Bag-of-Features based Activity Classification using Body-joints Data

Parul Shukla, K. K. Biswas and Prem K. Kalra
Indian Institiute of Technology, Hauz Khas, New Delhi, 110016, India

Keywords: Action Recognition, Kinect, Bag-of-Words, Body-joint.

Abstract: In this paper, we propose a Bag-of-Joint-Features model for the classification of human actions from body-joints data acquired using depth sensors such as Microsoft Kinect. Our method uses novel scale and translation invariant features in spherical coordinate system extracted from the joints. These features also capture the subtle movements of joints relative to the depth axis. The proposed Bag-of-Joint-Features model uses the well known bag-of-words model in the context of joints for the representation of an action sample. We also propose to augment the Bag-of-Joint-Features model with a Hierarchical Temporal histogram model to take into account the temporal information of the body-joints sequence. Experimental study shows that the augmentation improves the classification accuracy. We test our approach on the MSR-Action3D and Cornell activity datasets using support vector machine.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

The availability of depth-based sensors like Kinect has opened a new dimension for action recognition which finds application in surveillance, human-computer interaction, smart homes and content-based video search among others (Li et al., 2010), (Ni et al., 2011), (Sung et al., 2011), (Wang et al., 2012). These depth sensors provide depth maps which can be effectively used to estimate 3D joint positions of human skeleton. The maps provide 3D information of not only the human body but also the total scene, which is useful for recognition in situations where humans interact with other subjects or objects. However, this comes at a cost, the depth maps substantially increase the amount of data to be processed.

It is a well known fact that humans tend to recognize actions based on the variation in poses, where a pose is defined as the spatial configuration of body joints at a given point in time. However, lack of effective and efficient mechanism for estimation of joints resulted in earlier action recognition approaches relying on methods based on features extracted from color images and videos (Bobick and Davis, 2001), (Laptev, 2005), (Lv and Nevatia, 2006), (Laptev et al., 2008), (Niebles et al., 2008). With depth-based sensors such as Kinect having facilitated effective estimation of body joints, the interest in methods based on skeleton data has again taken off (Sung et al., 2011), (Jin and Choi,).

A wide variey of approaches have been used for the task of action recognition from conventional videos. Part-based approaches relying on extraction of local features around interest points and building bag-of-words model, have been widely used. This representation ignores the positional arrangement of the spatio-temporal interest points. Although the representation turns out to be simpler, the lack of spatial information provides little information about the human body. Further, the lack of long term temporal information does not permit modeling of more complex actions (Niebles et al., 2008). Besides, it still remains an open question as to how, if at all, a bag-of-words model can be constructed from body-joints data.

In this paper, we propose to extract novel set of scale and translation invariant features. We employ a Bag-of-Joint-Features (BoJF) model to represent an action sample using these features. Our idea is to describe an action sample as an encoded sequence of 'key' features for each joint. In order to take into account the temporal variation in movements of joints, we propose a hierarchical temporal-histogram (HT-hist) model which uses histograms to represent movement of a set of joints. The HT-hist tries to characterize the temporal variation of a joint in the action sample. Support vector machine is used to perform classification of various actions. The proposed features and models are evaluated on the benchmark MSR-Action3D dataset(Li et al., 2010) and Cornell Activity dataset(Sung et al., 2012).

Shukla P., Biswas K. and Kalra P..
Bag-of-Features based Activity Classification using Body-joints Data.
DOI: 10.5220/0005303103140322
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 314-322
ISBN: 978-989-758-089-5
Copyright © 2015 SCITEPRESS (Science and Technology Publications, Lda.)

The rest of the paper is organized as follows: Section 2 provides a brief review of some of the related work in this field. Section 3 introduces the proposed features based on skeleton data. Section 4 describes the action models used by us to represent an action sample. Section 5 gives the method used for classification followed by results of experimental study in section 6. Finally, section 7 presents conclusion and future extensions.

2 RELATED WORK

Human action recognition from images and video has been an active area of research for the past decade, with the focus being on recognition of actions in more challenging scenarios. The approaches primarily vary in the terms of the visual input and the recognition algorithm with complexity of actions and environmental settings being the driving factors. A wide variety of approaches can be found in the literature (Turaga et al., 2008), (Poppe, 2010).

Earlier approaches involved extraction of silhouettes from color videos. In (Bobick and Davis, 2001), Hu moments were computed from motion energy and motion history images, constructed by temporally accumulating the silhouettes. The Hu moments serve as action descriptors with Mahalanobis distance measure being used for classification. Recognition based on spatio-temporal interest points (Schuldt et al., 2004), (Laptev, 2005), (Laptev et al., 2008) and methods using spatio-temporal features with models such as pLSA (Niebles et al., 2008) for action recognition have shown good performance. In (Niebles et al., 2008), the authors build a bag-of-words model by first extracting local space-time interest regions and then by clustering them into a set of spatio-temporal words, called codebook. Bag-of-words model is also used in (Laptev et al., 2008), where authors use HoG and HoF features to describe interest points in a video. They cluster the features and assign the features to closest cluster centers to construct a bag of visual words. Eventhough the methods based on RGB data achieve good results, it may be noted that the RGB data is voluminous. In this paper, we show that instead of the exhaustive RGBD data, it is possible to get comparable results using only body-joints data.

Action recognition from skeleton data has been explored in (Lv and Nevatia, 2006), where the authors use 3D joint locations to construct a number of features. They further use Hidden Markov Model (HMM) and AdaBoost for classification. In (Yao et al., 2011), the authors use skeleton data to extract relational pose features such as joint velocity, plane feature between a joint and a plane, and joint distance feature as Euclidean distance between two joints. Further, they use Hough-transform voting method for classification. They suggest that pose-based features extracted from skeleton data indeed aid in action recognition.

With the introduction of depth sensors, the field of action recognition has received an impetus. In (Li et al., 2010), the authors use action graph to model the dynamics of action from depth sequences. They use bag of 3D points to characterize a set of salient postures corresponding to nodes in action graph. They propose a projection based sampling scheme to sample the bag of 3D points from depth maps. The authors in (Ni et al., 2011), use depth-layered multi-channel representation based on spatio-temporal interest points. They propose multi-modality fusion scheme, developed from spatio-temporal interest points and motion history images, to combine color and depth information.

Approaches based on skeleton data obtained from Kinect have been used in (Sung et al., 2011), (Wang et al., 2012), (Jin and Choi,). In (Sung et al., 2011), the authors consider a set of subactivities to constitute an activity. They use features extracted from estimated skeleton and use a two-layered Maximum-Entropy Markov Model (MEMM) where the top layer represents activities and the mid-layer represents subactivities connected to the corresponding activities in top-layer. Wang et al.(Wang et al., 2012) use depth maps data and skeleton data to construct novel local occupancy (LOP) feature. Each 3D joint is associated with a LOP feature which can be treated as depth appearance of a joint. They further propose fourier temporal pyramid and use these in mining approach to obtain actionlets where an actionlet is a combination of features for a subset of joints. They propose to consider an action as a linear combination of actionlets and their discriminative weights are learnt via multiple kernel learning (MKL). In (Jin and Choi,), the authors propose an encoding scheme to convert skeleton data into a symbolic representation and perform activity recognition using longest common subsequence method.

Our contributions include the proposal of Bag-of-Joint-Features (BoJF) from novel set of skeleton features. We also suggest a hierarchical histogram based scheme to take into account the temporal information.

3 FEATURES BASED ON BODY-JOINTS DATA

In this section, we describe our approach for obtain-

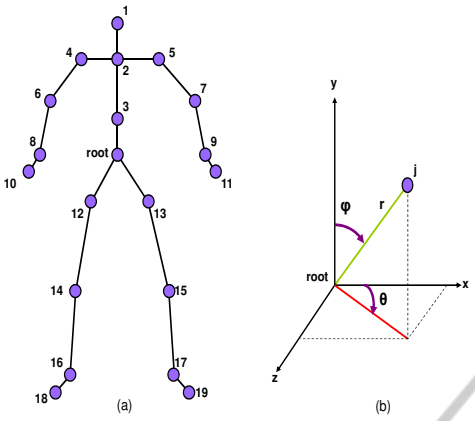


Figure 1: Features based on skeleton data: (a)Skeleton model (b)Features for joint j.

ing scale and translation invariant features from skeleton data. The output skeleton data from Kinect SDK provides 3D coordinates of 20 body joints. For each joint j in a frame, we have 3 values of the coordinates specifying the position, $p_j = (x_j, y_j, z_j)$. Using these 60 values directly as features for action classification may not give good results if persons are likely to move around while they perform the actions. To take care of this we can choose a reference point for the features. This may be done as follows:

a) Using the hip joint as reference (root joint in Figure 1), and computing the relative values of rest of the joints with respect to this one. We name this feature set as 'hip' set in our paper.

b) Using the mean of all the joints as the reference point. We term the feature set obtained this way as the 'mean' set.

c) Using spherical coordinates with the root joint serving as the origin. We term this set as'Spherical' set. Figure 1 illustrates the spherical set. The feature set now consists of joint-distance r_j, joint-angle ϕ_j and joint-angle θ_j given as:

$$r_j = \| p_j - p_{root} \| \quad (1)$$

$$\phi_j = \arccos(\hat{y}_j / r_j) \ , \ 0 \leq \phi_j \leq \pi \quad (2)$$

$$\theta_j = \arctan(\hat{z}_j / \hat{x}_j) \ , \ 0 \leq \theta_j \leq 2\pi \quad (3)$$

Feature θ_j is the angle between x-axis and the projection of vector from root to joint j onto the x-z plane while ϕ_j is the angle between y-axis and the vector from root to joint j. $(\hat{x}_j, \hat{y}_j, \hat{z}_j)$ are the components of vector from root joint to p_j .

Since these joint-features are computed with respect to the root joint, the features are invariant to translation and scale variations. Specifically, if a person moves by a small amount in a subsequent frame, resulting in all the joints' positions being shifted by same amount, the distance feature r_j would remain

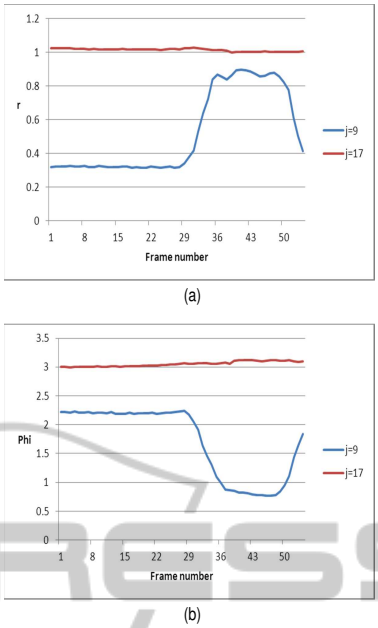


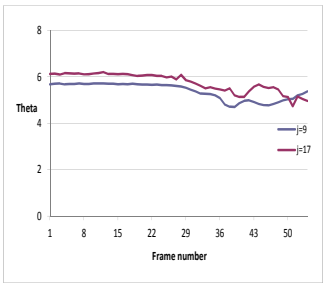
Figure 2: High Arm Wave action for joints $j = 9$ and $j = 17$ (a)Represents variation in r over time (b)Represents variation in ϕ over time.

invariant since it is computed with respect to the root joint in that frame. Likewise, θ_j and ϕ_j are invariant to changes in the scale or size of a person. In other words, scale invariance is achieved because angles remain the same even though height changes across different subjects.

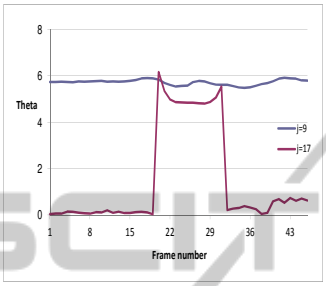
While the features r and ϕ capture the global variations of a joint in 3D coordinate system, θ captures variations with respect to depth. Figure 2(a) and (b) illustrate respectively, how r and ϕ vary for high arm wave action. We can infer that joint 17(corresponding to leg joint) shows little variation in comparison to joint 9(corresponding to hand joint) for high arm wave action. Figure 3 illustrates that for high arm wave action, the variation of θ is insignificant for both joints, whereas in case of forward kick, the variation in θ is significant for leg joint.

4 ACTION REPRESENTATION MODEL

The bag-of-words model is very popular in human action recognition (Niebles et al., 2008), (Laptev et al., 2008). We develop a Bag-of-Joint-Features (BoJF) model towards this end, so that the learning could be made more efficient. This is described in detail below. While the BoJF model is able to take care of the spatial arrangements of various joints for most of



(a)



(b)

Figure 3: Variations in θ for joints $j = 9$ and $j = 17$ (a)High arm wave (b)Forward kick.

the actions, it totally ignores the temporal variations in joint positions and may result in wrong classification. For example, consider an action where a person moves his hand forward to hit someone or something, and another action where the hand moves backwards to save himself or herself from being hit. The BoJF model would create similar clusters in both the cases. The lack of temporal information may affect recognition accuracy of complex actions. Towards this end, we propose a hierarchical temporal histogram model (HT-hist) later in this section.

4.1 Bag-of-Joint-Features Model (BoJF)

A feature point $f_j(t) = (r_j, \phi_j, \theta_j)$ describes the configuration of joint j in time frame t. We propose to locate 'key' configurations for each joint and to represent an action sample as a distribution of its feature points across the 'key' configurations. In particular, given an action sample, we first obtain the joint-feature $f_j(t)$, $j = 1,2,...19$; $\forall t$ as discussed in section 3. We, next, construct individual codebooks for each joint j by clustering the feature points f_j over all the actions using k-means clustering. The resulting k cluster centers represent the codewords or the 'key' configurations for a joint.

Each feature point can now be described by the codeword it is closest to, using Euclidean distance measure. Specifically, a feature point $f_j(t)$, belonging to an action sample, can be represented by

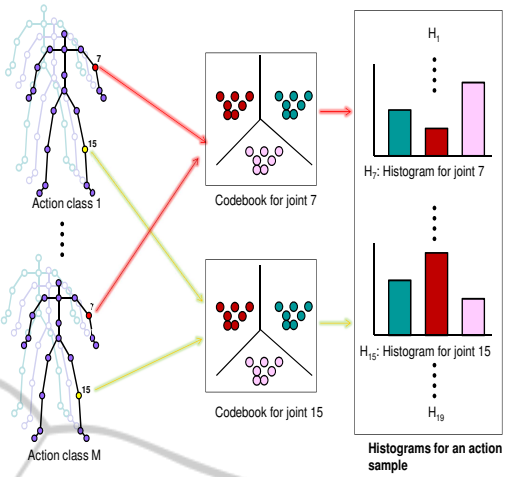


Figure 4: Bag-of-Joint-Features (BoJF) model.

$w_j(t)$ where $w_j(t)$ is a k-dimensional vector containing a single 1 specifying the closest codeword and rest zeroes. A typical vector would have the form: $\langle 0,0,\ldots,0,1,0,\ldots,0\rangle$. Thus, given a sample video, we obtain a histogram of 'key' configurations for each joint. Formally, a BoJF model for an action sample can be represented as:

$$V = \{H_j | j = 1,2,\ldots,19\} \quad (4)$$

$$H_j = \sum_{t=1}^{N} w_j(t) \quad (5)$$

where H_j is a k-dimensional vector representing histogram of codewords for joint j, V is the set of histograms for an action sample and N represents the number of frames in a sample video. Each histogram is normalized and finally, all the 19 histograms of an action sample are concatenated to form the BoJF representation. Figure 4 illustrates the process of constructing BoJF.

4.2 Hierarchical Temporal Histogram (HT-hist)

The BoJF model provides a simple and compact representation of an action video. However, it is possible that the temporal ordering for a set of joints may differ in two or more action sequences. More specifically, consider a complex action which can be broken down into subactions A, B, C and another complex action which has A, C, B as subactions. Both the actions would result in similar clusters since the timing of subactions is not taken into account while clustering. To differentiate between such actions, we develop a Hierarchical Temporal-histogram model (named HT-hist henceforth).

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Let us say there are N frames in an action video. We group these in a single set and call it the top layer. At the next layer, we partition the N frames in two halves resulting in two subsequences. At the bottom-most layer each of these subsequences are again partitioned in two equal parts resulting in 4 subsequences.

At this point we again revisit the set of joints needed for our study. It is observed that while Kinect provides data for 20 joints, most of the actions are carried out by hands or legs. As such, we group the joints of the left arm, right arm, left leg and right leg into separate joint sets (*JS*). Since the features are computed with respect to the root joint, changes in feature values of shoulder and hip joints are often very small in the temporal domain, and need not be considered further. This results in 4 joint sets (*JS*), one for each limb, with 3 joints in each *JS*. Joint sets JS_1 and JS_2 correspond to right and left arms while JS_3, JS_4 correspond to right and left legs respectively as illustrated in Figure 5(a).

$$JS_1 = \{j | j = 7,9,11\} \quad (6)$$

$$JS_2 = \{j | j = 6,8,10\} \quad (7)$$

$$JS_3 = \{j | j = 15,17,19\} \quad (8)$$

$$JS_4 = \{j | j = 14,16,18\} \quad (9)$$

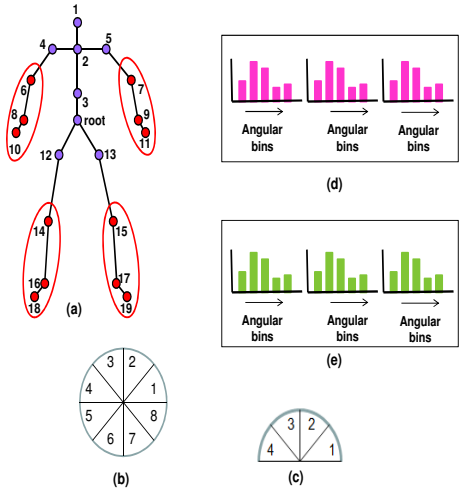


Figure 5: (a)Joint sets (b)Angular bins for θ (bin size of $\pi/4$ used for illustration) (c)Angular bins for ϕ (bin size of $\pi/4$ used for illustration) (d)θ-histograms for the three joints in a *JS* (e)ϕ-histograms for the three joints in a *JS*.

We restrict our attention to temporal variations in θ and ϕ by constructing separate angular bins with bin size of $\pi/12$. We propose $N_\theta = 24$ bins for θ and $N_\phi = 12$ bins for ϕ respectively, since the range for ϕ extends from 0 to π and that of θ extends from 0 to 2π. Figure 5(b) and 5(c) illustrate the angular bins for θ and ϕ respectively.

Given a frame *f*, for each joint *j* in a *JS*, we obtain a N_θ dimensional vector a_f^j of all but one zeros with the 1 corresponding to the angular bin for θ_j. Likewise, we obtain N_ϕ dimensional vector b_f^j of all but one zeros with the 1 corresponding to the angular bin for ϕ_j. Next, we construct histograms A_j and B_j for each subsequence as:

$$A_j = \sum_{f=n_l}^{n_u} a_f^j \quad (10)$$

$$B_j = \sum_{f=n_l}^{n_u} b_f^j \quad (11)$$

where n_l and n_u are the lower and upper frame indices of a subsequence. A θ-histogram is of dimension N_θ while a ϕ-histogram is of N_ϕ dimension. Further, these histograms are constructed for each joint *j* in a *JS* as illustrated in Figure 5(d) and 5(e) where the horizontal axis represents angular bins. The final step of constructing HT-hist model involves normalization and concatenation of the histograms.

5 CLASSIFICATION

We use support vector machine to perform action recognition by considering it as a multi-class classification problem. In particular, we use SVM implemented by LIBSVM (Chang and Lin, 2011) in the one-vs-one scheme mode. Since our representation of an action sample utilizes histograms, we use the histogram intersection kernel (Swain and Ballard, 1991) defined as:

$$I(X,Y) = \sum_{i=1}^{m} min(x_i, y_i) \quad (12)$$

where X and Y represent histograms consisting of m bins, x_i and y_i represent the i^{th} bin of X and Y respectively.

The HT-hist model consists of hierarchical representation, resulting in the number of matches at level *l* also being included in level *l+1*. Therefore, we use weighted histograms for intersection (Lazebnik et al., 2006) with the weights for histogram at level *l* being defined as:

$$q(l) = \begin{cases} 1/2^L & \text{for } l = 0 \\ 1/2^{L-l+1} & \text{for } 1 \le l \le L \end{cases} \quad (13)$$

where $L+1$ are the total number of hierarchical levels. Therefore, given a subsequence *c* of level *l* with weight q_l and histograms A_j and B_j for that subsequence, the final histograms are $q_l A_j$ and $q_l B_j$ respectively.

Table 3: Recognition accuracy comparison for three action subsets of MSR-Action3D dataset.

	Test Two				Cross-subject			
	BOP (Li et al., 2010)	Ours			BOP (Li et al., 2010)	Ours		
		BoJF	HT-Hist	**Combined**		BoJF	HT-Hist	**Combined**
AS1	93.4%	91.8%	86.3%	**94.5%**	72.9%	78.1%	78.1%	**85.7%**
AS2	92.9%	82.7%	85.3%	**92%**	71.9%	65.2%	77.7%	**77.7%**
AS3	96.3%	98.6%	91.9%	**98.6%**	79.2%	90.9%	79.3%	**90.1%**
Overall	94.2%	91.1%	87.8%	**95.1%**	74.7%	78.1%	78.4%	**84.5%**

Table 1: The three subsets of actions in MSR-Action3D dataset (Li et al., 2010).

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal 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

Table 2: Cross-validation results.

JS_1	JS_2	JS_3	JS_4
94.7%	93.3%	92.3%	90.8%
JS_1+JS_2	JS_3+JS_4	*allJS*	
94.01%	91.5%	94.4%	

We, now, have a BoJF and a HT-hist representation for an action sample. The final representation consists of the combined BoJF and HT-hist models.

6 EXPERIMENTS

In the absence of a standard data set consisting of complex actions (which may have overlapping subsequences but in different temporal order) it was not possible to test the usefulness of the HT-hist method over other existing methods in the literature. However, we evaluate the performance of the proposed method on the MSR-Action3D dataset (Li et al., 2010) and Cornell Activity dataset(Sung et al., 2012). The evaluation results are reported in terms of average accuracy and class-confusion matrix.

6.1 MSR-Action3D Dataset

The MSR-Action3D dataset (Li et al., 2010) consists of 20 action classes: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw. Each action is performed 2-3 times by 10 subjects. It consists of depth maps sequences as well as 3D joint positions. It has a total of 557 samples that are used for experiments by us.

We used 2 testing scenarios mentioned in (Li et al., 2010), namely, 'Test Two' and 'Cross-Subject' test. In 'test two' scenario, 2/3 samples are chosen randomly as training samples and the rest as testing samples. In 'cross-subject' testing scenario, half of the subjects are used as training and the rest are used as testing samples. We used the same subject splits as used by the authors in (Li et al., 2010), where subjects 1, 3, 5, 7, 9 were used for training and the rest for testing.

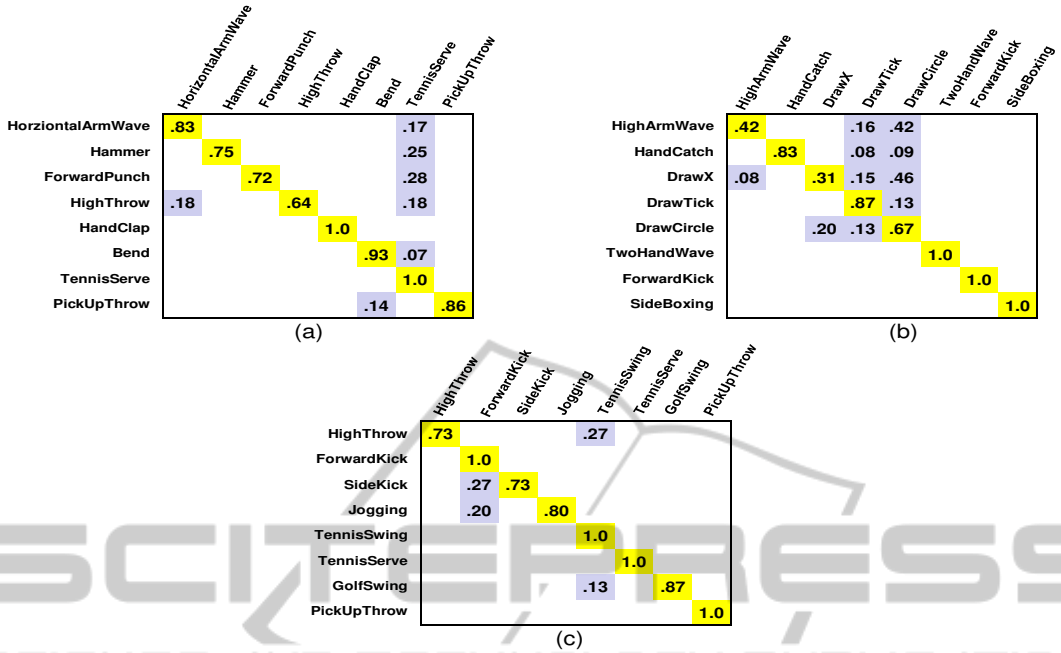
The authors in (Li et al., 2010) divide the 20 actions into three subsets, each having 8 actions as listed in Table 1. The AS1 and AS2, group similar actions with similar movements. AS3, on the other hand, consists of complex actions. We used the same settings as well to compare the effectiveness of our model.

We carried out 5-fold cross-validation on the training data of 'cross-subject' scenario to find the combination of *JS* for the final model. Tables 2 summarizes the results of cross-validation experiment. It turns out that simply using the subset of joints JS_1 yields higher accuracy since most actions are performed with right arm. One could construct the HT-hist model simply with JS_1 as well.

Figure 6 illustrates the class confusion matrix for the different Action sets(AS) of MSR-Action3D dataset using our proposed model. We observe that AS2 consists of very similar actions such as 'Draw X', 'Draw circle' and 'Draw tick'. Hence, most of the misclassification occurs between these classes.

In Table 3, we compare our method with (Li et al., 2010) for the three different action subsets. It is observed that while BoJF model and the HT-hist models individually do not fare too well, the combination gives better results for all the action subsets.

Table 4 shows the comparison for MSR-Action3D dataset with state-of-the art methods in cross-subject



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 6: Confusion matrices for MSR-Action3D dataset (a)Confusion matrix for AS1 (b)Confusion matrix for AS2 (c)Confusion matrix for AS3.

setup. We also compare our proposed models with baseline features. BoJFH-hip consists of our combined model applied to relative cartesian coordinates(joint positions) computed from hip joint in a frame. BoJF-mean consists of our combined model applied to relative cartesian coordinates(joint positions) computed from mean of joint positions in a frame. BoJFH-Spherical represents the combined model consisting of BoJF and HT-hist computed using the proposed features. While our method gives better results compared to other researchers, we observe that in (Wang et al., 2012), the authors achieve higher accuracy since they use lower order fourier coefficients which helps to reduce the noise inherent in skeleton data.

6.2 Cornell Activity Dataset

Cornell Activity dataset (CAD-60) (Sung et al., 2012) contains depth sequences, RGB frames and tracked skeleton joint positions captured with Kinect camera. It consists of 12 different actions: “rinsing mouth”, “brushing teeth”, “wearing contact lens”, “talking on the phone”, “drinking water”, “opening pill container”, “cooking (chopping)”, “cooking (stirring)”, “talking on couch”, “relaxing on couch”, “writing on whiteboard”, “working on computer”. The data consists of actions recorded in 5 different environments: office, kitchen, bedroom, bathroom, and living room wherein 4 different subjects perform these activities.

Table 4: Comparision for MSR-Action3D dataset (Cross-subject Testing).

Method	**Accuracy**
BoJFH-hip	56.4%
BoJFH-mean	61.7%
(Li et al., 2010)	74.7%
(Yang and Tian, 2012)	82.3%
(Wang et al., 2012)	88.2%
BoJFH-Spherical	**84.5%**

Table 5: Comparision for Cornell Activity dataset.

Method	**S-Person**	**C-Person**
(Sung et al., 2012)	81.15%	51.9%
(Koppula et al., 2013)	-	71.4%
(Wang et al., 2012)	94.12%	74.7%
BoJFH-Spherical	**100%**	**86.8%**

Besides these, it also includes “random” and “still” activities.

The recognition accuracy is shown in Table 5. We used the same experimental setup as (Sung et al., 2012). The ‘Have seen’ or the ‘S-Person’ setup uses half of the data of same person as training and the ‘New Person’ or the ‘C-Person’ setting uses leave-one-person-out cross-validation. We achieve 100 % accuracy for the “S-Person” setup and 86.8 % accuracy for ‘C-Person’ setup which are better than the state-of-the-art methods. Figure 7 illustrates the confusion matrix obtained for ‘C-Person’ setting.

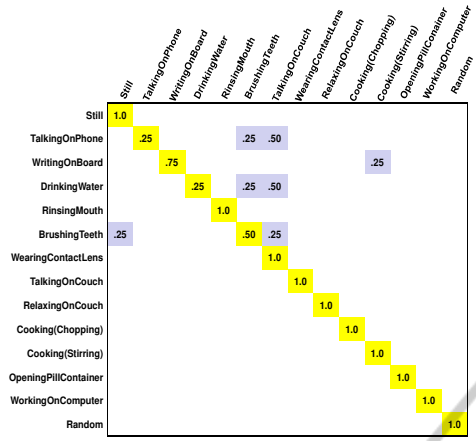


Figure 7: Confusion matrix for Cornell Activity Dataset.

7 CONCLUSION

In this paper, we presented a novel way of using the bag-of-words model to represent an action sample from noisy skeleton data. We have proposed a set of novel joints based features and used these in the proposed Bag-of-Joint-Features (BoJF) model. Further, to take into account temporal differences within and outside an action class, we have proposed the Hierarchical Temporal-histogram (HT-hist) model. We tested our approach on the MSR-Action3D and Cornell activity datasets and obtained results that are comparable with the other state-of-the-art methods. The key advantage of this approach is that it provides an efficient and simpler way of representing an action sample. However, there are some challenges to overcome. Actions involving interaction with environment may not be well represented using just the skeleton data. We may need data from other channels and appropriate methods to represent such actions. In future, we intend to use knowledge from color and depth maps as well to improve the recognition process.

REFERENCES

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267.

Chang, C. C. and Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3).

Jin, S. Y. and Choi, H. J. Essential body-joint and atomic action detection for human activity recognition using longest common subsequence algorithm. In *Computer Vision - ACCV 2012 Workshops*, volume 7729 of *Lecture Notes in Computer Science*, pages 148–159.

Koppula, H., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *IJRR*, 32(8):951–970.

Laptev, I. (2005). On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proceedings of the 2008 Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3D points. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using HMM and Multi-class Adaboost. In *Proceedings of the 9th European Conference on Computer Vision*, pages 359–372.

Ni, B., Wang, G., and Moulin, P. (2011). RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *ICCV Workshops*, pages 1147–1153. IEEE.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*, 79(3):299–318.

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

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, (ICPR'04)*, volume 3, pages 32–36.

Sung, J., Ponce, C., Selman, B., and Saxena, A. (2011). Human activity detection from RGBD images. In *Association for the Advancement of Artificial Intelligence (AAAI) workshop on Pattern, Activity and Intent Recognition*.

Sung, J., Ponce, C., Selman, B., and Saxena, A. (2012). Unstructured human activity detection from RGBD images. In *International Conference on Robotics and Automation (ICRA)*.

Swain, M. and Ballard, D. (1991). Color indexing. In *IJCV, 7(1):1132*.

Turaga, P. K., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 18(11):1473–1488.

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 '12, pages 1290–1297.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Yang, X. and Tian, Y. (2012). Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, pages 14–19.

Yao, A., Gall, J., Fanelli, G., and Van Gool., L. (2011). Does human action recognition benefit from pose estimation? In *BMVC*, pages 67.1–67.11.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS