Multi-Modality Fusion based on Consensus-Voting and 3D Convolution for Isolated Gesture Recognition

Jiali Duan[1], Shuai Zhou[*], Jun Wan[†], Xiaoyuan Guo[2], and Stan Z. Li[3]

[1,†,3]Center for Biometrics and Security Research & National Laboratory of Pattern Recognition

Institute of Automation, Chinese Academy of Sciences

[*] Macau University of Science and Technology

[2] School of Engineering Science, University of Chinese Academy of Sciences

[1,*,2]{jli.duan,shuaizhou.palm,xiaoyuanguo.ucas}@gmail.com, [†,3]{jun.wan,szli}@nlpr.ia.ac.cn

Abstract

Recently, the popularity of depth-sensors such as Kinect has made depth videos easily available while its advantages have not been fully exploited. This paper investigates, for gesture recognition, to explore the spatial and temporal information complementarily embedded in RGB and depth sequences. We propose a convolutional two-stream consensus voting network (2SCVN) which explicitly models both the short-term and long-term structure of the RGB sequences. To alleviate distractions from background, a 3d depth-saliency ConvNet stream (3DDSN) is aggregated in parallel to identify subtle motion characteristics. These two components in an unified framework significantly improve the recognition accuracy. On the challenging Chalearn IsoGD benchmark, our proposed method outperforms the first place on the leader-board by a large margin (10.29%) while also achieving the best result on RGBD-HuDaAct dataset (96.74%). Both quantitative experiments and qualitative analysis shows the effectiveness of our proposed framework and codes will be released to facilitate future research.

1. Introduction

Vision based gesutre recognition has drawn much attention from both the academic and industrial community for its widespread applications, such as human computer interaction and sign language translation and so on. Many methods have been proposed over the last few years [23, 14, 13, 36], which can be classified mainly into two categories: continuous gesture recognition and isolated gesture recognition. The former can be converted to the latter once continous gestures are segmented into seperate ones.

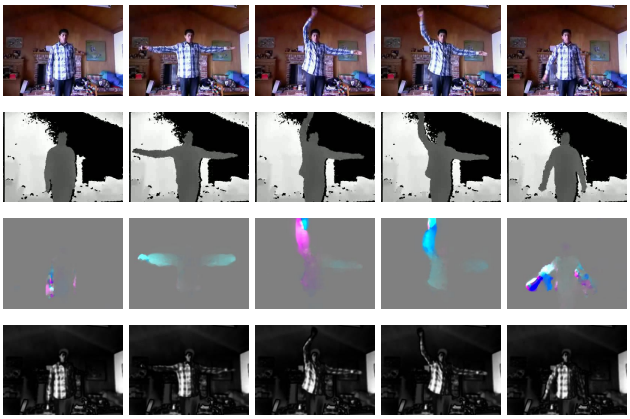


Figure 1. Examples of different types of modalities listed from up to bottom: RGB images, Depth images, Optical flow fields (magnitudes obtained from x and y direction are used as color channel), and Saliency

In this paper, we focus on isolated gesture recognition, especially for RGB-D video sequence, an area that has recently come into popularity due to the advancement and availability of depth-sensors such as Kinect. Although a significant amount of efforts have been made in the area of video recognition [17, 42, 37] for RGB modality, the complementary advantages inherently embedded in RGB-D sequences have largely been ignored. As shown in Fig.1, 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, thereby concentrating on the salient regions i.e. gestures. As were pointed out in recent works such as [43, 24, 34], depth cues could act as important supplement to the original RGB sequence, particularly for datasets that contain subtle inter-class variations. Besides depth information, we also include saliency for inspection. In fact,

[*]Joint first author

[†]Corresponding author

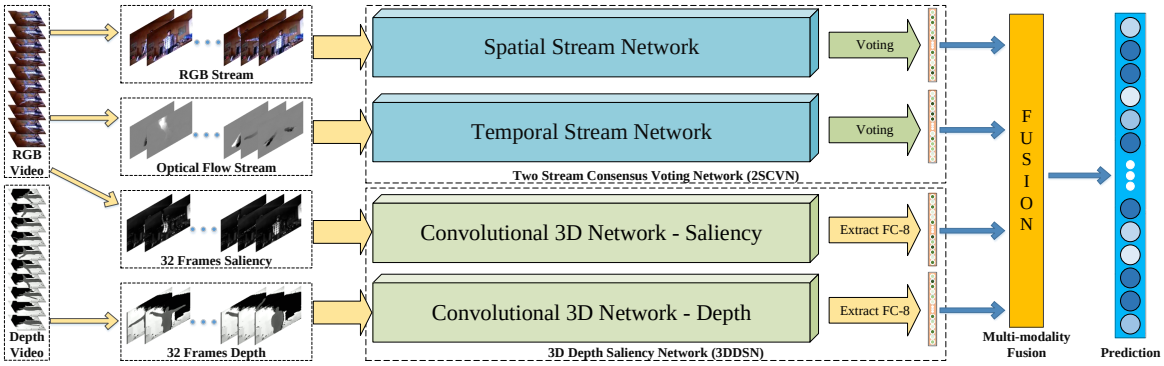


Figure 2. An overview of our approach. An input video is represented in different modalites, where the RGB stream and optical flow fields are handled by *Two Stream Consensus Voting Network (2SCVN)* while the saliency-stream and depth-stream are handled by *3D Depth-Saliency Network (3DDSN)*. 2SCVN takes into account consensus votes of the first two streams and outputs scores, which are later fused with another two streams from 3DDSN to give the final scores.

an impressive performance gain brought about by these two modalities as explained in Section 4 further consolidates our observation.

Our main motivations are: *1) how to reduce estimation variance when it comes to the decision of classifying videos of comprehensive inter-and-intra class vairations; 2) how to design a general 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 [14, 13, 23] which deal with a limited number of gestures can make very "biased" estimations when it comes to classifiying gesture datasets that involve comprehensive inter-and-intra class variations. Second, current main-stream approaches such as [25, 5] 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 network (2SCVN) based on the idea of consensus voting adapted from [25, 38, 7]. 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 video level predictions. Finally, these predictions are aggregated to reduce estimation variance.

For the second problem, we realize that as human motions are in essense 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. 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 3), demonstrating the effectiveness and generalization ability of our framework.

The proposed framework is shown in Fig.2 and 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

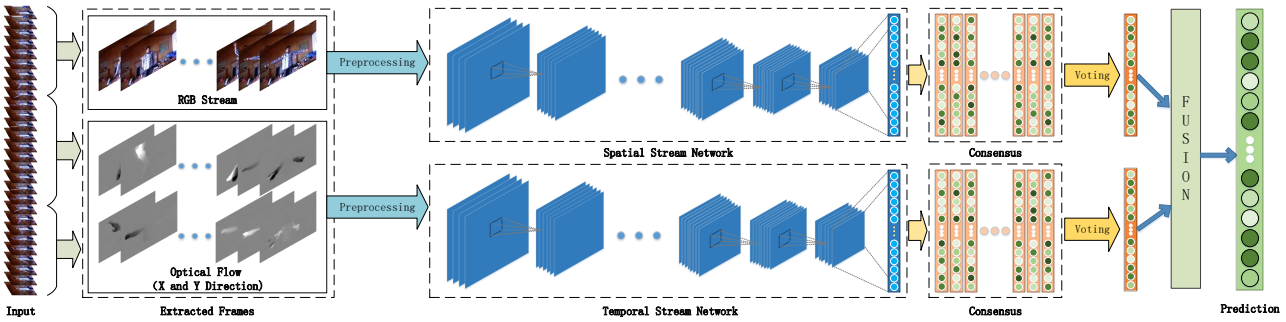


Figure 3. 2SCVN is based on the idea of consensus voting, where its spatial and temporal stream sample RGB images and its stacked optical flow fields from different segments of a sequence according to a uniform distribution. Consensus from these frames as well as their augmented counterparts are taken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

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 [36] and RGBD-HuDaAct [18] 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 gesture and action 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 [28, 39, 2, 31, 35]. For example, Wan et al [34] extracted a novel spatiotemporal feature named MFSK while [32] proposed to calculate SIFT-like descriptors on 3D gradient and motion spaces respectively for RGB-D video recognition. Dardas et al [2] recognized hand gestures via bag-of-words vector mapped from extracted keypoints using SIFT and a multi-class SVM was trained as gesture classifier.

Recently, the convolutinal neural networks [12] have been introduced to the field of gesture and action recognition due to its rich capacity for representation [13, 19, 14]. 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 structual information. This significantly reduces motion ambiguity when projecting the three-dimensional motion onto the two-dimensional image plane [18, 24, 33]. For example, Molchanov et al [13] proposes to use depth and intensity data with 3D convolutional networks for gesture recognition. Nishida et al [19] proposes a multi-stream recurrent neural network that can be trained end to end without domain-specific hand engineering while [14] combines 3DCNN with RNN for online gesture detection and classification. Ohn-Bar et al [20] first detects a hand in the region of interaction and then combines RGB and depth descriptor for classification. Neverova et al [16] proposes a multi-modal architecture that operates at 3 temporal scales corresponding to dynamic poses for gesture localization.

In the camp of action recognition, Karpathy et al [10] extended CNNs into video classification on a large-scale dataset of 1 million videos (Sports-1M). Donahue et al. [5] embraced recurrent nerual networks to explicitly model the complex temporal dynamics. Tran et al. [30] 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. [25] designed an architecture that captures the complementary information on appearance and motion between frames. Based on which, Feichtenhofer et al. [7] studied several levels of granularity in feature abstraction to fuse spatial and temporal cues.

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.

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 length. Here we base our method on top of mainstream approaches [25, 7] and adopts an *Consensus Voting Strategy* to reduce estimation variance.

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 funciton 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 fuction 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 [21].

Implementations: We conducted experiments on Inception [29] 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 over-fitting 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 [4]. To further mitigate the problem, we also adopted batch-normalization [8] and dropout [27] layer 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 [26] to involve not only jittering of scales but also aspect ratios.

The optical flow fields are acquired using [40]. We use Caffe [9] 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.

3.2. 3D Depth-Saliency Network

Network Architecture: We base our method on top of 3D convolutional kernel proposed by Tran et al [30] while getting rid of the orginal Linear SVM configuration to train in an end to end manner. Compared to previous deep architectures, 3D CNNs are capable of encoding the spaital and temporal information in the data without requiring additional temporal modeling. Fig. 4 shows the structure of 3DDSN.

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. 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

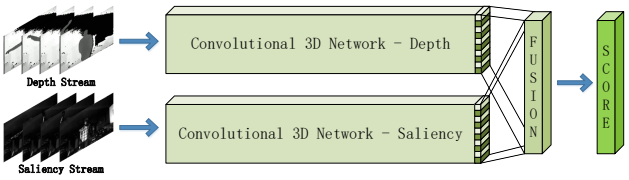


Figure 4. 3DDSN employs 3D convolution on depth and saliency stream respectively, then takes scores from each stream for late fusion.

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

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

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

$$Pool(x,y,t) = \max_{x,y,t}(\mathrm{Re}Lu(x,y,t)) \quad (9)$$

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

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

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 [1]. We first re-sampled each sequence to 32 frames using nearest neighbour interpolation by droping or repeating frames [15]. Given each frame F_t, volumes are constructed using its surrouding 32 frames ($F_{t-15:t+16}$) with the label being the gesture occuring at its central frame. The spatial-temporal kernel size is set to $3 \times 3 \times 3$ in our experiments and the sacle of the pooling is set to $2 \times 2 \times 2$ for all but the first layer. Addionally, 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 agumentation technique to incorporate this varaiability. 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[36], we also conducted experiments on RGBD-HuDaAct [18], 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 [6]. 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 exsiting 12 classes) performed by 30 volunteers with rich intra-class variations for each activity representation.

In the following subsections, we follow the pipeline of Algorithm 1 for testing of our proposed approach.

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 evaluted two kinds of functions, max and mean. Table 1 shows the accu-

Algorithm 1 Test pipeline of our proposed approach

Require:
Input and Model: 2SCVN and 3DDSN model;
RGB-D video I
Ensure: label
$I \xrightarrow{divide} S = \{S_1, S_2, S_3, ...S_k\}$
$V = \{\}$
for $S_t \; in \; S$ **do**
  $\tau_t \xleftarrow{sample} S_t$
  $V = V \cup \tau_t$
end for
$V \xleftarrow{augment} V$
$RGB, Flow, Depth, Sal \xleftarrow{compute} V$
$RGB \xrightarrow{2SCVN} Spatial\ Votes$
$Flow \xrightarrow{2SCVN} Temporal\ Votes$
$p\{2SCVN\} \xleftarrow{h} \{Sptial\ Votes \cup Temporal\ Votes\}$
$Depth \xrightarrow{3DDSN} Depth\ Scores$
$Saliency \xrightarrow{3DDSN} Depth\ Scores$
$p\{3DDSN\} \xleftarrow{h} \{Depth\ Scores \cup Depth\ Scores\}$
$p \xleftarrow{multi-modality} \{p\{2SCVN\} \cup p\{3DDSN\}\}$
$y = \arg\max_{i \in S_l}(p(c|V))$
return label

racies of the spatial and temporal stream of 2SCVN under different aggregation functions.

Table 1. Accuracies of different modalities and their combinations in the framework of 2SCVN on ChaLearn IsoGD, where Max and Mean aggregation functions have been tested. "-F" indicates optical flow fields

2SCVN	RGB	RGB-F	RGB+RGB-F
Max	45.65%	58.36%	62.72%
Mean	43.52%	56.74%	61.23%

From Table 1, we can draw the following conclusions: 1) Depth yields higher accuracies compared to RGB modality. This is perhaps that RGB modality encodes static appearance while depth incorporates 3-dimensional information and is less sensitive to possible background distractions and illumination changes as is shown in Fig.1; 2) The performance of temporal stream is higher compared to spatial stream, 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 improvement in performance.

4.3. Fusion Schemes

In this subsection, we focus on discussions related to 3DDSN and explored the following questions: 1) How do depth, saliency perform individually; 2) Whether feature fusion does better than score fusion; 3) Any need for preprocessing before fusion? We conduct experiments on the Chalearn IsoGD benchmark using the aforementioned network configuration for each stream and explored their combinations.

Table 2. Accuracies of different modalities on ChaLearn IsoGD, "-" indicates that softmax is not used while"+" indicates vice-versa. D is short for Depth while S for saliency.

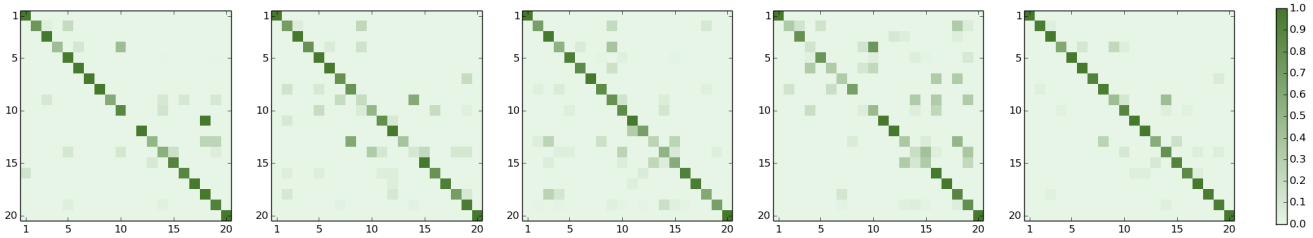
3DCNN	Depth	Saliency	D:Sal(2/1)
Softmax -	54.95%	43.36%	58.86%
Softmax +	54.95%	43.36%	56.37%

We trained the mainstream 3D + SVM approach [30] 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 [36] is 53.60%. Table 2 reports the highest accuracies on IsoGD benchmark of different modalities and their combinations according to the evaluation protocol. The following conclusions can be derived: 1) Depth seems to be the more effective and discriminative than saliency; 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".

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



(a) RGB (b) RGB-Flow (c) Depth (d) Saliency (e) proposed

Figure 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.

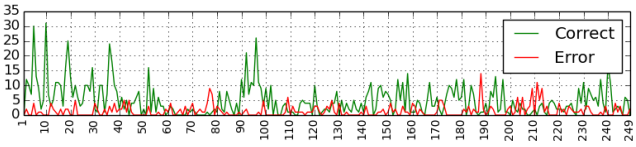


Figure 5. Changes after fusing the depth stream into RGB stream. The x-axis denotes the category ID while the y-axis represents the number of changes. For an individual ID, "Correct" means that rgb stream makes the wrong predictions but rgb+depth fusion are correct. Conversely, "Error" indicates that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

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 and RGB-Flow, 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.

The confusion matrix of *Depth* and *Saliency* stream from 3DDSN on ChaLearn IsoGD are shown in Fig.6(c) and Fig.6(d) respectively. Fig.6(e) 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 mis-classified 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.

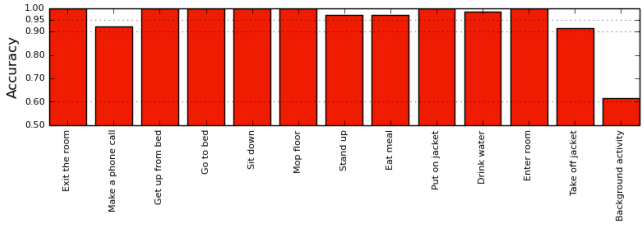


Figure 7. Qualitative recognition results of our proposed approach on RGBD-HuDaAct benchmark

4.7. Comparison with State of the Art

We compare our proposed approach with competitors ranking top on the leaderboard of ChaLearn IsoGD benchmark [6, 36] and state-of-the-art results on RGB-HuDaAct [18] datasets. We also tested each modality of our proposed framework as well as their combinations. Final results are summarized in Table 3.

On ChaLearn IsoGD, hand-crafted features such as MFSK [34] as well as its variant which combines DeepID feature [36] scores relatively low compared to deep learning based methods such as AMRL [22] which incorporates

Table 3. Comparison with state-of-the-art methods on ChaLearn IsoGD and RGBD-HuDaAct benchmarks(%)

ChaLearn IsoGD Dataset				RGBD-HuDaAct Dataset			
Method	Result	Method	Result	Method			Result
NTUST	20.33%	MFSK+DeepID [36]	23.67%	STIPs(K=512) [11]			79.77%
MFSK [34]	24.19%	TARDIS	40.15%	DLMC-STIPs(M =8) [18]			79.49 %
XJTUfx	43.92%	ICT_NHCI [41]	46.80%	DLMC-STIPs(K=512,SPM) [18]			81.48%
XDETVP-TRIMPS	50.93%	AMRL [22]	55.57%	3D-MHIs(Linear) [3, 18]			70.51%
FLiXT	56.90%	-	-	3D-MHIs(RBF) [3, 18]			69.66%
2SCVN-RGB (Ours)	45.65%	3DDSN-Sal (Ours)	43.35%	2SCVN-RGB	83.91%	2SCVN-Flow	95.32%
2SCVN-Flow (Ours)	58.36%	3DDSN-Depth (Ours)	54.95%	3DDSN-Depth	92.26%	3DDSN-Sal	92.06%
2SCVN-Fusion (Ours)	62.72%	3DDSN-Fusion(Ours)	56.37%	2SCVN-Fusion	96.13%	3DDSN-Fusion	93.68%
2SCVN-3DDSN (Ours)	**67.19%**	-	-	2SCVN-3DDSN	**96.74%**	-	-

three representations DDI, DDNI and DDMNI based on bidirectional rank pooling and ICT [41] 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-Fusion scores rather high, the performance gain brought about 3DDSN after integration is still remarkable (over 6%). This is in accordance with our observation that depth and saliency is supplementary to RGB modality. As can seen from Table 3, 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.

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. We also notice the possibility of employing 2SCVN on other modalities such as depth, depth-flow or saliency and we leave it as future work. Extensive experiments show the effectiveness of our framework and codes would be realsed to facilitate future research.

References

[1] R. Achanta, S. S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. 2009.

[2] N. H. Dardas and N. D. Georganas. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and Measurement*, 60(11):3592–3607, 2011.

[3] J. W. Davis and A. F. Bobick. The representation and recognition of action using temporal templates. *Proc of Cvpr*, pages 928–934, 2000.

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. Long-term recurrent convolutional networks for visual recognition and description. 2015.

[6] W. J. Escalante H J, Ponce-Lpez V. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. IEEE, 2016.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. 2016.

[8] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Computer Science*, 2015.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Feifei. Large-scale video classification with convolutional neural networks. 2014.

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

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] P. Molchanov, S. Gupta, K. Kim, and J. Kautz. Hand gesture recognition with 3d convolutional neural networks. 2015.

[14] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. 2016.

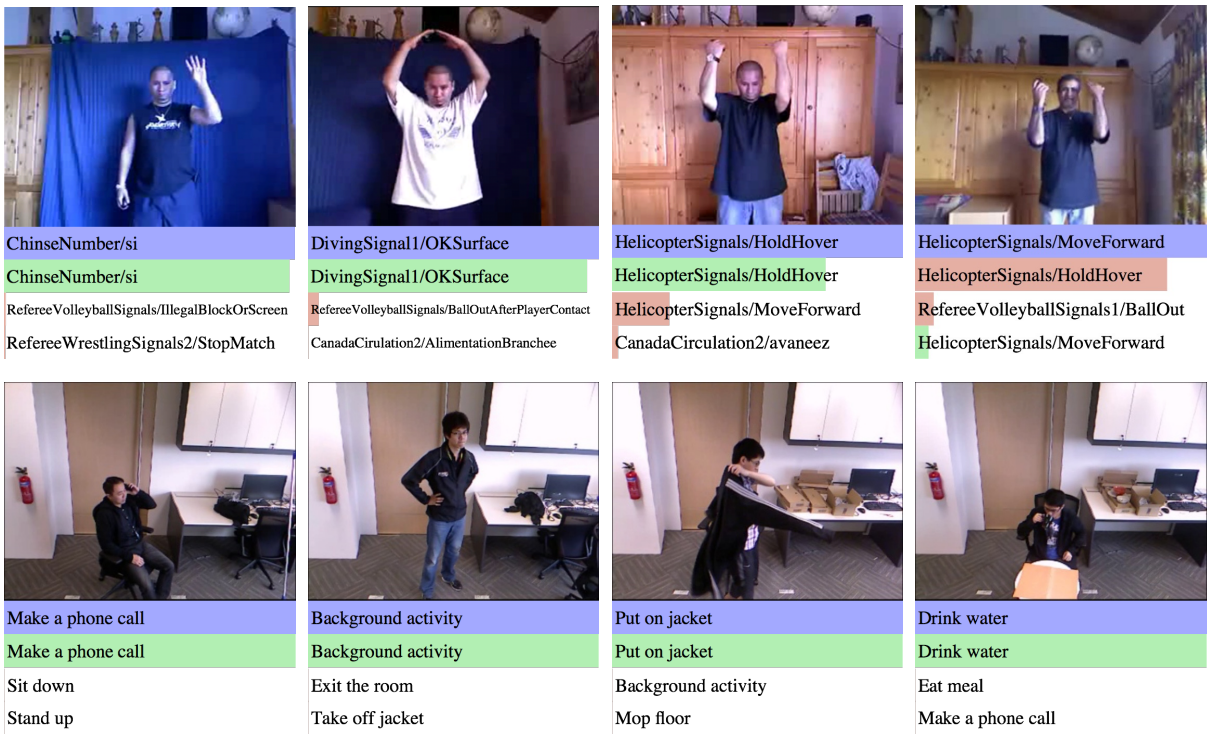


Figure 8. Qualitative recognition results of our proposed approach on ChaLearn IsoGD (1st row) and RGB-D HuDaAct (2nd row) benchmarks. Bars colored blue indicate ground truths while green indicate correct and red wrong. The length of the bar represents confidence.

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

[16] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. *Multi-scale Deep Learning for Gesture Detection and Localization*. 2015.

[17] J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. 2015.

[18] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. 2011.

[19] N. Nishida and H. Nakayama. *Multimodal Gesture Recognition Using Multi-stream Recurrent Neural Network*. Springer-Verlag New York, Inc., 2015.

[20] E. Ohn-Bar and M. M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 15(6):2368–2377, 2014.

[21] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. 2013.

[22] S. L. Z. G. C. T. Pichao Wang, Wanqing Li and P. Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. *ICPRW*, 2016.

[23] L. Pigou, A. V. Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, pages 1–10, 2015.

[24] A. Shahroudy, J. Liu, T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. 2016.

[25] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. 2014.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[28] T. Starner, A. Pentland, and J. Weaver. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *Computer Science*, 2015.

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. 2015.

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

[32] A. J. Wan, Q. Ruan, W. Li, G. An, and R. Zhao. 3d smosift: three-dimensional sparse motion scale invariant feature transform for activity recognition from rgb-d videos. *Journal of Electronic Imaging*, 23(2):1709–1717, 2014.

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

[34] J. Wan, G. Guo, and S. Z. Li. Explore efficient local features from rgb-d data for one-shot learning gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1626–1639, 2016.

[35] J. Wan, Q. Ruan, W. Li, and S. Deng. One-shot learning gesture recognition from rgb-d data using bag of features. *Journal of Machine Learning Research*, 14(1):2549–2582, 2013.

[36] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. 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*, June 2016.

[37] H. Wang and C. Schmid. Action recognition with improved trajectories. 2013.

[38] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*. Springer International Publishing, 2016.

[39] S. B. Wang, A. Quattoni, L. P. Morency, and D. Demirdjian. Hidden conditional random fields for gesture recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1521–1527, 2006.

[40] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers. An improved algorithm for tv-l1 optical flow. *Lecture Notes in Computer Science*, 5604(7):23–45, 2009.

[41] F. Y. Z. L. X. C. Xiujuan Chai, Zhipeng Liu. Two streams recurrent neural networks for large-scale continuous gesture recognition. *ICPRW*, 2016.

[42] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. 2016.

[43] Y. Zhu and S. Newsam. *Depth2Action: Exploring Embedded Depth for Large-Scale Action Recognition*. Springer International Publishing, 2016.