Gesture Control Using Active Difference Signatures and Sparse Learning

Andreas Savakis[1], Raviteja Rudra[2], Raymond Ptucha[1]
[1]Computer Engineering, [2]Electrical Engineering,
Rochester Institute of Technology,
Rochester, New York, USA

***Abstract*—With the introduction of low cost depth sensors, gesture recognition systems for computer interface control are becoming a reality and require accurate recognition in real time. In this paper, we introduce a salient feature for gesture recognition called active difference signature, obtained by robust processing of depth maps and kinematic joint information. This feature is classified with variants of semi-supervised Linear extension of Graph Embedding and sparse representations to demonstrate advanced gesture control using depth cameras. Our active difference signature technique delivers highly accurate gesture recognition that is invariant to subject distance to camera, subject size, subject location within the camera field of view, and speed at which the subject performs the gestures.**

Keywords— gesture recognition; depth maps; sparse representation; human computer interaction

I. INTRODUCTION

The introduction of low cost depth cameras along with advances in computer vision have spawned an exciting new era for Human Computer Interaction (HCI). Pervasive or ambient intelligence is based on the seamless diffusion of sensors into the environment, thereby shifting computer interfaces from the primarily purposeful foreground (mouse/keyboard) to the ambient background. This new paradigm relies on touchless interfaces as a primary input modality. As such, new gesture control algorithms and architectures are of critical importance.

Depth sensors provide more salient information than RGB cameras for gesture recognition, as depth facilitates the extraction of objects against complex backgrounds and simplifies the tracking of objects. Shotten et al. [1] have shown how Kinect depth images are segmented into body clouds, then converted to body parts, and finally to skeletal joints in real time. These depth cameras are capable of video resolution frame rates, and have given the gesture recognition community a revolutionary leap in controller-less capability [2].

There has been much research on improving the overall HCI experience [3-5]. Lew [5] argues that in order to achieve effective human to computer communication, the computer needs to interact with the human. The recognized gestures can be used to direct interactive large-scale displays for a satisfying user experience [6].

The ChaLearn [7] challenges have facilitated algorithm benchmarking and expedited algorithm development in the gesture recognition community. Recognizing gestures is often accomplished by feeding tracked skeletal joints into Hidden Markov Models (HMMs) to accurately model sequences of complex gestures [8, 9]. Spatial action representations, such as body models [10], body pose estimations [11], kinematic joint models [12], and stick figures [13] offer intuitive representations but may not adequately capture the human body's high degree of variability. Spatial parametric image features such as contour/silhouette representations [14], optical flow [15], and motion history images [16] don't require body part labeling or tracking, but are more computationally intensive.

The notion of Sparse Representations (SRs), or finding sparse solutions to underdetermined systems, has found applications in a variety of scientific fields. The resulting sparse models are similar in nature to the network of neurons in V1, the first layer of the visual cortex in the human, and more generally, the mammalian brain [17]. SR systems are comprised of an input sample $y \in \mathbf{R}^D$ along with an overcomplete dictionary Φ of m samples, $\Phi \in \mathbf{R}^{Dxm}$. SR solves coefficients $a \in \mathbf{R}^m$ that satisfy the ℓ^1 minimization of $|a|_1$ s.t. $\hat{y} = \Phi a$. It has been shown that under typical conditions, the minimal solution is the sparsest one [18]. There have been several studies combining both ℓ^1 minimization and the selection of dictionary elements [19].

Although designed for reconstruction purposes, the SR framework has been successfully adapted for classification problems. Wright et al. [20] passed the a coefficients directly into a minimum reconstruction error classifier for facial recognition. In this framework, the dominant signal always prevails, but it could produce some unintended effects. For example, when trying to extract facial identity, pose variation may contaminate or even dominate the sparse coefficients. Ptucha et al. [21] addressed the coefficient contamination problem in the context of expression recognition by preprocessing the data with supervised manifold learning.

Distinct advantages afforded by SR architectures include: 1) The class and numeric value of each non-zero coefficient $a \in \mathbf{R}^m$ can be used as salient input to a classifier; 2) Sparse systems accommodate both large and small training dictionaries; 3) The addition of new training samples does not require retraining, as required by many popular classification methods.

In this paper, we introduce active difference signatures as means to select temporal regions of interest based on both the depth map and the estimated kinematic joint positions [1]. The

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.680

skeletal joints are normalized to make the method invariant to body size, distance from camera, and location within the frame. The difference between the normalized joints and a canonical representation forms an active difference signature, a salient feature descriptor across the video sequence. This descriptor is dynamically warped to a fixed temporal duration. Semi-supervised Linear extension of Graph Embedding (LGE) manifold learning is used to convert active difference signatures to a low dimensional object. This low dimensional space is more computationally efficient and discriminative. Sparse coefficients, obtained by ℓ^1 minimization of this low dimensional object, are fed into a minimum reconstruction error engine to achieve state-of-the-art gesture recognition results. We contrast our technique to other popular techniques on the Microsoft Action 3D (MSR3D) Dataset [22].

The rest of this paper is organized as follows. After the introduction, Section 2 overviews manifold learning, sparse representation, and temporal technologies used in our model. Section 3 introduces our temporally invariant difference signature framework. Section 4 presents the experimental results, and Section 5 contains concluding remarks.

II. BACKGROUND

A. Manifold Learning

Manifold learning techniques reduce the dimensionality of input data by identifying a non-linear lower dimensional space where the data resides [23]. In order to support the extension of the manifold model to new examples, LGE linearized techniques, solve a linear approximation of the non-linear object [24].

The input feature space contains n samples, $x_1, x_2, \ldots, x_n$, with $x_i \in \mathbf{R}^D$. These n samples are projected onto a lower dimensional space, yielding $y_1, y_2, \ldots, y_n$, with $y_i \in \mathbf{R}^d$. LGE solves a $d \times D$ projection matrix U, such that $y_i = Ux_i$, and $d \ll D$. LGE creates an adjacency mapping of the top k neighbors for each feature point x_i by weighting each neighbor by distance to form a $n \times n$ adjacency matrix W with entries w_{ij}. W is defined similarly for X and Y, such that if neighbors x_i and x_j are close, y_i and y_j are also close to each other. LGE computes eigenvectors of the generalized eigenvector problem:

$$XLX^TU = \lambda XDX^TU \tag{1}$$

where D is a diagonal matrix of the column sums of W, L is the Laplacian matrix, $L=D-W$, and U is the projection matrix.

There are several strategies to set connection weights w_{ij} of W. Different choices of W yield a multitude of dimensionality reduction techniques. Locality Preserving Projections (LPP) [25] uses Gaussian kernel weighting. If nodes i and j are connected, LPP sets:

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \tag{2}$$

When supervised labels are available, a discriminative embedding can be achieved via a supervised kernel. Formally, W is initialized to all zeros, and then w_{ij} entries corresponding to the same classes are set to $1/k$, where k is the number of samples per class.

To achieve a discriminative embedding while maintaining the input topology, we utilize a convex combination of the Gaussian and supervised kernels:

$$W = \alpha W_{Supervised} + (1-\alpha)W_{Gaussian}, \quad 0 \le \alpha \le 1 \tag{3}$$

B. Sparse Representation Classification

Motivated by evidence of sparsity in the human brain and the success of SRs in recognition problems [20], we utilize a Sparse Representation Classification (SRC) engine to process our dimensionality reduced data. In the SRs framework, a test sample is represented as a sparse linear combination of exemplars from a training dictionary, Φ. Let the input signal be $y \in \mathbf{R}^d$ and the dictionary be $\Phi \in \mathbf{R}^{dxn}$. Finding the sparsest solution $\hat{y} = \Phi a$ in the presence of noise is called Basis Pursuit Denoising (BPDN):

$$\hat{a} = \text{argmin}\|a\|_1 \quad s.t. \ \|y - \Phi a\|_2 \le \varepsilon \tag{4}$$

Often (4) is approximated by loosening the error constraints and reconfigured to specifically include a regularization term, λ which encourages sparseness by incurring a penalty on the resulting coefficients:

$$\hat{a} = \text{argmin}\{\|y - \Phi a\|_2^2 + \lambda\|a\|_1\} \tag{5}$$

Given the sparse representation coefficients $\hat{a}$ of a test sample, a minimum reconstruction error estimates the class c^* of our test sample by comparing the error between all coefficients $\hat{a}$ with the coefficients a^c corresponding to each class c, one class at a time:

$$c^* = \text{argmin}\|y - \Phi a^c\|_2 \quad c \in 1 \ldots z \tag{6}$$

C. Temporal Representations

Gestures can occur at any point in time and are variable in length. For streaming video, we define sliding temporal windows W_l^θ of duration θ, where θ is the number of frames in a gesture sequence, and l is a window identifier. Each of these temporal windows can be used as input to a gesture classifier.

Motion History Images (MHI) were initially introduced as descriptors for human movement recognition [16]. MHI describes the motion over each sliding temporal window W_l^θ into a single frame called a MHI template. MHI evaluates the movement between all possible frames f and $f+1$ in W_l^θ, where $f=1,\ldots,\theta-1$. For each pair of frames $\{f, f+1\}$ in W_l^θ, we first calculate motion energy at the pixel level in a binary fashion:

$$d_f = \begin{cases} 1 & if \ |g(x,y,f) - g(x,y,f+1)| > \gamma \\ 0 & otherwise \end{cases} \tag{7}$$

where $g(x,y,f)$ is a Gaussian filtered version of frame f and γ is a noise threshold. These energy difference frames are morphologically filtered with an opening operation to remove isolated noise. Each sliding window produces a single MHI_l^θ template such that more recent movements are assigned higher weights:

$$MHI_l^\theta = \frac{1}{\theta-1}\max_f\{f d_f(x,y), \quad 0 \le f \le \theta - 1\} \tag{8}$$

III. Active Difference Signatures

Gesture recognition is a challenging problem due to the wide variation in performing gestures both in terms of manner and time duration. The automatic detection of gesture onset and offset is valuable but can be difficult. The active difference framework is developed to effectively deal with variations in gestures performed by the same person or across individuals.

Fig. 1 outlines the gesture extraction framework used in this work. The depth map is processed to find gesture start and end boundaries; skeleton joints for all frames within each gesture region are normalized; difference signatures are formed by comparing normalized skeletal joints to reference skeleton joints; the difference signature is dynamically time warped; and manifold learning along with Sparse Representation Classification (SRC) are used for gesture classification.

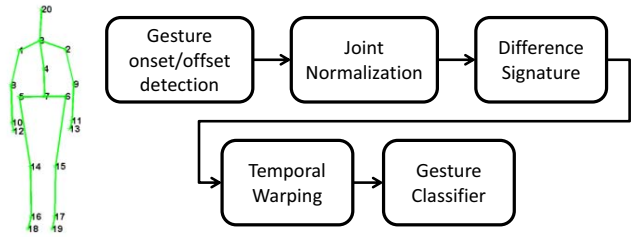

Fig. 1. Overview of the active difference signature framework.

Gesture boundary detection finds frames which indicate the onset and offset of a gesture. For example, Fig. 2 shows a time sequence of a user executing four gestures in a single video from the ChaLearn Gesture Challenge dataset [7]. Motion detection along with a measure of the difference between each frame and the resting position are good markers for gesture boundaries.

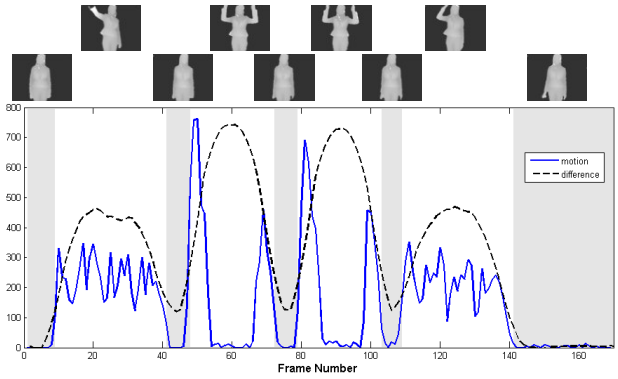

Fig. 2. Multi-gesture video sequence with four active gesture areas separated by five non-gesture regions indicated in gray areas. The solid blue curve is a frame to frame difference signature; the black dotted line is a frame-canonical depth image signature. The images at top show the depth frame at the center of the non-gesture and gesture regions.

The solid blue line in Fig. 2 is an indicator of frame to frame motion. The dashed black line records the difference of the current frame to a resting frame. The gray regions pictorially illustrate the non-gesture regions, or the resting regions. The white regions are active regions where a gesture is being performed. The formation of the blue motion and black difference curves is done via a variation of MHI. Specifically, depth frames are first resampled down to 60×80 pixels, and then converted to a difference frame using (7). The motion depth image signature (blue line in Fig. 2) is:

$$\upsilon_m = \sum d_f(x,y) \tag{9}$$

Higher values of υ_m indicate more motion from one frame to the next. Each time υ_m crosses $\eta \times max(\upsilon_m)$, $0<\eta<1$, this indicates the possible beginning or end of a gesture.

A canonical depth image, or canonical resting frame c_d, is formulated by averaging all start and end frames of a video sequence in which no motion is detected. Using c_d, a difference indicator is calculated using (7), replacing $g(x,y,f+1)$ with c_d:

$$d_{ref} = \begin{cases} 1 & if\ |g(x,y,f) - c_d| > \gamma \\ 0 & otherwise \end{cases} \tag{10}$$

The black dashed frame-canonical depth image signature in Fig. 2 is calculated as:

$$\upsilon_{ref} = \sum d_{ref}(x,y) \tag{11}$$

Higher values of υ_{ref} indicate more difference from the canonical resting frame. Each time υ_{ref} crosses $\kappa \times max(\upsilon_{ref})$, $0<\kappa<1$, it indicates the possible beginning or end of a gesture. Given υ_m and υ_{ref}, active regions where gestures are being performed are defined by $\upsilon_m > \eta max(\upsilon_m)$ OR $\upsilon_{ref} > \kappa max(\upsilon_{ref})$. Similarly, resting regions, or inter-gesture regions are defined by $\upsilon_m < \eta max(\upsilon_m)$ AND $\upsilon_{ref} < \kappa max(\upsilon_{ref})$.

The 20 XYZ skeletal joint coordinates in each frame of each active region are normalized akin to a Procrustes analysis in preparation for subsequent processing. This normalization makes the technique invariant to subject distance from the camera, subject size, and subject location within the frame. Setting s equal to a vector of the frame's XYZ skeleton joints, c equal to a vector of XYZ canonical skeleton joints (see Fig. 1), and n equal to the number of frames in the active region:

$$s' = s - \frac{1}{n}\sum s_i \qquad i \in 1 \ldots 20 \tag{12}$$

$$s'' = s'\left(\frac{size(c)}{size(s')}\right) \tag{13}$$

$$s''' = s'' + \frac{1}{n}\left(\sum_i c_i - \sum_i s_i''\right) \quad i \in [3,4,7,20] \tag{14}$$

Equation (12) shifts the skeleton joints such that the centroid is at the origin. The size of a skeleton is the joint-to-joint geodesic distance (i.e. the 3D length of segments connecting the joints in Fig. 1). After scaling in (13), the skeleton is shifted to the canonical skeleton location using a centroid calculation of only the head and spine joints (joints numbed 3, 4, 7, and 20). Omitting arm and leg joints enables the body mass to remain stationary even if a subject's arm or leg is fully extended.

After joint normalization, the active difference signature attribute is formed by differencing the 20 normalized skeleton joints s''' of each frame with the 20 canonical skeleton joint locations. Each frame in the active region yields a 20×3 feature $x \in \mathbf{R}^{60}$. All frames between gesture boundaries are

dynamically time warped to form a standard window of 25 frames. This is done separately for the X, Y, and Z directions by forming an *image* of size $\theta \times 20$, where $\theta=25$ is the number of frames in the active region and 20 is the number of skeletal joints. This procedure yields 25 frames with 20 skeletal joints per gesture, where the combination of the X, Y, and Z resampled *images* form our temporal joint attribute feature $x \in \mathbf{R}^{1500}$.

Not only does the dynamic time warping convert each active sequence to a fixed number of dimensions in preparation for classification, but it also removes high frequency noise between frames. Fig. 3 shows a sample video from the MSR [22] dataset with the corresponding difference signature in the middle and an active difference signature on the bottom. Each of the 20 lines represents the temporal movement of each of the 20 skeletal joints across the dynamically warped 25 frame timeline. The gesture in Fig. 3 did not start until after frame 10, and thus the middle signature does not generalize well with different start and end times. The dynamically warped active difference signatures at the bottom of Fig. 3 are invariant to the speed at which the subject is performing gestures.

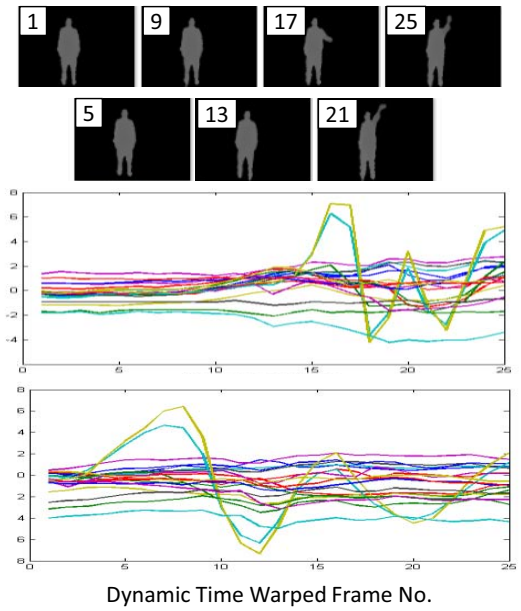

Fig. 3. Comparison of an difference signature (middle) vs. active difference signature (bottom) for the waving gesture from sample 1 of the MSR3D dataset (top). Each of the colored lines in the two figures shows the temporal displacement of one of the 20 skeletal joints from the canonical skeletal frame. Kinematic joints 11 and 13 (yellow and cyan repsectively) showed the most displacement from the canonical skeleton.

Another approach to representing the canonical resting frame is to consider a family of reference frames. After joint normalization, a family of active difference signature attributes are formed by differencing the 20 normalized skeleton joints s''' of each frame with the 20 skeleton joint locations of each reference frame. To select reference frames, K-SVD [26] dictionary learning can learn the top k dictionaries or k-means clustering with k set to the number of reference frames is repeated many times on the dataset using random initialization. The k-means class centers, or reference frames are chosen with the minimum cost:

$$J = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \mu c_i\|^2 \quad (15)$$

where μc_i is the cluster centroid to which sample x_i has been assigned.

Classification can be done using methods such as k-NN, artificial neural nets, or support vector machines (SVM). LGE manifold learning in conjunction with sparse representations offers the highest accuracy and provides moderate robustness to outliers and partial occlusions. In particular, we use equation (3) to solve for W, and LPP to generate dimensionality reduction matrix U using (1), to solve for the low dimensional sample, $y = Ux$. Sparse Representation techniques convert y into a gesture estimate by solving the ℓ^1 minimization of low dimensional sample y, using (5), and making a class estimate on sparse coefficients a using a minimum reconstruction error (6).

IV. Results

A. Dataset

The Microsoft Action 3D (MSR3D) dataset proposed by Li et al. [22] contains both depth maps and corresponding skeletal joint locations. It consists of depth map sequences with a resolution of 320×240 pixels recorded with a depth sensor at 15 FPS. There are ten subjects performing twenty actions two to three times for a total of 567 depth map sequences. The dataset actions are: high arm wave, horizontal arm wave, hammer, catch, tennis swing, forward punch, high throw, draw X, draw tick, tennis serve, draw circle, hand clap, two hand wave, side boxing, golf swing, side boxing bend, forward kick, side kick, jogging, and pick up and throw. No corresponding RGB information is available, however 3D kinematic joint positions are provided for each frame.

B. Experimental Methodologies

A leave-one-subject out cross-validation methodology was used to separate the MSR3D dataset into separate training and testing sets. Each test subject is validated against the remaining nine subjects and the process is repeated until all subjects have been used for training and testing. The results from each subject are averaged to give a final performance result. The dimensionality reduction techniques capture 99% of the data variance. The LPP method uses $\alpha=0.5$ in creation of W using (3). Sparse coefficients from test samples are generated using (5), setting $\lambda=0.15$. The low dimensional projection of all training samples in the cross-validation training split forms the training dictionary in the SR techniques. The corresponding sparse coefficients of test samples use (6) to make a final classification estimate.

C. Experimental Results

Table I shows the classification results of our method against other state-of-the-art gesture recognition techniques. The first three techniques are existing temporal techniques adopted for gesture recognition. Techniques '4' and '5' were published results on the MSR3D dataset, and technique '6' is the active difference signatures method proposed in this paper. SIFT flow [27] is an image alignment algorithm introduced to

register two similar images. Optical Flow of Skeletal Joints tracks the skeletal joints frame by frame, forming the difference between each joint coordinate and a canonical skeletal coordinate. Bag of Features [22] uses action graphs to model the dynamics of the actions and a bag of features to encode the action. Spatio-Temporal Joint Descriptor [28] encodes the difference between each skeletal joint and the centroid of the skeleton, and then uses dynamic time warping to generate gesture attributes.

TABLE I. CLASSIFICATION ACCURACY ON THE MSR3D DATASET FOR VARIOUS GESTURE RECOGNITION TECHNIQUES.

Method	*Classifier*	*% Accuracy*
MHI [29]	SRC	62.1
SIFT flow [27]	SRC	40.8
Optical Flow of Skeletal Joints	SVM	40.9
Bag of Features [22]	NERF	74.7
Spatio-Temporal Joint Desc. [28]	SRC	73.3
Active Difference Signatures	SRC	**82.5**

The results in Table I show the significant advantage of active difference signatures on final classification rates. In particular, by concentrating only on active regions, the attributes passed into the classifier are more discriminative. For example, the last two methods in Table I both used sparse representation classifiers, but only the latter used active regions. The MHI, SIFT flow, and Bag of Features methods used depth pixels as the primary feature, while the Optical Flow, Joint Descriptor, and Active Difference Signature methods used the 3D skeletal joint coordinates as the primary feature. It should be noted that the results for the Bag of Features method used half the subjects for training, the other half for testing, which is not directly comparable to the leave-one-subject-out cross validation used by all other methods. Nonetheless, we include this method as it introduced the MSR3D dataset. Under the classifier column, SRC is the Sparse Representation Classifier (6), and NERF is a fuzzy spectral clustering method that classified the test sample according to the training sample with the minimum Hausdorff distance.

To examine the value of using a family of reference frames: 1) K-SVD was used for dictionary selection and the top 12 dictionary elements were used as reference frames; 2) *k*-means clustering with k=12 was repeated 100 times to pick the class centers that minimized (15) (depicted in Fig. 4); and 3) To select the optimal 12 reference frames, we manually inspected all 8,136 frames of the MSR dataset and selected 12 salient body positions (e.g.: arms straight up, left arm straight out, etc.). For each salient body position, ten exemplars from different subjects were averaged to form each of the 12 reference frames.

Table II compares the active difference signature method using a single resting reference frame vs. a family of reference frames, and across standard SVM classification vs. sparse representation classification (SRC).

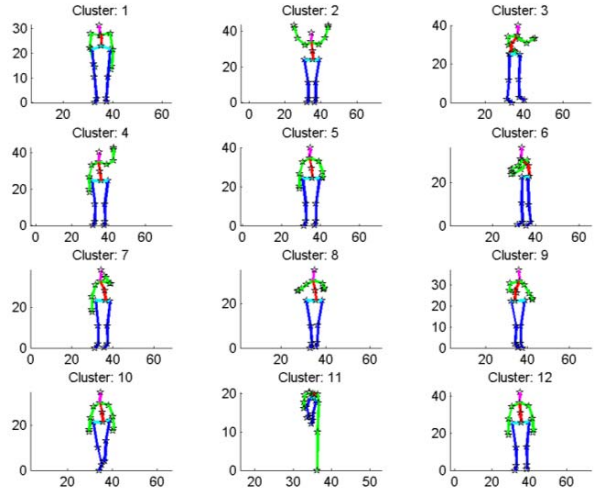

Fig. 4. Twelve cluster centers as selected by *k*-means.

TABLE II. CLASSIFICATION ACCURACY ON THE MSR3D DATASET FOR VARIOUS GESTURE RECOGNITION TECHNIQUES

Method	*SVM*	*SRC*
Single Canonical Resting Frame	79.3	82.5
Family of Reference Frames-KSVD	79.1	81.4
Family of Reference Frames: *k*-means	79.8	82.7
Family of Reference Frames: manual	79.1	**82.9**

SRC always outperforms SVM, and the addition of a family of resting frames does not necessarily increase classification accuracy over a single resting frame. The manual selection of the 12 salient body positions is quite an arduous task, and was only marginally better than the other methods.

V. CONCLUSIONS

This paper presents a new gesture recognition method which introduces active difference signatures, a novel salient descriptor for gesture control. The active difference signature attribute is dynamically time warped and converted to a gesture estimate using manifold based sparse representations to achieve state-of-the-art gesture estimation for HCI systems. We utilize information from depth maps to segment out active regions from video. The kinematic body joints for each frame within the active regions are normalized and then differenced from a canonical skeletal representation to obtain a difference signature. When these difference signatures are dynamically time warped across the active region, an active difference signature is formed. These signatures are invariant to subject speed of performing gestures, subject distance from the camera, subject size, and subject location within the frame. By utilizing semi-supervised LGE dimensionality reduction before sparse representation dictionary learning, we reduce compute overhead and make our gesture recognition system more robust to diverse test environments.

REFERENCES

[1] J. Shotton, *et al.*, "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *Computer Vision & Pattern Recognition*, 2011.

[2] E. Suma, B. Lange, A. Rizzo, D. Krum, and M. Bolas, "FAAST: The Flexible Action and Articulated Skeleton Toolkit," in *Virtual Reality*, 2011, pp. 247-248.

[3] S. Afzal, C. Morrison, and P. Robinson, "Intentional affect: an alternative notion of affective interaction with a machine," Proceedings of the 23rd British HCI Group Annual Conference on People and Computers, 2009.

[4] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, "Human computing and machine understanding of human behavior: a survey," in *Artifical Intelligence for Human Computing. ICMI 2006 and IJCAI 2007 International Workshops*.

[5] M. Lew, E. M. Bakker, N. Sebe, and T. S. Huang, "Human-computer intelligent interaction: A survey," in *4th IEEE International Workshop on Human - Computer Interaction,* 2007.

[6] B. Yoo, *et al.*, "3D user interface combining gaze and hand gestures for large-scale display," *28th international conference extended abstracts on Human factors in computing systems,* 2010.

[7] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hammer, and H. J. Escalente, "ChaLearn Gesture Challenge: Design and First Results," in *Computer Vision and Pattern Recognition Workshops*, 2012.

[8] R. Munozsalinas, R. Medinacarnicer, F. Madridcuevas, and A. Carmonapoyato, "Depth silhouettes for gesture recognition," *Pattern Recognition Letters,* vol. 29, pp. 319-329, 2008.

[9] K. Nickel and R. Stiefelhagen, "Pointing Gesture Recognition based on 3D-Tracking of Face , Hands and Head Orientation Categories and Subject Descriptors," pp. 140-146, 2003.

[10] H. Ghasemzadeh, V. Loseu, and R. Jafari, "Collaborative Signal Processing for Action Recognition in Body Sensor Networks: A Distributed Classification Algorihtm Using Motion Transcripts," in *9th ACM/IEEE Int. Conf. Inf. Process.*, 2010.

[11] K. Raja, I. Laptev, P. Perez, and L. Oisel, "Joint Pose Estimation and Action Recognition in Image Graphs," in *IEEE International Conference on Image Processing*, 2011.

[12] D. Weinland, E. Boyer, and R. Ronfard, "Action Recognition from Arbitrary Views Using 3D Exemplars," in *International Conference on Computer Vision*, 2007.

[13] S. Maji, L. Bourdev, and J. Malik, "Action Recognition from a Distributed Representation of Pose and Appearance," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[14] Y. Wang and Z. Zhang, "View Invariant Action Recognition in Surveillance Videos," in *Asian Conference on Pattern Recognition*, 2011.

[15] H. Imtiaz, U. Mahbub, and M. A. R. Ahad, "Action Recognition Algorithm Based on Optical Flow and RANSAC in Frequency Domains," in *SICE Annual Conference*, 2011.

[16] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 257-67, 2001.

[17] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Research,* vol. 37, pp. 3311-25, 1997.

[18] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory,* vol. 52, pp. 6-18, 2006.

[19] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *IEEE Conference on Computer Vision and Pattern Recognition,* 2011.

[20] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 210-27, 2009.

[21] R. Ptucha and A. Savakis, "Manifold Based Sparse Representation for Facial Understanding in Natural Images," *Image and Vision Computing,* vol. 31, pp. 365-378, 2013.

[22] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *IEEE conference on Computer Vision and Pattern Recognition*, 2010.

[23] A. Ghodsi, "Dimensionality Reduction A Short Tutorial," Univ. of Waterloo, 2006.

[24] C. Deng, H. Xiaofei, H. Yuxiao, H. Jiawei, and T. Huang, "Learning a spatially smooth subspace for face recognition," *IEEE Conference on Computer Vision and Pattern Recognition,* 2007.

[25] X. He and P. Niyogi, "Locality Preserving Projections," in *Advances in Neural Information Processing Systems*, 2003.

[26] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing,* vol. 54, pp. 4311-22, 2006.

[27] L. Ce, J. Yuen, and A. Torralba, "SIFT flow: dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, pp. 978-94, 2011.

[28] S. Azary and A. Savakis, "A Spatiotemporal Descriptor based on Radial Distances and 3D Joint Tracking for Action Classification," IEEE Int. Con. Image Processing, 2012.

[29] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 1940-54, 2010.