A Non-Linear Mapping for Representing Human Action Recognition Under Missing Modality Problem in Video Data

Aidin Gharahdaghi[a], Farbod Razzazi[a], Arash Amini[b]

[a]*Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[b]*EE department, Sharif University of Technology, Tehran, Iran.*

Abstract

Human action recognition based on standard video files is a well-studied problem in the literature. In this study, we assume to have access to single modality standard data of some actions (training data). Based on this data, we aim at identifying the action involved in target modality video data without the source-target relationship information. In this case, the training and test phases of the recognition task are based on different imaging modalities. Our goal in this paper is to introduce a mapping (a nonlinear operator) on both modalities such that the outcome shares some specific features. These common features were then used to recognize an action in each domain. Simulation results on MSRDailyActivity3D, MSRActionPairs and UTKinect-Action3D Dataset datasets showed that the introduced method outperforms state-of-the art methods with a success rate margin of 15% on average.

Keywords: Depth Information, Human Action Recognition, RGB Video, Transfer Learning, Non-Linear Mapping.

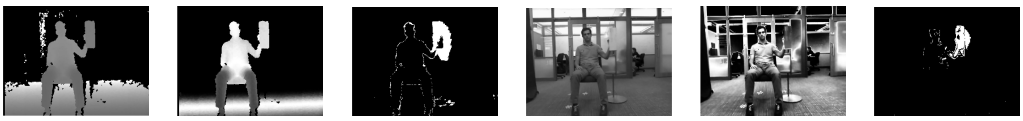
1. Introduction

In recent years, various approaches have been proposed for identifying human action. The goal in these approaches is to automatically analyze the activities from a video recordings (i.e. a sequence of image frames). While the RGB modality is the dominant imaging format, other modalities such as the skeleton-based and depth images are also used. The cost-effectiveness

and availability of Microsoft Kinect has led to popularity of depth imaging in parallel to RGB imaging. Indeed, Kinect sensors have found applications in consumer electronics including smart automobiles, health care, surveillance and activity recognition [1]. The simultaneous imaging of RGB and depth modalities is commonly referred to as RGB-D data format. The existence of the additional depth information, is likely to equip RGB-D data with a better human action recognition compared to the conventional RGB data [2]. Besides, we have witnessed remarkable progress in the field of machine learning in recent years which potentially enables us to take advantage of the additional information in the RGB-D format.

One of the challenging problems in RGB-D action recognition is the missing modality issue. This happens when one of the modalities becomes temporarily unavailable. For instance, during the night, RGB recordings become very noisy and could be even ignored in low-light situations. Now the challenge is to identify an action based on only one modality (e.g., depth during the night) by having access to a database of valid RGB-D data of multiple actions. The existence of RGB-D dataset could potentially increase the accuracy of the action recognition beyond that of a method based solely on a single modality. One possible approach is to relate two modalities and employ this relation to retrieve the other modality when missing [3], [4]. Another approach is the multimodal transfer learning based on an existing RGB-D dataset and extending the technique to the available problem [5], [6]. In transfer learning, since the training and test data belong to two separate domains, it is common to call them source and target domains, respectively. Supervised transfer learning is a subfield of transfer learning in which the target domain contains a limited part with labeled data. The use of transfer learning in the missing modality problem consists of two steps: 1) transferring the knowledge from source database to the target database, and 2) transferring the knowledge from the source modality to the target modality [7]. The outcome could be considered as multimodal transfer learning. Chengcheng et al, have proposed a transfer learning approach is presented for the missing modality problem in which the missing modality is treated as latent information in the target domain. Using a tensor-based framework, these latent information are recovered via rank-minimization. The above studies highlight that the representation of the two modalities and their connections have a great impact on the accuracy of the method [8].

In this work, we consider a somewhat more difficult problem. We assume to know a database of the actions only in one modality (i.e., either RGB



(a) (b) (c) (d) (e) (f)

Figure 1: Example of steps for RGB and depth sequence in proposal method a) normalized depth frame b) mapping c) adaptive background subtraction for rgb sequence d) gray scale frame e) mapping f) adaptive background subtraction for depth sequence

or depth, and not both); then, for a given recording of action in the other modality, we shall determine the action. In other words, unlike the transfer learning approach, we do not have access to a database of joint RGB-D modalities. Our approach is to map two modalities into a lower-dimensional subspace such that the representations in two modalities become similar. With this technique, we eliminate the need of having an auxillary RGB-D dataset. This mapping consists of finding the significant parts of the scene in which the action is taking place and evaluating the HOG features of the cropped videos. A KNN classifier is ultimately used to determine the action. To the best of our knowledge, the missing modality problem without an auxillary database is not studied in the past. For an illustrative example, we have plotted the procedures in Figure 1: the RGB and depth modalities of a frame are shown in (a) and (b) subfigures, respectively. The outcome of our method on these two modalities is shown in subfigures (h) and (d), respectively. It is evident that the two outcomes are very similar, which greatly facilitates the classification task.

1.1. Related works

The problem has been studied at two major levels of complexity: 1)actions and 2)activities. Actions are characterized by simple motion patterns typically executed by a single human. Activities are more complex and involve coordinated actions among a small number of humans [9]. Feature representation methods have been developed for recognizing actions from video sequences based on color scale cameras. Ahad et al [10], have proposed, silhouettes are temporally accumulated to form motion energy images (MEIs) and motion history images (MHIs). Alp et al [11] they were extracted from both MEIs and MHIs as action descriptors. Used Gaussian mixture models (GMM) to capture the distribution of the moments of silhouette sequences was proposed by Davis in [12]. Several other approaches utilize motion flow

patterns to represent human actions. Typically, optical flows are calculated for the entire image by matching consecutive video frames. In addition, series of Spatio-temporal interest points (STIPs) based methods have been proposed, which achieve state-of-the-art performances in activity recognition. In addition ,optical flow [13], context [14], HoG/HOF [15] and extended SURF [16], can be considered as a clustering of low-level features. There are a set of useful features for the depth data, such as depth map [17], RGB-D points [4], HON4D [3]. Compared with the RGB data, the depth data can separate the background and foreground according to the hierarchical distance to the camera. these are useful method, base on representing video as feature or descriptor. Another method was two feature representation methods has been introduced for color and depth information fusing for activity recognition, which is developed in a straight forward manner from two state-of-the-art action representation methods, i.e., spatial-temporal interest points (STIPs) and motion history images (MHIs). The idea which the fusion of color and depth information is suitable for the task that has these two modalities [18].

Other approaches belong to the transfer learning task. Missing modality problem in transfer learning is defined as the target modality is unavailable in the training stage, while only the source modality can be obtained in this stage. Recently, low-rank matrix constraint [19], [20] has been introduced into the transfer learning problem in the image processing task. It can be revealed that the subspace structure of both source and target data can be achieved through the locality aware reconstruction. This reconstruction keeps guiding the knowledge transfer in a latent shared subspace. Some transfer learning methods take advantage of unlabeled data by predicting their labels. There are two situations for the target labels, one in the categories of source and target data are the same. Ding et al [21] have proposed a deep transfer learning method in two domains and have predicted target labels. They combined labeled and unlabeled data, and predicted the labels each time to reconstruct the whole dataset. The second situation is in the case that the categories of source and target data are different. A semi-supervised model by transferring semantic attributes have proposed by Rohrbach et al [22]. They was exploiting the manifold structure of target data to improve the prediction of unlabeled data. Their method could predict the new categories of data in one domain. The important point in this method is the source domain data that can help in recognizing the target domain data.

LTSL[19] and LRDAP[20] are two typical transfer learning methods that

include the low-rank constraint. LTSL aims to find a common subspace where the source data can represent the target in the low-rank framework well. LRDAP aims to find a rotation on source domain data to be represented by the target domain in the low-rank framework. LRDAP considers that the rotated source data can be used to test the target data when the rotated source data can be reconstructed in the target domain. In both cases, the role of the auxiliary data is very important. Different from other studies, jia proposed method which have transfered common properties between two domains, and aligned source and target domains by learning a coefficient matrix, which has been optimized by two graphs. One graph reflects the manifold structure of labeled source data, while the other is generated by the neighbors of unlabeled target data. Since the two graphs are cooperated to update the coefficient matrix, the graph with unlabeled target data is guided by the other with label information, therefore they could obtain more reliable predicted labels of the target data. Most of these approaches need another dataset which can be time-consuming and may have a very different distribution between two domains. Also, we purpose a non-linear mapping to dismiss Malicious information and special attention to the area that has more information about the act.

1.2. Our Contributions

Different from the previous studies, our method aims to solve the missing modality problem by:

• proposing a nonlinear mapping between two domains .

• Using an adaptive background subtraction to bring both modalities closer together.

• Our method do not use the relation information between RGB and depth data by another auxiliary database.

For the classification task, we used the HoG feature for both new representations of depth and RGB and it was shown that is more efficient than other methods. To the best of our knowledge, we are the first to consider the missing modality problem in the human action recognition framework, by recovering the relation between two modalities. The result shows classification accuracy on average in all experiments 15% higher than other methods.

1.3. road map

In section II, we express the process on RGB and depth data by mapping and adaptive background subtraction. We introduced a non-linear mapping

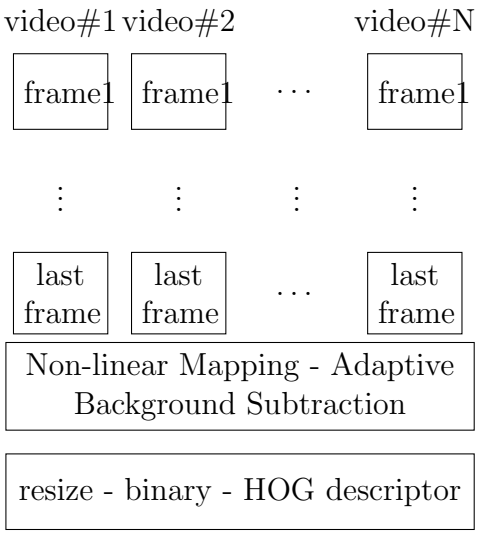


Figure 2: An overview of the process performed

and prepare the data for the next stage in this section. In section III, we introduce datasets and experimental settings. Then, we evaluate the accuracy of our proposed method and compare it with other methods. Finally, section IV concludes the paper.

2. Method

To address the Missing Modality Problem, we need to find a method to represent two modalities that are as similar as possible, so we could assume that we have just one modality. To this end, the non-linear mapping on both modalities and adaptive background subtraction could help to find more similarities between the two modalities.

we first resized the frames of depth and RGB data. This makes the video frames have the same size and normalizes the amplitudes of the images to the range [0,1]. Given action databases with N samples of the video in two modalities; we denote $\mathcal{X}$ for RGB and $\mathcal{Y}$ for depth representations respectively. $\mathcal{X} = \{X_1, X_2, ...X_M\}$ and $\mathcal{Y} = \{Y_1, Y_2, ...Y_M\}$ where M is the number of videos and each sample X_i or $Y_i \in \mathbb{R}^{a \times b \times N}$ where a, b are the width and length of each frame which are equivalent in both modalities and N is the number of frames in each sample.

The steps of the proposed method in both RGB and depth modalities are illustrated in figure 2. The most important step in our work is applying nonlinear mapping to the whole video. In video for both modalities, due to the detection of movements by the difference of frames, we can see which intensity level or in depth modality which distance, could have more information about the actions. To estimate it, we calculate the area that action happens with the help of consecutive frames. It yields the 3 dimensional binary mask, which helps us to estimate the most important position related to actions. Thus,

$$B_i = |X_i - X_{i-1}| > \theta_1, \quad i = 2, 3, ..., N \tag{1}$$

θ_1 is the threshold for RGB modality to consider the amount of difference in the image sequence. This value is obtained empirically. But for depth data, due to the nature of the depth images, some sort of noise like shot noise will appear. To achieve a result more similar to the RGB modality, we must overcome this type of noise as much as possible. Since the difference in images in two consecutive frames does not have sudden and large changes, we could consider changes smaller than a threshold like 0.7. It makes sense when the action happened, the difference between two consecutive frames is not large in amplitude and when the variation is too high, it is due to such noise, so we used an upper limit in depth modality to avoid this noise.

$$B_i = (|Y_i - Y_{i-1}| < 0.7) > \theta_1, \quad i = 2, 3, ..., N \tag{2}$$

After we calculated the binary mask for both modality, we multiplied it to whole video frames, if the input data is RGB, then:

$$V_i = B_i . * X_i, \quad i = 2, 3, ..., N \tag{3}$$

and if the input data is depth:

$$V_i = B_i . * Y_i, \quad i = 2, 3, ..., N \tag{4}$$

Now $V \in \mathbb{R}^{a \times b \times (N-1)}$ is zero valued for unimportant data which have no action was occurred there and have a non zero intensity level or distance value for the important part of the action video. This cube is now suitable to estimate the parameters of our non linear mapping in both modalities.

2.1. NON LINEAR MAPPING

To Emphasis and consider the important part of intensity for grayscale and distance for depth modalities, we must estimate the action variations intensity or distance. The approach to estimating these variations is to use a Gaussian mapping which has the mean and variance values of the intensity (in RGB modality) and distance(in depth modality) changes. The spatial-temporal mask which calculates in the previous step is now a good guide to find out these parameters. The mean value is selected as the most occurring level that appears in the movement of a non zero V. So, we calculated the average of the whole non zero elements in V. For the variation of action, we calculated the variance of non-zero elements in the action. A small amount of variance indicates small actions like pushing or pulling and big actions that have bigger variance is expressible as activities like walking around or push the chair, etc. So if we consider G as non zero elemnts of V, hence we can write:

$$\theta = \frac{1}{N-1}\sum_{i=2}^{N}(G_i) \quad , \quad \sigma^2 = \frac{1}{N}\sum_{i=2}^{N}(G_i - \mu)^2. \tag{5}$$

Where θ and σ^2 are the mean and variance of G, respectively. After calculating these parameters, now we employ non linear mapping. Thus:

$$D_i = \exp\big(-(1/\sigma) * (X_i - \theta)^2\big), \quad i = 2, 3, ...N \tag{6}$$

and for depth images:

$$D_i = \exp\big(-(1/\sigma) * (Y_i - \theta)^2\big), \quad i = 2, 3, ...N \tag{7}$$

This mapping converts the space of original videos into a space in which the intensity level or distance is emphasizes and in addition, we greatly reduced the effect of shot noise in depth modality. To express this issue, figure 3 shows the surface of one frame in depth modality before and after mapping.

2.2. Adaptive Background Subtraction

After non-linear mapping, we represented the action by the shape of changes in intensity or depth modalities of image sequences. as:

$$output_i = |S(i-1) - D(i)| > \theta_2, \quad i = 2, 3, ...N \tag{8}$$

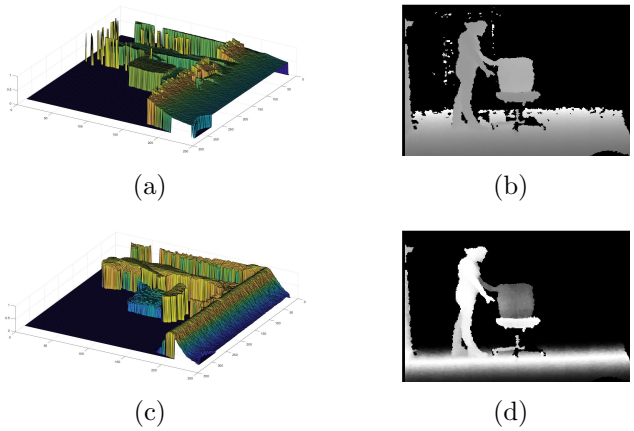


Figure 3: a)surface of one depth frame b)image of depth frame c)surface of same frame after mapping by Gaussian transform d)image after mapping

where D_i is the current frame and S_{i-1} is the result of the previous background subtraction frame and θ_2 is the threshold used to binarized the difference values. Adaptive background subtraction Helps to identify temporal changes and somehow preserve the history of actions. History of changes that will be preserve, depends on frame number of videos and level of action's movement. Thus, S_i defines as:

$$S_i = \alpha * D_i + \beta * S_{i-1}, \qquad i = 2, 3, ...N \tag{9}$$

The value of α and β were obtained empirically. these value determining the weighted effect of the previous frame and current frame in the formation of background subtraction framework. For the first step, S is equivalent to frame number 2. The process going from frame number 2 until the end for both modalities. The entire process can consider a non-linear operation too. The last step is, to extract 8 frames at specific intervals. This reduces the sensitivity to pick exactly one frame in each modality. In other words, the robustness of the algorithm was increased to timing misalignment of the frame in videos of two modalities. The reason for such resistance is the use of adaptive background subtraction for the final display of each video in both modalities. The details of the algorithm are outlined in Algorithm 1 for both modalities. After this processing for all train videos, we resize the videos and binarized them. The final threshold for binarization after resizing is determined experimentally.

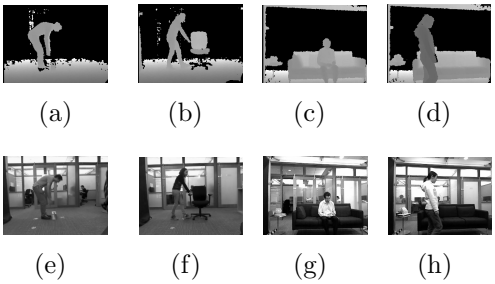


Figure 4: RGB and depth samples from the datasets: a,b) depth frames from MSR ActionPAIR, c,d) depth frames from MSRDailyActivity3D, e,f) RGB frames from MSR ActionPAIR, g,h) RGB frames from MSRDailyActivity3D

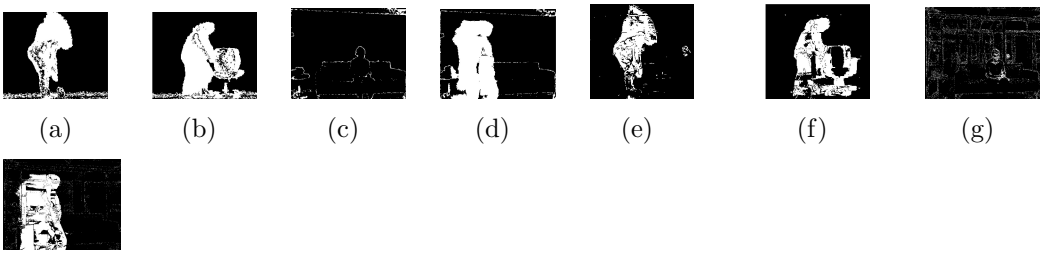


Figure 5: Proportional frames in the figure 4, after proccess. The shadow that appears is the result of adaptive background subtraction. in (c),(g) the actions is difficult to identify due to law movement and small in analogy or the whole scene.

3. Experiments

3.1. DATASET

For the purpose of experiments, we consider three action datasets: MSRDailyActivity3D[1][23], MSRActionPairs[2][3] and UTKinect-Action3D Dataset[3][24]. Each dataset contains the two modalities of RGB and depth. In the MSRDailyActivity3D dataset, there are 16 categories of actions performed by 10 subjects, each performing every action twice. There are 320 RGB samples as well as 320 depth samples. In the MSRActionPairs dataset, there are six pairs of actions performed by 10 subjects with three trials each.

[1]http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/
[2]http://www.cs.ucf.edu/ oreifej/HON4D.html
[3]http://cvrc.ece.utexas.edu/KinectDatasets/HOJ3D.html

In total, there are 360 RGB and 360 depth action samples. Finally, in the UTKinect-Action3D dataset, there are 10 subjects performing each of 10 action types twice. In Figure 4 we show RGB and depth samples from each of the mentioned datasets. In addition, we show the output of our non-linear mapping and background subtraction in Figure 5.

3.2. Competing methods

To better illustrate the results of the proposed method, we compare them with the results of LTSL, GFK [25], HOG-KNN [26], SALAD [27], LTTL [8] and SLRTL[8]. Some highlights of these methods are: LTSL learns a common subspace by transferring information between two domains via a low-rank constraint. In GFK, the common subspace is found by maximizing a correlation measure between source and target domains. The HoG-KNN is a benchmark without transfer learning, i.e., the source dataset is not used in the training phase. SALAD focuses on domain adaptation where the actions in the source and target datasets are the same with known source action labels. Latent tensors and low-rank constraints are incorporated in LTTL[8] to estimate the missing modality in a low dimensional subspace. SLRTL is the same as LTTL with the exception that the action labels are known.

3.3. Results and Discussions

To study the action recognition performance of the proposed method and other referenced approaches, we consider two main setups: training based on the RGB modality and testing on the depth modality (results in Table[1]) and training based on the depth modality and testing on the RGB modality. Moreover, for each test, once we use the first half of the subjects as training data and the other half as the test data, and then, we swap the train and test data and repeat the experiments again. We should highlight that the experiments are applied according to the provided settings in [8] and therefore, the results (besides the proposed method) are directly reported from this reference.

For the results of the proposed method, we recall that the last step in Figure 2 was to resize each frame and make a binary map by applying a thresholding scheme. To find the optimal resize dimensions, we have considered resizing to dimensions $[80, 50]$, $[60, 40]$, $[60, 30]$, $[50, 30]$, $[50, 25]$, and $[40, 30]$, with equal thresholds. Overall, the best dimension for highest accuracy (averaged over all 4 experiments) is observed to be $[50, 25]$. Therefore, resizing to $[50, 25]$ is consistently used in all the reported results in figure 6.

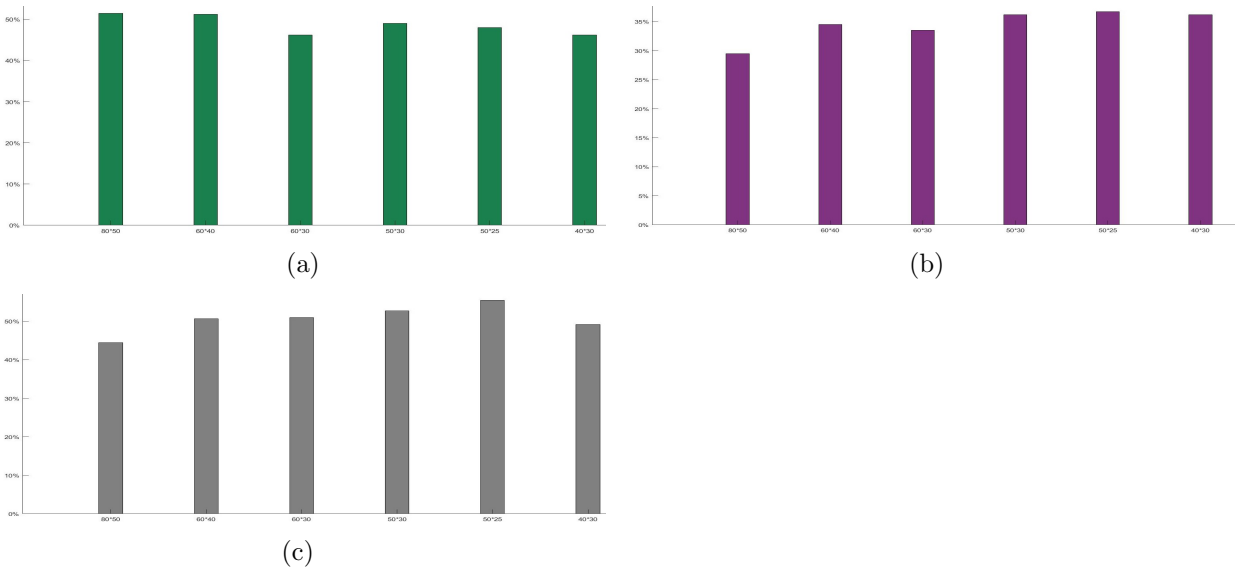


Figure 6: Average accuracy in all 4 experiments per data set based on frame size changes. a)UTKinect-Action3D Dataset b)MSRDailyActivity3D Dataset c)MSRActionPairs Dataset

It is worth mentioning again that we derive the HoG features of the resized frames and employ a traditional KNN for classifying the action.

Table 1 shows the results of all the compared methods when RGB modality is used for training to estimate the actions in depth modality (test data). The results confirm the superior accuracy of the proposed method compared to the other techniques in all experiments. Our method achieves its highest accuracy when applied to the MSRActionPairs dataset; meanwhile, MSRDailyActivity3D seems to be the most challenging dataset, as we observe the worst accuracy here. The poor performance of our method for this dataset has three main reasons

1. As explained earlier, the accuracy of recognition in our method is directly linked to the level of movements. Therefore, the actions of drinking, eating, talking-with-mobile, reading and writing are expected to be similar (potential miss classification between the actions); in the same way, due to the lack of significant movements, the actions of still and working with a laptop are also similar.
2. Our training dataset consists of a single subject performing an action; therefore, when a second subject (human or and object) makes a move-

Table 1: accuracy of test1: RGB-depth on three action databases

Dataset	MSRdaily Activity3D		MSRAction Pairs		UTKinect-Action	
Training data	1st half	2nd half	1st half	2nd half	1st half	2nd half
METHODES						
GFK	22.50	16.25	18.18	11.86	25.00	30.50
LTSL	06.88	05.00	10.80	07.34	11.50	21.00
HOG-KNN	28.75	26.25	22.73	24.29	17.50	25.50
SALAD	27.67	27.67	30.43	30.43		
LTTL	29.38	34.38	35.23	31.07	28.50	25.50
SLRTL	29.38	34.38	35.23	31.07	33.00	28.50
OURS	**39.37**	**38.75**	**65.55**	**55.55**	**40.00**	**49.00**

ment, we have a combination of movements and our method is likely to miss clasify the first subject's action. We observe that some of the test data in the MSRDailyActivity3D dataset are corrupted by second human / object movement; an example is when someone throws an object or type on a laptop while another person is walking around.

3. The MSRDailyActivity3D dataset is mainly used for object recognition and not action recognition [28]. This dataset consists of actions with very similar nature such as "talking on the phone while walking" and "just walking", which are difficult to distinguish particularly, in our method that employs only 8 frames from a video. We should highlight that the type of movements in MSRActionPairs dataset are more distinct and easier to distinguish.

Table 2: accuracy of test2: depth-RGB on three action databases

Dataset	MSRdaily Activity3D		MSRAction Pairs		UTKinect-Action	
Training data	1st half	2nd half	1st half	2nd half	1st half	2nd half
METHODES						
GFK	15.63	14.38	07.44	19.17	18.00	18.00
LTSL	05.63	05.00	07.44	12.50	19.00	18.00
HOG-KNN	17.50	16.88	12.40	22.50	33.00	28.00
SALAD	31.97	31.97	30.98	30.98		
LTTL	**35.00**	31.88	23.14	23.33	35.00	28.00
SLRTL	35.00	31.88	28.10	23.33	35.00	28.00
OURS	34.37	**34.37**	**58.00**	**45.00**	**58.00**	**43.00**

To justify the superior performance of our method compared to the competing methods is that these methods make use of transfer learning, which in turn requires a considerable overlap between the source and target datasets. This requirement is however, widely violated in our experiments; for instance, we have "pushing a chair" in the source dataset while the closest counter part in the target dataset is "sitting on a sofa". The difference between source and target image modalities also further complicates the task of transfer learning.

The UTKinect-Action dataset has short-length action videos with missing frames. Therefore, the tuned α and β parameters (in adaptive background subtraction stage) for the previous datasets no longer achieve desirable results here. To better accumulate the motions in our combined frames, we consider $\alpha = 0.9$ and $\beta = 0.1$ here, which amount to longer memory for keeping the actions (in contrast to $\alpha = 0.1$ and $\beta = 0.9$ for the previous datasets).

According to [28], the most challenging multimodal action recognition is when the image modality in the source dataset is depth and the image modality in the target dataset is RGB. The rationale is that the depth images are quite noisy and less informative that the RGB images. In our method, by applying a denoising stage, we have tried to decrease the information gap between the two modalities. As a result, in Table 2 we observe that the most and least challenging multimodal action recognition tasks have comparable accuracy levels in our method. It is worst highlighting that our method does not make use of any auxiliary dataset to enhance the accuracy level.

4. CONCLUSION

Unlike previous studies, our method aims to solve the missing modality problem by finding an intersection of the two modalities which is yet information-preserving. For this purpose, we proposed a non-linear mapping and adaptive background subtraction to form a binary representation. The representation actually captures the movements involved in an action. The important point was that the aimed representation is achieved without having access to an RGB-D dataset of actions. Through a number of experiments, we showed that the proposed method outperforms the existing techniques in terms of accuracy. In particular, it marks great improvements when trained on a depth dataset and tested on an RGB dataset.

Since our method solely takes the movements into account, it makes mistakes when two actions consist of similar movements. Therefore, as a future work, one can combine the movements with other features such as the involved objects to improve the accuracy.

[1] M. L. Gavrilova, Y. Wang, F. Ahmed, and P. Polash Paul, "Kinect Sensor Gesture and Activity Recognition: New Applications for Consumer Cognitive Systems," *IEEE Consumer Electronics Magazine*, vol. 7, no. 1, pp. 88–94, 2018.

[2] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust Part-Based Hand Gesture Recognition Using Kinect Sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[3] O. Oreifej and Z. Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition From Depth Sequences," in *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013.

[4] W. Li, Z. Zhang, and Z. Liu, "Action Recognition Based On a Bag Of 3D Points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 9–14.

[5] C. Sun, B.-K. Bao, and C. Xu, "Knowing Verb From Object: Retagging With Transfer Learning On Verb-Object Concept Images," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1747–1759, 2015.

[6] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning Personalized Models For Facial Expression Analysis And Gesture Recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 775–788, 2016.

[7] Z. Ding, M. Shao, and Y. Fu, "Missing Modality Transfer Learning via Latent Low-Rank Constraint," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4322–4334, 2015.

[8] C. Jia, Z. Ding, Y. Kong, and Y. Fu, "Semi-Supervised Cross-Modality Action Recognition by Latent Tensor Transfer Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.

[9] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[10] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, "Human Activity Analysis: Concentrating On Motion History Image And Its Variants," in *2009 ICCAS-SICE*, 2009, pp. 5401–5406.

[11] E. C. Alp and H. Y. Keles, "Action Recognition Using MHI Based Hu Moments With HMMs," in *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, 2017, pp. 212–216.

[12] J. W. Davis and A. Tyagi, "Minimal-latency Human Action Recognition Using Reliable-Inference," *Image and Vision Computing*, vol. 24, no. 5, pp. 455–472, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885606000461

[13] Y. Wan, Z. Miao, X. Zhang, Z. Tang, and Z. Wang, "Illumination Robust Video Foreground Prediction Based on Color Recovering," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 637–652, 2014.

[14] C. Yuan, X. Li, W. Hu, H. Ling, and S. J. Maybank, "Modeling Geometric-Temporal Context With Directional Pyramid Co-Occurrence for Action Recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 658–672, 2014.

[15] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic Model Vectors for Complex Video Event Recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 88–101, 2012.

[16] G. Willems, T. Tuytelaars, and L. Van Gool, "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 650–663.

[17] V. Megavannan, B. Agarwal, and R. V. Babu, "Human Action Recognition Using Depth Maps," in *2012 International Conference on Signal Processing and Communications (SPCOM)*, 2012, pp. 1–5.

[18] S. Zhu and L. Xia, "Human Action Recognition Based on Fusion Features Extraction of Adaptive Background Subtraction and Optical Flow Model," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–11, 2015.

[19] M. Shao, C. Castillo, Z. Gu, and Y. Fu, "Low-Rank Transfer Subspace Learning," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 1104–1109.

[20] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang, "Robust Visual Domain Adaptation With Low-Rank Reconstruction," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2168–2175.

[21] Z. Ding, M. Shao, and Y. Fu, "Deep Low-Rank Coding For Transfer Learning," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[22] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer Learning In a Transductive Setting," in *Advances in neural information processing systems*, 2013, pp. 46–54.

[23] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining Actionlet Ensemble For Action Recognition With Depth Cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[24] L. Xia, C. Chen, and J. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[25] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic Flow Kernel For Unsupervised Domain Adaptation," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2066–2073.

[26] C. Jia, Y. Kong, Z. Ding, and Y. R. Fu, "Latent Tensor Transfer Learning For RGB-D Action Recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 87–96.

[27] S. Schneider, A. S. Ecker, J. H. Macke, and M. Bethge, "Multi-Task Generalization And Adaptation Between Noisy Digit Datasets: An Empirical Study," in *Neural Information Processing Systems (NeurIPS), Workshop on Continual Learning*, 2018.

[28] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D Based Action Recognition Datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86 – 105, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316301029

Algorithm 1 New RGB/depth data representation

Require: $\mathcal{X} = \{X_1, X_2, ...X_M\}$ as the normalized RGB tensors, or $\mathcal{Y} = \{Y_1, Y_2, ...Y_M\}$ as the normalized depth tensors with M videos

N= the number of frames

α, β=weights for adaptive background subtraction

θ_1, θ_2= thresholds used in the nonlinear mapping and adaptive background subtraction

-Nonliner mapping

For $i = 2 : N$

if input is RGB: $B_i = |X_i - X_{i-1}| > \theta_1$

if input is depth: $B_i = (|Y_i - Y_{i-1}| < 0.7) > \theta_1$

end

if input is RGB: $v_i = B_i .* X_i$

if input is depth: $v_i = B_i .* Y_i$

Define: $G \leftarrow V_{\neq 0}$

Define: $\sigma^2 = \frac{\sum_{i=2}^{N}(G_i - \mu)^2}{N}, \theta = \frac{\sum_{i=2}^{N}(G_i)}{N-1}$

$for\ i = 1 : n$

for RGB data:

D(i)$= exp(-(1/2 * \sigma^2) * (Y_i - \theta)^2)$

for depth data:

D(i)$= exp(-(1/2 * \sigma^2) * (Y_i - \theta)^2)$

end

-Adaptive Background Subtraction

$for\ i = 2 : n, \quad S_1 = D_1$

volum(i) $\Longleftarrow |S_{i-1} - D_i| > \theta_2$

$B_i = \alpha * D_i + \beta * S_{i-1}$

end

output: volum$\in \mathbb{R}^{a \times b \times N-1}$