




In the format provided by the authors and unedited.

Discrimination of the behavioural dynamics of visually impaired infants via deep learning

Erping Long ^{1,8}, Zhenzhen Liu^{1,8}, Yifan Xiang^{1,8}, Andi Xu², Jialing Huang³, Xiucheng Huang², Xiaoyan Li¹, Zhuoling Lin¹, Jing Li¹, Jingjing Chen¹, Yan Zhang², Yi Zhu^{1,4}, Chuan Chen^{1,4}, Ziheng Zhou⁵, Xiaowei Ding⁵, Xiaohang Wu¹, Wangting Li¹, Hui Chen¹, Ruiyang Li¹, Yahan Yang¹, Weishi Zheng⁶, Weirong Chen¹, Haotian Lin ^{1,7*} and Yizhi Liu ^{1*}

¹State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-Sen University, Guangzhou, China. ²Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China. ³School of Public Health, Sun Yat-Sen University, Guangzhou, China. ⁴Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, FL, USA. ⁵VoxelCloud, Los Angeles, CA, USA. ⁶School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. ⁷Center for Precision Medicine, Sun Yat-sen University, Guangzhou, China. ⁸These authors contributed equally: Erping Long, Zhenzhen Liu, Yifan Xiang. *e-mail: gddlht@aliyun.com; yizhi_liu@aliyun.com

Supplementary Information

Table of Contents

1. Supplementary Figures	2
2. Supplementary Tables	5
3. Study Population Details	8
4. Definitions of the Behaviors	9
5. Detailed Dominance Analysis	10
6. Detailed Algorithm Information	11
7. Supplementary References	13

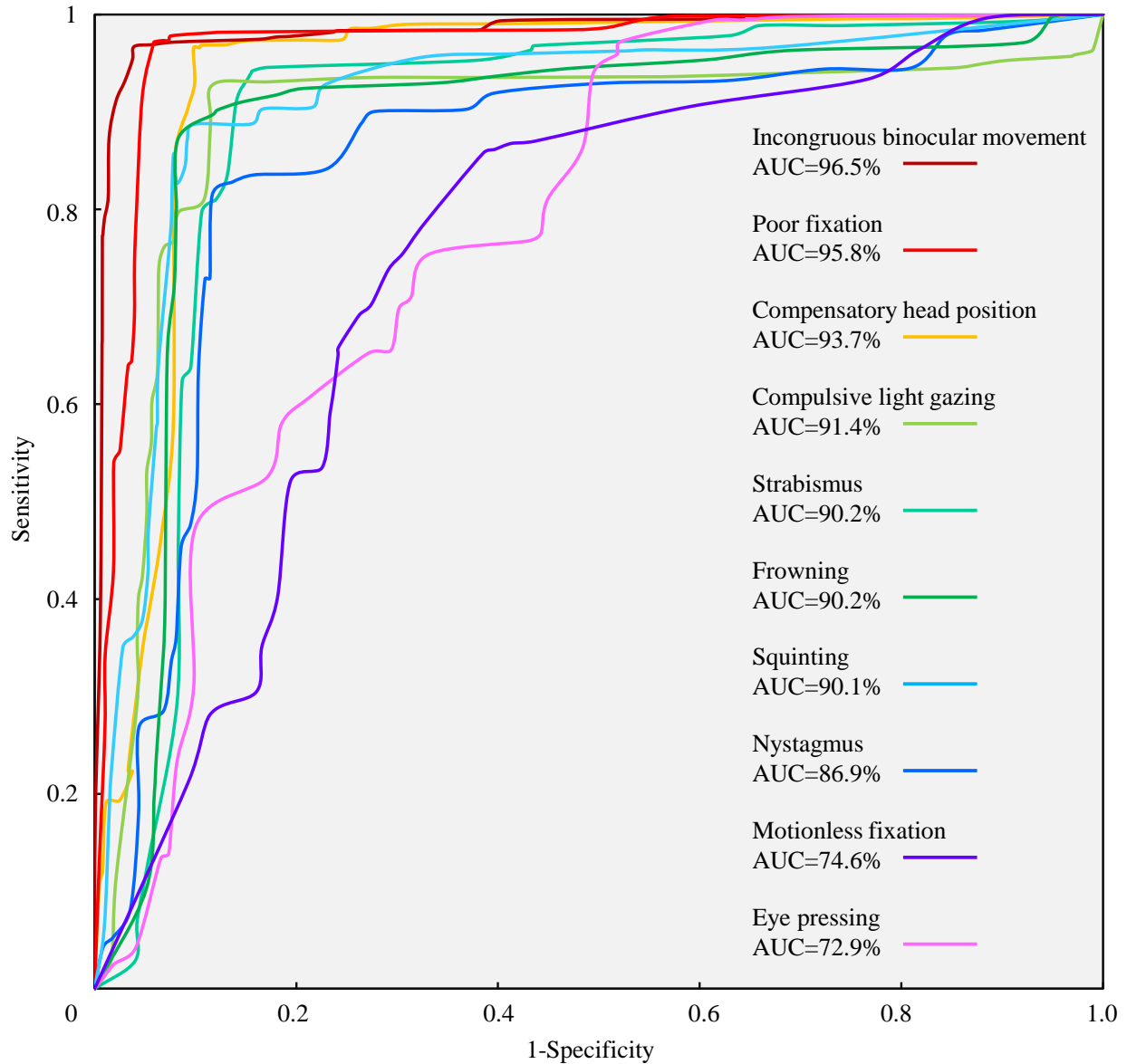


Figure S1. Algorithm performance for detecting behaviors with discriminated features.

Based on the minimum P value method, our algorithm can effectively detect patients with 8 abnormal behaviors from a healthy population with good performance (incongruous binocular movement, AUC 96.5%; poor fixation, AUC 95.8%; compensatory head position, AUC 93.7%; compulsive light gazing, AUC 91.4%; frowning, AUC 90.2%; strabismus, AUC 90.2%; squinting, AUC 90.1%; and nystagmus, AUC 86.9%), and detected 2 behaviors with acceptable performance (motionless fixation, AUC 74.6% and eye pressing, AUC 72.9%). **Notes:** AUC=area under the curve.

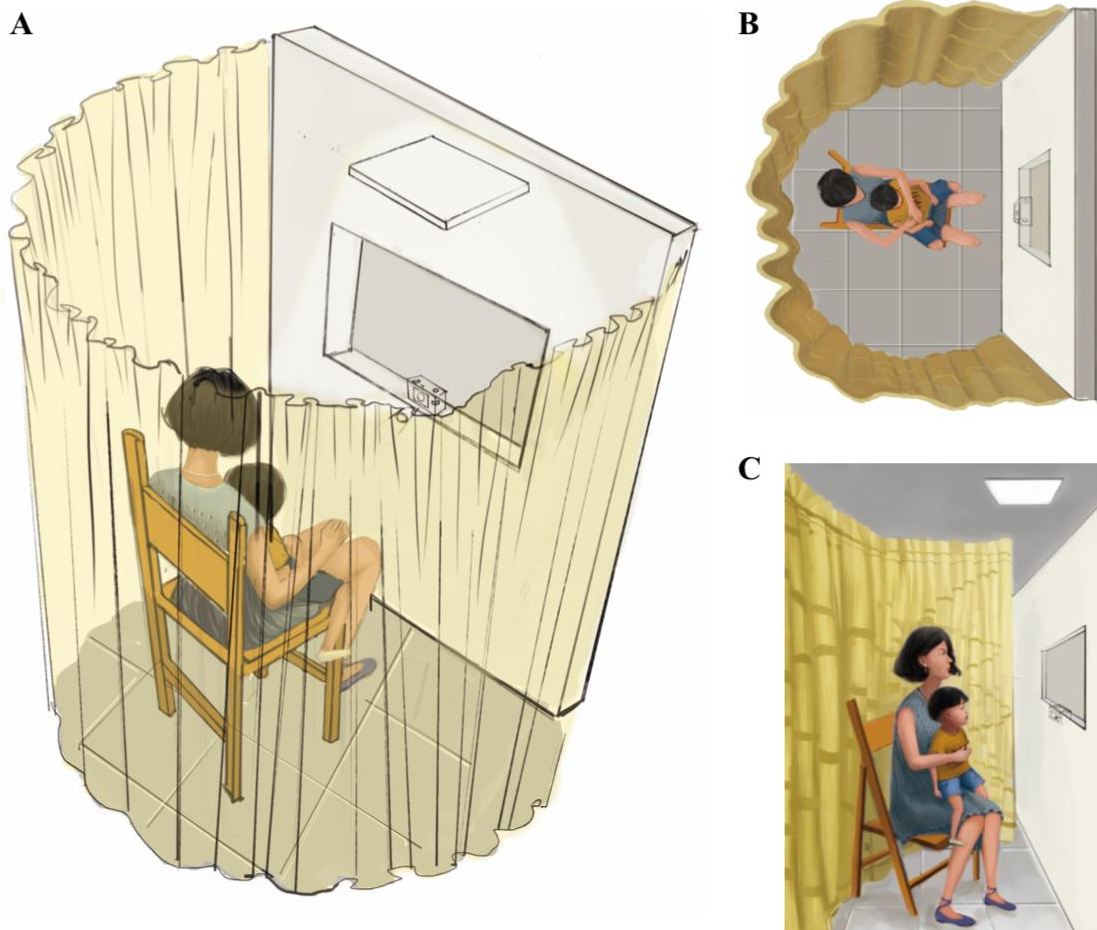


Figure S2. Standardized scenario for video recording.

A. The standardized apparatus consisted of a recording stage, a curtain, a chair, a light (10 candelas/m²) and a video recorder. Both the stage and the curtain were used to eliminate external interferences.

B. The video recorder was embedded in the middle of the stage with 1-meter ground height. The chair was fixed at a position of 0.55 meter facing the stage, to ensure that all the infants' actions can be fully recorded.

C. For each standardized procedure, the guardian sat in the chair, holding the infant facing the stage. Each infant was given a few minutes to adapt to the new surroundings and to be calm before recording. No hints or simulations were permitted during the process. The recording process lasted for more than 5 minutes to ensure that the behavioral phenotypes could be completely and repetitively recorded.

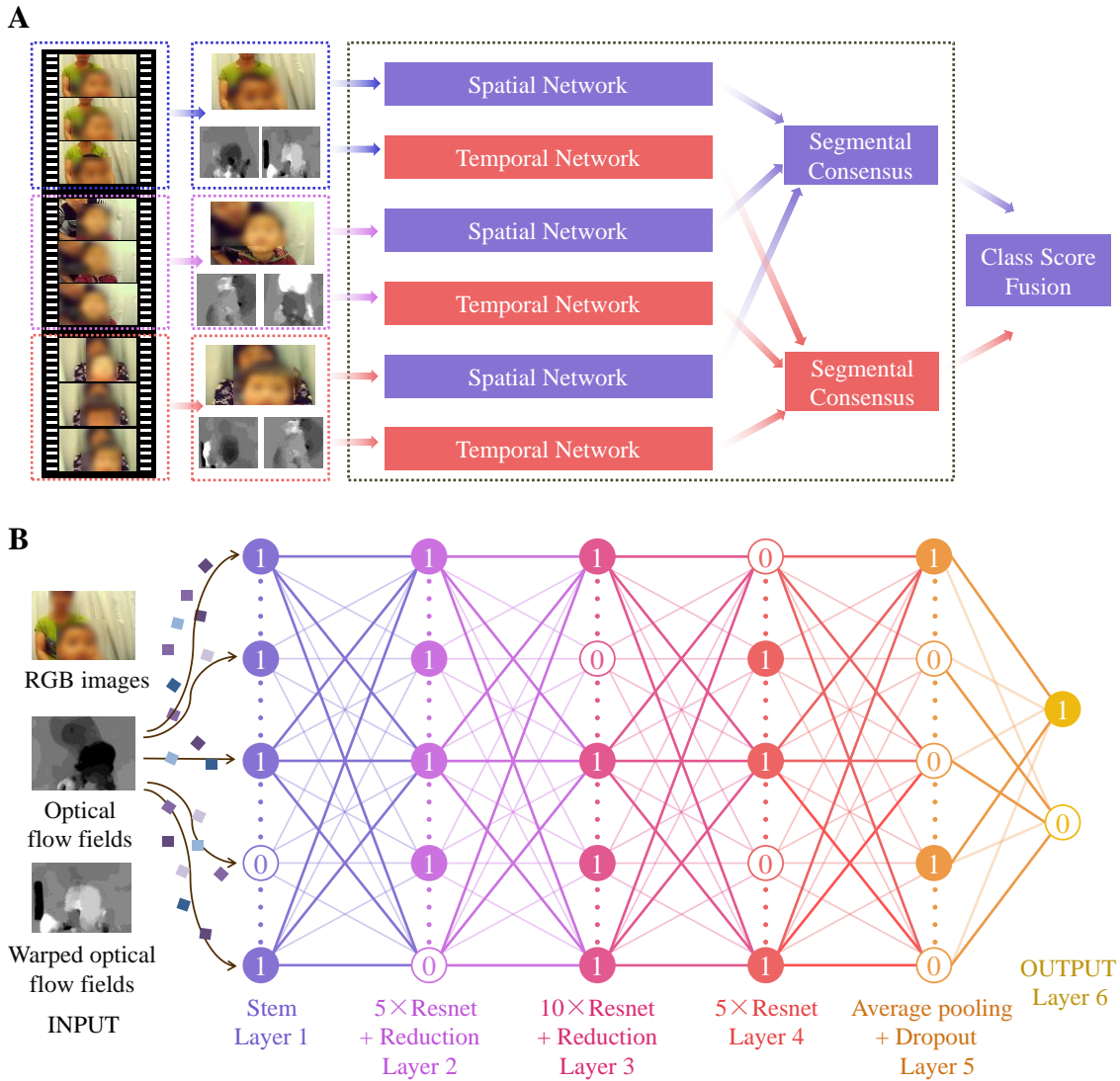


Figure S3. The detailed architecture of the temporal segment network.

A. For the temporal segment network, 1 input video was divided into 3 segments, and then, a short snippet was randomly selected from each segment. A two-stream network was arranged as follows and the class scores of different snippets were fused by the segmental consensus function. Predictions from all modalities were then fused to produce the final possibilities for classification.

B. We adapted the BN-Inception-resnet-v1 architecture to the design of the two-stream network. The spatial stream network operated on RGB images, and the temporal stream network took a stack of consecutive optical flow fields as input. The warped optical flow field was applied to enhance the discriminative power. We did not use RGB differences as one of the inputs because our videos are in relatively lower contrast. For detailed network structure, we settled the stem as the first layer, followed by three parts consisting of 20 Resnets and 2 reduction modules. In the following layer, we incorporated the average pooling and dropout technique before outputting.

Notes: RGB=Red-Green-Blue; BN=Batch Normalization.

Supplementary Tables

Table S1. Ophthalmological conditions of the infants based on structural examinations.

Diagnosis	Sample size	Percentage
Pupillary membrane	10	0.24%
Aphakic eye	1518	36.18%
Retinal detachment	41	0.98%
Microphthalmia	171	4.08%
Microcornea	104	2.48%
Congenital ptosis	4	0.10%
PHPV	101	2.41%
Lens dislocation	122	2.91%
Congenital cataract	1542	36.75%
Traumatic cataract	12	0.26%
Healthy	571	13.61%

The summarized characteristics of the ophthalmological conditions of the infants based on structural examinations are tabulated. **Notes:** PHPV=Persistent Hyperplastic Primary Vitreous.

Table S2. Overall data with age distribution based on visual conditions.

Age	Healthy	Mild	Severe	Overall
0-3 m	41	73	87	201
3-6 m	31	83	81	195
6-9 m	56	117	142	315
9-12 m	85	151	178	414
1-1.5 y	107	318	411	836
1.5-2 y	121	406	551	1078
2-2.5 y	81	459	143	683
2.5-3 y	49	306	119	474
Overall	571	1913	1712	4196

We recruited 4,196 infants 0 to 3 years old. The summarized statistics with age distribution based on visual conditions are tabulated. **Notes:** m=months; y=years.

Table S3. Summary of the statistical indices of the algorithm performance.

Primary screening	AUC	Sensitivity (%)	Specificity (%)
Mild vs. healthy	0.852 (0.804, 0.898)	83.7 (77.6, 88.7)	86.3 (73.7, 94.3)
Severe vs. mild	0.819 (0.775, 0.858)	87.9 (81.9, 92.4)	80.4 (74.0, 85.9)
Congenital cataract vs. healthy	0.930 (0.881, 0.960)	92.9 (87.3, 96.6)	92.2 (81.1, 97.8)
Microcornea vs. healthy	0.899 (0.804, 0.964)	91.7 (61.5, 99.8)	86.3 (73.7, 94.3)
Microphthalmia vs. healthy	0.864 (0.757, 0.936)	80.0 (51.9, 95.7)	88.2 (76.1, 95.6)
Lens dislocation vs. healthy	0.862 (0.765, 0.944)	83.3 (51.6, 97.9)	84.3 (71.4, 93.0)
Aphakic eyes vs. healthy	0.816 (0.756, 0.866)	81.3 (74.3, 87.1)	82.4 (69.1, 91.6)
Leber's amaurosis vs. healthy	0.909 (0.836, 0.962)	87.8 (73.8, 95.9)	92.2 (81.1, 97.8)
Duane syndrome vs. healthy	0.891 (0.808, 0.955)	88.9 (70.8, 97.7)	94.1 (83.8, 98.8)
Metabolic cataract vs. healthy	0.899 (0.805, 0.958)	89.5 (66.9, 98.7)	90.2 (78.6, 96.7)
Disease-level diagnosis	F1-measure	Precision (%)	Recall (%)
Congenital cataract	0.900 (0.873, 0.922)	89.7 (85.7, 92.8)	90.3 (86.4, 93.3)
Microcornea	0.878 (0.738, 0.959)	90.0 (68.3, 98.8)	85.7 (63.7, 97.0)
Microphthalmia	0.849 (0.746, 0.922)	81.6 (65.7, 92.3)	88.6 (73.3, 96.8)
Lens dislocation	0.870 (0.737, 0.951)	90.9 (70.8, 98.9)	83.3 (62.6, 95.3)
Aphakic eyes	0.870 (0.874, 0.924)	90.4 (86.5, 93.5)	89.8 (85.8, 93.0)
Behavior detection	AUC	Sensitivity (%)	Specificity (%)
Incongruous binocular movement	0.965 (0.920, 0.989)	96.7 (90.7, 99.3)	96.1 (86.5, 99.5)
Poor fixation	0.958 (0.931, 0.979)	97.2 (94.2, 98.9)	94.1 (83.8, 98.8)
Compensatory head position	0.937 (0.900, 0.967)	96.6 (92.7, 98.7)	90.2 (78.6, 96.7)
Compulsive light gazing	0.914 (0.876, 0.943)	92.8 (88.9, 95.7)	88.2 (76.1, 95.6)
Frowning	0.902 (0.855, 0.940)	89.7 (83.9, 94.0)	88.2 (76.1, 95.6)
Strabismus	0.902 (0.876, 0.924)	94.2 (92.0, 96.0)	84.3 (71.4, 93.0)
Squinting	0.901 (0.860, 0.933)	85.7 (80.5, 90.0)	92.2 (81.1, 97.8)
Nystagmus	0.869 (0.827, 0.904)	82.8 (77.7, 87.2)	86.3 (73.7, 94.3)
Motionless fixation	0.746 (0.678, 0.813)	70.3 (61.3, 78.2)	72.6 (58.3, 84.1)
Eye pressing	0.729 (0.646, 0.803)	72.6 (59.8, 83.2)	70.6 (56.2, 82.5)

The 95% CIs on the metrics are provided in parentheses. **Notes:** AUC=area under the curve.

Study Population Details

All 4,196 infants were recruited from 3 populations with various settings: 1) The National Visual Screening Project (NVSP) is a population-based study focused on children in communities; 2) the Childhood Blindness Project of South China (CBP-SC) is a multicenter collaboration consisting of tertiary hospitals and primary clinics; 3) the Vision of Infants in Guangzhou (VI-GZ) is conducted at Zhongshan Ophthalmic Center, the largest specialized eye hospitals in China. These projects have covered populations from communities, clinics, tertiary hospitals, and specialized centers.

Definitions of the Behaviors

Strabismus and Nystagmus were defined according to the 11th Revision of the International Classification of Diseases (ICD-11) Beta Draft.

Incongruous binocular movement was presented as interocular incompatibility of eye movement.

Eye rubbing was defined as smoothly rubbing using the back of the hand; **Eye pressing** was defined as aggressively pressing using fingers; **Eye poking** was defined as forced poking using fingertips directly to eyeball.

Compulsive light gazing was defined as gazing directly at the light for more than 5 seconds.

Compensatory head position was defined as the head out of the normal primary straight head position, including chin up, chin down, tilting of the head to the right or left, face turns to the right or left, or a combination of any of these head positions.

Motionless fixation was defined as fixation that lasted longer than 5 seconds, accompanied with vacant expressions.

Poor fixation was indicated by irregular sequential saccadic or rotary eye movements with no obvious objectives.

Frequent blinking was defined as forced blinking with an interval of less than 2 seconds between blinks.

Squinting was indicated by constant and forced squint in one or both eyes.

Frowning was defined as eyebrows becoming drawn together, presenting a forced watching action.

The magnitudes of behaviors were defined as follows: strabismus and nystagmus are given in terms of the occurrence pattern (intermittent or persistent); incongruous binocular movement, compulsive light gazing, compensatory head position, motionless fixation, and poor fixation are given in terms of the average persistent period; frequent blinking is given in terms of the mean frequency during the blinking process; eye rubbing, pressing, poking, squinting, and frowning are given in terms of the mean frequency during the recording process.

Detailed Dominance Analysis

Dominance analysis, as proposed by Azen and Traxel¹, was applied to determine the relative importance of each predictor in a multiple regression problem. This analysis is especially important in the presence of a large dataset that includes intercorrelation components, with 3 unique advantages: 1) dominance analysis considers a pairwise fashion for measuring relative importance; 2) all relative subset models are considered when comparing predictors; 3) this analysis provides dominant and alternative predictors to be ranked from “most important” to “least important”. Dominance analysis has been identified a particularly useful approach^{2,3}.

Specifically, dominance analysis defines the additional contribution of any given predictor to a given subset model as the change in R^2 (in our regressions, McFadden R^2) when the predictor is added to the model. The general dominance weight for each predictor is calculated from the McFadden R^2 statistic. The measure of McFadden R^2 varies naturally between 0 and 1 and is independent of the units of measurement of the variables. The general dominance weight of a variable represents its contribution to variance explained and importance in multivariable regression, including the direct effect and effect when combined with other variables in the regression³.

Detailed Algorithm Information

Overall architecture. For the temporal segment network (TSN), 1 input video was divided into 3 segments, and then, a short snippet was randomly selected from each segment. The samples were distributed uniformly along the temporal dimension. The spatial stream network operates on a single red-green-blue (RGB) image, and the temporal stream network takes a stack of consecutive optical flow fields as input. The warped optical flow field was applied to enhance the discriminative power of TSN. To mitigate the overfitting problem, several strategies (cross modality pretraining, regularization techniques, and data augmentation) were designed during the training process.

Two-stream network. The batch normalization (BN)-inception-Inception-resnet-v1 architecture was employed in the design of the two-stream networks. We settled the stem as the first layer, followed by three main parts of the hidden layers. In the first part, we applied 5 Resnets and 1 reduction module. In the second part, we additionally settled 10 Resnets with 1 reduction module. In the third part, we used 5 Resnets followed by average pooling and dropout techniques. For output, we utilized the final classification possibility for further analyses instead of any computational classifier. All details of the Inception-resnet-v1 can be found in ref⁴.

Cross modality pretraining. For spatial networks (RGB images as input), the model was pretrained on the ImageNet as initialization. For the optical flow field, a cross modality pretraining technique was utilized due to their different visual aspects and distributions. We did not use RGB differences as one of the inputs because our videos are in relatively lower contrast.

First, optical flow fields were discretized into the interval (0 to 255) by a linear transformation. This step made the range of optical flow fields the same as the RGB images. Then, we modified the weights of the first convolution layer of the RGB models to handle the input of the optical flow fields. Specifically, we averaged the weights across the RGB channels and replicated this average by the channel number of the temporal network input.

Regularization techniques. After the initialization with pretrained models, the mean and variance parameters of all BN layers (except the first layer) were frozen. As the distribution of the optical flow is different from the RGB images, the mean and variance of the activation value of the first convolution layer was re-estimated accordingly. Additionally, an extra dropout layer was added after the global pooling layer in BN-Inception architecture, with the dropout ratio set as 0.8 for spatial stream networks and 0.7 for temporal stream networks.

Data augmentation. The corner cropping and multiscale cropping were exploited for data augmentation.

In the corner cropping technique, the extracted regions were selected from only the corners or the center of the image. In the multiscale cropping technique, the size of the input image or optical flow fields was fixed at 256×340 , and the width and height of the cropped region were randomly selected from $\{256, 224, 192, 168\}$. These cropped regions were resized to 224×224 for network training.

Supplementary References

1. Azen, R., & Traxel, N. Using Dominance Analysis to Determine Predictor Importance in Logistic Regression. *Journal of Educational and Behavioral Statistics*, **34**(3):319-347 (2009).
2. Grömping, U. Estimators of relative importance in linear regression based on variance decomposition. *Am Stat*, **61**(2):139–147 (2007).
3. Johnson, JW. & Lebreton, JM. History and use of relative importance indices in organizational research. *Organ Res Meth*. 7:238–257 (2004).
4. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. <https://arxiv.org/abs/1602.07261>