations than the compared pose-based approaches.

Motivated with the presented results, in a future scope we will explore some kind of temporal modelling in order to deal with changes of order of motion execution and will test them on-line in real world scenarios.

ACKNOWLEDGEMENTS

This work has been done under project IPRES, DPI2011-29660-C04-01 of the Spanish National Research Program, with partial FEDER funds.

REFERENCES

Aggarwal, J. and Ryoo, M. (2011). Human activity analysis: A review. *ACM Computing Survey*, 43(3):16:1–16:43.

Chen, C., Zhuang, Y., Xiao, J., and Liang, Z. (2009). Perceptual 3D pose distance estimation by boosting relational geometric features. *Computer Animation and Virtual Worlds*, 20(2-3):267277.

Ellis, C., Masood, S. Z., Tappen, M. F., Laviola, Jr., J. J., and Sukthankar, R. (2013). Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 101(3):420–436.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Gu, J., Ding, X., Wang, S., and Wu, Y. (2010). Action and gait recognition from recovered 3-d human joints. *Trans. Sys. Man Cyber. Part B*, 40(4):1021–1033.

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

Lu, Y., Cohen, I., Zhou, X. S., and Tian, Q. (2007). Feature selection using principal feature analysis. In *Proceedings of the 15th international conference on Multimedia*, pages 301–304, New York, NY, USA. ACM.

Luo, X., Berendsen, B., Tan, R. T., and Veltkamp, R. C. (2010). Human pose estimation for multiple persons based on volume reconstruction. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, pages 3591–3594, Washington, DC, USA. IEEE Computer Society.

Matikainen, P., Hebert, M., and Sukthankar, R. (2010). Representing pairwise spatial and temporal relations for action recognition. In *Proceedings of the 11th European conference on Computer vision: Part I*, pages 508–521, Berlin, Heidelberg. Springer-Verlag.

Matthias Straka, Stefan Hauswiesner, M. R. and Bischof, H. (2011). Skeletal graph based human pose estimation in real-time. In *Proceedings of the British Machine Vision Conference*, pages 69.1–69.12, Aberystwyth, Wales. BMVA Press.

Miranda, L., Vieira, T., Morera, D. M., Lewiner, T., Vieira, A. W., and Campos, M. F. M. (2012). Real-time gesture recognition from depth data through key poses learning and decision forests. In *SIBGRAPI*, pages 268–275, Washington, DC, USA. IEEE Computer Society.

Müller, M., Baak, A., and Seidel, H.-P. (2009). Efficient and robust annotation of motion capture data. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 17–26, New York, NY, USA. ACM.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990.

Raptis, M., Kirovski, D., and Hoppe, H. (2011). Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 147–156, New York, NY, USA. ACM.

Schwarz, L., Mateus, D., Castaneda, V., and Navab, N. (2010). Manifold learning for tof-based human body tracking and activity recognition. In *Proceedings of the British Machine Vision Conference*, pages 80.1–80.11, Aberystwyth, Wales. BMVA Press.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, Washington, DC, USA. IEEE Computer Society.

Sung, J., Ponce, C., Selman, B., and Saxena, A. (2012). Unstructured human activity detection from rgbd images. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 842–849, Washington, DC, USA. IEEE.

Uddin, M. Z., Thang, N. D., Kim, J. T., and Kim, T.-S. (2011). Human activity recognition using body joint-angle features and hidden markov model. *ETRI Journal*, 33(4):569–579.

Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z., and Campos, M. F. M. (2012). Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In *Proceedings of the 17th Iberoamerican Congress*, pages 252–259, Berlin, Heidelberg. Springer-Verlag.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, Washington, DC, USA. IEEE Computer Society.

Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, pages 20–27, Washington, DC, USA. IEEE.

Yao, A., Gall, J., Fanelli, G., and Van Gool, L. (2011). Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference*, pages 67.1–67.11, Aberystwyth, Wales. BMVA Press.

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T., and Samaras, D. (2012). Two-person interaction detection using body-pose features and multiple instance learning. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 28–35. IEEE.