A Random Forest Approach to Segmenting and Classifying Gestures

Ajjen Joshi[1], Camille Monnier[2], Margrit Betke[1] and Stan Sclaroff[1]
[1]Department of Computer Science, Boston Univeristy, Boston, MA 02215 USA
[2]Charles River Analytics, Cambridge, MA 02138 USA

Abstract— **This work investigates a gesture segmentation and recognition scheme that employs a random forest classification model. Our method trains a random forest model to recognize gestures from a given vocabulary, as presented in a training dataset of video plus 3D body joint locations, as well as out-of-vocabulary (non-gesture) instances. Given an input video stream, our trained model is applied to candidate gestures using sliding windows at multiple temporal scales. The class label with the highest classifier confidence is selected, and its corresponding scale is used to determine the segmentation boundaries in time. We evaluated our formulation in segmenting and recognizing gestures from two different benchmark datasets: the NATOPS dataset of 9,600 gesture instances from a vocabulary of 24 aircraft handling signals, and the ChaLearn dataset of 7,754 gesture instances from a vocabulary of 20 Italian communication gestures. The performance of our method compares favorably with state-of-the-art methods that employ Hidden Markov Models or Hidden Conditional Random Fields on the NATOPS dataset.**

I. INTRODUCTION

The problem of spotting and recognizing meaningful gestures has been an important research endeavor in the fields of computer vision and pattern recognition. Research in this domain has a broad scope of applications such as recognizing sign-language symbols, enabling video surveillance, and developing new modes of human-computer interaction, among others.

A common approach in solving the gesture segmentation and classification problem involves separating them into two-subproblems where the task of segmentation precedes the task of recognition. In this method [1], [2], [3], the focus is on first finding the gesture segmentation boundaries in time. The candidate gestures produced by the segmentation algorithm is then classified. One of the limitations of this approach is the dependence of classification on segmentation: a good gesture classification algorithm will fail to yield desirable results if the segmentation algorithm is inaccurate. Another disadvantage of this method is the difficulty in distinguishing contiguously occurring gestures.

Ours is a unified approach that simultaneosly performs the tasks of segmentation and classification. In methods such as ours [4], [5], gesture intervals for which above-threshold scores are given by the classifier are deemed to be the labeled and segmented gesture. Thus, we attempt to design a framework capable of automatically and accurately spotting and classifying gestures present in a set of test videos, given a training set of RGBD videos and 3D joint locations with multiple examples of all gestures in a gesture vocabulary.

We take a random forest approach to creating a fast and accurate classifier. Random forests are an example of an ensemble method, where multiple classifiers engage in a voting strategy to provide the final prediction.They have been applied to good effect in real-time human pose recognition [6], object segmentation [7], and image classification [8] among others. Many works dealing with spatiotemporal signals, such as gestures, employ graphical models such as Conditional Random Fields (CRFs) [9], [10], and Hidden Markov Models (HMMs) [5], [11] in order to model relationships and variations in both the temporal and spatial domains. We show that taking a random forest approach in gesture classification tasks can be beneficial because they are often simpler to implement and easier to train than graphical models, while providing a comparable (and in some cases, better) accuracy in recognition.

The key contributions of this work are: (1) the design of a simple framework that employs a single multi-class random forest classification model to distinguish gestures from a given vocabulary in a continuous video stream, (2) the fusion of 3D joint-based features with color and appearance-based features to create an accurate feature representation of gestures that is robust to variations in user height, distance of user to sensor and speed of execution of gesture, and (3) the creation of a uniform feature descriptor for gestures to account for the variability in their length by dividing gestures into a fixed number of temporal segments followed by the concatenation of the representative feature vectors of each temporal segment.

II. RELATED WORK

Here, we list and briefly explain some of the important methods that have been used in gesture recognition and are relevant to our work. A more comprehensive survey of gesture recognition techniques can be found elsewhere [12].

Nearest neighbor models are often used in gesture classification problems. Malassiotis et al. [13] used a k-NN classifier to classify static sign language hand gestures. A normalized cross-correlation measure was used to compare the feature vector of an input image with those in the k-NN model. Dynamic Time Warping (DTW) can be used to compute a matching score between two temporal sequences, a variant of which was used by Alon et al. [4]. A drawback of k-NN models is the difficulty in defining distance measures that clearly demarcate different classes of time series observations.

A Hidden Markov Model (HMM) is another widely used tool in temporal pattern recognition, having been implemented in applications of speech recognition, handwriting recognition, as well as gesture recognition. Starner et al. [11]

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

employed an HMM-based system to recognize American Sign Language symbols. One difficulty while implementing HMMs is to determine an appropriate number of hidden states, which can be domain-dependent.

The Conditional Random Field (CRF), introduced by Lafferty et al. [14] is a discriminative graphical model with an advantage over generative models, such as HMMs: the CRF does not assume that observations are independent given the values of the hidden variables. Hidden Conditional Random Fields (HCRF) use hidden variables to model the latent structure of the input signals by defining a joint distribution over the class label and hidden state labels conditioned on the observations [15]. HCRFs can model the dependence between each state and the entire observation sequence, unlike HMMs, which only capture the dependencies between each state and its corresponding observation. Song et al. used a Gaussian temporal-smoothing HCRF [9] to classify gestures that combine both body and hand signals. They also presented continuous Latent Dynamic CRFs [10] to classify unsegmented gestures from a continuous input stream of gestures.

Randomized decision forests have been applied in a variety of ways in problems related with classifying gestures. Miranda et al. [16] used a gesture recognition scheme based on decision forests, where each node in a tree in the forest represented a keypose, and the leaves of the trees represented gestures corresponding to the sequence of keyposes that constitute the gesture as one traverses down a tree from root to leaf. Gall et al. [17] used Hough forests to perform action recognition. In Hough forests, a set of randomized trees is trained to perform a mapping from a densely-sampled d-dimensional feature space into corresponding votes in Hough space. Demirdjian et al. [18] proposed the use of temporal random forests in order to recognize temporal events. Randomized decision forests have been shown to be robust to the effects of noise and outliers. Moreover, they generalize well to variations in data [19]. Thus, random forests are suitable for classification tasks involving data such as gestures because data collected by image and depth sensors can be sensitive to noise and their execution can exhibit a high level of variance.

III. System Overview

A. Training

An overview of how our gesture recognition framework is trained is shown in Figure 1. Here, we explain in detail the elements of our framework:

1) Input: The input to our framework consists of RGB images, depth maps and 3D skeletal joint data for every frame of the videos in the datasets. Each input video contains several in-vocabulary gestures and is labeled with ground truth temporal segmentation as well as class labels. Let c be the number of different gestures that are present in the gesture vocabulary. We used all instances of each of the c different gestures and created additional examples to represent a non-gesture class by randomly selecting intervals between two gestures of varying length. This set of examples was used to train a $c+1$-class random forest classifier.

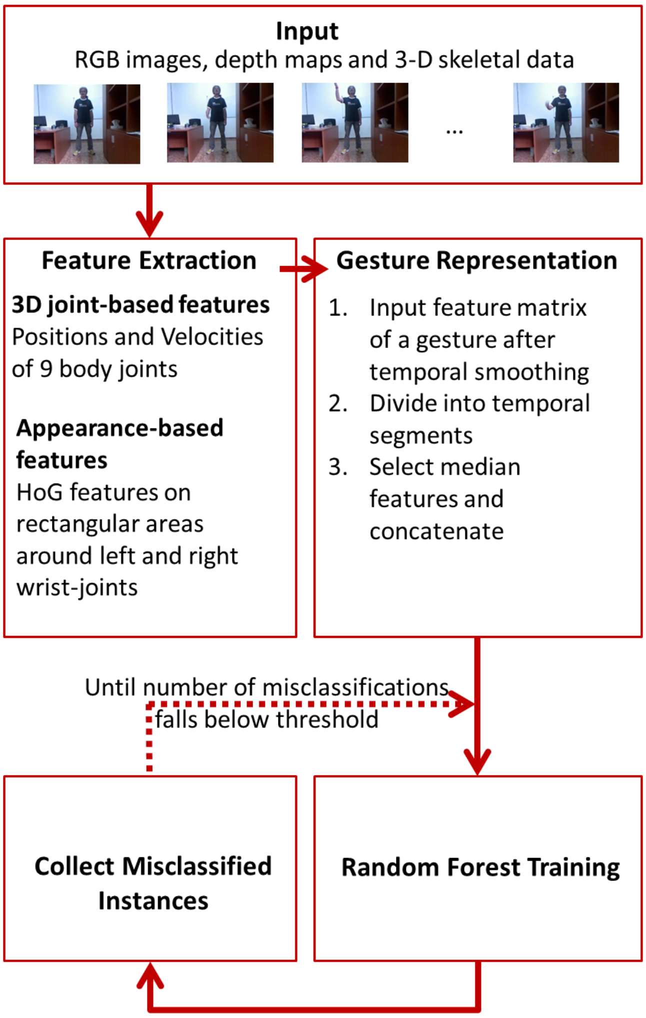

Input
RGB images, depth maps and 3-D skeletal data
...
Feature Extraction
3D joint-based features
Positions and Velocities of 9 body joints
Appearance-based features
HoG features on rectangular areas around left and right wrist-joints
Gesture Representation
1. Input feature matrix of a gesture after temporal smoothing
2. Divide into temporal segments
3. Select median features and concatenate
Until number of misclassifications falls below threshold
Collect Misclassified Instances
Random Forest Training

Fig. 1. Pipeline view of training our gesture recognition framework

2) Feature Extraction: Each training example consists of a varying number of frames, each of which is described by a feature descriptor. Our system computes normalized positional and velocity features for nine different skeletal body joints (left and right shoulders, elbows, wrists and hands, as well as the head joint). Since gestures are performed by subjects with different heights, at different distances from the camera sensor, we first normalized the positional coordinates of the users' joints using the length of the user's torso as a reference. The normalized position vector for joint j at time t is:

$$\mathbf{W}_j(t) = \frac{\mathbf{W}^r_j(t) - \mathbf{W}^r_{hip}(t)}{l}, \qquad (1)$$

where $\mathbf{W}^r_j(t)$ is the raw position vector for joint j at time t, $\mathbf{W}^r_{hip}(t)$ is the raw position vector for the hip joint at time t, and l is the length of the torso defined as:

$$l = \|(\mathbf{W}_{head} - \mathbf{W}_{hip})\|. \qquad (2)$$

Our system uses the normalized positional coordinates (W_x, W_y, W_z) of these nine joints along with their rotational values (R_x, R_y, R_z, R_w), which are provided with the dataset, and computes values for their velocities $(W'_x, W'_y, W'_z, R'_x, R'_y, R'_z, R'_w)$.

$$OFdense(t) = \sum_{v \in V(t)} \sum_{u \in U(t)} \frac{\sqrt{u^2 + v^2}}{H(t) \times W(t)}, \tag{3}$$

Thus, there are 126 feature descriptors extracted from 3D skeletal data for every frame. In addition, we found that augmenting our skeletal feature vector with Histogram of Oriented Gradients (HOG) features [20] on 32x32 pixel squares centered on the left and right hands help improve classification accuracy. The HOG features for each of the two squares can be represented as a 324-dimensional vector. We obtain a dimensionality-reduced representation by performing Principal Component Analysis (PCA) and using the first 20 principal components for each hand. Thus, every frame of every instance in our training set is represented by a 166 dimensional feature descriptor.

3) Gesture Representation: In order to remove the effects of noisy measurments, we first smoothed all features using a moving average filter spanning 5 frames. Smoothing features slightly improved classification accuracy. Because instances of gestures and non-gestures in our training set are temporal sequences of varying length, there arises the need to represent every gesture with a feature vector of the same length. We achieved this by dividing the gesture into 10 equal-length temporal segments, and representing each temporal segment with a vector of the median elements of all features. Using 10 temporal segments provided a balance between keeping the feature representation concise, while encapsulating enough temporal information useful in discerning the gesture classes. Using the median elements of all features provided better performance than using the mean feature value, or the feature value correspondng to the median frame of the temporal segment. The representative vectors of each temporal segment were then concatenated into a single feature vector.

4) Random Forest Training: We defined the training set as $\mathscr{D} = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)\}$. Here, $(\mathbf{X}_1, ..., \mathbf{X}_n)$ corresponds to the uniform-length feature vector representing each gesture or non-gesture, and $(Y_1, ..., Y_n)$ represents their corresponding class labels.

A random forest classification model consists of several decision tree classifiers $\{t(\mathbf{x}, \phi_k), k = 1, \dots\}$ [19]. Each decision tree $t(\mathbf{x}, \phi_k)$ in the forest is constructed until they are fully grown. Here $\mathbf{x}$ is an input vector and ϕ_k is a random vector used to generate a bootstrap sample of objects from the training set $\mathscr{D}$. The ideal number of trees in our random forest model was determined to be 500 by studying the Out-of-Bag (OOB) error rate in the training data.

Let d be the dimensionality of the feature vector of the inputs. At each internal node of the tree, m features are selected randomly from the available d, such that $m < d$. m = $\sqrt{d}$ provided the highest accuracy among other common choices for m (1, $0.5\sqrt{d}$, $2\sqrt{d}$, d). From the m chosen features, the feature that provides the most information gain is selected to split the node. Information gain (I) can be defined as:

$$I_j = H(S_j) - \sum_{k \varepsilon (L,R)} \frac{|S_j^k|}{|S|} H(S_j^k), \tag{4}$$

where S_j is the set of training points at node j, $H(S_j)$ is the Shannon entropy at node j before the split, and S_j^L and S_j^R are the sets of points at the right child and left child respectively of the parent node j after the split.

The Shannon entropy can be defined as:

$$H(S) = -\sum_{c \varepsilon C} p_c log(p_c), \tag{5}$$

where S is the set of training points and p_c is the probability of a sample being class c.

We trained and saved a random forest classification model based on the features that we extracted. There is a need to strengthen the classifier's ability to accurately detect intervals of non-gestures because the randomly chosen intervals of non-gestural examples fail to fully model the class of non-gestures. In order to achieve this, we applied the random forest model on continuous input of the training set and collected false positives and false negatives, which are examples of intervals from the training set that the classifier fails to classify correctly. The set of false positives and false negatives instances is then added to the original training set, and the random forest is re-trained using the new extended set of training examples. This process of bootstrapping, as performed by Marin et al. [21], is performed iteratively until the number of false positives gets reduced below a threshold, which was empirically determined to be one false positive per training sample.

B. Testing

The task during testing is to use our trained random forest model to determine the temporal segmentation as well as class labels of gestures in a continuous video. We performed multi-scale sliding window classification to predict the class labels of the gestures, as well as their start and end points.

For each input video, gesture candidates were constructed at different temporal scales. Let f_s be the number of frames in the shortest gesture in the training set and f_l be the number of frames in the longest gesture in the training set. Then, the temporal scales ranged from length f_s to length f_l, in increments of 5 frames. Let, $\mathscr{G} = \{g_1, ...g_n\}$ be the set of gesture candidates at different temporal scales. At each scale, a candidate gesture g_i was constructed by concatenating the feature vectors at an interval specified by the temporal scale, so that the dimensions of the feature vector matched those of the gestures used to train the classification model.

Within a buffer of length larger than the longest temporal scale, a sliding window was used to construct gesture candidates at each temporal scale. For a buffer of size b, the number of gesture candidates at scale s_i is equal to $b - s_i + 1$.

We chose b to be 100 frames, which is marginally greater than the maximum length of a gesture in the training set. Gesture candidates generated by the sliding window within the temporal neighborhood defined by the buffer at each scale were classified by our trained random forest model and competed to generate a likely gesture candidate G_{s_i} at that scale. Since gesture candidates at the neighborhood of where the gesture is truly temporally located tend to be classified as the same gesture, we performed Non-Maxima Suppression to select the most likely gesture candidate.That is, for each scale s_i, $b - s_i + 1$ gesture candidates were generated and the one classified with the highest confidence (G_{s_i}) within a temporal neighborhood was selected. The confidence score is the percentage of decision trees that vote for the predicted class. Finally, the likely gesture candidates at the various scales competed to generate the final predicted gesture within the buffer.

Therefore, within the buffer, the scale of the final predicted gesture helps determine the segmentation boundaries of the gesture, whereas its class label is that which is predicted by the random forest classifier. The end point of the predicted gesture was chosen to be the start point of the new buffer. This process was then repeated until the end of the test video was reached.

IV. Datasets

Here, we describe the nature of the datasets we have used to test our gesture recognition system.

A. NATOPS

The Naval Air Training and Operating Procedures Standardization (NATOPS) gesture vocabulary comprises a set of gestures used to communicate commands to naval aircraft pilots by officers on an aircraft carrier deck. The NATOPS dataset [22] consists of 24 unique aircraft handling signals, which is a subset of the set of gestures in the NATOPS vocabulary, performed by 20 different subjects, where each gesture has been performed 20 times by all subjects. An example gesture is illustrated in Figure 2. The samples were recorded at 20 FPS using a stereo camera at a resolution of 320 x 240 pixels. The dataset includes RGB color images, depth maps, and mask images for each frame of all videos. A 12 dimensional vector of body features (angular joint velocities for the right and left elbows and wrists), as well as an 8 dimensional vector of hand features (probability values for hand shapes for the left and right hands) collected by Song et al. [22] was also provided for all frames of all videos of the dataset.

B. ChaLearn

The ChaLearn dataset was provided as part of the 2014 Looking at People Gesture Recognition Challenge [23]. The focus of the gesture recognition challenge was to create a gesture recognition system trained on several examples of each gesture category performed by various users. The gesture vocabulary contains 20 unique Italian cultural and anthropological signs. Figure 3 shows an example gesture being perfomed.

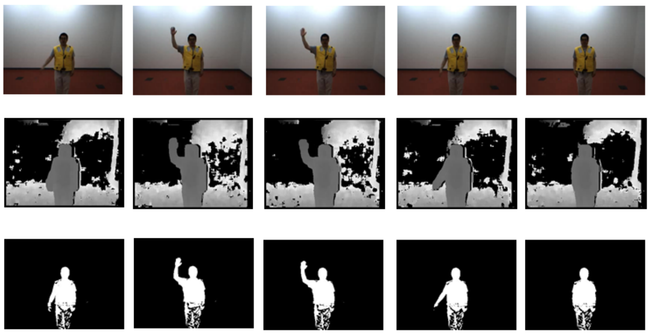


Fig. 2. RGB, Depth, and User-Mask Segmentation of a subject performing gesture 1 'I Have' in the NATOPS dataset

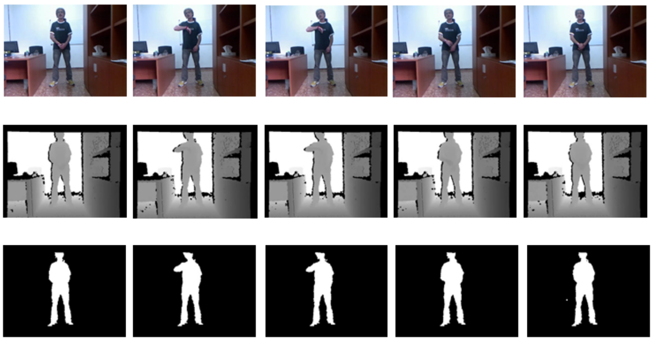


Fig. 3. RGB, Depth, and User-Mask Segmentation of a subject performing gesture 1 'sonostufo' in the ChaLearn dataset

The data used to train the recognition system contains a total of 7,754 manually labeled gestures. Additionally, a validation set with 3,363 labelled gestures was provided to test the performance of the trained classifier. During the final evaluation phase, another 2,742 gestures were provided. The dataset includes RGB and depth video along with 3D skeletal joint positions for each frame.

V. Experiments

Here we describe the experiments performed to evaluate our gesture recognition system on the two datasets. We used the NATOPS dataset to evaluate our gesture classification system in a non-continuous setting. We used a set of gesture samples to train our gesture classifier, and tested its performance on a test-set of segmented gestures. The ChaLearn dataset consists of training and test videos where the user performs both in-vocabulary and out-of-vocabulary gestures, with intervals of gestural silence or transitions. Thus, we used the ChaLearn dataset to test the performance of our system on continuous input.

From the NATOPS dataset, we trained our gesture recognition model with the following features sets in order to formulate a good feature representation:

(a) 3D skeletal joints and hand-shape based feature set (SK+HS): This feature set [9] consists of 20 unique features for each timeframe for every gesture. The

TABLE I

AVERAGE CLASSIFICATION ACCURACY ON ALL 24 GESTURES OF THE NATOPS DATASET

Feature set	Average Classification Accuracy
Feature set a (SK+HS)	84.77%
Feature set b (EOD)	76.63%
Feature set c (EODPCA)	67.74%
Feature set d (SK+HS+EODPCA)	87.35%

TABLE II

PERFORMANCE COMPARISON ON PAIRS OF SIMILAR GESTURES WITH OTHER APPROACHES (THE HMM, HCRF, AND LINKED HCRF) PRESENTED BY SONG ET AL. [24].

Classifier	Average Classification Accuracy
HMM	77.67%
HCRF	78.0%
Linked HCRF	87.0%
Random Forest (our)	88.1%

extracted features are angular joint velocities for the right and left elbows and wrists, as well as probability values of hand shapes for the left and right hands. Since each gesture instance is described by a single feature descriptor obtained by concatenating 10 representative feature vectors, the feature vector representing a gesture instance is of length 200.

(b) Appearance-based feature set (EOD): Each frame of the gesture instances is represented by a 400 dimensional feature vector, which was calculated using randomly pooled edge-orientation and edge-density features. Each gesture example is represented by a single-dimension feature vector of length 4000.

(c) EODPCA: In this feature representation, we reduced the above 4,000-d feature space into a 200-d feature space via Principal Component Analyis (PCA).

(d) SK+HS+EODPCA: This feature set was obtained by concatenating the 200-d 3D skeletal joints and hand-shape based (SK+HS) feature descriptor of a gesture with the corresponding dimensionality-reduced edge orientation and density (EOD-PCA) feature descriptor to form a 400-d feature vector for every gesture.

For each feature set described above, we trained random forests with 500 trees on 19 subjects and tested on the remaining subject in a leave-one-out cross-validation approach.

We computed the average recognition accuracy (averaged across all subjects and all gestures) of the random forest classifier on the four different feature sets (a) - (d) of the NATOPS dataset for all 20 test subjects each performing the 24 gestures in the vocabulary (Table I). The feature set containing 3D skeletal joints and hand-shape features (SK+HS) is correctly classified 84.77% of the time, whereas the feature set containing features based on edge density and orientation is correctly classified 76.63% of the time. This suggests, in our case, that 3D joint-based based features encode more class-discerning information than features based on edge density and orientation. However, the highest classification accuracy of 87.35% is achieved on the feature set that combines joint-based features with appearance-based features, suggesting the benefit of combining the two approaches of collecting features.

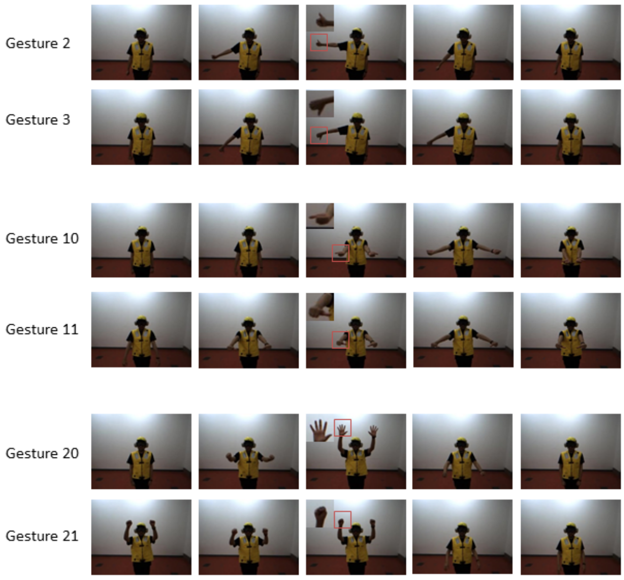
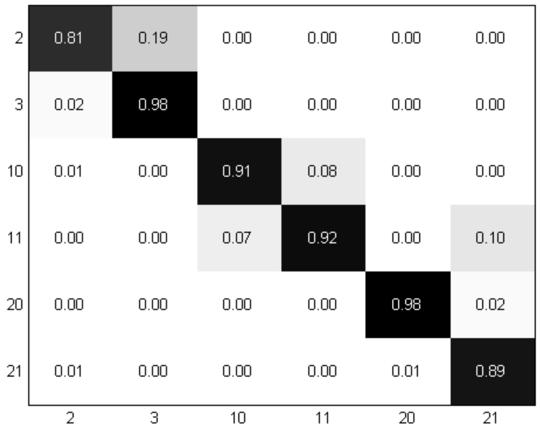


Fig. 4. Some pairs of similar gestures in the NATOPS dataset



	2	3	10	11	20	21
2	0.81	0.19	0.00	0.00	0.00	0.00
3	0.02	0.98	0.00	0.00	0.00	0.00
10	0.01	0.00	0.91	0.08	0.00	0.00
11	0.00	0.00	0.07	0.92	0.00	0.10
20	0.00	0.00	0.00	0.00	0.98	0.02
21	0.01	0.00	0.00	0.00	0.01	0.89

Fig. 5. Confusion Matrix for pairs of similar gestures in the NATOPS dataset

Gesture pairs (2,3), (10, 11) and (20, 21) were confused, often getting misclassified as the other (Figure 4). Figure 5 uses a confusion matrix to illustrate the misclassifications between these pairs of similar gestures.

We compared the classification performance of our random forest classifier with the performance of other classifiers that have been used on this dataset (Table II). Our random forest approach on the challenging subset of similar gestures, tested on samples from 5 subjects as specified by Song et al. [24], yields results that exceeds those produced by the state-of-the-art (Linked HCRF) (Table II). The graphical models presented by Song et al. [24] were trained using feature set a (SK+HS), whereas we use feature set d (SK+HS+EODPCA) to train our gesture recognition model.

From the ChaLearn dataset, we trained our gesture recognition model with the following feature sets:

a Raw 3D skeletal joint data (RAW): Features contain unedited raw skeleton data, that is, each frame consists of 9 values for all 20 joints. The feature vector per frame has 180 dimensions, and per gesture has 1800 dimensions.

b Normalized skeletal joint positions and velocities (SKPV): This feature set contains normalized positional and velocity data for 9 joints. The feature vector per frame has 126 dimensions, and per gesture has 1260 dimensions.

c Normalized skeletal joint positions, velocities and accelerations (SKPVA): This feature set contains positional, velocity, and acceleration data for 9 joints. The feature vector per frame has 189 dimensions, and per gesture has 1890 dimensions.

d Appearance-based features (HOG): This feature set contains Histogram of Oriented Gradients (HOG) data for 32x32 pixel boxes around 9 joints (head, left shoulder, left elbow, left wrist, left hand, right shoulder, right elbow, right wrist, right hand). The feature vector per frame has 2916 dimensions, and per gesture has 29,160 dimensions.

e SK+HOGPCA: This feature set was obtained by concatenating the 1260-d feature vector of normalized skeletal joint positions and velocities (SK) with the 400-d feature vector of HOG data for 32x32 pixel squares around the left and right hands whose dimensionality has been reduced by PCA. The resultant feature vector per gesture example is 1660-d.

For each feature set described above, we trained random forests with 500 trees on gesture instances from the training and validation sets, and tested the performance of our classifier on the test dataset. The division of the data into training, validation and test sets has been described earlier [23].

The feature set that combines the normalized positional and velocity information (SKPV), with HOG features of the hands (HOGPCA), is correctly classified correctly 88.91% of the time (Table III), which is the highest average classification accuracy of all feature sets.

The iterative procedure of training a random forest improves its capacity to correctly classify and segment gestures. This is evident in the increase in average classification accuracy in the test set (Figure 6).

TABLE III

AVERAGE CLASSIFICATION ACCURACY ON ALL 20 GESTURES OF THE CHALEARN DATASET

Feature set	Average Classification Accuracy
Feature set a (RAW)	81.45%
Feature set b (SKPV)	88.12%
Feature set c (SKPVA)	83.50%
Feature set d (HOG)	54.65%
Feature set e (SK+HOGPCA)	88.91%

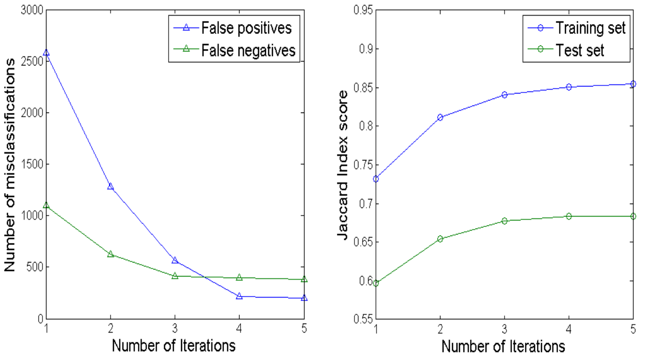


Fig. 6. Plot of number of misclassifications and Jaccard index score with number of iterations of training classifier

TABLE IV

JACCARD INDEX SCORES ON CHALEARN GESTURE RECOGNITION CHALLENGE 2014 [23]

Method	Jaccard Index Segmentation and Classification Score
Deep Neural Network [25]	0.84
Our Score	0.68
Competition Baseline [23]	0.37

Table IV shows the Jaccard score of our method compared with the baseline and winning scores of the ChaLearn gesture recognition challenge. The competition winner, a team from Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS), used features extracted from skeleton joints and a deep neural network classifier to achieve a Jaccard score of 0.84 [25]. Our classifier achieves a good recognition accuracy of 88.91% on pre-segmented gestures. Our Jaccard score of 0.68 underlines the difficulty of achieving optimal results in classification tasks where temporal segmentation is not provided.

VI. CONCLUSION

We have presented a random forest framework for a multi-gesture classification problem. The method consists of first creating a uniform fixed-dimensional feature representation of all gesture samples, and then using all training samples to train a random forest. On a challenging subset of the NATOPS dataset, our approach yields results comparable to those produced by graphical models such as HCRFs. Although a random forest classifier does not explicitly model the inherent temporal nature of gestural data as done by graphical models, its performance in accuracy on this particular dataset exceeds that achieved by graphical models such as HMMs, and different variants of HCRFs, which are presented by Song et al. [24]. Additionally our experiments also show that classification accuracy was improved by combining 3D skeletal joint-based features with appearance-based features, thus underlying the importance of a well-chosen feature set for a classification task.

On the ChaLearn dataset, our classifier yields an average accuracy of 88.91% when tested on a set of segmented gestures. However, the task of simultaneously detecting and classifying gestures is a more difficult challenge than classifying accurately segmented gestures.

The strengths of our framework lie in its simplicity, speed, its capacity to generalize well to variations in user size, distance to the sensor, speeds at which gestures are performed, as well as its robustness to the effects of sensor noise. One area of the framework that can be improved is the process of selecting and creating better feature sets. Many additional features, such as joint-pair distances used by Yao et al. [26], can be experimented with in order to improve the accuracy of our framework. Additionally, selecting a small group of features over an interval of frames to split a node in a decision tree, instead of selecting a single feature at a single frame, might be better suited to the purpose of learning complex spatio-temporal objects such as gestures.

Acknowledgments

The authors would like to thank the reviewers for their constructive feedback. This work was supported in part by National Science Foundation grants IIS-0910908, MRI-1337866 and US Navy contract N00014-13-P-1152.

References

[1] J. Alon, V. Athitsos, and S. Sclaroff, "Accurate and efficient gesture spotting via pruning and subgesture reasoning," in *Proceedings of the 2005 IEEE International Conference on Computer Vision (ICCV 2005)*, 2005, pp. 189–198.

[2] H. Junker, O. Amft, P. Lukowicz, and G. Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognition*, vol. 41, no. 6, pp. 2010–2024, 2008.

[3] M. Betke, O. Gusyatin, and M. Urinson, "Symbol design: A user-centered method to design pen-based interfaces and extend the functionality of pointer input devices," *Universal Access in the Information Society*, vol. 4, no. 3, pp. 223–236, 2006.

[4] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "Simultaneous localization and recognition of dynamic hand gestures," in *Proceedings of the 2005 IEEE Motion Workshop on Application of Computer Vision (WACV/MOTIONS 2005)*, vol. 2. IEEE, 2005, pp. 254–260.

[5] H.-K. Lee and J.-H. Kim, "An HMM-based threshold model approach for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 1999)*, vol. 21, no. 10, pp. 961–973, 1999.

[6] 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.

[7] F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using random forests," in *Proceedings of the 2008 British Machine Vision Conference (BMVC 2008)*, 2008, p. 10pp.

[8] A. Bosch, A. Zisserman, and X. Muoz, "Image classification using random forests and ferns," in *Proceedings of the 2007 IEEE International Conference on Computer Vision (ICCV 2007)*. IEEE, 2007, pp. 1–8.

[9] Y. Song, D. Demirdjian, and R. Davis, "Multi-signal gesture recognition using temporal smoothing Hidden Conditional Random Fields," in *Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 388–393.

[10] Y. Song, D. Demirdjian, and R. Davis, "Continuous body and hand gesture recognition for natural human-computer interaction," *ACM Transactions on Interactive Intelligent Systems (TIIS 2012)*, vol. 2, no. 1, p. 5, 2012.

[11] T. Starner and A. Pentland, "Real-time American Sign Language recognition from video using Hidden Markov Models," in *Motion-Based Recognition*. Springer, 1997, pp. 227–243.

[12] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Proceedings of the 2007 IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311–324, 2007.

[13] S. Malassiotis, N. Aifanti, and M. G. Strintzis, "A gesture recognition system using 3D data," in *Proceedings of the 2002 IEEE International Symposium on 3D Data Processing, Visualization and Transmission*. IEEE, 2002, pp. 190–193.

[14] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data," *Proceedings of the 2001 International Conference on Machine Learning (ICML 2001)*, pp. 282–289, 2001.

[15] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden Conditional Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 2007)*, vol. 29, no. 10, pp. 1848–1852, 2007.

[16] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, A. W. Vieira, and M. F. Campos, "Real-time gesture recognition from depth data through key poses learning and decision forests," in *Proceedings of the 2012 IEEE SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI 2012)*. IEEE, 2012, pp. 268–275.

[17] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 2011)*, vol. 33, no. 11, pp. 2188–2202, 2011.

[18] D. Demirdjian and C. Varri, "Recognizing events with temporal random forests," in *Proceedings of the 2009 ACM International Conference on Multimodal Interfaces (ICML 2009)*. ACM, 2009, pp. 293–296.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 1. IEEE, 2005, pp. 886–893.

[21] J. Marin, D. Vázquez, A. M. López, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV 2013)*. IEEE, 2013, pp. 2592–2599.

[22] Y. Song, D. Demirdjian, and R. Davis, "Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database," in *Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 500–506.

[23] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon, "ChaLearn looking at people challenge 2014: Dataset and results," in *Proceedings of the 2014 IEEE European Conference on Computer Vision (ECCV 2014) ChaLearn Workshop on Looking at People*. IEEE, 2014.

[24] Y. Song, L. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*. IEEE, 2012, pp. 2120–2127.

[25] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Proceedings of the 2014 IEEE European Conference on Computer Vision (ECCV 2014) ChaLearn Workshop on Looking at People*, 2014.

[26] A. Yao, J. Gall, G. Fanelli, and L. J. Van Gool, "Does human action recognition benefit from pose estimation?." in *Proceedings of the 2011 British Machine Vision Conference (BMVC 2011)*, vol. 3, 2011, p. 6.