Human Daily Action Analysis with Multi-view and Color-Depth Data

Zhongwei Cheng[1], Lei Qin[2], Yituo Ye[1], Qingming Huang[1,2], and Qi Tian[3]

[1] Graduate University of Chinese Academy of Sciences, Beijing 100190, China
[2] Key Lab of Intelli. Info. Process., ICT CAS, Beijing 100190, China
[3] University of Texas at San Antonio, TX 78249, U.S.A.
{zwcheng,lqin,ytye,qmhuang}@jdl.ac.cn, qitian@cs.utsa.edu

Abstract. Improving human action recognition in videos is restricted by the inherent limitations of the visual data. In this paper, we take the depth information into consideration and construct a novel dataset of human daily actions. The proposed ACT4[2] dataset provides synchronized data from 4 views and 2 sources, aiming to facilitate the research of action analysis across multiple views and multiple sources. We also propose a new descriptor of depth information for action representation, which depicts the structural relations of spatiotemporal points within action volume using the distance information in depth data. In experimental validation, our descriptor obtains superior performance to the state-of-the-art action descriptors designed for color information, and more robust to viewpoint variations. The fusion of features from different sources is also discussed, and a simple but efficient method is presented to provide a baseline performance on the proposed dataset.

Keywords: Daily action, Multi-View, RGB-D, Depth descriptor.

1 Introduction

Human action analysis has attracted sustaining attentions in computer vision community. Along with the huge demands in intelligent surveillance, advanced human-computer interaction and smart-house/e-monitoring, it inspires more and more interests to tackle practical problems by adopting action analysis techniques these years. Recognizing daily activities may have comparatively greater potential in real applications. To imagine if computers could promptly detect people's fall or correctly perceive the behavior of taking medicine, it would have significant merits for the digitized guardianship of the kids and olds. However, existing methods of modeling and understanding human daily activities are far from satisfactory in practical applications. The main challenges come from two aspects: 1) Traditional vision based approaches [1–4] depend on single data source of color cameras, which has inherent limitations and insuperable defects, such as illumination changes, looks and clothing variations, and cluttered background patterns. 2) Due to the viewpoint variations, which are common in real conditions, the observed actions have conspicuous intra-class diversities.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 52–61, 2012.

Introducing depth modality is likely helpful for action analysis, as motion characteristics are critical to represent actions and temporal depth data reflect the motion properties directly. Additionally, depth information has an advantage in privacy preserving, for that only distances between targets and sensors are captured (as visualized in Fig. 2). Therefore in this paper, we construct a multi-view action dataset with both color and depth sources. It focuses on human daily activities to innovate technology facilitating real applications. We also propose a novel method to represent actions utilizing depth information, and investigate the fusion of heterogeneous features from color and depth sources to present a baseline performance on the proposed dataset. Our contributions are two-fold:

1) We build a large dataset of human daily actions, which is well organized with multi-view and color-depth information.

2) We propose a new descriptor to represent depth information for action recognition, which is efficient and relatively robust to viewpoint changes.

The remainder of the paper is organized as follows. We introduce related work in Section 2. The construction of the ACT4^2 dataset is detailed in Section 3, and the new depth descriptor is proposed in Section 4. We present the experimental results and discussions in Section 5, and finally conclude the paper in Section 6.

2 Related Work

Some benchmark datasets have been published in the last decade, and they promote the development of human action analysis. In Table 1, we summarize some action datasets related to our work. The KTH [2] dataset is extensively used in action recognition domain, while it only contains simple and basic actions such as "running" and "waving". The IXMAS dataset [5] takes viewpoint changes into consideration and provides action frames observed from 5 views, but the action content is still about basic movements. Some relatively complex actions are presented in the Activities of Daily Living dataset (UR ADL) [6], which are semantic actions like "dialing a phone". The MuHAVi [7] is a large multi-view dataset for the evaluation of action recognition methods. All of above databases adopt only the conventional color cameras as data source. With the emergence of Microsoft Kinect, human action datasets employing depth information sprout in past two years. The dataset in [8] supplies the sequences of depth maps for actions focused on human-computer interaction in games. Furthermore, Sung et al. [9] construct a RGBD dataset for human activity detection and recognition in the living environment, and the RGBD-HuDaAct [10] is aiming to encourage application of assisted living in health-care. The last two are most related to our dataset, but there are clearly different emphases on activity categories and specific targets. More importantly, as shown in Table 1, the proposed ACT4^2 is the first multi-view RGBD human action dataset.

Action representations with traditional color information often exploit global features such as silhouettes or shapes [1], and local features like space-time interest points [3]. For the newly introduced depth information, there are also some similar methods proposed. Li et al. [8] sample 3D points from the depth maps

Table 1. Benchmark databases for human action recognition

Database	*Data Source*	*Multi-View*	*Action Classes*	*Samples*
KTH [2]	Color	No	6	2391
IXMAS [5]	Color	Yes	13	2340
UR ADL [6]	Color	No	10	150
MuHAVi [7]	Color	Yes	17	1904
MSR ADS[8]	Depth	No	20	4020
CU HAD [9]	Color+Depth	No	12	N/A
HuDaAct [10]	Color+Depth	No	12	1189
ACT4^2	Color+Depth	Yes	14	6844

within human contours, and propose a bag of 3D points representation to characterize the key postures in actions. Shotton et al. [11] present a human pose recognition method based on body part expressions, with predicting the body joints from single depth images. However, these pose-level representations may encounter problems of occlusions. When part of human body is shaded by background objects or occluded by other body parts, the observed representations become unstable and noisy. On the other hand, Zhang et al. [12] consider the actions of color and depth information as 4D hyper cuboids to extract local feature points. Ni et al. [10] employ local features from color space and utilize the depth as a reference channel in feature pooling. These two methods are straightforward extensions of local representations for color data, and therefore they are not very appropriate to capture the characteristics of depth information. In this paper, we design a novel comparative coding descriptor, aiming to extract inherent properties from the depth data.

3 The ACT4^2 Dataset

To break the bottleneck of existing action recognition approaches, we construct a new dataset. We aim to provide an infrastructure for investigating both color and depth information in human action analysis and handling action variations over viewpoints. Furthermore, we expect it to facilitate practical applications in real life, like smart house or e-healthcare. Consequently, the action categories in ACT4^2 mainly focus on the activities of daily living. Other than simple atomic movements as 'wave' or 'clap', these actions are complete activities with clear semantics in real life. For additional consideration of multiple views and indoor settings, we finally choose 14 actions: *Collapse, Drink, MakePhonecall, MopFloor, PickUp, PutOn, ReadBook, SitDown, SitUp, Stumble, TakeOff, ThrowAway, TwistOpen* and *WipeClean*. Intuitive samples of these daily activities are illustrated in the PNG figure in supplementary material. *Collapse* and *Stumble*, which are seldom introduced in previous datasets, are interesting activities especially for Homecare applications. *Collapse* stands for people falling by inner factors, such as hurt or giddiness, while *Stumble* means body dropping caused by outside effects such as tripped by an obstacle.

3.1 Data Acquisition

To simultaneously capture color and depth data from different views, we install four Microsoft Kinects which are connected to four computers respectively. These devices are placed in a common living room with different heights and diverse angles to the action area, as illustrated in Fig. 1. Totally 24 people are invited to perform each of the selected 14 activities several times. The recorded data are synchronized in the capturing procedure. For individual views, the frames from color and depth cameras are spatio-temporally aligned in pixel level by the Kinect drivers. Cross-view synchronization is achieved by the time stamps of frames. Although it would not assure precise frame-to-frame alignment, it can provide satisfying quality for cross-view action analysis. Finally, about 34 hours data are recorded with the resolution of 640×480 at 30 FPS, in the form of 8-bit 3-channel color images and 16-bit 1-channel depth images.

3.2 Data Preprocessing

Due to the mechanism of depth camera in Kinect, it will obtain zero value which may generate "black holes" in depth images, when the measure procedure fails. Especially, when four Kinects work simultaneously within the same area, the "black hole" effect is magnified as shown in Fig. 2(a), because of the interference of each other. To reduce its impact, we produce clean background depth images from the four viewpoints individually. Then all of the raw depth frames are refined by filling "black holes" with respect to the depth values at the same coordinates in adjacent frames and those background depth images. As shown in Fig. 2(b), the quality of processed depth data is improved obviously. Notice that the refinement is not to enhance but to restore depth data to the original quality which should be obtained by depth sensor individually.

Sequential data are segmented manually by labeling the start/stop points of single actions. With dropping the irrelevant interlude frames, complete action instances are extracted. The durations of action samples vary from seconds to tens of seconds. Finally, ACT4^2 contains 6844 action clips with both color and depth information, which are collected from 4 viewpoints. To the best of our knowledge, the proposed dataset is the largest multi-view color-depth human action dataset to date.

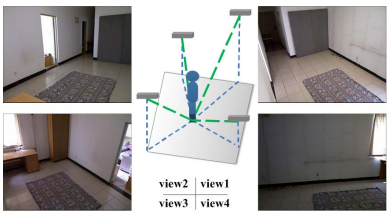

Fig. 1. The environment setting

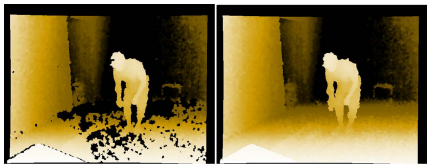

Fig. 2. Depth data preprocessing. (a) is the raw frame; (b) is the refined frame

4 Depth Information Representation

Data captured by depth camera are spatial distances between targets and camera. Therefore, the spatial structure of targets is inherently embedded in depth information. We believe it provides complementary information to visual color data. Depth can be straightforwardly used as auxiliaries to improve foreground segmentation in color images, or quantization weights of ordinary color features. But focused on the characteristics of depth information itself, we propose a new descriptor to explore the embedded structural relations for action analysis.

4.1 Comparative Coding Descriptor

The new depth descriptor is designed to capture the spatial geometrical relations and related variations over time. For differences of values in depth images can properly reflect the relative distances between positions, and also inspired by the idea of comparative description in LBP [13], we propose Comparative Coding Descriptor (CCD) to represent the depth information for action analysis.

Depth data of an action instance is regarded as a spatiotemporal volume of depth values. Small cuboids can be extracted from the volume with selected reference points as centers. For action representation, the reference points can be chosen as salient points or spatiotemporal corners. Cuboid with size of $3\times3\times3$ is treated as atomic cuboid, based on which CCD feature is extracted. The value on the reference point is compared with that of the other 26 points respectively, and the differences are coded. The coding scheme is noted in Eq. 1,

$$\Psi(p_i) = \begin{cases} 1, & D(p_i) - D(p_r) \geq \gamma \\ -1, & D(p_i) - D(p_r) \leq -\gamma \\ 0, & else \end{cases} \tag{1}$$

where p_r denotes the reference point, and p_i is other point in cuboid, $i = 1, ..., 26$. $D(p)$ indicates the depth value of p, and γ is a comparison threshold. Then the 26 codes are sequentially composed to form the CCD feature vector, according to the order of a spiral line over the atomic cuboid.

Via describing the structure of the depth cuboid by sequential codes, CCD properly presents the spatiotemporal constraints within actions. Benefitting from coding with qualitative comparisons and preserving the neighborhoods, CCD could have some degree of robustness to perspective variation. Furthermore, it can be easily expanded by introducing a spatial scalar s and a temporal scalar t. To express every $s \times s \times t$ sub-cuboid with a single value, for example the mean value, the expanded $(3 \times s) \times (3 \times s) \times (3 \times t)$ cuboid can be converted to an atomic cuboid. Fig. 3 illustrates the generation of CCD features. Colored slices present depth frames over time, and the red dot presents the reference point. The coding scheme is demonstrated in the rightmost chart.

4.2 Distance Metric

CCD is in the form of sequential codes with values restricted to [-1, 0, 1]. To measure the similarity between CCD features, we design a corresponding

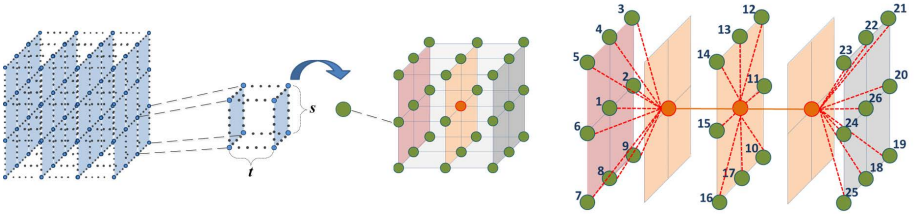

Fig. 3. Comparative Coding Descriptor. Blue cuboid on the left denotes the depth volume. The middle cuboid with green vertexes is the atomic cuboid for CCD extraction. The numbers in the right chart indicate the coding order. (best viewed in color).

distance metric. As defined in Eq. 2, this metric consists of two parts, considering both code pattern and sequential structure.

$$M(f_i, f_j) = \alpha_1 M_p(f_i, f_j) + \alpha_2 M_s(f_i, f_j) \tag{2}$$

$f_i = (c_{i1}, ..., c_{in})$ is a CCD feature for the i-th reference point, c_{in} denotes the code at position n and α_1, α_2 are the weights of the metrics for code pattern (M_p) and sequential structure (M_s). The code pattern is measured by the consistency of the codes in different CCD features, which is defined as Eq. 3,

$$M_p(f_i, f_j) = \sum_{k=1}^{k=n} \|c_{ik} - c_{jk}\|_2 \tag{3}$$

The sequential structure of CCD holds the spatiotemporal relations within action volumes. To consider this specialty in distance metric, the discrimination power can be well retained. We employ the first-order derivatives of the features to present the characters of their sequential structures, as formalized in Eq. 4, 5.

$$M_s(f_i, f_j) = M_p(\nabla f_i, \nabla f_j) = \sum_{k=1}^{k=n-1} \|d(c_{ik}, c_{i(k+1)}) - d(c_{jk}, c_{j(k+1)})\|_2 \tag{4}$$

$$d(c_{ik}, c_{i(k+1)}) = \begin{cases} 1, & c_{ik} < c_{i(k+1)} \\ 0, & c_{ik} = c_{i(k+1)} \\ -1, & c_{ik} > c_{i(k+1)} \end{cases} \tag{5}$$

5 Experiments

We evaluate the performance on ACT4[2] by considering color and depth modalities independently and jointly. All experiments are performed on a subset of the database for efficiency. With randomly choosing 2 samples per person per action from all 4 views, the subset consists of 2648 instances in total. 8 out of 24 persons' data are used for training, and the rest are for testing. Every experiment is performed 10 times by random sampling training set, and the average performance is reported. We adopt a popular, local feature based action recognition

approach as the baseline method. It extracts action features based on the space-time interest points (STIP) [3], and then represent actions as histograms with the Bag of Visual Words (BoVW) method [4]. Finally, one-against-all Support Vector Machine is introduced for classification. With this broadly employed approach, we expect to present a baseline performance on ACT4^2 and demonstrate the characteristics of this dataset.

5.1 Color Information versus Depth Information

Color data provide visual appearance clues of actions, on the other hand, depth data supply the structural information. In this experiment, we extract local features from both modalities respectively, and evaluate their discrimination capabilities. The image sequences of actions are encoded into video clips. For depth data, the 16-bit images are firstly converted to 8-bit intensity images with min-max normalization. The detector in [3] is applied on the videos to generate STIPs. Then features of the Histograms of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF) are extracted according to color and depth STIPs respectively. Codebooks are generated by k-means with varied size from 200 to 2000. As illustrated in Fig. 4, the recognition precision of features in depth is consistently superior to features from color data. It is interesting that with exactly the same recognition approach, the performance gap between depth and color features is beyond 10%. This reveals depth information is more effective than color data for presenting actions on ACT4^2 dataset. There might be two reasons: 1) It is more "direct" to get motion clues from depth data, so depth features are more accurate. Foreground location variations can be obtained immediately from the distance changes in depth maps. But for color data, motions are estimated implicitly with visual appearance matching. 2) Depth features are comparatively less noisy. The distance information in depth data is robust to environment variations. However, visual information in color data suffers from the appearance diversity. And inevitable consideration of unrelated visual clues also impacts the quality of color features.

5.2 Evaluation of Depth Descriptor

Depth information has shown inherent advantages for action recognition in the previous experiment. It motivates us to design a new depth descriptor to explore its distinct characters. And we evaluate the proposed CCD descriptor here. The STIPs extracted from depth videos are considered as the reference points. And we generate CCD features with s and t varying from 3 to 11. The used depth values are from original 16-bit depth images. To maintain the properties of CCD features, k-medoids clustering with the distance metric in Section 4.2 is adopted in codebook generation. The metric weights α_1 and α_2 are equally set to 0.5. Related experimental results are shown in Fig. 5.

The recognition precision is increasing with the spatiotemporal scales of CCD expanding, while the rate of increase is declining. This may be related to how the cuboid scale matches the movement range. Therefore, CCD provides a flexibility

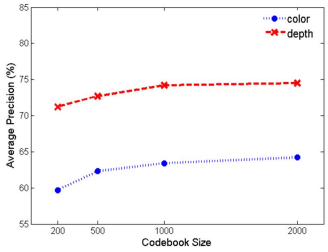

Fig. 4. Performance comparison of color and depth information

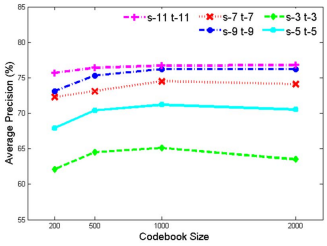

Fig. 5. Recognition performance of CCD features in various scales

to represent depth characteristics for different scales. The validity of CCD can be proved by comparing recognition performance with HOGHOF, which is a state-of-the-art descriptor for action recognition[14]. On depth data, HOGHOF achieves its best performance of 74.5% at k=2000, while CCD obtains superior performance of 76.2% at k=1000. It should be noted that dimension of CCD is only 26 and it requires simple calculation. In contrast, dimension of HOGHOF is 144 and it requires complex computing for gradients and optical flows. That is, CCD gains better performance with fewer dimensions and lower computation cost. This advantage may be attributed to the effective description of the motion structure in actions with depth information.

5.3 Action Recognition Fusing Two Modalities

Though the previous results suggest depth information is superior to color information in representing human actions, it should be beneficial to jointly use the features from the heterogeneous modalities. Color and depth actually present actions from different perspectives in essence. Thus a united representation can be enhanced with introducing broader characteristics. It is not trivial to fuse color and depth information in action recognition tasks. We have attempted a Cascade BoVW method with hierarchically clustering color and depth features in different levels, but its performance is disappointing.

Combining different feature vectors into a united vector is proved to be a valid method in existing works. For example, Spatial Pyramid Matching [15] concatenates features vectors from different pyramid levels. Thus in this paper, we adopt such idea and utilize a Super Feature Representation (SFR) to fuse features from the heterogeneous sources. For each action instance, the BoVW histogram vectors of different features are concatenated together and normalized. Additionally, we implement another fusion method, the DLMC-STIPs, with applying the same setting in [10]. The recognition performances of different approaches are shown in Table 2.

Comparing with HOGHOF features on color data, the performance gain of 16.3% is achieved by our SFR, whereas only 2.1% by DLMC-STIPs. This is probably because DLMC-STIPs just applies primitive combination by treating depth as a reference channel in color feature pooling, while we describe the characteris-

Table 2. Comparison of different features

Feature	*Avg. Precision*
Color-HOGHOF	64.2%
Depth-HOGHOF	74.5%
Depth-CCD	76.2%
DLMC-STIPs (SPM) [10]	66.3%
SFR	**80.5%**

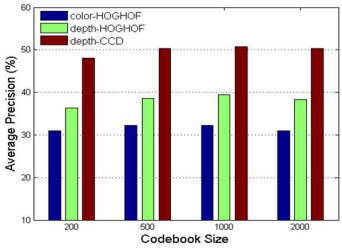

Fig. 6. Cross-view performances

tics of depth directly and fuse the discriminative features. Notice that although our fusion method is algorithmically simple, its efficiency and effectiveness are satisfactory. In addition, the performance of our fusion representation is superior to that of any single feature. This may indicate color and depth are complementary information for modeling human actions. To further investigate joint models over the heterogeneous feature spaces is promising for improving action analysis.

5.4 Cross-View Performance on ACT4²

All previous experiments do not distinguish action samples from different views. In this section, we investigate the cross-view action recognition by choosing 2 views for training and the other 2 views for testing. The codebook generation and classifier construction are restricted to the training views. All 6 cases of train/test views combinations are evaluated, and the average of recognition precisions is reported. Fig. 6 shows the cross-view performances of different features.

Compare to previous results, the cross-view performance decreases by 25.4% for CCD, 31.9% and 35% for HOGHOF in color and depth data. The performance drop is reasonable, as the model trained on selected views is hard to handle the missing variations in other views. It should be pointed out that HOGHOF loses its half precision in cross-view evaluation, while CCD loses only one third. And CCD achieves its best performance of 50.8%, which is more than 11% superior to HOGHOF. These results show that CCD has a better robustness to the view variations. This advantage may be benefited from preserving the relative sequential restrictions in both feature description and distance metric.

6 Conclusion

In this paper, we construct a large multi-view and multi-source benchmark dataset of human actions. Both color and depth information is provided to innovate better action representations. We also propose a novel descriptor to depict the characteristics of depth information. The experimental results prove the effectiveness and efficiency of the proposed feature. With fusion of color and depth features, a baseline performance on the proposed dataset is demonstrated. In the future work, we plan to investigate representations for the joint space of color and depth information and action models adapting to viewpoint variations.

Acknowledgments. This work was supported in part by National Basic Research Program of China (973 Program): 2009CB320906, in part by National Natural Science Foundation of China: 61025011, 61035001, 61003165 and 61133003, and in part by Beijing Natural Science Foundation: 4111003.

References

1. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV 2005, vol. 2, pp. 1395–1402. IEEE (2005)
2. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR 2004, vol. 3, pp. 32–36. IEEE (2004)
3. Laptev, I.: On space-time interest points. IJCV 64(2), 107–123 (2005)
4. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV 79(3), 299–318 (2008)
5. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. In: ICCV 2007, pp. 1–7. IEEE (2007)
6. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: ICCV 2009, pp. 104–111. IEEE (2009)
7. Singh, S., Velastin, S., Ragheb, H.: Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: International Conference on Advanced Video and Signal Based Surveillance, pp. 48–55. IEEE (2010)
8. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: CVPRW, pp. 9–14. IEEE (2010)
9. Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from rgbd images. In: AAAI Workshop on PAIR (2011)
10. Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: IEEE Workshop on Consumer Depth Cameras for Computer Vision in conjunction with ICCV (2011)
11. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR, vol. 2, p. 3 (2011)
12. Zhang, H., Parker, L.: 4-dimensional local spatio-temporal features for human activity recognition. In: IROS, pp. 2044–2049. IEEE (2011)
13. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
14. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009 (2009)
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR 2006, vol. 2, pp. 2169–2178. IEEE (2006)