A Study on the Learning Based Human Pose Recognition

Faisal Sajjad, Adel F. Ahmed, Moataz A. Ahmed
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
{g201409220, adelahmed, moataz}@kfupm.edu.sa

***Abstract*—Human pose recognition is considered a well-known process of estimating the human body pose from a single image or a series of video frames. There exist many applications that can benefit from human pose technology e.g. activity recognition, human tracking, 3D gaming, character animation, clinical analysis of human gait and other HCI applications. Due to its many challenges, such as illumination, occlusion, outdoor environment and clothing, it is considered one of the active areas in computer vision. For the last 15 years, Human pose recognition problem significantly gained interest of many researchers and therefore, many techniques were proposed in order to address the challenges of human pose recognition. In this study, we review the recently progressed work in human pose recognition using computer vision feature extraction and machine learning classification techniques. Accordingly, we identify gaps in existing work and give direction for future work.**

***Index Terms*—Human pose recognition, human pose estimation, human pose surveys, classification of human pose, time of flight camera, body part detection, human pose traditional methods.**

I. Introduction

Human pose plays an important role in the human communication process. The human posture is used to represent different emotions. A recent study [1] shows that human body poses express emotions better than facial expression. Birdwhistlell [2] described the human communication process. According to him, spoken words represent only 7% of human communication while non-verbal actions, such as posture and facial expressions, represent 55% of the overall communication process. Human pose is a non-verbal communication method, which is realized by recognizing the pose of a human. A pose can be extracted from any action such as eating, walking, sitting, waiting, and discussion to mention a few. Human pose can be recognized by localizing joints on human body and dividing the body around these joints into body parts. One such division may include head, neck, shoulders, chest, arms, elbow, thighs, legs, ankle and foot. Once such segmentation is accomplished, the human pose can be recognized accurately using segmented body parts.

Throughout the last few decades, most of computer vision problems such as object recognition and scene recognition were solved in parts. An image or a video is segmented into parts in order to recognize the objects in them. Human pose recognition is a kind of part-based computer vision problem. The body parts are recognized from the whole body. Then, using these recognized parts the exact human pose can be predicted. Many studies [3], [4], [5], [6], [7], [8], [9], [10], [11] used this method for estimation of human pose.

The pose recognition problem gained significant attention by researchers ever since the Microsoft Kinect was announced [12]. Human pose recognition found its way into an increasing number of new range of applications such as human gesture based gaming, sign language interaction with deaf people, human-robot interaction, sports performance examination and human gait analysis. In spite of numerous research, human pose recognition is still a tough and unsolved problem. Furthermore, there exist some challenges in human pose recognition that were not completely addressed by existing methods such as human pose recognition of subjects wearing traditional culture-specific clothes. However, as Microsoft Kinect devices spread and became easily available, data acquisition for this problem became easy and human pose recognition, once again, regained focus in the literature. Researchers started proposing new methods and algorithms to solve the pose recognition challenges using the Microsoft Kinect device.

Figure 1 shows the general framework for recognizing human pose using classification techniques. The purpose of this study is to review the latest advancement of human pose recognition in light of different feature extraction and machine learning classification techniques. We also list the datasets available on-line that are used in human pose recognition studies. As a conclusion, we identify gaps in existing work and give the direction for future work.

The rest of the paper is ordered as follows. The next section presents a brief survey and related work on human pose recognition. The different stages of the general framework shown in Figure 1 will be discussed. Then, details of sensors available for human pose recognition will be given in Section II. Section III presents preprocessing techniques that pave the way for feature extraction. Section IV lists feature extraction techniques proposed in literature for human pose recognition. Section V describes the machine learning classification techniques used in human pose recognition. Section VI describe the post processing techniques used in literature. Then the results of existing work will be shown in Section VII. Finally, we conclude in Section VIII.

978-1-5386-2756-3/17/$31.00 ©2017 IEEE

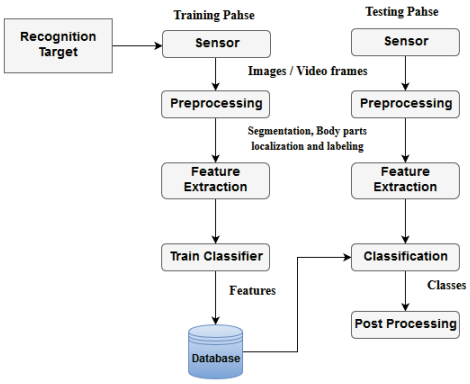

Figure 1. General Framework for Learning Based Human Pose Recognition[13]

II. Related Work

Human pose recognition has gained lots of attention in the last decade. During this time, more than a few surveys or reviews have been published to sum up all related work on human pose in one place. Most recently Nikolaos *et al.* [12] summed up recent published work on 3D human pose recognition extracted from RGB and sequence of images. They categories the methods based on the input and arrangements of key properties. They also conducted experiments on state of the art approaches for human pose recognition by producing their own synthetic dataset.

In the same year, 2016, Gong *et al.* [14] published a comprehensive survey on human pose recognition from monocular images. They also conducted a validation on frequently used methods by collecting 26 available datasets. The survey was comprehensive and covered techniques proposed up to the year 2015. While in 2016 many more human pose recognition techniques were proposed [15]. Liu *et al.* [16] presented a vision based human pose estimation. The authors identified gaps in the existing human pose approaches and also listed down the available public dataset up to 2015.

Roanna *et al.* [17] conducted a survey on human motion recognition and its applications using Microsoft Kinect depth sensor. Perez [18] presents a survey on a model based approaches of human pose recognition. In 2014 Guo *et al.* [19] focused their research on still images and published a survey paper on human action recognition. Chen *et al.* [20] conducted a survey on human motion using depth images. The authors discussed all of the states-of-art work done previously on human motion analysis and its methods.

Another vision based survey was presented by Weinland [21] in 2010. The survey was based on action recognition of human using segmentation techniques. The author precisely focused on full body motion approaches such as weaving, punching, and kicking. In the survey, they identify how the recognition techniques model the temporal and spatial structure of an action. They also classify how human action is segmented using visual data from input stream. There are other surveys that are specifically focused on vision-based human pose recognition [22], [23], [24], [25]. Hotle *et al.* [26] cover model-based methods in the review for 3D human pose recognition, whereas Sminchisescu [27], [28] review the reconstruction of 3D human motion using monocular sequences images.

One drawback of these surveys is that they are too specific and the conducted work in those surveys is at least 5 years old. In this text, we attempt to include all human pose recognition techniques up to the end of the year 2016. For completeness purposes, Table I shows the different parameters each sensor provides, which can be considered for recognizing human pose.

Table I
Available Sensor for Recognizing Human Pose [29]

Sensor	Type	3D Resolution	RGB Resolution	Frame Rate
Microsoft Kinect 2.0	Time of flight	512x424	1920x1080	30-fps
Asus Xtion Pro	Structured light	640x480	1280x1024	30-fps
Intel RealSense R200	Stereo and pattern projector	640x480	1920x1080	60-fps
IFM Efector	Time of flight	176x132	N/A	25-fps
Stereolabs ZED	Embedded stereo	2208x1242	2208x1242	15-fps
Carnegie Robotics	Embedded stereo	2048x1088	2048x1088	15-fps
Ensenso	Structured light	1280x1024	1280x1024	10-fps
SICK 3visitor -T	Time of flight	144x176	N/A	30-fps
e-Con System Tara Stereo	Embedded stereo	752x480	N/A	60-fps
Narian SPI	FPGA Stereo	640x480	N/A	30-fps

III. Preprocessing

A. Human Body Model

Selection of a human body model is one of the major factors in recognizing the human pose. The body model encloses information such as human texture and shape. In the literature, we found three types of human body models, namely, cylindrical human model, pictorial structure human model and kinematic human model. Human pose recognition is heavily dependent on these models whether it is used for full body pose recognition or specifically for recognition of upper body pose.

1) Cylindrical Human Model: Also called a volumetric model. It is used to represent both human pose and human body parts. In this model the human body parts are represented as fixed cylinders. Each cylinder consists of joints. For example, a single human arm represents three joints and these three joints represent one cylinder. The cylinder is further connected to other cylinders in order to form human body structure. The meshes with cylindrical model can also be used to represent human body and its parts. Ganapathi *et al.* [11] represent the human body via meshes. Siddiqui *et al.* [30] proposed an approach that represents human body as a skeleton. The skeleton is then mapped with cylinders with fixed width.

Ling *et al.* [31] proposed a similar cylindrical technique for tracking lower body parts such as the thigh, leg, calf, and foot.

2) Pictorial Structure Human Model: This model is also a very famous model for recognizing human pose. The model represents human body parts as rectangular shape. Mykhaylo *et al.* [32] in 2009 used pictorial structure to predict human pose. Eichner *et al.* [33] [34] used the same model for human pose recognition.

3) Kinematic Human Model: With the announcement of depth sensor this model is frequently used nowadays. This model represents the human body as a set of joints. The human body model generated by the depth sensors consist of 30 to 32 joints depending on the depth sensor used. Using the 3D coordinates of human joints, the human pose is easily estimated. Shotton *et al.* [7] used Kinect sensor to compute human skeleton. Zequn *et al.* [3] used skeleton data for human pose recognition. Youness *et al.* [4] and Ishan *et al.* [5] also used the same model for human pose recognition.

B. Localization of Human Body, Joints, and Parts

Localization of human body, joints, and parts is one of the key steps in the preprocessing phase. Localization is the process of locating the position of human body, its parts and joints from a given image. Many authors [7], [3], [4], [5] use depth sensors like Microsoft Kinect to localize the human body and joints. The depth sensors provide the skeleton information of human. Using this information some techniques [3] [4] find the relative distance between joints and localize the human body parts. Localization of human body is also done through motion sensor devices. Marta *et al.* [15] used Vicon motion sensor to locate human joints.

C. Segmentation of Human Body Parts and Labeling

Ganapathi *et al.* [11] segmented the human body model into 15 rigid parts. Shotton *et al.* [7] divided the human body into 31 parts. These body parts are then labeled with unique colors in order to distinguish each body from another. Zequn *et al.* [3] divided the body parts into three regions, namely, body part, arm part and leg part Similarly, Youness *et al.* [4] also divided the whole body into 20 joints. Mingyuan *et al.* [35] divided the upper body section into 8 parts and labeled each part with a unique color. Ishan *et al.* [5] identified joints from Kinect and then with help of these joints the author segmented the body parts for pose estimation.

However, there are many public datasets available that use synthesized data and label the data with unique colors in order to represent different human body parts.

D. Background Separation

Background subtraction is another important step in the preprocessing phase. It is used to eliminate irrelevent details from the image by removing unwanted pixels. It was found that nearly every study found in literature [7], [36], [37], [38], [39] uses background separation.

IV. Feature Extraction

A. Global Descriptor Based Feature Extraction

Histogram of Oriented Gradient (HOG) is a computer vision based global feature extraction technique. It is applied to the whole image and computes both vertical and horizontal gradients orientation and magnitude. This technique is normally used to detect humans in images. However, there is a vast amount of literature available that uses this technique for recognizing human pose.

In HOG, the image is divided into blocks, then the histogram of the gradient is computed for each block and finally, all histograms are concatenated to form a final feature vector. Sanzari *et al.* [15] estimate human 3D pose using Pyramid Histogram of Oriented Gradients (PHOG) visual features. They divide the human skeleton joints into groups and generate a dictionary of idiosyncratic motion snaps for each group. Each group contains the visual features while the groups are connected hierarchically. The purpose of a dictionary is to evaluate the probability of the group based on its visual features.

Wang *et al.* [40] used pose tree structure and applied HOG features on the human body. Similarly, Sun *et al.* [10] and Yang *et al.* [9] also applied HOG for features extraction. Eichner *et al.* [33] used pictorial structure-based model for recognizing human pose and used Edge HOG feature techniques on it. Fathi *et al.* [41] also used the same Edge HOG techniques to encounter feature vector using Hidden Markov Model. Fathi *et al.* [41] recognize human pose through video frames.

HOG is a global descriptor that is applied to the whole image instead of individual parts of the image. However, the human pose can only be predicted by first localizing the body parts. Therefore, this HOG technique may perform well for detecting human but may not give good accuracies for human poses. To solve this issue there is another technique called Deformable Part Multiscale Model (DPM) [42]. The technique is HOG based applied to individual body parts instead of the whole image.

B. Local Descriptor Based Feature Extraction

A local descriptor like SIFT (Scale Invariant Feature Transform) and LBP (Local Binary Patterns) can be very effective for human pose recognition. These techniques are applied to each body part. The SIFT is gradient-based techniques. It calculates the orientation histogram for each cell and the resultant feature vector is the concatenation of all computed histograms. Ganapathi *et al.* [11] used local descriptor and Holt *et al.* [8] used local binary features for recognizing human pose.

C. Skeleton-based Feature Extraction

Skeleton-based features are normally calculated from depth sensors like the Microsoft Kinect sensor. The depth sensor gives the 3D coordinates of human joints. Using the relative distance between the joints, the feature vector can be computed. Youness *et al.* [4] calculate the set of 20 features from

Table II
CLASSIFIER RELATED WORK

Classifier	Relatedwork
Random Forest (RF)	[5][7][8][40]
Support Vector Machine (SVM)	[34][3][4][31][10][43]
Bayesian and Naive Bayesian (NB)	[15][11][30][33][4][5]
Artificial Neural Network and Deep Learning	[6][35]

each pose using Microsoft Kinect skeleton data. The features are invariant with respect to position and size. Similarly, Zhang *et al.* [3] identify 9 features from Kinect depth sensor. The authors calculate the features by computing the relative distances between joint pairs. The features include left forearm, right forearm, left upper arm, right upper arm, left thigh, right thigh, left crus, right crus and finally the spine. In 2015 Ishan *et al.* [5] also calculates the feature with the help of the Kinect sensor. The authors acquire skeleton information from the sensor and use velocity, position, and acceleration in feature computation. Siddiqui *et al.* [30] also use skeleton joints as features.

D. Depth Based Feature Extraction

Depth feature is calculated from depth images produced by the depth sensors. Shotton *et al.* [7] use depth features for recognizing human pose using random forest classifier. Contrary to analyzing the color data of an acquired image, the method extracts features by analyzing the depth information collected in the depth image by the sensors.

V. CLASSIFICATION

Machine learning techniques are used to train classifiers for human pose recognition. The classifier is first trained with training feature dataset then test feature set are used with the trained classifier for prediction. Table II summarizes the work done with each classifier. Furthermore, studies exist in literature [32], [4], [41] that used AdaBoost, K-nearest neighbor classifier, and Hidden Markov Model respectively. Table V presents a comparison of classifier accuracies in human pose recognition.

VI. POST PROCESSING

After successful classification of features, the post-processing phase defines the classes for the human pose. A pose can be classified as eating, walking, sitting, discussion and waiting to mention a few. Each pose mentioned belongs to a separate class.

Youness *et al.* [4] recorded total 18 poses and each pose consists of 20 features. Marta *et al.* [15] recorded 15 poses. Ishan *et al.* [5] define their own poses in order to detect emotions of designer team member. They used, engage, frustration, boredom and neutral as poses. Similarly, Zequn *et al.* [3] recorded a total 22 different human poses. They divide these 22 pose into 3 categories. The categories are body part, arm part, and leg part. The body part category has 7 different poses, while the arm part has 8 different poses and finally the leg part has 7 poses. Some of the authors recognized upper and lower body pose based on dataset they used.

VII. FINDINGS AND DISCUSSION

A. Datasets

There are many datasets available for human pose recognition. Some of the datasets are specific to upper body parts, others are specific to lower body parts and yet others cover the whole human body parts. Table III lists down the available dataset for human pose.

B. Evaluation Metrics

1) Percentage of Correct Parts (PCP): This performance evaluation metric is normally used by many researchers for detecting human body parts or for recognizing human pose. This metric evaluates the correctly localized body parts in percentage.

2) Precision, Recall, and Accuracy: These metrics are also used to evaluate the performance of human pose. They are calculated as follows.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

C. Comparisons and Analysis

Youness *et al.* [4], in 2016, recognized human pose in real-time using Microsoft Kinect sensor. The features were extracted from skeleton data provided by Kinect. These features were then fed into different machine learning classification techniques in order to identify which classifier performs best in real time. Figure 2 shows the average accuricies of different classifier. The authors tested the performance of the classifier by initially using 22% of the data as training data and then increase the training data up to 88% of the entire dataset. As can be noticed from Figure 2, all the classifiers achieved almost 99% accuracy except naïve Bayesian classifier. The naïve Bayesian classifier reached the accuracy of 98.21% only.

Similarly, Ishan *et al.* [5] applied different machine learning classification techniques to detect the emotional states of each individual subject in a team using Microsoft Kinect. Table IV shows the different classifiers' accuracies. Each classifier performance was tested with up to 4 subjects, i.e. team members. As can be seen from the Table IV, when emotions are recognized with one team member in the frame, the Random Forest classifier outperformed others followed by the IBK classifier. But when the number of team members increased in the frame, the performance of all classifier decreased slightly except for the naïve Bayesian classifier who's accuracy dropped significantly from 98.44% to 53.44%.

Table III
THE AVAILABLE DATASETS

Dataset	Content	Type
HumanEva [44]	50,600 training frames, 26,400 testing frames	Whole body
EVAL [45]	24 sequences	Whole body
LSP [46]	1000 training images, 205 testing images	Whole body
Parse [47]	100 training images, 276 testing images	Whole body
SMMC-10 [11]	6 performers, 28 sequences	Whole body
PDT [48]	26,400 testing frames, 40 sequences	Whole Full
FLIC [49]	3987 training images, 1016 testing images	Whole body
PASCAL 12 [50]	Total 20 classes, 11,530 images for training and validation	Whole body
Buffy [51]	748 frames from "Buffy the vampire slayer" TV show	Upper body
MPII [52]	410 activities of different color and sizes	Whole body
Poses in the wild [53]	30 sequences of different color and sizes	Upper body
Human 3.6M [54]	3.6 million images with 17 scenarios	Whole body
CMU [55]	23 actions, 109 subjects and 2605 videos	Whole body
MPII Cooking Activities [56]	65 actions, 12 subjects, and 44 videos	Upper body
UMPM [57]	Multiple people with 30 subjects and 36 videos	Whole body
TUM Kitchen [58]	4 actions, 4 subjects, and 20 videos	Whole body
KTH Multiview Football [59]	Total 8307 images with 2D and 3D dataset	Whole body
Video Pose [60]	Total 1289 images from 44 short clips	Upper body

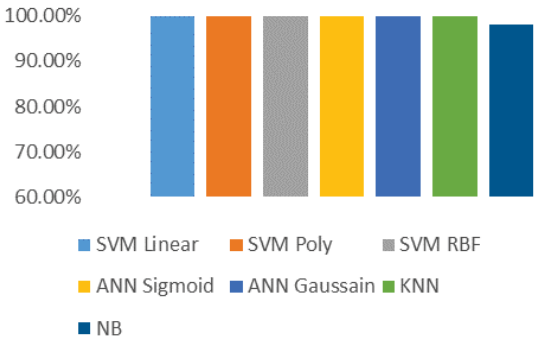

Figure 2. Youness *et al.* [4] Classifiers Average Accuracies

Table IV
ISHAN *et al.* , [5] CLASSIFIERS ACCURACIES

Members	C4.5	RF	IBK	NB
1	99.59 %	100%	99.92%	98.44%
2	98.25%	99.85%	98.97%	81.45%
3	98.36%	99.94%	99.04%	62.78%
4	98.40%	98.85%	99.51%	53.44%
AVG	**98.65%**	**99.66%**	**99.36%**	**74%**

Jiu *et al.* [35] applied deep learning technique on body parts using depth images. The author created the ground truth image by labeling the upper body parts. The author used Convolutional Neural Network based deep learning technique in addition to the Logistic Regression (LR) classifier. These two classifiers were then trained with spatial relationships. The author compared their results with random forest technique and Dimensionality Reduction by Learning an Invariant Mapping (DrLIM) [61] without spatial learning method. The method proposed by Jiu *et al.* [35] produced better results as compared to Random Forest and DrLIM. Furthermore, it is apparent from result comparison that Even DrLIM results are better than Random Forest.

Table V describe the accuracies of previously reported work with other parameters such as proposed techniques, feature extraction techniques, classification techniques and sensor/ datasets. We conclude that using skeleton features the Random Forest classifier performed best followed by SVM, Artificial Neural Network, C4.5, and KNN. The naïve Bayesian classifier performed worst with skeleton features. However, in part based features techniques the Convolutional Neural Network based deep learning technique, the Deep Neural Network performed

VIII. CONCLUSION AND FUTURE WORK

Human pose recognition is an important research topic nowadays. There are many techniques proposed to accuratly recognize human pose. In this paper, we reviewed recently published literature in learning based human pose recognition. We listed the available sensors used for human pose recognition. We described the feature extraction and classification techniques used in literature. We also listed the publicly available datasets that are used in research for human pose recognition. Finally, we discussed and compared the results of different classifiers.

There are many limitations in existing work. As we can see from Table V, many authors used HOG features for human pose recognition. There are some limitations of the HOG technique. For example, it is can only be applied on the whole image. However, in pose recognition, we need to apply feature extraction technique on individual body parts

Table V
RELATED WORK SUMMARY

Methods	Proposed Techniques	Features	Classifiers	Source	Sensor /Datasets	Accuracy
Marta *et al.* 2016 [15]	Construct human pose using representation of the idiosyncratic motionof human body parts	PHOG	Hierarchical Bayesian	Images	Human 3.6 M	-NA-
Youness *et al.* 2016 [4]	Part based approach and Recognize pose using Kinect skeleton data	20 features using relative distance	SVM, ANN, KNN and Naïve Bayes	Images	Real data from Kinect	See Fig 2
Ishan *et al.* 2015 [5]	Part based approach and Recognize design team member emotion using Kinect skeleton data	Velocity, acceleration and position	C4.5, Random Forest, IBK, Naïve Bayes	Images	Real time frames from Kinect	98%
Zequn *et al.* 2014 [3]	Part based approach and recognize user defined pose using Kinect	9 features using relative distance	SVM	Images	Real time frames from Kinect	99.14%
Toshev *et al.* 2014 [6]	Part based approach	-NA-	Deep Neural Networks	Images	LSP, FLCIC and Image Parse	69%
Jiu *et al.* 2014 [35]	Part based approach	Energy function	Deep Learning	Images	CDC4CV Poselets dataset	66.92%
Wang *et al.* 2013 [40]	Pose tree structure	HOG	Tree style	Images	PARSE, LSP	62.8%
Shotton *et al.* 2013 [7]	Part based approach using depth image	Depth features	Random Forest	Images	Own created dataset	60.30%
Eichner *et al.* 2012 [34]	Pictorial Structure	HOG, Shapes Edges	SVM	Images	PASCAL 08, Buffy	-NA-
Sapp *et al.* 2011 [43]	Sub Gradient approach	Geometry, Color optical	SVM	Video	Video Pose 2.0	68.3%
Sun *et al.* 2011 [10]	Part based model	HOG	SVM	Images	PASCAL 07	64.2%
Yang *et al.* 2011 [9]	Mixture of parts based approach	HOG	SVM	Images	LSP, Image Parse and Buffy	55.1%
Holt *et al.* 2011 [8]	Part based approach	Binary features	Random Decision Forest	Images	CDC4CV Poselets	67%
Ganapathi *et al.* 2010 [11]	Part based approach	Local descriptors	Dynamic Bayesian Model	Images	MOCAP from Phase Space System	-NA-
Siddiqui *et al.* 2010 [30]	MCMC which is data driven approach and used fixed cylinders	Skeleton joints	Bayesian	Range images	Real time frame from SR300 sensor by MESA	0.930
Eichner *et al.* 2010 [33]	Pictorial structure and Probability based approaches	Edgelet HOG	Probability based	Images	Own dataset	-NA-
Mykhaylo *et al.* 2009 [32]	Pictorial structure based approach	Shape Context	AdaBoost	Images	TUD-Pedestrians, TUD-Upright People	55.2%
Fathi *et al.* 2007 [41]	Hidden Markov Model	Edges HOG	HMM	Video	CMU MoBo	-NA-

instead of the entire human body. That is why the accuracies of these technique are not good enough as can be seen in Table V. Therefore, for human pose recognition we need computer vision based local descriptors that can be applied on individual body parts. Such local descriptor can be SIFT, LBP and DPM.

Similarly, for classification techniques, the probability-based classifiers such as AdaBoost and naïve Bayesian are unable to give satisfactory results see Figure 2, Table IV and Table V. The drop in the accuracy of these classifiers is due to class conditional independence. Therefore, there is a need to use alternative classifiers that exhibit good training and testing accuracies. From Table V we can observer that Random Forest, SVM, Neural Networks and Deep Learning were among the top performers in human pose recognition.

Nowadays many authors are proposing new techniques with the help of skeleton features, which is provided by the depth sensors. Again, from Table V, we can see that techniques using skeleton features outperformed other techniques in terms of accuracy. Furthermore, these techniques in conjunction with the available sensors are good enough for recognizing poses and skeleton of subjects wearing non-cultural clothes where the limbs are easily and clearly distinguished.

Furthermore, we observed that all of the human pose recognition techniques were applied subjects wearing pants and shirts, where all four limbs are clearly distinguishable and visible. However, there is still an opportunity to work on human pose recognition for subjects wearing draped clothes, like long skirts, dresses and kimonos, where limbs are partially or fully covered and non-distinguishable.

ACKNOWLEDGMENT

We desire to acknowledge King Fahd University of Petroleum and Minerals (KFUPM) for utilizing the various facilities in carrying out this research.

REFERENCES

[1] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[2] R. L. Birdwhistell, *Kinesics and context: Essays on body motion communication*. University of Pennsylvania press, 2010.

[3] Z. Zhang, Y. Liu, A. Li, and M. Wang, "A novel method for user-defined human posture recognition using kinect," in *Image and Signal Processing (CISP), 2014 7th International Congress on*, pp. 736–740, IEEE, 2014.

[4] C. Youness and M. Abdelhak, "Machine learning for real time poses classification using kinect skeleton data," in *Computer Graphics, Imaging and Visualization (CGiV), 2016 13th International Conference on*, pp. 307–311, IEEE, 2016.

[5] I. Behoora and C. S. Tucker, "Machine learning classification of design team members' body language patterns for real time emotional state detection," *Design Studies*, vol. 39, pp. 100–127, 2015.

[6] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014.

[7] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[8] B. Holt, E.-J. Ong, H. Cooper, and R. Bowden, "Putting the pieces together: Connected poselets for human pose estimation," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1196–1201, IEEE, 2011.

[9] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1385–1392, IEEE, 2011.

[10] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 723–730, IEEE, 2011.

[11] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 755–762, IEEE, 2010.

[12] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, vol. 152, pp. 1–20, 2016.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[14] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-h. Zahzah, "Human pose estimation from monocular images: A comprehensive survey," *Sensors*, vol. 16, no. 12, p. 1966, 2016.

[15] M. Sanzari, V. Ntouskos, and F. Pirri, "Bayesian image based 3d pose estimation," in *European Conference on Computer Vision*, pp. 566–582, Springer, 2016.

[16] Z. Liu, J. Zhu, J. Bu, and C. Chen, "A survey of human pose estimation: the body parts parsing based methods," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 10–19, 2015.

[17] R. Lun and W. Zhao, "A survey of applications and human motion recognition with microsoft kinect," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, no. 05, p. 1555008, 2015.

[18] X. Perez-Sala, S. Escalera, C. Angulo, and J. Gonzalez, "A survey on model based approaches for 2d and 3d visual human pose recovery," *Sensors*, vol. 14, no. 3, pp. 4189–4210, 2014.

[19] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343–3361, 2014.

[20] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013.

[21] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[22] Y. Li and Z. Sun, "Vision-based human pose estimation for pervasive computing," in *Proceedings of the 2009 workshop on Ambient media computing*, pp. 49–56, ACM, 2009.

[23] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[24] S. J. Krotosky and M. M. Trivedi, "Occupant posture analysis using reflectance and stereo image for" smart" airbag deployment," in *Intelligent Vehicles Symposium, 2004 IEEE*, pp. 698–703, IEEE, 2004.

[25] R. Poppe, "Vision-based human motion analysis: An overview," *Computer vision and image understanding*, vol. 108, no. 1, pp. 4–18, 2007.

[26] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of selected topics in signal processing*, vol. 6, no. 5, pp. 538–552, 2012.

[27] C. Sminchisescu, "3d human motion analysis in monocular video: techniques and challenges," in *Human Motion*, pp. 185–211, Springer, 2008.

[28] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*

(Applications and Reviews), vol. 34, no. 3, pp. 334–352, 2004.

[29] "Sensor survey.." http://rosindustrial.org/news/2016/1/13/3d-camera-survey. Accessed: 2017-01-10.

[30] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 1–8, IEEE, 2010.

[31] R. Z. L. Hu, "Vision-based observation models for lower limb 3d tracking with a moving platform," 2011.

[32] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1014–1021, IEEE, 2009.

[33] M. Eichner and V. Ferrari, "We are family: Joint pose estimation of multiple persons," in *European Conference on Computer Vision*, pp. 228–242, Springer, 2010.

[34] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International journal of computer vision*, vol. 99, no. 2, pp. 190–214, 2012.

[35] M. Jiu, C. Wolf, G. Taylor, and A. Baskurt, "Human body part estimation from depth images via spatially-constrained deep learning," *Pattern Recognition Letters*, vol. 50, pp. 122–129, 2014.

[36] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Ieee iccv*, vol. 99, pp. 1–19, 1999.

[37] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, pp. 246–252, IEEE, 1999.

[38] P. D. Z. Varcheie, M. Sills-Lavoie, and G.-A. Bilodeau, "A multiscale region-based motion detection and background subtraction algorithm," *Sensors*, vol. 10, no. 2, pp. 1041–1061, 2010.

[39] C. Guillot, M. Taron, P. Sayd, Q. C. Pham, C. Tilmant, and J.-M. Lavest, "Background subtraction adapted to ptz cameras by keypoint density estimation," in *Proceedings of the British Machine Vision Conference*, pp. 34–1, 2010.

[40] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 596–603, 2013.

[41] A. Fathi and G. Mori, "Human pose estimation using motion exemplars," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[42] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[43] B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1281–1288, IEEE, 2011.

[44] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International journal of computer vision*, vol. 87, no. 1-2, p. 4, 2010.

[45] "Eval dataset.." http://www.comp.leeds.ac.uk/mat4saj/lsp.html. Accessed: 2017-01-10.

[46] "Lsp dataset.." http://www.comp.leeds.ac.uk/mat4saj/lsp.html. Accessed: 2017-01-10.

[47] "Parse dataset.." https://computing.ece.vt.edu/~santol/projects/zsl_via_visual_abstraction/parse/index.html. Accessed: 2017-01-10.

[48] T. Helten, "Processing and tracking human motions using optical, inertial, and depth sensors," 2013.

[49] "Flic dataset.." http://bensapp.github.io/flic-dataset.html. Accessed: 2017-01-10.

[50] "Pascal dataset.." http://host.robots.ox.ac.uk/pascal/VOC/. Accessed: 2017-01-10.

[51] "Buffy dataset.." http://www.robots.ox.ac.uk/~vgg/data/stickmen/. Accessed: 2017-01-10.

[52] "Mpii dataset.." http://human-pose.mpi-inf.mpg.de/. Accessed: 2017-01-10.

[53] "Mixing body-part sequences for human pose estimation dataset.." http://lear.inrialpes.fr/research/posesinthewild/. Accessed: 2017-01-10.

[54] "Human 3.6h dataset.." http://vision.imar.ro/human3.6m/description.php. Accessed: 2017-01-10.

[55] "Cmu-mocap dataset.." http://mocap.cs.cmu.edu/. Accessed: 2017-01-10.

[56] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1194–1201, IEEE, 2012.

[57] "Umpm benchmark: A multi-person dataset." https://www.projects.science.uu.nl/umpm/. Accessed: 2017-01-10.

[58] "Tum kitchen dataset." https://ias.cs.tum.edu/software/kitchen-activity-data. Accessed: 2017-01-10.

[59] "Kth multiview football dataset.." http://www.csc.kth.se/cvap/cvg/?page=software. Accessed: 2017-01-10.

[60] "Video pose dataset.." http://bensapp.github.io/videopose-dataset.html. Accessed: 2017-01-10.

[61] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.