A Unified Framework for Multi-Modal Isolated Gesture Recognition

Jiali Duan, CBSR & NLPR, Institute of Automation, University of Chinese Academy of Sciences
Jun Wan*, CBSR & NLPR, Institute of Automation, University of Chinese Academy of Sciences
Shuai Zhou, Macau University of Science and Technology
Xiaoyuan Guo, School of Engineering Science, University of Chinese Academy of Sciences
Stan Z. Li, CBSR & NLPR, Institute of Automation, University of Chinese Academy of Sciences

In this paper, we focus on isolated gesture recognition and explore different modalities by involving RGB stream, depth stream and saliency stream for inspection. Our goal is to push the boundary of this realm even further by proposing a unified framework which exploits the advantages of multi-modality fusion. Specifically, a spatial-temporal network architecture based on consensus-voting has been proposed to explicitly model the long term structure of the video sequence and to reduce estimation variance when confronted with comprehensive inter-class variations. In addition, a 3D depth-saliency convolutional network is aggregated in parallel to capture subtle motion characteristics. Extensive experiments are done to analyze the performance of each component and our proposed approach achieves the best results on two public benchmarks–ChaLearn IsoGD and RGBD-HuDaAct, outperforming the closest competitor by a margin of over 10% and 15% respectively. We will release our codes to facilitate future research.

CCS Concepts: •**Computing methodologies** → **Activity recognition and understanding;** •**Human-centered computing** → Walkthrough evaluations;

Additional Key Words and Phrases: Multi-Modal, Consensus-Voting, 3D Convolution, Isolated Gesture Recognition

ACM Reference Format:
Jiali Duan, Jun Wan, Shuai Zhou, Xiaoyuan Guo, Stan Z. Li, 2017. A Unified Framework for Multi-Modal Isolated Gesture Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 4, Article 39 (January 2017), 16 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Gesture recognition is a fast expanding field with applications beyond gaming, consumer electronics UI, automotive, sports training and etc. Continuous gesture recognition [Yin and Davis 2014; Evangelidis et al. 2014; Molchanov et al. 2016] and isolated gesture recognition [Molchanov et al. 2015; Wan et al. 2016; Wan et al. 2014] are the two main tasks of gesture recognition in computer vision and the former can be converted to the latter once temporal segmentation is performed on continuous gestures.

In this paper, we concentrate on isolated gesture recognition, especially for RGB-D video input, see Fig.1 for an illustration. As is noted in [Zhu and Newsam 2016; Shahroudy et al. 2016; Wan et al. 2016], depth sequence contains structural information from the depth channel and are more capable of dealing with noises from background, clothing, skin color and other external factors, therefore acting as a supple-

Author's addresses: Jiali Duan (jli.duan@gmail.com) and Jun Wan* (Corresponding Author: jun.wan@ia.ac.cn) and Stan Z. Li (szli@nlpr.ia.ac.cn), Institute of Automation, University of Chinese Academy; Shuai Zhou, Macau University of Science and Technology; Xiaoyuan Guo, School of Engineering Science, University of Chinese Academy of Sciences.

DOI: 0000001.0000001

Fig. 1. Examples of different types of modalities listed from up to bottom: RGB images, Depth images, Optical flow fields (magnitudes obtained from x and y direction are used as color channel), and Saliency

ment to the original RGB sequence. Unlike previous works [Ng et al. 2015; Zhang et al. 2016; Wang and Schmid 2013] which focus purely on RGB modality, we try to build a unified framework that exploits the inherent advantages of multi-modality modelling, namely RGB stream, depth stream and saliency stream.

Our main motivations are: *1) how to reduce estimation variance when it comes to classifying videos of comprehensive inter-and-intra class variations; 2) how to design a unified and effective framework that is able to take advantage of different modalities.*

For the first problem, we notice that unlike other video recognition tasks such as action recognition, which contains relatively rich contextual information of body correlations and interactions, the task of gesture recognition usually involves only the motion of hands and arms. In other words, existing gesture recognition methods [Molchanov et al. 2016; Molchanov et al. 2015; Pigou et al. 2015] which deal with a limited number of gestures can make very "biased" estimations when it comes to classifying gesture datasets that involve comprehensive inter-and-intra class variations like ChaLearn IsoGD [Wan et al. 2016]. Second, current main-stream approaches such as [Simonyan and Zisserman 2014a; Donahue et al. 2015] usually deal with short-term motions, possibly missing important information from actions that span over a relatively long time. For example, some gestures such as "OK" or number signals involve only motions of a short period while gesticulations denoting forced landing, diving signals or slow motions require temporal modeling of a relatively long sequence.

To solve the aforementioned issue, we propose a novel two stream convolutional network (2SCVN) based on the idea of consensus voting adapted from [Simonyan and Zisserman 2014a; Wang et al. 2016; Feichtenhofer et al. 2016b]. It first takes frames sampled from different segments of the sequence according to uniform distribution and stacks their corresponding optical flow fields as input. Compared to dense sampling or pre-defined sampling interval which may be highly redundant, this leads to less computations and ensures that videos which are short or those which involve multiple stages can be completely covered fairly well. These frames are then combined to cover more diversity before being fed into the spatial and temporal streams of 2SCVN for

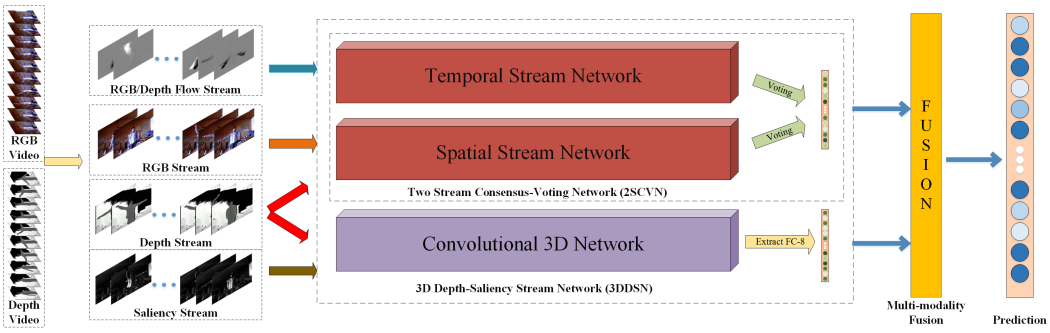


Fig. 2. An overview of our approach. An input video is represented in different modalities, where the RGB stream and depth stream are handled by *Spatial Stream Network* while the RGB/Depth Flow Stream are handled by *Temporal Stream Network*. Saliency stream and depth stream are fed into *Convolutional 3D Network*.

video level predictions. Finally, these predictions are aggregated to reduce estimation variance.

For the second problem, we realize that as human motions are in essence three-dimensional, the information loss in the depth channel could cause degradations to the discriminative capability of feature representation. On the other hand, saliency helps eliminate ambiguity from possible distractions of color camera. We show in our experiment that saliency information helps boost the overall performance further. To the best of our knowledge, we are the first to perform investigations that highlight spatial and temporal combinations from these two modalities, based on which 3D depth-saliency (3DDSN) fusion scheme is proposed. Eventually, predictions from both 2SCVN and 3DDSN are taken into consideration as the final score. What's worth noticing is that our proposed approach also works surprisingly well for other tasks of video recognition (See Table III), demonstrating the effectiveness and generalization ability of our framework.

The overall pipeline is shown in Fig.2. It mainly consists of two network architectures–Two Stream Convolutional Network (2SCVN) and 3D Depth-Saliency Convolutional Network (3DDSN) which would be further elaborated in Section 3.1 and Section 3.2 respectively. The 2SCVN tries to model the spatial-temporal information from a given modality through two streams. The spatial stream takes into account RGB and depth sequences while RGB-Flow and Depth-Flow sequences are fed as input to the temporal stream. Scores from both the spatial and temporal stream are taken into consideration as the prediction of 2SCVN. Additionally, we absorb 3D convolution to implicitly capture both spatial and temporal information. In our network architecture, we input depth and saliency modality to 3DDSN and do late score fusion for outputs from these two streams. Finally, scores from 2SCVN and 3DDSN are further aggregated as the eventual video-level prediction. The main contributions of our paper are:

(1) We proposed a novel framework that combines the merits of 2SCVN and 3DDSN for multi-modality fusion. It absorbs depth and saliency streams as important constituents to capture subtle spatial-temporal information supplementary to RGB sequence.
(2) A convolutional network design (2SCVN) based on the idea of consensus voting is proposed to explicitly model the long term structure of the whole sequence, where video-level predictions from each frame and its augmented counterparts are aggregated to reduce possible estimation variance.

(3) We are the first to perform an integration of 3D depth-saliency stream to address the loss of three-dimensional structural information and distractions from backgrounds, noises and other external factors.
(4) Our approach performs particularly well not only for RGB-D gesture recognition but also for human daily action recognition, achieving the best results on ChaLearn IsoGD [Wan et al. 2016] and RGBD-HuDaAct [Ni et al. 2011] benchmarks.

The remainder of this paper is organized as follows: Section 2 is an review of related works. Our unified framework is illustrated in Section 3 and validated in Section 4 respectively. Section 5 concludes the paper.

2. RELATED WORK

In this section, we first introduce some works for gesture recognition deploying hand-crafted features, then we review works related to ours both in the field of action and gesture recognition that conduct research on different modalities and convolutional networks.

Many hand-crafted features have been proposed for video analysis in the area of gesture recognition [Starner et al. 1998; Wang et al. 2006; Dardas and Georganas 2011; Trindade et al. 2012; Wan et al. 2013]. For example, Wan et al [Wan et al. 2016] extracted a novel spatiotemporal feature named MFSK while [Wan et al. 2014] proposed to calculate SIFT-like descriptors on 3D gradient and motion spaces respectively for RGB-D video recognition. Dardas et al [Dardas and Georganas 2011] recognized hand gestures via bag-of-words vector mapped from extracted key-points using SIFT and a multi-class SVM was trained as gesture classifier.

In the camp of action recognition, Karpathy et al [Karpathy et al. 2014] extended CNNs into video classification on a large-scale dataset of 1 million videos (Sports-1M). Donahue et al. [Donahue et al. 2015] embraced recurrent neural networks to explicitly model the complex temporal dynamics. Tran et al. [Tran et al. 2015] proposed to simultaneously extract the spatio-temporal features with deep 3D Convolutional Neural Networks (3D-CNN) followed by a SVM classifier for classification. Simonyan et al. [Simonyan and Zisserman 2014a] designed an architecture that captures the complementary information on appearance and motion between frames. Based on which, Feichtenhofer et al. [Feichtenhofer et al. 2016b] studied several levels of granularity in feature abstraction to fuse spatial and temporal cues. In contrast, Ng et al. [Ng et al. 2015] utilizes LSTM+CNN structure to model temporal information without resorting to additional optical flow information. Most recently, Feichtenhofer et al. [Feichtenhofer et al. 2016a] employs two powerful residual networks to learn stronger spatiotemporal features for action recognition.

In the other camp, the convolutional neural networks [Lecun et al. 1998] have been introduced to the field of gesture and recognition due to its rich capacity for representation [Molchanov et al. 2015; Nishida and Nakayama 2015; Molchanov et al. 2016]. For example, Nishida et al [Nishida and Nakayama 2015] proposes a multi-stream recurrent neural network that can be trained end to end without domain-specific hand engineering while [Molchanov et al. 2016] combines 3DCNN with RNN for online gesture detection and classification. Additionally, the rapid emergence of depth-sensor has made it economically feasible to capture both color and depth videos, providing motion information as well as three-dimensional structural information. This significantly reduces motion ambiguity when projecting the three-dimensional motion onto the two-dimensional image plane [Ni et al. 2011; Shahroudy et al. 2016; Wan et al. 2014]. For example, Molchanov et al [Molchanov et al. 2015] proposes to use depth and intensity data with 3D convolutional networks for gesture recognition. Ohn-Bar et al [Ohn-Bar and Trivedi 2014] first detects a hand in the region of interaction and then

combines RGB and depth descriptor for classification. Neverova et al [Neverova et al. 2015] proposes a multi-modal architecture that operates at 3 temporal scales corresponding to dynamic poses for gesture localization. However, our work is different from previous ones mainly by the following two difficulties. First, compared to previous gesture benchmarks which contain relatively few categories, the latest ChaLearn IsoGD dataset is larger and more challenging, covering 249 gesture labels with comprehensive inter-and-intra class variations; Second, video sequences of variable lengths are from different modalities, it is essential to design an effective way to capture and combine merits from both spatial, temporal and multi-modal information.

3. OUR METHOD

Fig.2 is an overview of our proposed approach. It mainly consists of Two Stream Consensus Voting Network (2SCVN) and 3D Depth-Saliency Network (3DDSN). Votings from 2SCVN and Fc-8 outputs from 3DDSN represent predictions from different modalities. These scores are further fused as the eventual label for isolated gesture recognition. In the following subsections, we describe in detail how 2SCVN and 3DDSN work.

3.1. Two Stream Consensus Voting Network

As is pointed out in Introduction, the bottleneck for improving the performance of large-scale gesture recognition lies in: 1) comprehensive inter-and-intra class variations; 2) long-term modeling of motions from sequences of variable lengths. Here we base our method on top of mainstream approaches [Simonyan and Zisserman 2014a; Feichtenhofer et al. 2016b] and adopts an *Consensus Voting Strategy* to reduce estimation variance.

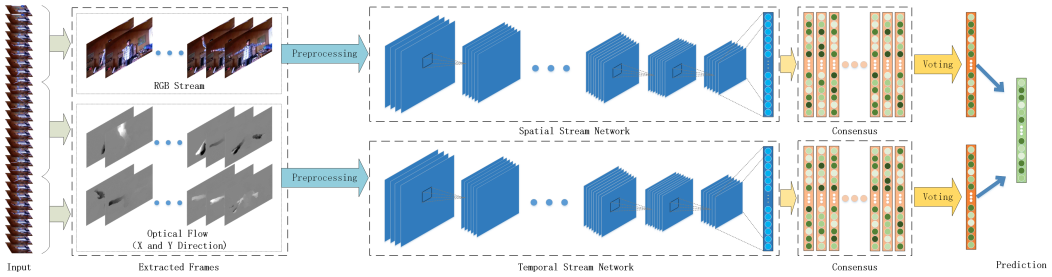


Fig. 3. 2SCVN is based on the idea of consensus voting, where its spatial and temporal stream sample RGB images (take rgb input as an example) and its stacked optical flow fields from different segments of a sequence according to a uniform distribution. Consensus from these frames as well as their augmented counterparts are taken as "votes" for the predictions.

Consensus Voting Strategy: The structure of 2SCVN is illustrated in Fig.3. We formalize the operations by convolutional networks as F parameterized by θ:

$$F : \Re^{h\times w\times t\times m} \rightarrow \Re^{l}, \boldsymbol{f} = F(\tau;\theta) \tag{1}$$

where an input snippet τ of sequential length $m \geq 1$ with t channels of size $h \times w$ pixels is transformed into a vector $\boldsymbol{f}$. Then, we apply softmax function $g : \Re^{l} \rightarrow \Re^{l}$ on top of vector $\boldsymbol{f}$

$$[g(\boldsymbol{f})]_i = e^{\boldsymbol{f}_i} / \sum\nolimits_k e^{\boldsymbol{f}_k} \tag{2}$$

where the i^{th} dimension indicates the probability of the snippet belonging to class i. Therefore, given an input video V of T snippets, we can calculate $[p(c_1|\tau_j), p(c_2|\tau_j)...p(c_l|\tau_j)]^T$, the probability with respect to each category for snippet τ_j. By stacking these predictions together, we get the following matrix:

$$\begin{bmatrix} p(c_1|\tau_1) & p(c_1|\tau_2) & \cdots & p(c_1|\tau_T) \\ p(c_2|\tau_1) & p(c_2|\tau_2) & \cdots & p(c_2|\tau_T) \\ \cdots & \cdots & \cdots & \cdots \\ p(c_l|\tau_1) & p(c_l|\tau_2) & \cdots & p(c_l|\tau_T) \end{bmatrix} \xrightarrow{h} \begin{bmatrix} p(c_1|V) \\ p(c_2|V) \\ \cdots \\ p(c_l|V) \end{bmatrix}$$

where each column is the class predictions of each snippet and each row being the class-specific predictions from T snippets. The aggregation function (*voting*) $h : \Re^{l \times T} \rightarrow \Re^l$ then combines the predictions from each snippet along the horizontal axis to output the probability of the whole video V with respect to each class. Therefore, the predicted label for video V is

$$y = \arg\max_{i \in S_l}(p(c|V)) \tag{3}$$

Note that the choice of h is still an open question and is determined by each specific task, here we have tried out Max and Mean function in Section 4.2.

Using the prediction of video V for each class, we deploy the standard categorical cross-entropy loss to train our network:

$$L(\boldsymbol{y}, \boldsymbol{p}) = -\sum_{i=1}^{l} y_i (p_i - \log \sum_{j=1}^{l} e^{p_j}) \tag{4}$$

where l is the number of categories and y_i the ground truth label concerning class i. Each network parameter with respect to the loss function is updated by stochastic gradient descent with a momentum $\mu = 0.9$. Each parameter in the network $\theta \in \omega$ is updated at every iteration step t by

$$\theta_t = \theta_{t-1} + \nu_t - \gamma\lambda\theta_{t-1} \tag{5}$$

$$\nu_t = \mu\nu_{t-1} - \lambda\eta \left(\left\langle \frac{\delta L}{\delta \theta} \right\rangle_{batch} \right) \tag{6}$$

where λ is the learning rate, γ is the weight decay parameter and $< \delta L/\delta\theta >_{batch}$ is the gradient of cost function L with respect to parameter θ averaged over the mini-batch. To prevent gradient explosion, we apply a soft gradient clipping operation η [Pascanu et al. 2013].

Implementations: We conducted experiments on Inception [Szegedy et al. 2015] with respect to the choice of ConvNet architecture due to its good balance between efficiency and accuracy. However, training deep networks is challenged by the risk of overfitting as current datasets for video recognition are relatively small compared to other computer vision tasks such as image classification. A common practice is to initialize the weights with pre-trained models on ImageNet [Deng et al. 2009]. To further mitigate the problem, we also adopted batch-normalization [Ioffe and Szegedy 2015] and dropout [Srivastava et al. 2014] layer (dropout ratio: 0.7) for regularization. Data augmentation is also employed to cover the diversity and variability of training samples. Besides random cropping and horizontal flipping, we also adapted the scale-jittering

cropping technique [Simonyan and Zisserman 2014b] to involve not only jittering of scales but also aspect ratios.

The optical flow fields are acquired using [Wedel et al. 2009]. We use Caffe [Jia et al. 2014] to train our networks. The learning rate is set to 0.1 and decreases to its $1/10$ for every 1500 iterations, lasting for over 20 epochs. It takes about 6 hours and 22 hours for training the spatial and temporal stream respectively on ChaLearn IsoGD with 2 TITANX GPUs. The parameter settings are the same when applied to RGB and depth modality.

3.2. 3D Depth-Saliency Network

Network Architecture: We base our method on top of 3D convolutional kernel proposed by Tran et al [Tran et al. 2015] while getting rid of the original Linear SVM configuration to train in an end to end manner. Compared to previous deep architectures, 3D CNNs are capable of encoding the spatial and temporal information in the data without requiring additional temporal modeling. Fig. 4 shows the structure of 3DDSN.

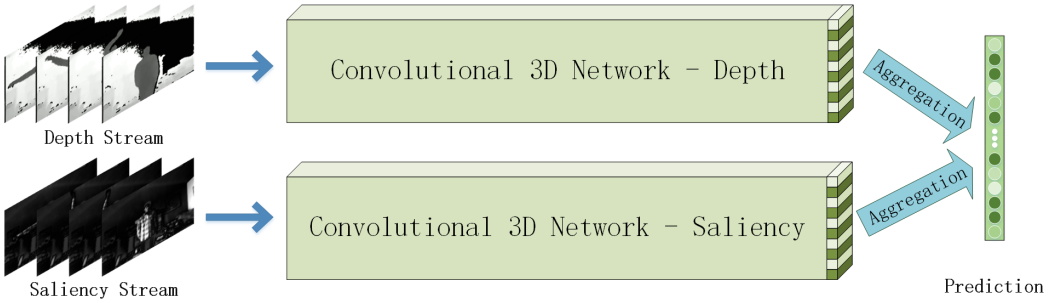


Fig. 4. 3DDSN employs 3D convolution on depth and saliency stream respectively, then takes scores from each stream for late fusion.

Specifically, we propose to combine depth and saliency stream, based on the observation that depth incorporates 3-dimensional structural information that RGB doesn't while saliency helps reduce the influence from backgrounds and other noises so as to focus on the salient regions, see Fig.1 for illustration. We didn't incorporate 3D-RGB stream for its performance lags far behind than that of 2SCVN-RGB. Each stream consists of eight 3D convolutional layers, each with a nonlinear Relu layer followed by five 3D Max Pooling layer. More formally, the 3D convolutional layer of the spatial-temporal CNN is defined as

$$\sum_{\delta_t}\sum_{\delta_y}\sum_{\delta_x} F_{t+\delta_t}(x+\delta_x, y+\delta_y) \times \omega(\delta_x, \delta_y, \delta_t) \tag{7}$$

where x and y define the pixel position for a given frame F_t. Then, nonlinearities are injected with Rectified linear unit, followed by the 3D pooling layer, defined as follows

$$\mathrm{ReLu}(x,y,t) = \begin{cases} Conv(x,y,t) & Conv > 0 \\ 0 & otherwise \end{cases} \tag{8}$$

$$Pool(x,y,t) = \max_{x,y,t}(\mathrm{ReLu}(x,y,t)) \tag{9}$$

Let $\chi = \{V_0, V_1, ..., V_B\}$ be a mini-batch of training samples and ω be the network's parameters. During training, we append a softmax prediction layer to the last fully-connected layer and finetune back-propagation with negative log-likelihood to predict classes from individual video V_i

$$L(\omega, \chi) = -\frac{1}{|\chi|} \sum_{i=0}^{|\chi|} \log(p(c^{(i)}|V^{(i)}; \omega)) \tag{10}$$

where $p(c^{(i)}|V^{(i)})$ is the probability of class label $c^{(i)}$ given video $V^{(i)}$ as predicted by 3DCNN. Finally, the predictions from the depth and saliency stream are fused to give the eventual label.

Implementations: Saliency images are extracted using [Achanta et al. 2009]. We first re-sampled each sequence to 32 frames instead of 16 originally used in [Tran et al. 2015] for better performance, using nearest neighbour interpolation by dropping or repeating frames [Molchanov P 2015]. Given each frame F_t, volumes are constructed using its surrounding 32 frames ($F_{t-15:t+16}$) with the label being the gesture occurring at its central frame. The spatial-temporal kernel size is set to $3 \times 3 \times 3$ in our experiments and the scale of the pooling is set to $2 \times 2 \times 2$ for all but the first layer. Additionally, the generalization ability of deep learning methods relies heavily on the data it trains on. In the specific task of gesture recognition, we observe that users might randomly choose their left or right hands while performing a gesture without changing the meaning, thus we adopt horizontal flipping as augmentation technique to incorporate this variability. The network is finetuned on sports-1M model with base learning rate of 0.0001 (decrease to its $1/10$ every 5000 stepsize) for 100K iterations. It needs about 2 days to finetune and update parameters and takes about 8G graphic memory for each modality.

4. EXPERIMENTS

To tap the full potential of our unified framework for RGB-D gesture recognition, we have explored extensively with various settings to examine how each component influences the final performance and experimented a number of good practices in terms of data augmentation, regularization and model fusion. We also visualize the confusion matrix, to give an intuitive analysis.

4.1. Datasets

RGB-D gesture recognition datasets suitable for evaluation of deep-learning based methods are very rare. Therefore, besides evaluations on ChaLearn IsoGD gesture recognition dataset[Wan et al. 2016], we also conducted experiments on RGBD-HuDaAct [Ni et al. 2011], one of the largest RGB-D action recognition datasets, where our proposed approach beat other methods, achieving state-of-the-art results.

Chalearn IsoGD: The CharLearn LAP RGB-D Isolated Gesture Dataset (IsoGD) contains 47933 RGB-D two-modality video sequences manually labeled into 249 categories, of which 35878 samples belong to the training set. Each RGB-D video represents one gesture instance, having 249 gesture labels performed by 21 different individuals. The IsoGD benchmark is one of the latest and largest RGB-D gesture recognition benchmarks and has a clear evaluation protocol, upon which the 2016 ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge has been held [Escalante H J 2016]. For the following evaluations, we conduct our experiments and report our accuracies on this dataset if not specifically mentioned.

RGBD-HuDaAct: The RGBD-HuDaAct database aims to encourage research efforts on human activity recognition on multi-modality sensor combination and each video

is synchronized with color and depth streams. It contains 1189 samples of 13 activity classes (including background videos which are added to the existing 12 classes) performed by 30 volunteers with rich intra-class variations for each activity representation.

4.2. Aggregation Function Discussion

In this subsection, we focus on discussions related to 2SCVN. As is mentioned in Section 3.1, aggregation function used for "voting" h is an open problem and is determined by a specific task. Here we empirically evaluated two kinds of functions, max and mean. Table I shows the accuracies of the spatial and temporal stream of 2SCVN under different aggregation functions. "-F" indicates corresponding optical flow fields while "(2:1)" means the combination ratio between the two streams. The ratio is chosen according to the accuracy of separate stream. For a specific modality, if optical flow stream is much higher than that of spatial stream, we will give more credence (i.e. weight) to the optical flow stream. On the validation set, the ratio given below yields almost the optimal performance.

Table I. Accuracies of different aggregation functions

2SCVN	RGB	RGB-F	RGB+RGB-F (1:2)
Max	45.65%	58.36%	62.72%
Mean	43.52%	56.74%	61.23%
	Depth	Depth-F	Depth+Depth-F (1:1)
Max	50.31%	49.85%	57.81%
Mean	47.14%	47.44%	54.31%

From Table I, we can draw the following conclusions: 1) Considering only spatial stream, depth modality is more accurate than RGB modality. However, when combined with their corresponding flow fields, RGB spatio-temporal fusion yields higher accuracies compared to Depth modality. This is mainly due to the fact that RGB-F stream encodes more temporal information which are essential for enhancing discriminative capability; 2) The performance of temporal stream is equivalent or higher compared to spatial stream, especially in RGB modality, which is reasonable because the spatial stream only captures actions at a fixed frame while the temporal stream takes into consideration motions at different time steps; 3) In terms of "voting", max aggregation seems to be more effective than mean aggregation and we leave other aggregation approaches as future work for related fields; 4) Compared to each modality, the combination of spatial and temporal stream leads to a large improvement in performance, indicating that both spatial and temporal information are essential for video recognition.

4.3. Fusion Schemes

In this subsection, we focus on discussions related to 3DDSN and explored the following questions: 1) How do RGB, depth, saliency perform individually; 2) Whether feature fusion does better than score fusion; 3) Any need for pre-processing before fusion? We conduct experiments on the Chalearn IsoGD benchmark using the aforementioned network configuration for each stream and explored their combinations. Table II reports the highest accuracies on IsoGD benchmark of different modalities and their combinations according to the evaluation protocol. "-" indicates that softmax is not used while"+" indicates vice-versa. D is short for Depth while S for saliency. "(2:1)" means the combination ratio between the two streams. It is also chosen according to separate modality performance.

Table II. Accuracies of different fusion schemes

3DDSN	RGB	Depth	Saliency	D:Sal(2/1)
Softmax -	46.08	54.95%	43.36%	58.86%
Softmax +	46.08	54.95%	43.36%	56.37%

We trained the mainstream 3D + SVM approach [Tran et al. 2015] as our baseline and used the same network architecture mentioned above, except that the spatial-temporal features from depth and saliency streams were concatenated to train the SVM classifier. The training of SVM takes about 6 hours and the accuracy on IsoGD [Wan et al. 2016] is 53.60%.

The following conclusions can be derived: 1) 3DDSN-Depth seems to be the more effective and discriminative than 3DDSN-saliency and 3DDSN-RGB; 2) Although for one modality, whether or not using softmax to convert the output to range $[0-1]$ yields the same accuracy, it generally reports higher accuracies for modality fusion without the "softmax" pre-processing. This is perhaps that the conversion reduces variance of features, thus abating the discriminative ability of model ensemble; 3) Compared with 3D + SVM baseline which employes feature-level fusion, score fusion seems to be more preferable, since features from different modalities may have very different distributions, therefore simple concatenation is not valid.

4.4. How does depth matter?

Besides recognition accuracies, to get a full appreciation of the potential from depth information, we compared RGB and RGB + Depth model trained using the architecture mentioned in Section 3.2 and counted the changes after fusing the depth into rgb stream and depth bring changes to "Correct" and "Error".

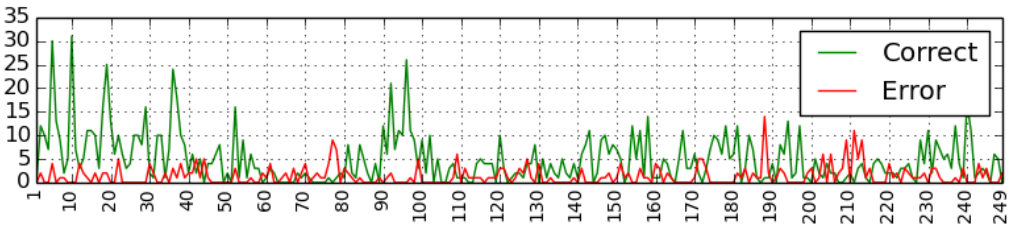


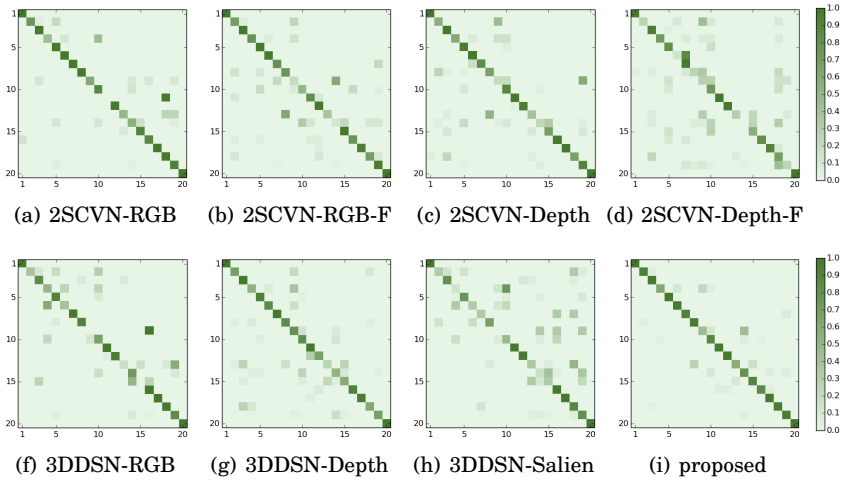
Fig. 5. Changes after fusing the depth stream into RGB stream. The x-axis denotes the category ID while the y-axis represents the number of changes. For an individual ID, "Correct" means that rgb stream makes the wrong predictions but rgb+depth fusion are correct. Conversely, "Error" indicates that the number of changes when the vice-versa is true.

A big "Correct/Error" means that depth brings positive/negative effect on RGB stream while zero means depth has no effect on the final prediction of that class. As shown in Fig.5, RGB stream works well in the range of $110-140$, however there are some class ranges such as 90 to 100, we have seen a huge improvement in terms of correct changes brought about by depth stream. The higher the green line ("Correct"), the more samples which have originally been predicted wrong are now correct. As the height of the green line is generally higher than that of red line, it confirms that depth indeed provides important supplementary information to RGB stream. Note that as the RGB stream of 3DDSN performs worse than that of 2SCVN, therefore this stream is not adopted in our final framework.

4.5. Visualization of Confusion Matrix

Fig.6 displays the confusion matrix of RGB, Depth, Saliency and overall approach.

From Classes such as 9 in RGB stream (Fig.6(a)) are not misclassified while there exist some confusions in RGB-Flow (Fig.6(b)). On the other hand, classes such as 11 which are confused in RGB stream perform relatively well in RGB-Flow. Thus, the spatial and temporal information actually supplements each other.



(a) 2SCVN-RGB (b) 2SCVN-RGB-F (c) 2SCVN-Depth (d) 2SCVN-Depth-F

(f) 3DDSN-RGB (g) 3DDSN-Depth (h) 3DDSN-Salien (i) proposed

Fig. 6. Performance confusion matrix of 2SCVN for RGB and RGB-F and 3DDSN for Depth and Saliency as well as the proposed fusion model on ChaLearn IsoGD dataset. The first 20 categories are used for visulization due to page size and the whole confusion matrix can be inferred from supplementary materials.

The confusion matrix of *Depth* and *Saliency* stream from 3DDSN on ChaLearn IsoGD are shown in Fig.6(g) and Fig.6(h) respectively. Fig.6(i) shows the confusion matrix of our proposed approach after modality fusion, which is obviously better than seperate streams. Note that we only displayed the first 20 categories due to page size and the whole confusion matrix are available in supplementary materials.

4.6. Qualitative Results

Example recognition results are shown in Fig.8 where the prediction distribution together with its confidence is displayed. We also show the ground-truth and top-3 predicted labels of each recognition result. As can be seen from the figure, our proposed approach correctly recognizes most of the gestures and attains pretty good accuracy even under challenging scenarios. However, the forth video in the first row is misclassified because the first prediction (4th video) is very similar to ground-truth (3rd video).

Fig.7 displays the recognition result of each category in RGBD-HuDaAct, where the first nine classes achieve an average accuracy of over 90%. For accuracies of different classes on ChaLearn IsoGD, please infer our supplementary material.

4.7. Comparison with State of the Art

We compare our proposed approach with competitors ranking top on the leaderboard of ChaLearn IsoGD benchmark [Escalante H J 2016; Wan et al. 2016] and state-of-the-art results on RGB-HuDaAct [Ni et al. 2011] datasets. We also tested each modality of our proposed framework as well as their combinations. Final results are summarized in Table III. “(S)” represents spatial stream while “(T)” means temporal stream. Small

number marked on the top right of a specific stream is used to represent this stream for indexing. For example, 2SCVN-3DDSN (5+6+8) means the model ensemble uses modality 5, 6, and 8 model indexed in Table III.

On ChaLearn IsoGD, hand-crafted features such as MFSK [Wan et al. 2016] as well as its variant which combines DeepID feature [Wan et al. 2016] scores relatively low compared to deep learning based methods such as AMRL [Pichao Wang and Ogunbona 2016] which incorporates three representations DDI, DDNI and DDMNI based on bidirectional rank pooling and ICT [Xiujuan Chai 2016] which trains a two-stream RNN for RGB and depth stream respectively.

2SCVN-RGB achieves an accuracy of 45.65% which is pretty good considering that it only uses one modalitity and that it only encodes static information. 2SCVN-Flow acquires a huge gain in performance as it captures motion information through stacked optical flow fields, which is reasonable because the accuracy of video recognition relies on the extent of understanding of the whole sequence. This is also what motivates us to explore consensus voting, a strategy that models the long term structure of the whole sequence to reduce estimation variance. The accuracy of 2SCVN-Fusion is further boosted as it combines the merits from spatial (2SCVN-RGB) and temporal stream (2SCVN-Flow). The performance of 3DDSN-Depth and 3DDSN-Sal are really impressive as they all score high compared to competing algorithms, due to rich representation capability of 3D convolution. Finally, although 2SCVN-RGB-Fusion scores rather high, the performance gain brought about 3DDSN after integration is still remarkable. This is in accordance with our observation that depth and saliency is supplementary to RGB modality. As can seen from Table III, our proposed approach outperforms other competing algorithms by a large margin with over 10% and 15% accuracy on ChaLearn IsoGD and RGB-D HuDaAct respectively.

Table III. Comparison with state-of-the-art methods on ChaLearn IsoGD and RGBD-HuDaAct benchmarks(%)

ChaLearn IsoGD Dataset			
Method	Result	Method	Result
NTUST	20.33%	MFSK+DeepID [Wan et al. 2016]	23.67%
MFSK [Wan et al. 2016]	24.19%	TARDIS	40.15%
XJTUfx	43.92%	ICT_NHCI [Xiujuan Chai 2016]	46.80%
XDETVP-TRIMPS	50.93%	AMRL [Pichao Wang and Ogunbona 2016]	55.57%
FLiXT	56.90%	-	-
2SCVN-RGB[1] (S)	45.65%	3DDSN-RGB (S+T)	46.08%
2SCVN-Depth[2] (S)	50.31%	3DDSN-Depth[3] (S+T)	54.95%
2SCVN-RGB-Flow[4] (T)	58.36%	2SCVN-Depth-Flow[5] (T)	49.85%
2SCVN-RGB-Fusion (1+4)	62.72%	2SCVN-Depth-Fusion (2+5)	57.81%
3DDSN-Fusion (3+6)	56.37%	3DDSN-Sal[6] (S+T)	43.35%
2SCVN-3DDSN (1+2+3+4+5+6)	**67.26%**	-	-

RGBD-HuDaAct Dataset			
Method			Result
STIPs(K=512) [Laptev and Lindeberg 2003]			79.77%
DLMC-STIPs(M =8) [Ni et al. 2011]			79.49 %
DLMC-STIPs(K=512,SPM) [Ni et al. 2011]			81.48%
3D-MHIs(Linear) [Davis and Bobick 2000; Ni et al. 2011]			70.51%
3D-MHIs(RBF) [Davis and Bobick 2000; Ni et al. 2011]			69.66%
2SCVN-RGB[1](S)	83.91%	3DDSN-RGB (S+T)	94.23%
2SCVN-Depth[2] (S)	88.19%	3DDSN-Depth[3] (S+T)	92.26%
2SCVN-RGB-Flow[4](T)	95.32%	2SCVN-Depth-Flow[5] (T)	90.84%
2SCVN-RGB-Fusion (1+4)	96.13%	2SCVN-Depth-Fusion (2+5)	93.89%
3DDSN-Fusion (3+6)	93.68%	3DDSN-Sal[6] (S+T)	92.06%
2SCVN-3DDSN (1+2+3+4+5+6)	**97.83%**	-	-

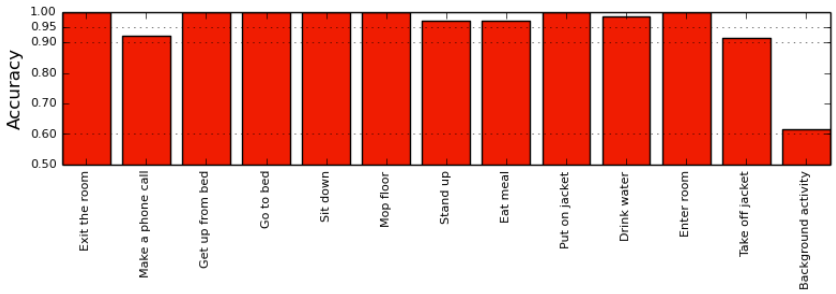


Fig. 7. Qualitative recognition results of our proposed approach on RGBD-HuDaAct benchmark

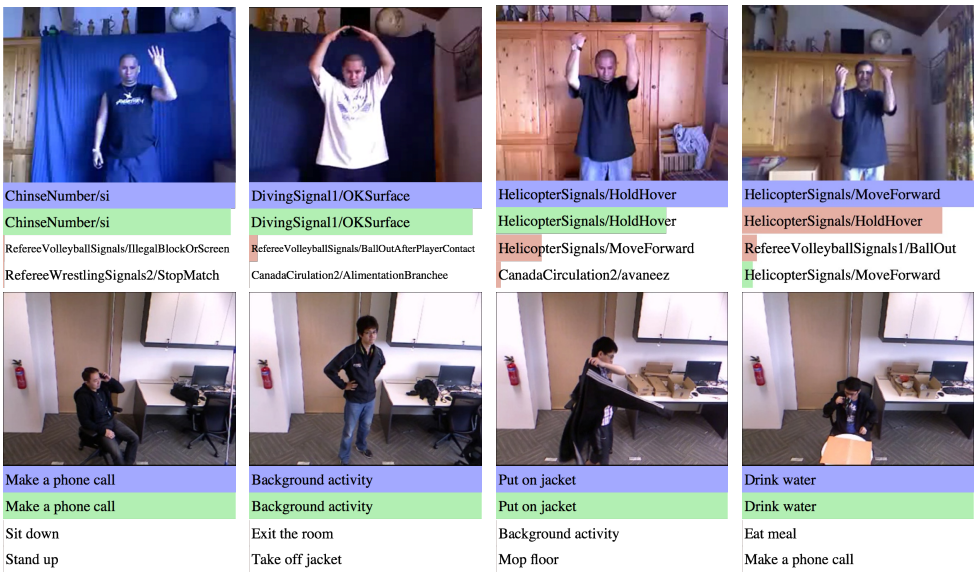


Fig. 8. Qualitative recognition results of our proposed approach on ChaLearn IsoGD (1st row) and RGB-D HuDaAct (2nd row) benchmarks. Bars colored blue indicate ground truths while green indicate correct and red wrong. The length of the bar represents confidence.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a multi-modality framework for RGB-D gesture recognition that achieves superior recognition accuracies. Specifically, 2SCVN based on the strategy of consensus voting is employed to model long term video structure and reduce estimation variance while 3DDSN composed of depth and saliency streams are aggregated in parellel to capture embedded information supplementary to RGB modality. 3D-RGB stream is not adopted as it is inferior to 2SCVN. Extensive experiments show the effectiveness of our framework and codes would be realsed to facilitate future research.

APPENDIX

In the paper, some of the figures and statistics are not complete due to the limit of page size. For the sake of completeness, in this appendix we first present the recognition result on ChaLearn IsoGD with respect to each of the 249 categories to supplement

Section 4.6. Then, we give the complete confusion matrices for ChaLearn IsoGD and RGBD-HuDaAct respectively to supplement Section 4.5.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects #61672521, #61473291, #61572501, #61502491, #61572536, NVIDIA GPU donation program and AuthenMetric R&D Funds.

REFERENCES

Radhakrishna Achanta, Sheila S Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. (2009).

N. H. Dardas and Nicolas D. Georganas. 2011. Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. *IEEE Transactions on Instrumentation and Measurement* 60, 11 (2011), 3592–3607.

James W. Davis and Aaron F. Bobick. 2000. The Representation and Recognition of Action Using Temporal Templates. *Proc of Cvpr* (2000), 928–934.

Jia Deng, Wei Dong, R. Socher, Li Jia Li, Kai Li, and Fei Fei Li. 2009. ImageNet: A large-scale hierarchical image database. 248–255.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. (2015).

Wan Jun Escalante H J, Ponce-Lpez V. ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge. https://competitions.codalab.org/competitions/10331. (????).

Wan Jun Escalante H J, Ponce-Lpez V. 2016. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. *International Conference on Pattern Recognition* (2016).

Georgios D Evangelidis, Gurkirt Singh, and Radu Horaud. 2014. Continuous gesture recognition from articulated poses. In *Workshop at the European Conference on Computer Vision*. Springer, 595–607.

Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. 2016a. Spatiotemporal Residual Networks for Video Action Recognition. In *Advances in Neural Information Processing Systems*. 3468–3476.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016b. Convolutional Two-Stream Network Fusion for Video Action Recognition. (2016).

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Computer Science* (2015).

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Feifei. 2014. Large-Scale Video Classification with Convolutional Neural Networks. (2014).

I. Laptev and T. Lindeberg. 2003. Space-time interest points. In *International Conference on Computer Vision*. 432 – 439.

Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. 2015. Hand gesture recognition with 3D convolutional neural networks. (2015).

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. 2016. Online Detection and Classification of Dynamic Hand Gestures With Recurrent 3D Convolutional Neural Network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

Kim K et al Molchanov P, Gupta S. 2015. Multi-sensor system for driver's hand-gesture recognition. *Automatic Face and Gesture Recognition (FG), 11th IEEE International Conference and Workshops* 1 (2015), 1–8.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. 2015. *Multi-scale Deep Learning for Gesture Detection and Localization*. 474–490 pages.

Joe Yuehei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. (2015).

Bingbing Ni, Gang Wang, and Pierre Moulin. 2011. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. (2011).

Noriki Nishida and Hideki Nakayama. 2015. *Multimodal Gesture Recognition Using Multi-stream Recurrent Neural Network*. Springer-Verlag New York, Inc.

Eshed Ohn-Bar and Mohan Manubhai Trivedi. 2014. Hand Gesture Recognition in Real Time for Automotive Interfaces: A Multimodal Vision-Based Approach and Evaluations. *IEEE Transactions on Intelligent Transportation Systems* 15, 6 (2014), 2368–2377.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training Recurrent Neural Networks. (2013).

Song Liu Zhimin Gao Chang Tang Pichao Wang, Wanqing Li and Philip Ogunbona. 2016. Large-scale Isolated Gesture Recognition Using Convolutional Neural Networks. *ICPRW* (2016).

Lionel Pigou, Aaron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. 2015. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *International Journal of Computer Vision* (2015), 1–10.

Amir Shahroudy, Jun Liu, Tiantsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. (2016).

Karen Simonyan and Andrew Zisserman. 2014a. Two-stream convolutional networks for action recognition in videos. (2014).

Karen Simonyan and Andrew Zisserman. 2014b. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2014).

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

Thad Starner, Alex Pentland, and Joshua Weaver. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 12 (1998), 1371–1375.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *Computer Science* (2015).

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. (2015).

Pedro Trindade, Jorge Lobo, and Joao P. Barreto. 2012. Hand gesture recognition using color and depth images enhanced with hand angular pose data. In *Multisensor Fusion and Integration for Intelligent Systems*. 71–76.

Author: Jun Wan, Qiuqi Ruan, Wei Li, Gaoyun An, and Ruizhen Zhao. 2014. 3D SMoSIFT: three-dimensional sparse motion scale invariant feature transform for activity recognition from RGB-D videos. *Journal of Electronic Imaging* 23, 2 (2014), 1709–1717.

J. Wan, V Athitsos, P Jangyodsuk, H. J. Escalante, Q Ruan, and I Guyon. 2014. CSMMI: class-specific maximization of mutual information for action and gesture recognition. *IEEE Transactions on Image Processing* 23, 7 (2014), 3152–3165.

Jun Wan, Guodong Guo, and Stan Z Li. 2016. Explore Efficient Local Features from RGB-D Data for One-Shot Learning Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1626–1639.

Jun Wan, Qiuqi Ruan, Wei Li, and Shuang Deng. 2013. One-shot learning gesture recognition from RGB-D data using bag of features. *Journal of Machine Learning Research* 14, 1 (2013), 2549–2582.

Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z. Li. 2016. ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Heng Wang and Cordelia Schmid. 2013. Action Recognition with Improved Trajectories. (2013).

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*. Springer International Publishing.

Sy Bor Wang, Ariadna Quattoni, Louis Philippe Morency, and David Demirdjian. 2006. Hidden Conditional Random Fields for Gesture Recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1521–1527.

Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. 2009. An Improved Algorithm for TV-L1 Optical Flow. *Lecture Notes in Computer Science* 5604, 7 (2009), 23–45.

Fang Yin Zhuang Liu Xilin Chen Xiujuan Chai, Zhipeng Liu. 2016. Two Streams Recurrent Neural Networks for Large-Scale Continuous Gesture Recognition. *ICPRW* (2016).

Ying Yin and Randall Davis. 2014. Real-time continuous gesture recognition for natural human-computer interaction. In *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 113–120.

Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. 2016. Real-time Action Recognition with Enhanced Motion Vector CNNs. (2016).

Yi Zhu and Shawn Newsam. 2016. *Depth2Action: Exploring Embedded Depth for Large-Scale Action Recognition*. Springer International Publishing.

Received February 2007; revised March 2009; accepted June 2009

Online Appendix to: A Unified Framework for Multi-Modal Isolated Gesture Recognition

Jiali Duan, CBSR & NLPR, Institute of Automation, University of Chinese Academy of Sciences
Jun Wan*, CBSR & NLPR, Institute of Automation, University of Chinese Academy of Sciences
Shuai Zhou, Macau University of Science and Technology
Xiaoyuan Guo, School of Engineering Science, University of Chinese Academy of Sciences
Stan Z. Li, CBSR & NLPR, Institute of Automation, University of Chinese Academy of Sciences

A. CHALEARN ISOGD

The ChaLearn IsoGD benchmark was proposed in the workshop of CVPR 2016 and was later used as the evaluation dataset for ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge [Escalante H J 2016], on the platform of CodaLab [Escalante H J]. For results and details of the competition, please refer [Escalante H J 2016].

The benchmark is one of the largest and latest $RGB + D$ dataset with more than 50,000 gestures, manually labelled into 249 categories [Wan et al. 2016]. Fig.9 shows the number of videos per category i.e. distribution of different gestures in ChaLearn IsoGD and Fig.10 displays the recognition result of each category on ChaLearn IsoGD mentioned in Section 4.6.

B. CONFUSION MATRIX

In Section 4.5, we visualize the confusion matrix of different modalities on ChaLearn IsoGD, but only list the first 20 categories due to page size. Here, we present the confusion matrix of complete 249 categories associated with each graph of Fig.6 in the paper (See Fig.13 - Fig.18).

As can be seen from the figures, depth modality is very discriminative, as it has higher accuracy compared to RGB and saliency. The overall accuracy after combining 2SCVN and 3DDSN brings an additional 5% gain in performance compared to 2SCVN alone, demonstrating the effectiveness of modality-fusion. The final accuracy of our proposed approach is 67.26%, outperforming the best result on leaderboard of ChaLearn IsoGD by a large margin.

For integrity, we also list the overall confusion matrix of our proposed approach on RGBD-HuDaAct (See Fig.19). Our method also achieves the best performance with an accuracy of 97.83%.

DOI: 0000001.0000001

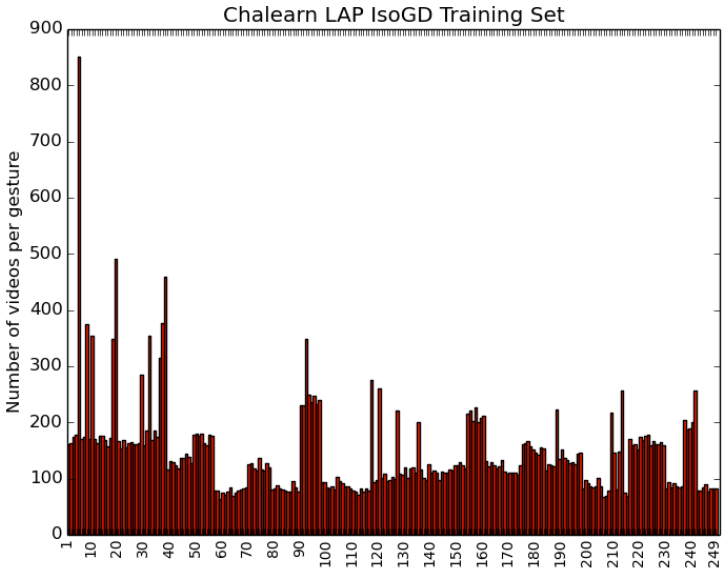


Fig. 9. The number of videos per class in ChaLearn IsoGD

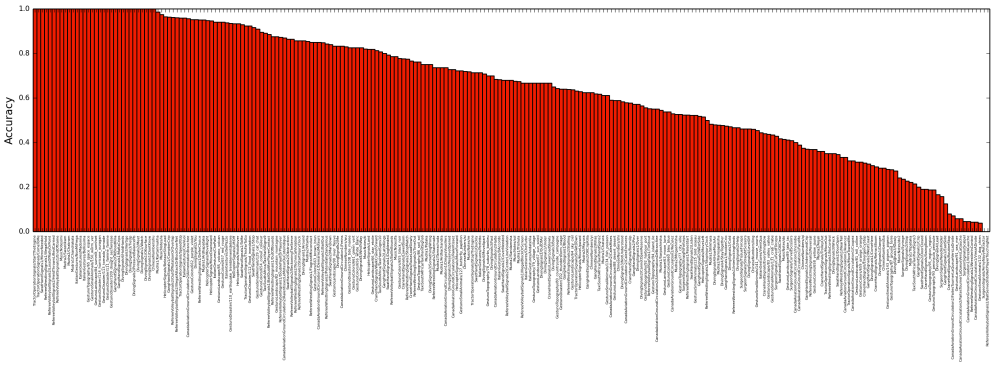


Fig. 10. Per class accuracy of our proposed approach on ChaLearn IsoGD benchmark

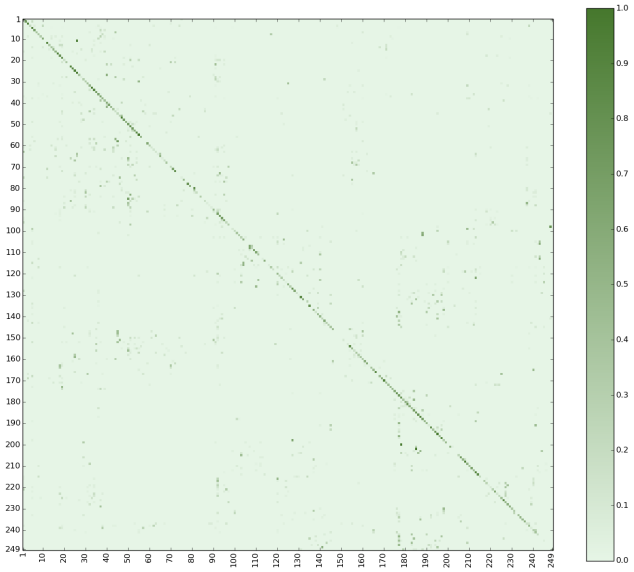


Fig. 11. Performance confusion matrix of 2SCVN-RGB on ChaLearn IsoGD (Accuracy: 45.65%)

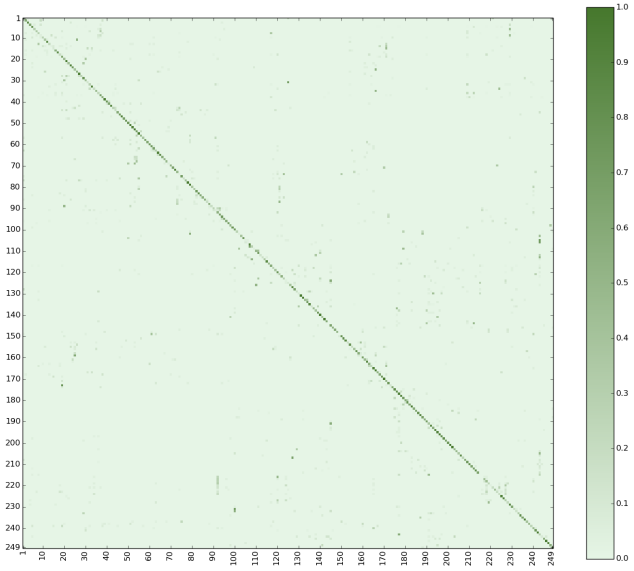


Fig. 12. Performance confusion matrix of 2SCVN-RGB-Flow on ChaLearn IsoGD (Accuracy: 58.36%)

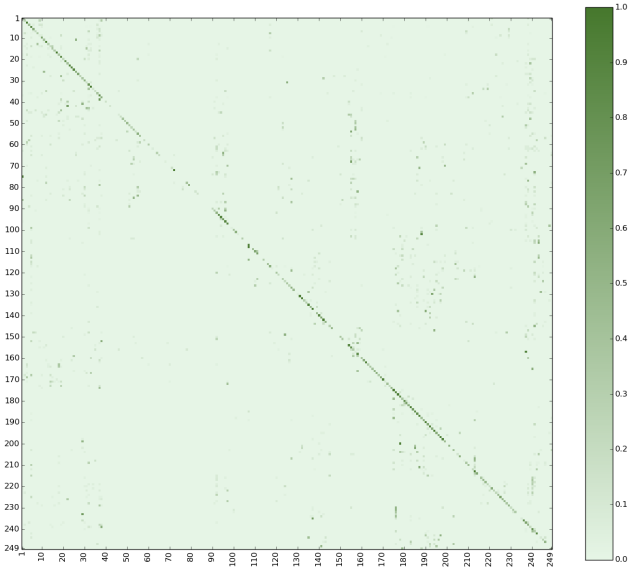


Fig. 13. Performance confusion matrix of 2SCVN-Depth on ChaLearn IsoGD (Accuracy: 50.31%)

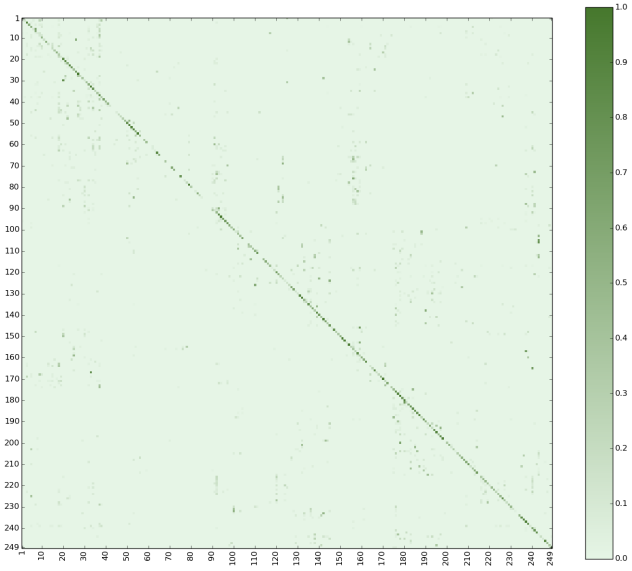


Fig. 14. Performance confusion matrix of 2SCVN-Depth-Flow on ChaLearn IsoGD (Accuracy: 49.85%)

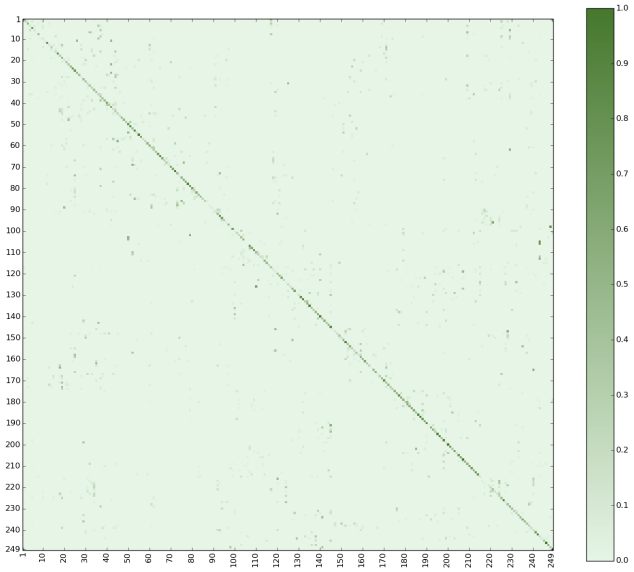


Fig. 15. Performance confusion matrix of 3DDSN-RGB on ChaLearn IsoGD (Accuracy: 46.08%)

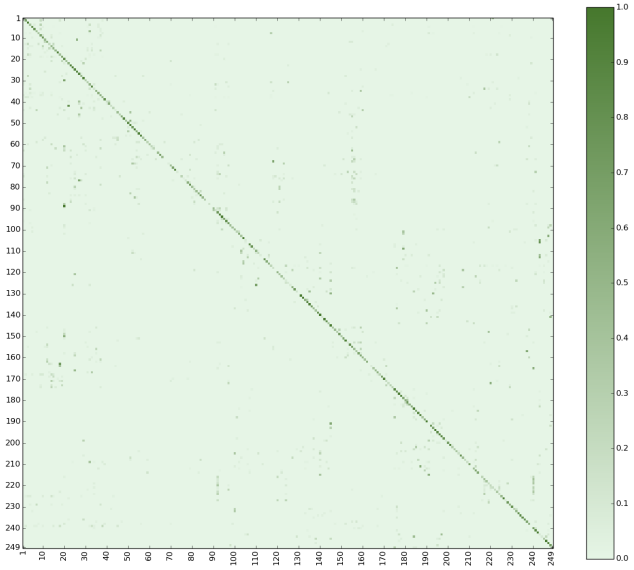


Fig. 16. Performance confusion matrix of 3DDSN-Depth on ChaLearn IsoGD (Accuracy: 54.95%)

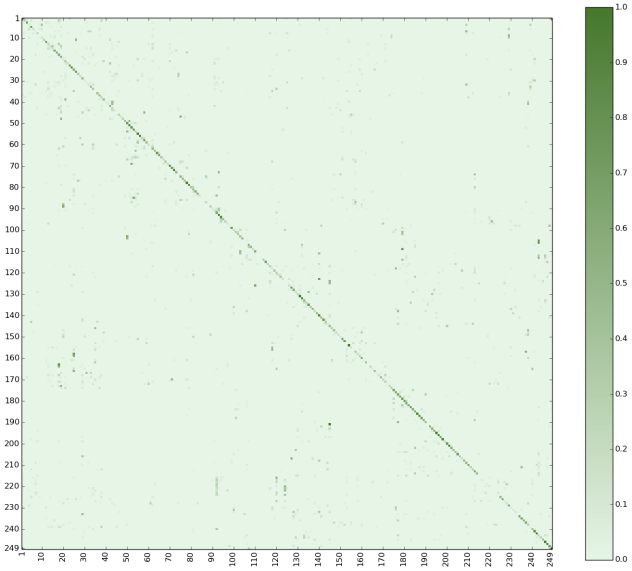


Fig. 17. Performance confusion matrix of 3DDSN-Saliency on ChaLearn IsoGD (Accuracy: 43.35%)

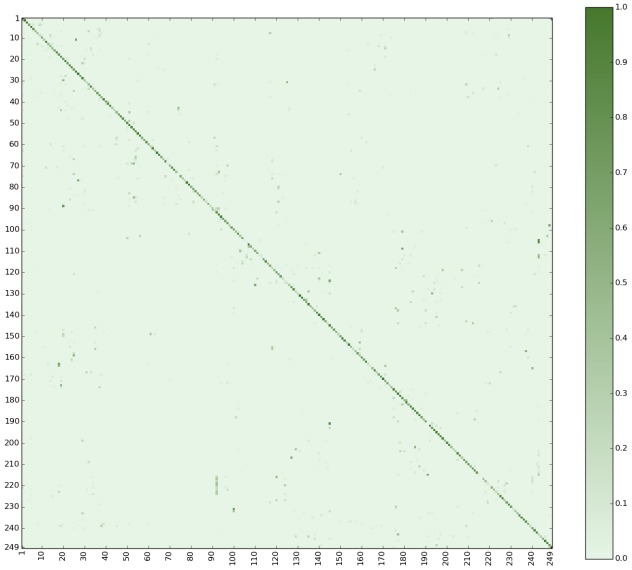


Fig. 18. Performance confusion matrix of the proposed approach on ChaLearn IsoGD (Accuracy: 67.26%)

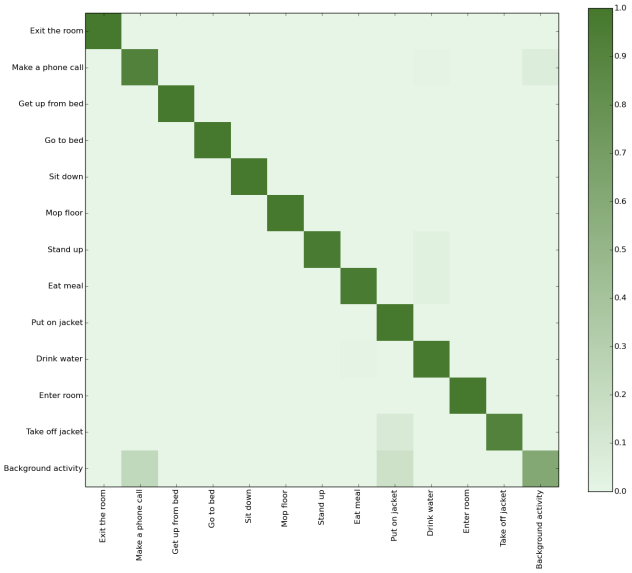


Fig. 19. Overall performance confusion matrix of our proposed approach on RGBD-HuDaAct (Accuracy: 97.83%)