One Shot Learning Gesture Recognition from RGBD Images

Di Wu, Fan Zhu, Ling Shao
The University of Sheffield
Sheffield, UK
{elp10dw,fan.zhu,ling.shao}@sheffield.ac.uk

Abstract

We present a system to classify the gesture from only one learning example. The inputs are duo-modality, i.e. *RGB and depth sensor from Kinect. Our system performs morphological denoising on depth images and automatically segments the temporal boundaries. Features are extracted based on Extended-Motion-History-Image* (Extended-MHI) *and the Multi-view Spectral Embedding* (MSE) *algorithm is used to fuse duo modalities in a physically meaningful manner. Our approach achieves less than 0.3 in Levenshtein distance in CHALEARN Gesture Challenge validation batches [1].*

1. Introduction

In this paper, we focus on the CHALEARN Gesture Challenge [1]. There are some unique distinctions in this dataset from other action/gesture recognition datasets [17, 2]. We reinstate the major easy/difficult aspects of the dataset and present our analysis and reasoning to solve/circumvent the problems as follows:

1.Availability of depth camera: depth cameras significantly reduce the huge color and texture variability induced by clothing, hair and skin. However, some imperfection/noise of various sources still exists [15] in current depth sensors: *e.g.* reflectance and mismatched patterns. *c.f.* to Figure 1, strong existence of salt and pepper noise is detected as real motion information. A spatial filtering and a morphological preprocessing step are adopted for noise reduction in Section 2.1.1.

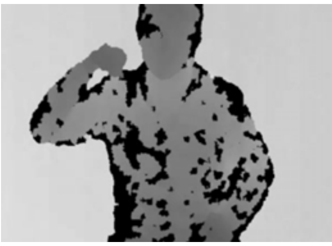

Figure 1. Noise in depth image

2.Multiple gestures in testing set: temporally unsegmented action sequences are real-world scenario. However, present action/gesture recognition datasets almost universally dodge this difficulty by providing training/testing sequences in a manually segmented manner. In the dataset of [1], however, the number of gestures contained in a testing video sequence varies from 1 to 5. Therefore, temporal segmentation is a precondition for gesture recognition. We argue that because of the unique property of this dataset, *i.e.* hands return to a resting position between each pair of neighboring gestures, the temporal segmentation as a preprocessing step is more effective than the action localization [16] approach. [24] presents a semi-supervised action recognition system that breaks down action sequences into primitive actions based on a motion history volume descriptor and automatically discovers the action taxonomies. Similarly as suggested by [1] that because hands return to a resting position between each pair of neighboring gestures, segmentation points occur near the peaks of hand motions in the lower part of an image. In our system, instead of using motion information for action segmentation, we adopt the appearance-based approach as in Section 2.1.2 and achieve 5% error in the metrics of Levenshtein distance for the verification of segmentation. Also note that the accuracy for our whole gesture recognition system is upper bounded by this temporal segmentation performance.

3.One-shot-learning: only one training example of each class is considered as the unique trait of this challenge whilst using more examples per sign typically improves accuracy (see, *e.g.* [9, 27]). The standard tools of statistical machine learning, *e.g.* classification and regression, have a chance to be equally matched to modeling purposeful behavior in a poor manner; an agent's goals often succinctly, but implicitly, encode a strategy that would require tremendous amounts of data to learn. Consequently, in the case of

978-1-4673-1612-5/12/$31.00 ©2012 IEEE

insufficient training data being available, complex models that have the demand of extensive parameters to learn are very likely to encounter with the over-fitting problem. To avoid such over-fitting problem in the one-shot-learning scenario, we experimentally find the Maximum Correlations Coefficient approach most suitable to be applied.

4.Depth & RGB camera decision fusion: how to effectively utilize multiple inputs to generate an informed decision is sometimes under-appreciated. Currently, the most commonly adopted approach when encountering different types/spectra of features is to concatenate multiple features into a long vector before the classification stage and feed this long feature vector into a classifier [12, 23, 20]. Mostly, for the sake of simplification, different view features are treated independently and have been ignored by their intrinsic relationships [13]. We argue that the interleaving relationship between different feature vectors is lost during this brute-concatenation process, and the interdependent relationship between different feature decisions could be better incorporated in an ensemble system. Moreover, the benefit of multi-spectrum video fusion always comes with a certain cost and complexity in the analysis process due to the fact that the involved modalities have different characteristics. On one hand, the more pronounced the independence between difference modalities, the more complementary information can be gleaned from each of them. On the other hand, there need to be a sufficient amount of correlations in order to be able to link features in one modality. We study the Multiview Spectral Embedding (MSE) in [26] and its derivative of spectral clustering. Then we present our discovery of the intrinsic property during the embedding process. With the brief theoretical analysis in Section 2.1.4, we demonstrate the effectiveness of the proposed approach by embedding information acquired from both depth and RGB cameras to further improve the recognition rate.

2. Experiments

In this section we detail our approach towards solving the general four issues in Section 1 and present both quantitative results and qualitative evaluations of our method on the CHALEARN dataset [1].

2.1. Methods

Error Metrics: We quantify our recognition rate by computing the Levenshtein distance between the list of predicted labels R and the corresponding list of true labels T, that is the minimum number of edit operations (substitution, insertion or deletion) that one has to perform to go from R to T (or vice versa). This error metrics measurement is also in accordance with the Leaderboard in [1] and we refer this error metrics as $\mathfrak{LD}$ from now on.

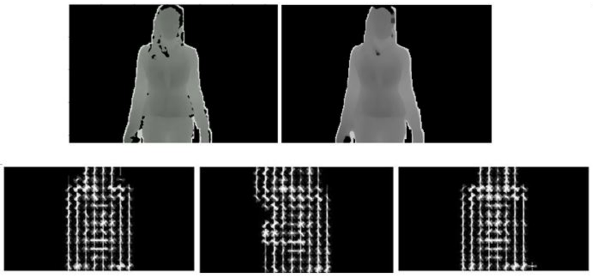

Figure 2. Top left: background segmentation; top right: depth image after noising; Bottom row: HOG descriptor for temporal segmentation. As it can be seen that the starting frame(bottom left) and the ending frame(bottom right) are quite similar to each other whereas in the midst of action (bottom middle), there is a substantially spatial difference.

2.1.1 Preprocessing: Background Separation and Noise Reduction for Depth Images

Taking advantage of the unique property of the depth sensor, from which human silhouettes can be easily segmented, we firstly segment human bodies from the background using Otsu's method of global image threshold [14] as shown in Figure 2 (top left). The resulting noise pattern in depth images resembles salt and pepper noise. We then use a spatial filtering and a morphological process for noise reduction. A median filter provides excellent salt and pepper noise reduction with considerably less blurring. As in [15], we adopt a 5×5 aperture median filter. Then, morphological process is used for further noise reduction. Specifically, we use opening operation which consists of erosion followed by dilation to smooth the outers, split the narrow region and remove the thine perimeter. Thus, the opening operation removes randomly generated noise and smooths the original image. The resulting depth image is shown in Figure 2 (top right). When the noise reduction method is applied to the motion image generated from the depth sensor, the resulting motion description is less prone to faulty defects from the depth sensor. These operations are highly effective for the depth image noise reduction especially if the action descriptor is motion-based as in our system. Experimental result shows that the noise reduction method can improve the performance in terms of $\mathfrak{LD}$ as much as 9%.

2.1.2 Temporal Segmentation

For temporal segmentation, we adopt the appearance based approach. Because hands return to a resting position between each pair of neighboring gestures, we aspire to find the frames that are similar to the beginning and ending frame in the unsegmented testing video sequence and define them as the interval frames between two gestures in a video sequence. A simple but effective option to retrieve similar frames is to divide a single frame into a $N \times N$ lattice and use the histogram of oriented gradients (HOG) [7] as the cell descriptor with B number of bins. Hence, a single frame can be represented as a feature vector of $N \times N \times B$ dimension. Then we use the k nearest neighbor approach to search for frames that are similar to the beginning and ending frame. Some implementation details worth mentioning here: first is how many similar frames should we search in this unsegmented video? Our solution is to first store the training example's average frame number L and when there is a test sequence, we make an rough estimation of gesture number as the quotient Q of test sequence frame number F and the average training tokens' frame number L. Then, the estimated frame number to retrieve is $N_r \times Q$. In our implementation, N_r is chosen as 8 by cross validation. After the similar frames being retrieved, a max pooling approach [18] is used to aggregate interval frames. We then merge minimal segmented sequences if the total action tokens segmented exceed the number of 5, which is the maximum gesture number for one test sequence in this dataset. Generally, the more lattices that one frame picture is divided into, the more accurate it is to segment action sequence. However, through our experiments, the varying of HOG grid size has little impact on temporal segmentation performance. Consequently, taking the computational cost into consideration, there is no significant point to set the size of the HOG grid into small values. In our experiment, bin number B is 9 and two lattice types were tested, *i.e.* 8×8 and 16×16. In the case of 8×8, $\mathfrak{LD}$ is 6.764% and for 16×16 is 5.235%. Note that as we mentioned in Section 2, accuracy for our whole gesture recognition system is upper bounded by this temporal segmentation performance.

2.1.3 Motion Descriptors and Scheme for Classifier

We experiment extensively on different motion descriptors and classifiers and via comparison we discuss our methodological insights. Our final adopted approach is *Extended-MHI* for action descriptor and *Maximum Correlation Coefficient* for classifier. The results are reported on the first 20 development batches unless otherwise we explicitly state on the validation dataset.

Cons for local method: Spatio-temporal features [8, 11] have shown success for many recognition tasks where preprocessing methods such as foreground segmentation and tracking are not possible. However, their computational complexity hinders their applicability in real-time applications. Wang et al. [23] showed that the average time for spatio-temporal feature extraction varies from 0.9 FPS to 4.6 FPS, which makes the STIP features too time-consuming in computation. Another major limitation of the local feature based methods is that the sparse representation such as bag-of-visual-words (BoVW) discards geometric relationship of the features and hence is less discriminative.

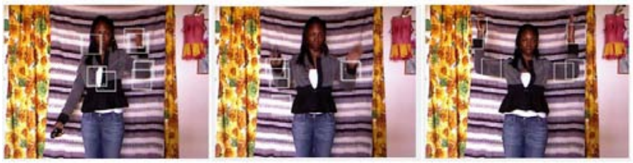

Figure 3. Spatial temporal interest points in white bounding box of three different gesture tokens.

We experiment on depth image using Dollar's method [8] for STIP detection, HOG3D [10] for cuboid descriptor, kernel codebook [21] for encoding and SVM [4, 6] for classifying BoVW model. The result shows that the $\mathfrak{LD}$ is merely 0.7232 and is even worse than the baseline of 0.5998. We argue that the reasons behind local BoVW method's ineffectiveness in gesture recognition lie in the following two aspects: 1) low interclass variation between different gestures makes local methods and their corresponding descriptors less discriminative. *c.f*. Figure 3, although motion interest points have been successfully detected around arms and hands area, similarity of interest points around bending elbows could hinder the discriminative power of local patch; 2) one-shot-learning renders it difficult to distinguish the most informative local patch in a BoVW model, especially temporal sequence has been discarded through the construction process of histogram. Insufficient training example would be very likely to lead to the failure in this histogram based approach.

Cons for generative models and others: Under the one-shot-learning configuration, traditional generative models, *e.g*. HMM, would be very likely to fail due to lack of training data and the consequently caused overfitting problem. In the meanwhile, for some discriminative models, one-shot-learning also restrains their discriminative power: *e.g*., for SVM, a single training example can not effectively define its hyperplane for discriminating multi-class; for Adaboost, certain quantity of positive and negative examples are needed to train the weak classifiers; the decision trees methods, *e.g*. Random Forest [5], require hundreds of thousands of training samples to avoids overfitting [19]. Comparatively, nonparametric methods, *e.g*. nearest neighbor, maximum correlation coefficient, *etc*. work surprisingly well for one-shot-learning because they are intrinsically template matching metrics and will not suffer from overfitting problems.

Our approach: *Extended-MHI* and Maximum Correlation Coefficient

Motion Templates: motion energy images (MEI) and motion history images (MHI) proposed by Davis and Bobick [3] are used to represent the motions of an object in video. All frames in a video sequence are projected onto one image across the temporal axis. As where and how motion happens are recorded in the images, MHI captures the temporal information of the motion in a sequence. Assume $I_t = (I_1, I_2, \dots, I_{nFrames}) \in \Re^3$ is a gray scale image sequence and let $B_t = (B_1, B_2, \dots, B_{nFrames-1}) \in \Re^3$ be a binary image sequence indicating regions of motion, which can be obtained from image differencing and thresholding:

$$B_t = \begin{cases} 1 & \text{if} \quad (I_{t+1} - I_t) > Threshold, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where threshold is defined as:

$$Threshold = \sqrt{\sum_{t}^{nFrames} \sigma_t/(h \times w \times nFrames)} \tag{2}$$

where σ_t is the second moment (variance) of a single frame I_t; $h, w, nFrames$ are the height, width and frame number of that video sequence.

The motion history image (MHI) $H(t;\tau)$ is used to represent how the motion image is moving, and is obtained with a simple replacement and a decay operator:

$$H(t;\tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ max(0, H((t-1);\tau) - 1) & \text{otherwise.} \end{cases} \tag{3}$$

We observe that the larger τ, the more information is encoded. Therefore, we set τ as the duration of the whole action to preserve the whole sequence motion trail. The redefined version of MHI is:

$$\widetilde{H}(t;\tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \widetilde{H}(t-1;\tau) - 1 & \text{otherwise.} \end{cases} \tag{4}$$

Note that there is no maximum operator in front of $\widetilde{H_\tau}$ *c.f.* Eq. (3) because setting τ as the sequence duration will lead to non-negativity of $\widetilde{H}(t;\tau)$.

We further extend motion templates to *Extended-MHI* as the early fusion of *MHI* with two more elements: *gait energy information (GEI)* and *inversed recording (INV)*:
Gait energy information (GEI) is to compensate for the non-moving regions and the multiple-motion-instants regions of the action. The summation of all image pixels and normalization of the pixel value define *GEI*:

$$G = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t \tag{5}$$

Inversed recording (INV) is used to recover the loss of initial frames' action information when setting τ as the whole action duration and is defined as follows:

$$\widetilde{I}(t;\tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \widetilde{I}(t+1;\tau) - 1 & \text{otherwise.} \end{cases} \tag{6}$$

Note that its subtle difference to Eq. (4) is the time variable becomes $t+1$ instead of $t-1$ from which we get the name *Inversed Recording*. The *extended-MHI* has been applied to action recognition in [25] and proved to outperform the original MHI.

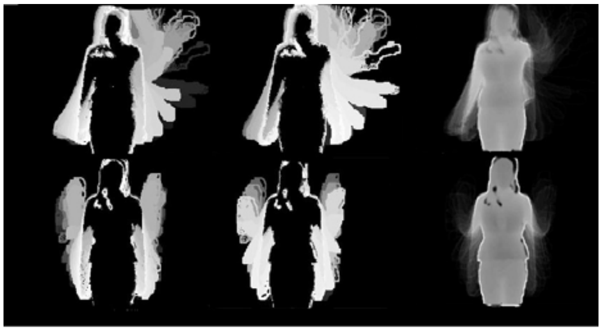

Figure 4. Illustration of the *MHI*, *INV* and *GEI* in two tokens (top row and bottom row). The projection images show that *MHI* emphasizes recent motion, *i.e.* ending frames whilst *INV* the beginning frames. *GEI* encodes the average gait information and is supplementary in repetitive actions where both *MHI* and *INV* are poor at representing.

We reason the complementary property of our *extended-MHI* as *MHI* is poor at representing repetitive actions and *INV* provides complementary information by emphasizing (assigning larger value) at initial motion frames instead of the last motion frames. Figure 4 illustrates the similarities and differences between *MHI*,*GEI* and *INV* of two gesture tokens. The first columns are the *MHI* projections, second are the *INV* projections and the last are the *GEI* projections. Again, the projection graphs show that *MHI* emphasizes recent motion, *i.e.* ending frames whilst *INV* the opposite. Hence the combination of the two is complementary. Furthermore, *GEI* encodes the supplementary information in repetitive actions where both *MHI* and *INV* are poor for the representation. Then, we reduce the dimensionality of each projection by dividing the projection into a 16×16 lattice using HOG as the feature descriptor and concatenate three vectors into a long feature vector. Supervised linear discriminant analysis (LDA) is adopted for the final stage of dimensionality reduction. The experimental results in Table 1 prove the viability of our conjecture. Note that in order to have a fairer comparison between different algorithms, we use the action boundaries provided by [1] for development batch instead of using the temporal segmentation results in Section 3.2 and *MSE* in Section 3.4 is also used for RGB and depth camera fusion so that irrelevant influences can be reduced to a minimum.

For the matching metric, nonparametric methods is more advantageous by avoiding the issue of overfitting. In our experiment, Maximum Correlation Coefficient works best. The correlation coefficient is defined as:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{7}$$

Methods	*GEI*	*MHI*	*INV*	*Extended-MHI*
$\mathfrak{LD}$	0.2761	0.3010	0.3022	**0.2600**

Table 1. Performance comparison of three elements in *Extended-MHI*

where σ_{xy} is the covariance of two feature vectors x and y, and σ_x, σ_y are the variances.

2.1.4 Multiview Spectral Embedding (*MSE*) for Data Fusion

To effectively and efficiently learn the complementary nature of different views, we adopt the spectral methods in [26] to search for a low dimensional representation and sufficiently smooth embedding over all views simultaneously. [22] elegantly presents the intuition behind why spectral clustering works. We briefly state the core algorithm in *MSE* and further cast light on the unaddressed dimensionality problem in [26] by Graph Cut point of view. For notational details, please refer to the paper [26]. Firstly, we construct the graph Laplacian L^i for each view i. The normalized graph Laplacians we choose for the system is L_{sys} as it is a symmetric matrix. Then we introduce a weight α_i to encode the significance for each view i. We try to find the low-dimensional embedding by solving:

$$\underbrace{argmin}_{Y,\alpha} \sum_{i=1}^{m} \alpha_i^r tr(YL^iY^T) \qquad (8)$$

$$\text{s.t. } YY^t = I; \qquad \sum_{i}^{m} \alpha_i = 1, \qquad \alpha_i \geq 0. \qquad (9)$$

where Y is the multiview fused embedding feature vector in a dimension of d, exponent r is the coefficient for controlling the interdependency between different modalities/views and should satisfy $r \geq 1$. Pronounced independence between difference modalities prefers smaller r while rich complementary prefers larger r. In our system, the value r has trivial influence over low dimensional embedding and is set to be 1.5. In our system, we only fuse RGB and depth camera, hence the number of views m is 2.

Eq. (9) is a nonlinearly constrained nonconvex optimization problem and an expectation-maximization (EM) like iterative algorithm can be used to obtain a local optimal solution. The alternating optimization iteratively updates Y and α in an alternating fashion. By introducing Lagrange multiplier λ to take the constraint $\sum_i^m \alpha_i = 1$ into consideration, we get the Lagrange function

$$L(\alpha, \lambda) = \sum_{i=1}^{m} \alpha_i^r tr(YL^iY^T) - \lambda(\sum_{i}^{m} \alpha_i - 1) \qquad (10)$$

By setting the derivative of $L(\alpha, \lambda)$ with respect to α_i and λ to zero, we have

$$\alpha_i = \frac{(1/tr(YL^iY^T))^{1/(r-1)}}{(\sum_{i=1}^{m} \alpha_i tr(YL^iY^T))^{1/(r-1)}} \qquad (11)$$

Here, we cast light on the choice of lower embedding dimension d and the interpretation of weights α_i dispatched to different views where the original paper [26] fails to accomplish. In the paper of [26], the value of d is acquired by cross validation. However, we argue that the low dimension d should be fixed to be the number of gesture *class-1*. According to the Graph Cut theorem, the multiplicity k[1] of the eigenvalue 0 of Graph Laplacian L equals the number of connected components in the graph. Similarly, *MSE* finds d smallest eigenvalues in the spectrum of L which corresponds to the smallest variation of the cluster. The smallest eigenvalue of L is always 0 [22] and the corresponding eigenvector is the constant one vector *1*. Therefore, the veritable number of d should be the number of cluster/gesture *class-1*. And the experiments in [26] are in agreement with our reasoning. Secondly, we explicitly express the physical meaning of the weights α_i as a measurement of the "closeness" of intra-class distance from each individual view. From Eq. (11), we can see that α_i is proportional to the inverse trace of YL^iY^T, and

$$tr(YL^iY^T) = \sum \lambda_i \qquad (12)$$

where λ_i are the eigenvalues of the Graph Laplacian L^i. Hence, $\alpha_i \propto 1/(\sum \lambda_i)$. In Spectral clustering [22], a small eigenvalues (closer to 0) represent the the "closeness" of intra-class distance from each individual view. A well clustered view, *i.e.*, easier to be classified, is more significant than other views. So a larger α_i assigns larger significance to that view.

We then use the low dimensional multiview fused representation Y as the feature vector for Correlation Coefficient comparison. Note that this approach unsupervisedly clusters the test set, however it does not violate the competition rule that allows using unlabeled examples for training the system. We compare the performance between our approach against the approach which directly concatenates the RGB and depth camera feature vector and there is a consistently 4% improvement in $\mathfrak{LD}$.

2.2. Performance Evaluation

By the time we are writing this paper, the performance of our system on validation data batch is **0.29685** and among the top entries on the public leader board in [1] with $\mathfrak{LD}$ less than 0.3. Figure 5 shows our system's performance on the first 20 development batches. It can be observed that our system performs well when there is large amount of motion

[1] multiplicities: the number of eigenvectors belonging to λ_i

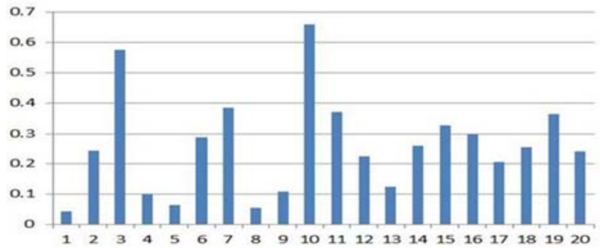

Figure 5. Performance on the 20 development data batches

presents in a gesture token, *e.g*. batch $01, 05, 08, 09$ whereas the performance suffers if the gestures are rather static, *e.g*. batch $03, 10$. We reason that our gesture descriptor is motion based so that little motion and subtle appearance differences in gesture tokens will degenerate our system's discriminative power.

The experiments were done on a Intel 2-core $3.0GHz$, $4GB$ memory desktop in a single thread running MATLAB and the average training and testing time for a single batch is around 220 seconds (approximately 20 fps) which is faster than real time requirement.

3. Discussion and Future Work

We proposed a one-shot-learning gesture recognition system that utilizes both RGB and depth information from Kinect sensor. We utilized depth sensor's unique property to segment human silhouettes and perform a morphological denoising on depth images. Temporal segmentation was performed on the appearance-based approach and an *extended-MHI* representation was adopted as the motion descriptor. We explored the intrinsic property between different spectra and made a physically meaningful embedding of multiviews through *Multiview Spectral Embedding*. Our approach achieves the $\mathfrak{LD}$ less than 0.3 in the [1] competition and performs at the speed of 20 fps. In the future work we would like to utilize motions for more accurate temporal segmentation and state-of-the-art skeleton tracker [19], which could better assist our system to conquer its ineffectiveness in discriminating static gestures by relying on a more advanced appearance-based descriptor.

References

[1] Chalearn gesture dataset. *CGD2011, ChaLearn, California*, 2011. 1, 2, 4, 5, 6

[2] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. The american sign language lexicon video dataset. In *Computer Vision and Pattern Recognition Workshops*, 2008. 1

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2001. 3

[4] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992. 3

[5] L. Breiman. Random forests. *Machine learning*, 2001. 3

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. 3

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, 2005. 3

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, 2005. 3

[9] T. Kadir, R. Bowden, E. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *Proc. BMVC*, 2004. 1

[10] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 2008. 3

[11] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003. 3

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition*, 2008. 2

[13] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *Computer Vision and Pattern Recognition*, 2007. 2

[14] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 1975. 2

[15] S. Park, S. Yu, J. Kim, S. Kim, and S. Lee. 3d hand tracking using kalman filter in depth space. *EURASIP Journal on Advances in Signal Processing*, 2012. 1, 2

[16] M. Ryoo and J. Aggarwal. Stochastic representation and recognition of high-level group activities. *International journal of computer vision*, 2011. 1

[17] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR 2004*, 2004. 1

[18] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition*, 2005. 3

[19] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 3, 6

[20] X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *Computer Vision and Pattern Recognition Workshops*, 2009. 2

[21] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2010. 3

[22] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 2007. 5

[23] H. Wang, M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. *BMVC*, 2009. 2, 3

[24] D. Weinland, R. Ronfard, and E. Boyer. Automatic discovery of action taxonomies from multiple views. In *Computer Vision and Pattern Recognition*, 2006. 1

[25] D. Wu and L. Shao. Silhouette analysis based action recognition via exploiting human poses. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2012. 4

[26] T. Xia, D. Tao, T. Mei, and Y. Zhang. Multiview spectral embedding. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 2010. 2, 5

[27] J. Zieren and K. Kraiss. Robust person-independent visual sign language recognition. *Pattern Recognition and Image Analysis*, 2005. 1