Edge Enhanced Depth Motion Map for Dynamic Hand Gesture Recognition

Chenyang Zhang and Yingli Tian
Department of Electrical Engineering
The City College of New York
{czhang10, ytian}@ccny.cuny.edu

Abstract

In this paper, we propose a novel approach to recognize dynamic hand gestures from depth video by integrating Edge Enhanced Depth Motion Map together with Histogram of Gradient descriptor. The novelty of this paper has two aspects: first, we propose a novel feature representation, Edge Enhanced Depth Motion Map (E^2DMM), balancing the information weighing between shape and motion, which is more suitable for hand gesture recognition; second, we further employ a dynamic temporal pyramid to segment the depth video sequence to address temporal structure information of dynamic hand gestures. Histogram of Gradient is applied on E^2DMM to generate vectored representation. Comparison study has been conducted with the state-of-the-art approaches and demonstrates that our approach can achieve better and more stable performance while keeping a relative simple model with lower complexity as well as higher generality.

1. Introduction

In recent years, the ease of using depth cameras together with the promising application potential of depth cameras has attracted a lot of research efforts into it. As the representative successful debut of Xbox-Kinect [7] by Microsoft in Human Computer Interaction (HCI) and Entertainment, it has caused substantial revolutionary affects both in marketing as well as academia areas such as computer vision and image processing. Human action and gesture recognition, as a significant component of computer vision, naturally has benefited and evolved obviously.

1.1. Related Work

Action and gesture recognition in depth videos has its endowed advantages over that in traditional color or grayscale videos. First, the background is relative clean since the depth sensor implicitly ignores the complex clutter pattern on the background, which is often the major headstream where noise comes from. Second, human body or other parts become easier to be segmented since the 3D spatial information is captured and visualized. Last but not least, the new type of data enables a different area of information which has rarely been touched by traditional action and gesture recognition research on color or grayscale videos. Although action and gesture recognition based on depth cameras is still a relatively new topic, many researchers have paved pebbles to its promising future.

One successful direction is to discover the correlation between action categories and body part joints, which uses estimated body joints [12] [13]. to obtain a reduced representation of human body structure. Shotton *et al.* in 2011 and 2012 [12] [13], proposed to model the body joints estimation problem from a single depth frame. The authors found modes from census of per-pixel classification and solved it utilizing Random Forest and Conditional Regression Forests. On one hand, their work has enabled efficient human body joint extraction from a depth video; on the other hand, it provided many of other researchers a powerful tool to manipulate this kind of raw representation to help to solve their specific human action recognition problems. Simple features computed from body joints solely are proved to be effective in human action recognition problem from Human Computer Interaction (HCI) perspective [15] and Activity of daily living (ADL) perspective [17]. In [15], Wang *et al.* made a very interesting observation that, comparing with using the universe of body joints, using an action-category specific subset of them makes more sense.

However, since methods which are relying on pose estimation are vulnerable to the failure of such pre-processing by self-occlusion, twisted gesture, or unknown human body layout. All the difficulties, which are common in real life, are summed up to some researchers getting rid of using estimated joints [6] [16] [14]. Li *et al.* selected representative points on the contours of three orthogonal projections of the 3D point cloud of human body. In [14], the authors proposed a novel random sampling mechanism using class separability measure together with a novel feature called Random Occupancy Pattern (ROP). This method performed effectively with sparse coding. Motivated by the idea of

Motion History Image [1], our previous work, Depth Motion Map with Histogram of Gradients (DMM-HoG) [16], was proposed to model an action as an energy distribution map over time and also reached good performance in several public datasets [8]. These methods have bypassed the obstacle of joint-based methods because body part estimation sometimes is not accurate and ambiguous for some subtle actions. These methods have also enabled gesture recognition from some part of the whole body, for example, hand gesture recognition.

Hand gesture recognition is a distinct and significant component of human action and gesture recognition since the information hand gestures convey is more sophisticated and linguistic than traditional activities. For example, American Sign Language (ASL) expresses more complicated information than jumping, hand-waving, running, *etc*. Since the substantial difficulty and complexity beneath the problem, related research work, especially those using depth cameras, is still on its infancy but a lot of work shows a promising potential. One approach is to recognize hand gestures using static depth frame as in [10] [11] [18]. In [10], the authors treated each static depth frame as a regular grayscale image and used a bank of Gabor filters to capture gradient information and solved the classification problem using random forests. Different from [10], the authors of [11] focused on a different type of information: contour; and a different application area: hand digits recognition. Other than using gradients and contours, Zhang *et al*. [18] proposed a new descriptor to model hand gesture using histogram of 3D normals. For dynamic hand gesture recognition, ROP in [14] also achieved promising results.

1.2. E^{2}DMM and Dynamic Temporal Pyramid (DTP)

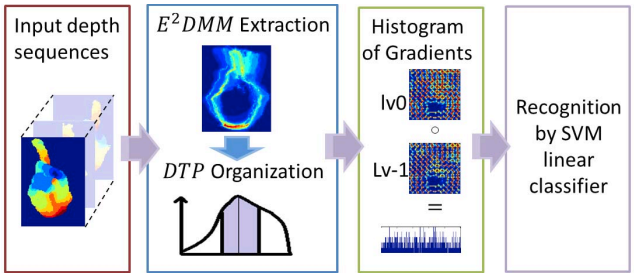


Figure 1. Flowchart of proposed approach. After edge enhanced depth motion map extraction and dynamic temporal pyramid (DTP) organization, we apply Histogram of Gradients (HoG) to generate vectored representation of the two levels of E^{2}DMM (lv0 and lv-1). Finally, a SVM classifier is trained and utilized for classification.

In this paper, we propose a new representation: Edge Enhanced Depth Motion Map (E^{2}DMM), to recognize dynamic hand gestures based on depth video. Moreover, to capture more temporal structure information of hand gestures, we further propose a saliency prior Dynamic Temporal Pyramid (DTP) representation. The framework of proposed approach is as shown in Figure 1. By extracting (E^{2}DMM) and organizing using DTP, the input depth frame sequence transforms to two-layer motion maps. Then Histogram of Gradients (HoG) feature is extracted from the two-layer motion maps and concatenated to generate a vectored representation. A SVM classifier is used to tackle the classification task. There are two benefits of such an approach. First, enhancing the edges provides more information for visually characterizing hand shapes. Second, salience prior temporal pyramid captures more accurate temporal layout of the depth frame sequence. Compared to the state-of-the-art methods [16] [15] [14] [3] [4], the proposed method can achieve higher accuracy in a public hand gesture recognition dataset [9] without more complicated decision models or any sparse favored dictionary learning, which further manifests the efficiency of proposed method. The rest of the paper is organized as follows. Proposed approach is described in details in Section 2. Experimental results and the comparisons with the state-of-the-art are summarized in Section 3. Section 4 concludes the paper.

2. Extraction of E^{2}DMM

In this section, we firstly rephrase DMM in a more general form. We then propose the formulation and computation method of Edge Enhanced DMM (E^{2}DMM). A novel saliency prior dynamic temporal pyramid structure is also proposed and compared to traditional temporal pyramid.

2.1. Depth Motion Map (DMM)

Depth Motion Map (DMM) [16] is a visual representation of human activities by accumulating the motion of each frame in a depth video. DMM is a global descriptor mainly focusing on modeling the spatial energy distribution of human actions. A DMM of a depth video can be given as:

$$f(X_{i,j}) = \sum_{t=1}^{T-1} (\delta(|x_{i,j}^{t} - x_{i,j}^{t+1}| - \epsilon) + \sigma_{i,j}) \qquad (1)$$

where X is a depth video given as a collection of depth frames $X = x^1, ..., x^T$ and ϵ is a parameter to determine the strength of motion, which is named *"penetration threshold"*; $\sigma_{i,j} = -e_{i,j}$ is a penalty term to suppress the energy accumulation in the edge pixels.

DMM accumulates the motion between each pair of consecutive depth frames to generate an energy distribution map to discriminatively represent an action. Usually Histogram of Gradients (HoG) operator is applied on a DMM to generate a concrete feature vector, as shown in Figure 2 (b) and (d). In this paper, based on such modeling, we first discuss the difference and similarity between human

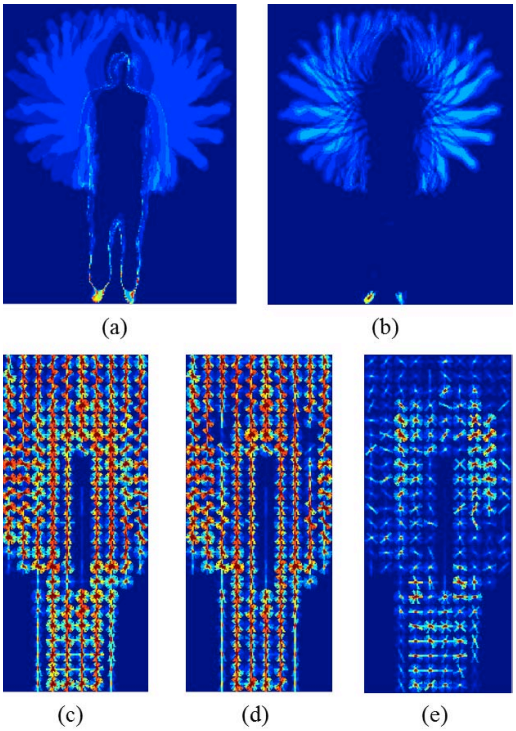


Figure 2. Illustration of edge suppression in DMM [16] computation of action "Two Hands Wave" in [8]. (a) DMM without edge suppression. (b) DMM after edge suppression in Equation 1. (c) and (d) are Visualizations of HoG Representations of (a) and (b), respectively. (e) is the difference in feature space of (c) and (d) . The contour of human body can be obviously observed without edge suppression, which causes ambiguity as shown in (e).

action recognition and hand gesture recognition in depth video and propose edge enhancement process to mutate DMM to E^{2}DMM to adapt to hand gesture recognition. Then a new temporal pyramid based on temporal saliency is proposed to capture more temporal structure information.

2.2. Edge Enhanced DMM (E^{2}DMM)

According to [16], edge suppression is suitable for dramatic human action recognition since the contours do not provide useful information to help distinguishing between different actions, but may introduce variance between different subjects.

However, static pose or gesture plays a different role in the domain of hand gesture recognition. In fact, static pose or gesture conveys a significant portion of information and a lot of work has been done in this direction [18] [11] [10]. In the perspective of dynamic hand gesture recognition, static gesture together with motion provides an integral description of gesture category. Thus, we formulate Edge Enhanced Depth Motion Map (E^{2}DMM) computation by changing the edge suppression term to an edge enhancement term:

$$g(X_{i,j}) = \sum_{t=1}^{T-1}(\delta(|x_{i,j}^{t} - x_{i,j}^{t+1}| - \epsilon) + \rho * e_{i,j}) \quad (2)$$

where parameter ρ is a weight to tune the degree of edge enhancement, as shown in Figure 3. We will demonstrate the effectiveness of the edge enhancement term and the effect of different selections of ρ in Section 3.1.

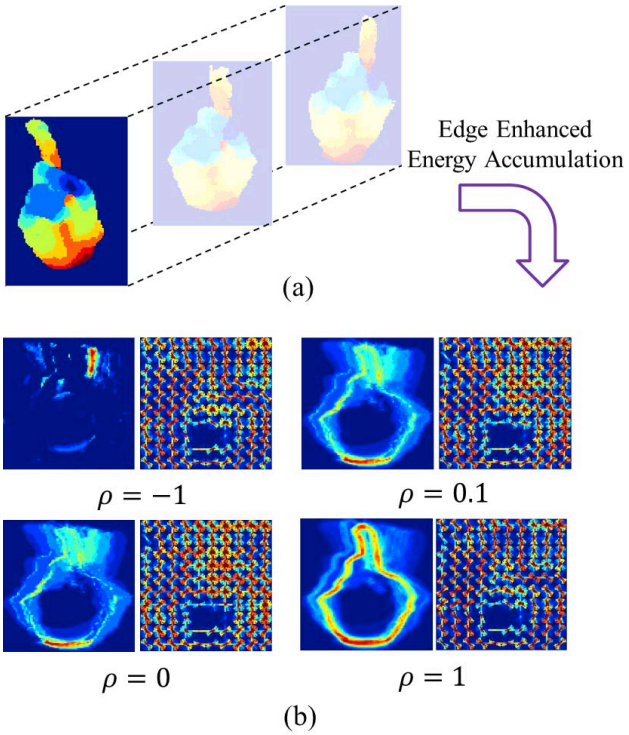


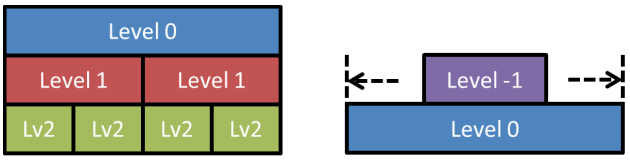
Figure 3. (a) A depth video showing the gesture"Where" in American Sign Language (ASL). (b) Accumulative Motion Maps and corresponding visualizations of HoG of different selection of ρ. When ρ = -1, it degenerates to DMM and when $\rho > 0$, it is in the form of E^{2}DMM.

2.3. Dynamic Temporal Pyramid

In this section, we will show how to model the temporal structure using saliency prior dynamic temporal pyramid.

Traditionally, to capture the temporal structure or layout of an action, one may use temporal pyramid, which is very similar to the idea of Spatial Pyramid [5]. As shown in Figure 4 (a), temporal pyramid evenly divides the features into 2 or 4 or more buckets in the temporal dimension. Intuitively, since an action may usually have several phases: onset, apex, and offset while each phase contributes differently to the distinguish power of the final feature representation. Therefore, dividing the features in the time domain may help to address such temporal structure information.

Moreover, the dimension of final feature representation is proportional to the total buckets the pyramid uses, *e.g.*, in Figure 4 (a), it uses a final feature vector with dimension 7 times to that only uses level 0 feature.



(a) (b)

Figure 4. Illustrative comparison between (a) traditional temporal pyramid and (b) our saliency preferred dynamic temporal pyramid structure. Compared to the fixed branching in (a), the branching in (b) is dynamically determined based on the energy distribution.

Therefore, we propose to use a new temporal organization method taking advantage of the energy distribution, which can be easily computed by modifying Equation 2:

$$E(X) = \times_{t=1}^{T-1} \sum_{(i,j)\in M_t \times N_t} (\delta(|x_{i,j}^t - x_{i,j}^{t+1}| - \epsilon) + \rho * e_{i,j}) \tag{3}$$

where E(.) is a vector computed as the Cartesian product of sums of non-zero entries in frame t. E(.) is plotted as blue curve (the magnitude of the curve at a certain time represents the spatial integral of E(.) in a certain frame) in Figure 5 and its integral over time is plotted as red curve.

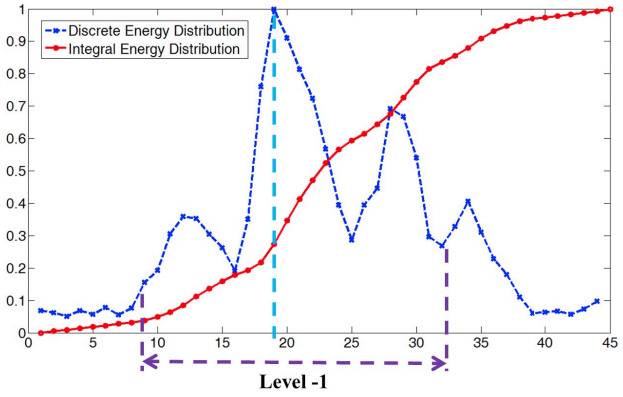


Figure 5. Illustration of how to dynamically select "level -1" based on the energy distribution over time. Blue curve shows the energy of each frame. Red energy shows the integral energy over time. Cyan dashed line shows the frame with highest energy and purple dashed line shows the chosen window for "level -1". X-axis is frame index and Y-axis is the relative magnitude of each curve.

As a result, other than trying to use more space to store different representations of different temporal segments, we tend to use a separate feature space to summarize those frames with more significant information in terms of energy. Experimental results show that this organization is more suitable for temporal information capture and saves space which means less computation cost.

Detailed parameter discussion as well as settings will be conducted in Section 3. HoG is applied as in [16] to generate feature vectors. To evaluate the distinguish power we simply use linear Support Vector Machine (SVM) classifier without any sophisticate manipulating such as dictionary learning.

3. Experimental Results

In this section, we first introduce the public dataset we use for evaluation as well as its statistics. Then evaluation of E^2DMM and comparison with the state-of-the-art methods will be described.

3.1. Experimental Setup

In the experiments, we use MSRGesture3D dataset [9], which was captured using a Kinect camera. The dataset is for dynamic American Sign Language (ASL). There are 12 gesture categories from two letters z and j to words: "Z", "J", "Where", "Store", "Pig", "Past", "Hungary", "Green", "Finish", "Blue", "Bathroom", and "Milk". There are 10 different subjects involved and each person performs each category 2 to 3 times. There are 336 samples in total, each of which is a depth sequence. The dataset is pre-segmented with only hand appeared in the depth videos

To extract E^2DMM, the input is a depth sequence and the output is a feature vector with a fixed length of dimension. In all our experiments, we normalize the patch sizes of E^2DMMs to 100×100. We use the off-the-shelf HoG generator in [2] with patch size 8, orientation bins 8, and all four normalization methods [1]. The dimension of final feature vector is 6400 after applying our proposed dynamic saliency prior temporal pyramid.

Following the benchmark setting as in [14], the performance evaluation is processed by using leave-one-subject-out strategy, which means each time one subject is chosen for test while the SVM classifier is trained on the data composed by the remaining 9 subjects. The performance is calculated as the average classification accuracy after all subjects are tested.

3.2. Evaluation of the Proposed Method

In this section, we explore the discriminative power of proposed descriptor for dynamic hand gesture recognition and the effects of several parameters: penetration threshold and degree of edge enhancement (ϵ, ρ). Our default setting of (ϵ, ρ) is (10, 0.1), where the ϵ has unit of mm.

[1] For more details related to parameters, please refer to [2].

The effects of different parameter settings can be visualized in term of classification accuracy as shown in Figure 6. From Figure 6, we can observe that with proper edge

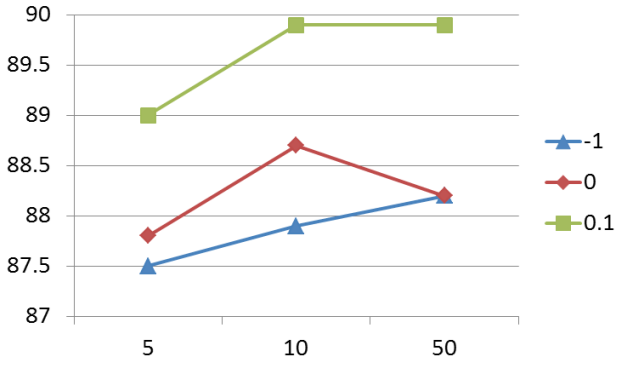


Figure 6. Performance evaluation of E^2DMM on dynamic hand gesture dataset [9] of different parameter settings. X axis: penetration threshold ϵ; Y axis: performance in term of classification accuracy. Blue, red and green colored curves indicate degree of enhancement -1, 0, and 0.1 respectively. Notice that when degree of enhancement is -1, it actually represents traditional DMM [16].

enhancement ($\rho = 0.1$), the performance is better (1.5%) than no-enhancement ($\rho = 0.1$) and edge suppression [16] ($\rho = -1$). Thus, we select a stable penetration threshold ϵ = 10 and edge enhancement degree ρ = 0.1, as default parameters.

While adding level -1 for dynamic temporal pyramid representation, we firstly compute the frame index which has the highest peak in the energy distribution, then we search to the left till: 40% of total energy is included OR reaching the starting index, as left bound of level -1 window; we search to the right till: 40% of total energy is included OR reaching the ending index, as right bound of level -1. The level -1 uses another 3200 dimension feature vector which is then concatenated with level 0. In total, the dimension of final feature vector is 6400. We empirically set different weights for level 0 and -1 with 2 and 1. The overall accuracy after applying this is 90.5% for our proposed method, which is used for comparison with the state-of-the-art approaches. Comparison of proposed dynamic temporal pyramid framework and traditional temporal pyramid framework (level 0 and level 1 as shown in Figure 4(a)) is shown in Table 1 with the same weights setting. Proposed DTP outperforms traditional temporal pyramid (TP) with 1.6% and 1.8%, respectively, in overall accuracy both in un-weighed and weighed cases.

3.3. Comparison with the State-of-the-arts

In this section, we compare our method with several other state-of-the-art methods in term of accuracy. The confusion matrix is as shown in Figure 7. Our method performs well in most classes and the worst one reaches 74%.

Table 1. Comparison between proposed Dynamic Temporal Pyramid (DTP) and traditional Temporal Pyramid (TP).

	Proposed DTP	Traditional TP
un-weighted	88.07%	86.45%
weighted	**90.53%**	88.70%

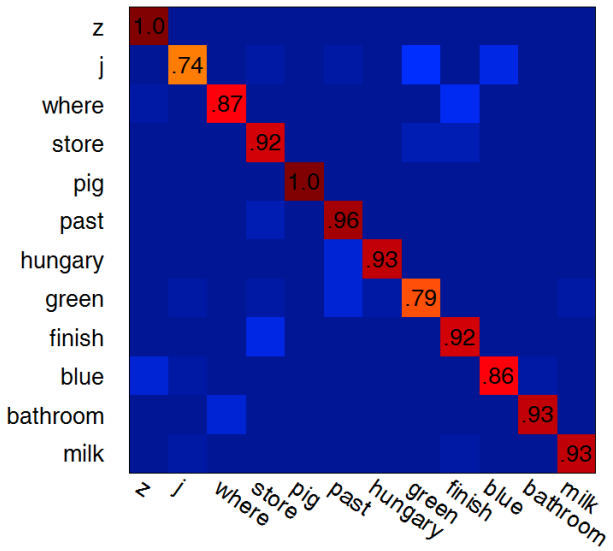


Figure 7. The confusion matrix of proposed method on dataset Gesture3D.

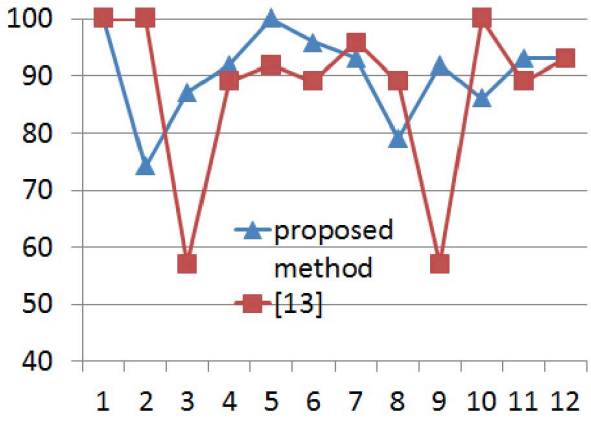


Figure 8. Class-wise accuracy comparison of proposed method and Random Occupancy Pattern with Sparse Coding (ROP-SC) in [14]. Obviously, besides that our method outperforms method in [14] with 2%, proposed method is more stable, which means our representation is more general for different action classes.

We compare the proposed method with several state-of-the-art methods such as [3] [4] [14] [16] as listed in Table 2. Our proposed method performs best (90.5%) in term of

averaged classification accuracy. Compared with DMM in [16] without edge enhancement, proposed E^2DMM together with dynamic temporal pyramid outperforms the previous one with 2% accuracy.

Compared with Random Occupancy Pattern feature together with Sparse Coding in [14], our approach outperforms the performance in [14] with about 2% and more details can be seen from Figure 8. Besides, our method performs rather stable while the worst accuracy in "j" is 74%, which is much higher than the worst cases in [14], 57% (in class "green" and "where"). The results demonstrate that our method is more general and suitable to represent dynamic hand gestures of different categories.

Table 2. Comparison of proposed method and other methods.

Method	Accuracy
High Dimensional Convolutional Network [3]	0.69
Action Graph on Occupancy Features [4]	0.805
Action Graph on Silhouette Features [4]	0.877
Random Occupancy Pattern (ROP) [14]	0.868
ROP+Sparse Coding [14]	0.885
Depth Motion Map (DMM) [16]	0.882
Proposed Method without any TP	0.899
Proposed Method with traditional TP	0.887
Proposed Method with DTP	**0.905**

4. Conclusion

In this paper, we have proposed an edge enhanced depth motion map framework to model different hand gestures from their visual effects. Then we have designed a new dynamic temporal pyramid organization approach to capture temporal structure to compensate the information loss due to building energy map by integrating discrete energy from each frame along temporal dimension. For classification, we apply a Support Vector Machine with linear kernel.

Experiments demonstrate that our method achieves better performance compared to the state-of-the-art methods while using relative simple classifier rather than involving complicate dictionary learning techniques. In addition, our proposed method is more general among different hand gesture categories.

Acknowledgment

This work was supported in part by NSF IIS-0957016 and Microsoft Research.

References

[1] G. Bobick and J. Davis. The representation and recognition of human movement using temporal templates. *IEEE Trans. on PAMI*, 2001. 2

[2] P. Dollar. Piotr's image and video Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html. 4

[3] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. In *ICML*, 2010. 2, 5, 6

[4] A. Kurakin, Z. Zhang, and Z. Liu. A real-time system for dynamic hand gesture recognition with a depth sensor. In *EUSIPCO*, 2012. 2, 5, 6

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE, 2006. 3

[6] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *IEEE Workshop on CVPR for Human Communicative Behavior Analysis*, 2010. 1

[7] Microsoft Corporation. Kinect for Xbox 360. http://www.xbox.com/en-US/kinect, 2010. 1

[8] Microsoft Research. MSR Action Recognition Datasets. http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/default.htm. 2, 3

[9] Microsoft Research. MSR Gesture3D Dataset. http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/gestureData.tar.bz2. 2, 4, 5

[10] N. Pugeault and R. Bowden. Spelling it out: Real-time asl fingerspelling recognition. In *ICCV Workshops*, 2011. 2, 3

[11] Z. Ren, J. Yuan, and Z. Zhang. Robust hand gesture recognition based on finger-earth mover's distance with commodity depth camera. In *International Conference on ACM Multimedia*, 2011. 2, 3

[12] J. Shotton, A. Fitzgibbon, , M. Cook, T. Sharp, M. Finocchio, M. moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011. 1

[13] M. Sun, P. Kohli, and J. Shotton. Conditional regression forests for human pose estimation. In *CVPR*, 2012. 1

[14] J. Wang, Z. Liu, Z. Chorowski, Z. Chen, and Y. Wu. Robust 3D action recognition with random occupancy patterns. In *ECCV*, 2012. 1, 2, 4, 5, 6

[15] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012. 1, 2

[16] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histogram of oriented gradients. In *International Conference on ACM Multimedia*, 2012. 1, 2, 3, 4, 5, 6

[17] C. Zhang and Y. Tian. RGB-D Camera-based Daily Living Activity Recognition. *Journal of Computer Vision and Image Processing*, 2(4), 12 2012. 1

[18] C. Zhang, X. Yang, and Y. Tian. Histogram of 3D facets: A characteristic descriptor for hand gesture recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013. 2, 3