Two-Stream SR-CNNs for Action Recognition in Videos

Yifan Wang[1]
yifan.wang@student.ethz.ch
Jie Song[1]
jsong@inf.ethz.ch
Limin Wang[2]
07wanglimin@gmail.com
Luc Van Gool[2]
vangool@vision.ee.ethz.ch
Otmar Hilliges[1]
otmar.hilliges@inf.ethz.ch

[1] Advanced Interactive Technologies Lab
ETH Zurich
Zurich, Switzerland

[2] Computer Vision Lab
ETH Zurich
Zurich, Switzerland

Abstract

Human action is a high-level concept in computer vision research and understanding it may benefit from different semantics, such as human pose, interacting objects, and scene context. In this paper, we explicitly exploit semantic cues with aid of existing human/object detectors for action recognition in videos, and thoroughly study their effect on the recognition performance for different types of actions. Specifically, we propose a new deep architecture by incorporating human/object detection results into the framework, called *two-stream semantic region based CNNs* (SR-CNNs). Our proposed architecture not only shares great modeling capacity with the original two-stream CNNs, but also exhibits the flexibility of leveraging semantic cues (e.g. scene, person, object) for action understanding. We perform experiments on the UCF101 dataset and demonstrate its superior performance to the original two-stream CNNs. In addition, we systematically study the effect of incorporating semantic cues on the recognition performance for different types of action classes, and try to provide some insights for building more reasonable action benchmarks and developing better recognition algorithms.

1 Introduction

Due to its importance in many application areas such as video surveillance, content driven retrieval, robotics and HCI, human action recognition has received a lot of attention recently [20, 24, 25, 26, 27, 28, 29]. However, robustly recognizing arbitrary free-form activities in real videos (cf. HMDB51 [14] and UCF101 [23]) is still a challenging task. The main difficulties stem from large intra-class variations caused by background clutter, scale and viewpoint changes, and drastically diverse dynamics in the observed motion. In addition, human action is inherently a high-level concept and often the exact semantic meaning is not well defined. Any given action is related to several semantic visual cues such as human poses, interacting objects, and scene context. Therefore it is expected that action recognition

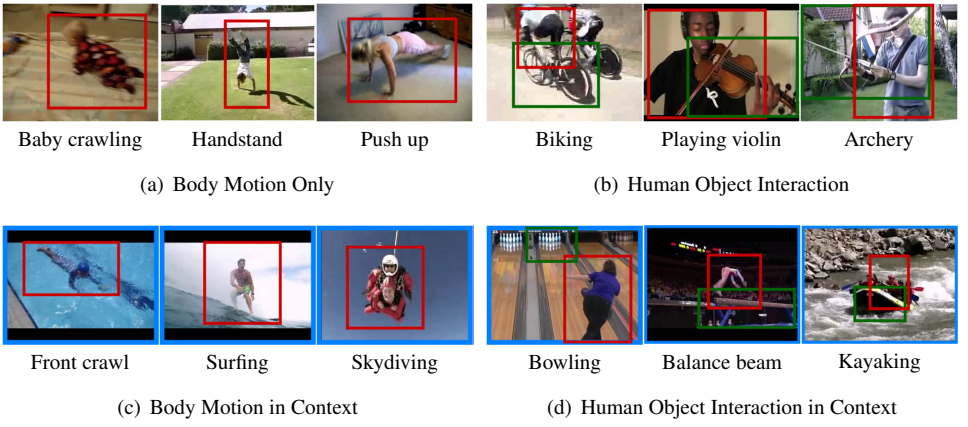


Figure 1: Action classes can be grouped into several types. The recognition of those actions is contributed by different semantics, like human body (red box), interacting objects (green box), and global context (blue box). Intuitively, different semantic cues impact the understanding and interpretation of different types of actions with different weights.

may benefit from the integration and fusion of the outputs of computer vision algorithms including human detection, object recognition, human pose estimation, and scene understanding.

Current action recognition benchmarks [14, 22] contain action classes derived from daily life, which can be roughly grouped into the following four types (see Figure 1): (1) **body motion only**: actions fully described by human movement like "crawling", (2) **human object interaction**: actions involving specific objects such as "playing violin", (3) **body motion in context**: body movement taking place in a specific environment like "skydiving", (4) **human object interaction in context**: actions containing representative objects and occurring in certain context, such as "balance beam" (see supplement for detailed categorization). Clearly different action types do not benefit equally from different semantic representations. However, current state-of-the-art action recognition methods [20, 27, 29] do not differentiate action types nor select or weight individual semantic channels based on action's properties. Hence there lies potential in explicitly exploring the semantic cues for robust action recognition algorithms with a special focus on selecting and fusing the various available cues in an optimal way.

Recently Convolutional Neural Networks (CNNs) [15] have led to great improvements in many vision problems such as image classification [13], object detection [6, 18], and scene recognition [30]. In particular, the region-based CNN (R-CNN) [6] detector is becoming an effective approach for image-based object localization. In this paper, we investigate how existing object detection methods can be leveraged to improve action recognition in videos and how to explicitly incorporate high level semantic cues, derived from object proposals, into the action understanding framework. We aim to answer the question of *whether state-of-the-art human/object detectors are advantageous for action recognition* and *how these detection results contribute to the recognition of different types of actions*.

Specifically, we propose an integrated deep neural network framework, called *two-stream semantic region based CNNs* (SR-CNNs), which incorporate various detection results (i.e. semantic cues) into the two-stream CNN architecture. We use the recently proposed Faster

R-CNN [18] as human and common object detectors, and adapt them to the video domain by leveraging temporal constraints among a sequence of detection results. These temporally coherent detection results provide semantic information about the activities portraited in the videos, such as the locations of a person and the relations of person and objects over time. We incorporate these semantic cues (detection results) into our proposed SR-CNN architecture, where detection bounding boxes guide the localization of salient regions in the video stream and enable CNNs to select discriminative features with ROI (Region Of Interest) pooling [7]. A final concern is then to fuse several semantic channels in an optimal fashion. To this end we propose and evaluate different fusion methods and training strategies to optimize the weights of our deep models. We empirically study the effect of the various settings and make recommendations for the best choice.

We perform a number of experiments on the UCF101 dataset and show that the proposed two-stream SR-CNNs outperform the original two-stream CNNs [20]. In addition, we systematically study the effect of incorporating semantic cues with respect to different types of actions. Our empirical study indicates that: (1) For current action recognition benchmarks, scene context acts as a very strong cue for action recognition. However, the use of R-CNNs in the video domain for person and object detection can still improve the overall performance of action recognition. (2) For the action types involving only body motion and human object interaction, incorporating human detections provides a significant boost in accuracy. (3) For action classes in a specific context (e.g., human object interaction in context), scene context again plays a dominant role for action recognition and incorporating semantic detection results does not yield big improvements. (4) Integrating semantic cues enhances the performance of spatial and temporal SR-CNNs differently due to the disparate properties of RGB images and optical flow fields.

2 Related Work

In this section we review previous works related to ours from two aspects: (1) deep learning for action recognition and (2) semantic cue augmentation for action recognition.

Deep learning for action recognition. There are a variety of works including 3D CNNs [11, 23], Deep CNNs [12], Two-Stream CNNs [20], and Temporal Segment Networks [29]. Two-Stream CNNs [20] was the first to successfully demonstrate competitive performance compared to the hand-crafted features [25]. Recently, Temporal Segment Networks [29] further advanced the action recognition performance by modeling long-term temporal structure and introducing good practices for training CNNs. These methods are supposed to capture the holistic but abstract information directly from the raw videos. Our method differs from them in important aspects. We explicitly incorporate several action-related semantic cues that go beyond the full image scene context, by proposing a new architecture (SR-CNN) with multiple semantic channels.

Exploring Semantic cues for action recognition. The perception and cognition of an action requires high-level abstraction and inference, leveraging and weighing of different semantic hints. For this reason, efforts have been made to exploit information for better action recognition. [16] jointly modeled the human and object interaction in a belief network. This idea was deepened in the work by [7], inferring human object reaction in a graphical model, which in turn can be used to discriminate confused action classes. Recently, [9] represented each video frame with object classification scores from ImageNet models [19] for action recognition. In [8], human and object detection results were explicitly incorporated

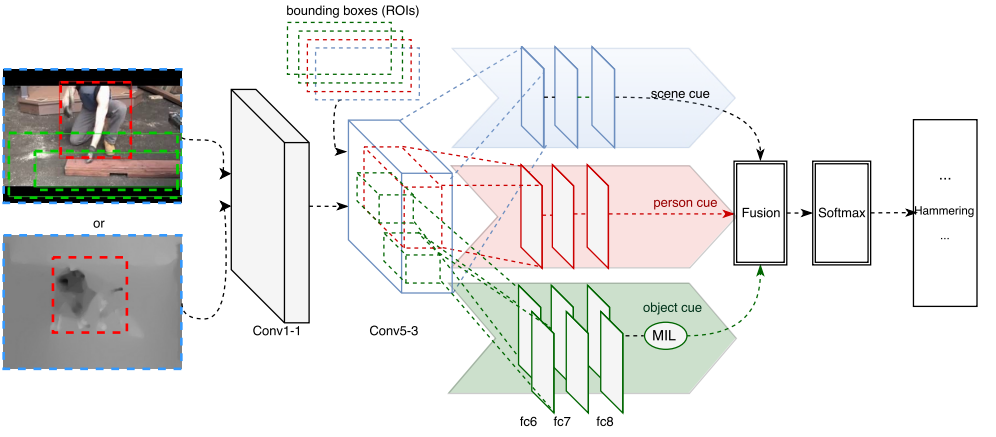


Figure 2: **Two-Stream SR-CNNs.** We incorporate the semantic regions detected by Faster R-CNN into the framework of two-stream CNNs for action recognition, and propose a new architecture, called as two-stream semantic region based CNNs. As spatial SR-CNNs and temporal SR-CNNs share similar architectures, we only plot one SR-CNNs for visual clarity.

into the recognition pipeline. However, their approach is based on traditional features. The most related to our work is R*CNN [5] for static image-based action recognition. Our work differs from R*CNN mainly in two folds: (1) no ground truth bounding boxes of human and objects are provided and we leverage advanced detectors to make our system more flexible and general, (2) action recognition in video domain is more challenging and our proposed network has a new stream with optical flow input.

3 Approach

In this section we describe our method for action recognition. First, we propose our network architecture of two-stream SR-CNNs. Then, we describe the extraction of inputs for semantic channels and discuss implementation details of our method.

3.1 Two-Stream SR-CNNs

Following the success of two-stream CNNs [20] and R*CNN [5], we propose a new architecture, called *two-stream semantic region based CNNs* (SR-CNNs). Figure 2 schematically illustrates our method. Two independent networks, spatial and motion SR-CNNs, respectively take in RGB images and optical flow fields as inputs.

The inputs are first passed through standard convolutional and pooling layers. We replace the last pooling layer with a RoiPooling [2] layer, which separate features for different semantic cues into parallel fully connected layers (called channels) using bounding boxes proposed from a Faster R-CNN [18] object detector (see subsection 3.2). Each channel produces a score vector independently. Specifically, to handle for multiple objects we adopt the MIL (Multiple Instance Learning) layer proposed in [5], which combines the scores via a max operation aiming to pick the most informative object for the task at hand. Finally, the scores obtained from different channels are merged by a fusion layer which is followed by a softmax layer for final action prediction.

Different from R*CNN [5], the parameters of these three branches are not shared because they represent very different information. This framework is still efficient, since every subsequent channel shares the computationally expensive convolutions which are performed at image-level in the beginning.

Regarding the fusion operation, four different strategies are designed as shown in Table 1. Max fusing takes the maximum score value among all channels for each class, essentially picking the strongest channel. Sum fusion directly adds up the scores from different channels, i.e. each channel is treated equal. Category-wise weighted fusion (Weighted-1) combines channel scores via weighted sum, aiming to represent varied relative contribution of each channel for different classes using their corresponding weights. As for correlation-wise weighted fusion (Weighted-2), the scores of other classes are also taken into consideration, implicitly encoding the correlation information between classes. Given L classes and C channels, the number of weights for Weighted Sum-1 and Weighted Sum-2 are $L \times C$ and $L \times L \times C$ respectively. All weights are trained together with the main parameters of networks through back-propagation process.

Fusion	Max	Sum	Weighted-1	Weighted-2
$s_l =$	$\max_c \{s_l^c\}$	$\sum_c s_l^c$	$\sum_c w_{ll}^c \cdot s_l^c$	$\sum_c \sum_m w_{lm}^c \cdot s_m^c$

Table 1: Various fusion methods. s_l^c is the score for class l obtained from channel c.

3.2 Semantic channels

In this section we describe our method to extract the bounding boxes of *action relevant* semantic cues with a generic object detector. We focus on person and object channels only since scene channel is simply the whole frame.

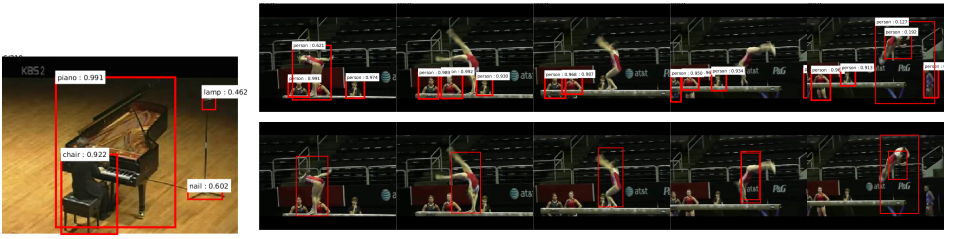
Detector. We extend the original Faster R-CNN model [18] from 20 object categories to 118 categories (listed in supplement) selected from ILSVRC2014 [19] (200 categories) and VOC 2007+2012 [1] detection challenge (20 categories), excluding categories such as small objects, food and most of animals. The complete training data is comprised of 196,780 images.

Objects detection. Objects detection in video dataset is challenging due to low resolution and motion blur as shown in Figure 3(a). To effectively filter out wrong or irrelevant detection results, we remove object proposals, whose (1) prediction confidence (from Faster R-CNN) is lower than λ (set as 0.1), (2) length is smaller than ε (set as 20 pixels), (3) overlapping with actors (if any) is zero.

Person detection. Human detection mostly fails in video data because of motion blur or large pose variations as shown in the first row of Figure 3(b). In general, our method needs to (1) filter out incorrect and irrelevant detection results; (2) recover missing detection results in individual frames; (3) refine locations of bounding boxes. Actors exhibit high motion salience, temporal consistency, and relatively larger size compared to "bystanders". Hence the actor in video can be identified by finding a "tube" maximizing the aforementioned properties, which can be reformulated as a Shortest Path Problem and solved efficiently using dynamic programming (DP).

Formally, a video is divided into N segments yielding N sets of raw detection results $\{\mathcal{B}_n\}_{n=1}^N$, where $\mathcal{B}_n = \{b_n^i\}_{i=1}^{N_n}$. The connection score for bounding boxes b_{n-1}^i, b_n^j in two consecutive segments is defined as:

$$s_{ij}\left(b_{n-1}^i, b_n^j\right) = s_{prob}\left(b_n^j\right) + s_{motion}\left(b_n^j\right) + s_{IoU}\left(b_{n-1}^i, b_n^j\right) + s_{size}\left(b_n^j\right), \quad (1)$$



(a) Object Detection (b) Action Performer Extraction

Figure 3: (a) Raw object detection results. The classification confidence and the detected object size usually provide good indication on the detection quality, while the relevance of an object is also strongly hinted by its distance to human. (b) Raw person detection (upper) and filtered detection (lower). Our method is able to find true action performer from noisy detections.

where s_{prob}, s_{motion}, s_{IoU} and s_{size} are person prediction probability, normalized magnitude of optical flow, intersection over union and size penalty for small persons respectively. By using dynamic programming, we can efficiently solve the optimization problem for person detection as follows:

$$\max_{ij} \left(\sum_{n=1}^{N} s_{ij}\left(b_{n-1}^{i}, b_{n}^{j}\right) \right), \forall b_{n}^{j} \in \mathcal{B}_{n} \tag{2}$$

Only one bound box is selected for each segment (N bounding boxes in total). We interpolate or extrapolate bounding boxes for other missing frames to obtain a refined consistent "person tube". Our method has several advantages over existing methods like [4] First, we do not require any ground truth annotation or prior action-specific knowledge, which makes our method more generic. Second, our method not only addresses temporal consistency as in [4], but also recovers and refines human bounding boxes in some frames. A sample result is shown in the bottom of Figure 3(b).

Evaluation. We quantitatively evaluate our method on JHMDB [10] and an annotated subset of UCF101 (24 categories) [6]. As metric, we compute the recall when the IoU of detected person tube with the ground truth is above a threshold ξ. With at most two tube proposals, we achieve recall of 82.65% ($\xi = 0.3$) and 36.45% ($\xi = 0.5$), and recall of 84.70% ($\xi = 0.3$) and 73.78% ($\xi = 0.5$) on JHMDB and UCF101 respectively. The lower recall in UCF101 is due to the fact all actors in multi-actor action classes (e.g. VolleyballSpiking) are annotated, while only the top-2 actors are considered in our case .

3.3 Implementation details

We choose the VGG16 network [21] as the basic network structure, which is trained on ImageNet data and has been used as initialization network for various tasks. We follow the two-stream settings as in [20, 29]. Both inputs are resized to 256×340. For spatial stream, the input is the single RGB image, while for temporal stream the input is 10-frame stacking of optical flow fields. For data augmentation, we adopted the corner cropping scheme suggested in [29], which introduces variance in both scale and ratio.

Training. In all experiments, we use the public VGG16 model [21] for initialization and set dropout rate to be 0.8 for both fully connected layers. Spatial stream is trained for 10,000 iterations; the learning rate is initialized to be 1×10^{-3} and is decreased every 4,000

iterations. For temporal stream, we train 19,000 iterations and set the initial learning rate to be 5×10^{-3}, which is reduced every 8,000 iterations. The batch size is set to 256 for both streams.

Testing. For testing, we follow the same routine as [20], which selects 10 samples from 5 crops and 2 flips for each frame. The final classification result for one video is given by averaging the classification scores of 25 evenly sampled frames with their all valid crops.

4 Evaluation

4.1 Dataset and evaluation protocol

We choose UCF 101 dataset because it is a good trade off between number of action classes and the variety of actions. This dataset contains 101 action classes and there are at least 100 video clips for each class. The whole dataset contains 13,320 video clips, which are divided into 25 groups for each action category. We first perform experiments on the split 1 of UCF101 dataset and then compare with the state-of-the-art algorithms over three splits.

4.2 Evaluation on fusion methods

In this subsection we investigate different fusion methods for multiple semantic channels proposed in Table 1. For training efficiency, we use spatial SR-CNNs with two semantic channels (i.e.scene, person) as our exploration architecture.

The results of four fusion strategies are reported in Table 2. We first observe that Max performs considerably lower than any kind of Sum methods. We believe this is because max fusion does not exploit the complementary contribution among channels; on the other hand, due to its selective update, max fusion would keep reinforcing the stronger channel, thus making the model sensitive to initialization. Second we see that the weighted fusion methods do not bring substantial improvement. We think it is because the training set is still not big enough to capture the real relative importance of different semantic cues for each individual action and also to reflect the true inter-class relation. Therefore, in the following experiments, we apply the Sum fusion.

Fusion	Max	Sum	Weighted Sum-1	Weighted Sum-2
Accuracy (%)	78.95	**80.46**	79.77	80.20

Table 2: Exploration of fusion methods with for spatial SR-CNNs on split1 of UCF101.

4.3 Evaluation on semantic channels

Models	S	P	S+P (two nets)	S+P (integrated)	S+P+O (integrated)
Spatial	79.42	73.82	80.16	**80.46**	79.71
Temporal	85.27	87.02	**87.85**	87.63	–

Table 3: Evaluation on SR-CNNs with different semantic channels on split1 of UCF101.

In this subsection, we study the effectiveness of incorporating different semantic cues for action recognition on both streams. Specifically, we compare five settings: (1) SR-CNNs with only "scene" channel, (2) SR-CNNs with only "person" channel, (3) late fusion of the scores produced by previous two SR-CNNs, (4) SR-CNNs jointly learned with two semantic

Split	Split1		Split2		Split3		Avg	
Model	Baseline	S+P	Baseline	S+P	Baseline	S+P	Baseline	S+P
Spatial	79.42	**80.46**	**77.14**	76.53	77.25	**77.97**	77.93	**78.32**
Temporal	85.27	**87.63**	88.13	**89.33**	86.96	**88.86**	86.79	**88.29**
Two Stream	90.98	**92.75**	91.45	**92.14**	91.05	**92.91**	91.15	**92.60**

Table 4: We propose using jointly trained "scene" + "person" model architecture (S+P) for spatial and temporal network and compare our proposed model with baseline over three splits on the UCF101 dataset, whereas baseline is the in-house implementation of two-stream CNNs [20]. Two Stream results are obtained by fusing spatial and temporal classification scores.

channels (5) SR-CNNs jointly learned with three semantic channels. The results are summarized in Table 3.

First, person and scene cues exhibit unequal relative performance in spatial and temporal streams. Specifically, for spatial net, recognition based on the whole scene is more accurate than based on person (79.42% vs 73.82%), the result is reversed in temporal net (85.27% vs. 87.02%). Secondly, compared to single cue settings, combination of the "person" and "scene" cues boosts the recognition performance for both spatial and temporal streams. Furthermore, this boost can be transferred to the proposed integrated model. For spatial network the classification accuracy of the integrated model even exceeds that of the late fusion of two separate models (80.46% vs. 80.16%).

However, when incorporating "object" into the joint framework for spatial stream, the accuracy drops down to 79.71%. Our explanation is: on one hand, in the current implementation, the initialization network is trained on object classification task, so "scene" channel inherently contains strong "object" information. Consequently, the integration of all three channels together implicitly adds more weight on scene/object channel, weakening the proved effectiveness of "person" channel. On the other hand, the dropped-down performance might be caused by the imperfect selection of relevant objects even though the concept of multiple instance learning has already been incorporated into the deep framework.

Based on previous empirical study, we are ready to build up our final action recognition method: sum-fused model with semantic channels "scene" and "person". The final results on all 3 splits are summarized in Table 4.

4.4 Evaluation on the effect of person tube detection

In this section we study the effect of the person detection quality on the model performance. For this purpose, we focus on the JHMDB [10] dataset which has actor bounding box annotations. The spatial "scene" + "person" SR-CNN model is evaluated with three different sources of person ROIs: (1) person ROIs from raw Faster R-CNN object detector, (2) filtered actor ROIs based on our proposed method as described in Section 3.2 and (3) ground truth person ROIs.

	Baseline	S+P (raw)	S+P (filtered)	S+P (GT)
Accuracy (%)	51.16	52.01	53.77	54.25

Table 5: Recognition performance of JHMDB (split 1) on spatial stream with different sources of person ROIs. Baseline is trained with VGG16 without semantic separation.

As shown in Table 5, the quality of extracted person ROIs directly affects action recognition performance. For JHMDB dataset (most of the actions are performed by single per-

Action type	P			P-O			P-S			P-O-S		
Model	S	S+P	S+P+O	S	S+P	S+P+O	S	S+P	S+P+O	S	S+P	S+P+O
Spatial	62.41	**65.48**	64.54	**82.52**	82.16	82.45	81.22	**82.48**	80.46	91.03	**91.33**	91.09
Temporal	79.98	**82.50**	-	85.16	**87.28**	-	87.11	**90.46**	-	88.62	**89.95**	-

Table 6: Model performance evaluated on four different action types: Body Motion Only (P), Human Object Interaction (P-O), Body Motion in Context (P-S), Human Object Interaction in Context (P-O-S). The evaluation is conducted on UCF101 split 1.

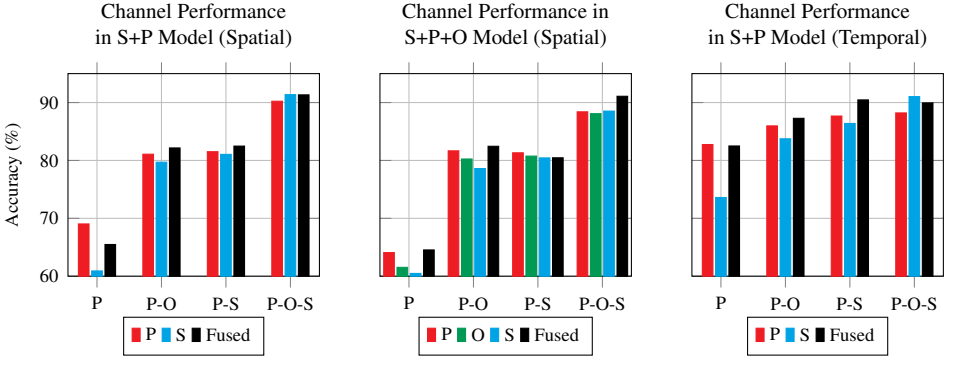


Figure 4: Individual channel performance evaluated using respective channel scores in the same model before fusion for four action types: Body Motion Only (P), Human Object Interaction (P-O), Body Motion in Context (P-S), Human Object Interaction in Context (P-O-S). Person, object and scene channels are represented in red, green and blue respectively, while the sum-fused result is in black. The evaluation is conducted on UCF101 split 1.

son), SR-CNN model with additional "person" channel could boost the accuracy even with raw person detection; furthermore, with filtered person ROIs, the recognition performance is raised evidently to the level of having ground truth annotations. This suggests that our method is robust against sub-optimal ROI extraction.

4.5 Evaluation on action types

In this section, we focus on our initial motivation and look into the effectiveness of SR-CNNs on different types of action classes introduced in Section 1 (Figure 1). We first compare the average accuracy of each action type in the settings with single scene channel (our baseline) and with multiple semantic channels.

As shown in Table 6, for both spatial and temporal streams the baseline model (S) significantly underperforms other SR-CNNs for "Body Motion Only" actions. This suggests that the baseline approach is strongly biased towards the global context. Hence for action classes where "scene" is a variable and uninformative, the recognition performance is influenced by fitting the context of training data'. By adding "person" and "object" channels, a substantial performance boost can be observed in this action type for both streams, proving that our proposed method effectively improves the robustness towards uninformative variances. Moreover, introducing "person" channel also induces a noteworthy improvement for "Body Motion in Context" type in both streams. This owes to the finer pooling-grids after RoiPooling, which encourages the model to capture more subtle differences in body pose between similar actions.

To further identify the source of improvement and understand how semantic channels

Method	iDT+FV [24]	MoFAP [28]	Object [9]	Two Stream [20]	[20]+LSTM [17]	TDD+FV [27]	Deep Two Stream [29]	Ours
Accuracy (%)	85.9	88.3	88.5	88.0	88.6	90.3	91.4	**92.6**

Table 7: Comparison with the state-of-the-art methods on the UCF101 dataset.

complement each other, in Figure 4, we list the accuracy of each channel by evaluating on respective channel score *before* fusion. As clearly shown, the performance gain for "Body Motion Only" type is empowered by "person" channel. Also it is evident that the performance of each semantic channel is aligned with their importance for the corresponding action type. For example, "person" channel is dominating in "Body Motion Only" action type, while "scene" is more indicative for "Human Object Interaction in Context". Furthermore, the fused result can be significantly raised by the stronger channel (for "Body Motion Only" and 'Human Object Interaction in Context" action types). When channels are equally decisive (for action type "Human Object Interaction" and "Body Motion in Context"), the fused result even exceeds both individual channels, demonstrating the desired synergic effect.

On the other hand, the individual channel performance in S+P+O model (the right figure of Figure 4) provides some hints for the question: why is "object" channel not helpful. Unlike "person" channel, which greatly complements "scene" channel, "object" channel does not show any distinct contribution but mostly aligns with the "scene" channel. This confirms our argument in Section 4.3.

4.6 Comparison with the state of the art

In this section we compare our proposed approach with other state-of-the-art methods on the UCF101 dataset. The results are summarized in Table 7. Specifically, we compare our method with both traditional approaches such as improved trajectories (iDTs) [24], MoFAP representations [28], and deep learning representations such as Trajectory-pooled Deep convolutional Descriptors (TDDs) [27], Two-Stream CNNs [20], Deep Two-Stream CNNs [29], and CNN+LSTM [17]. For fair comparison, we select the recognition performance of VGGNet architecture from [29]. Our best performance is better than those previous approaches by 1.2%. This superior performance demonstrates the effectiveness of explicitly incorporating semantic cues in our SR-CNN framework.

5 Conclusion and Future work

In this paper, we proposed an integrated framework that explicitly leverages multiple semantic cues for CNN-based action recognition in videos. Our empirical study shows that our method enhances robustness and generalizability of conventional frame-wise CNN-based approaches, allowing us to achieve the state of the art performance. Additionally, our thorough evaluation on different semantic channels also provides an insight for a more comprehensive understanding of action. Regarding to future work, we are interested in applying recurrent structure on the person channel to model the long term human dynamics, which is ignored in current settings.

Acknowledgements

This work is supported by ERC Advanced Grant *VarCity* (No.273940).

References

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[2] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[4] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *CVPR*, pages 759–768, 2015.

[5] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with R*CNN. In *ICCV*, pages 1080–1088, 2015.

[6] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://www.thumos.info/, 2015.

[7] Abhinav Gupta and Larry S Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, pages 1–8, 2007.

[8] Nazli Ikizler-Cinbis and Stan Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. *ECCV*, pages 494–507, 2010.

[9] Mihir Jain, Jan C van Gemert, and Cees GM Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, pages 46–55, 2015.

[10] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013.

[11] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, 2013.

[12] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[14] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] Darnell J Moore, Irfan A Essa, and Monson H Hayes III. Exploiting human actions and object context for recognition tasks. In *ICCV*, pages 80–86, 1999.

[17] Joe Yue-Hei Ng, Matthew J. Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, pages 4694–4702, 2015.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[20] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[23] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[24] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013.

[25] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[26] Limin Wang, Yu Qiao, and Xiaoou Tang. Latent hierarchical model of temporal structure for complex activity classification. *IEEE Trans. Image Processing*, 23(2):810–822, 2014.

[27] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015.

[28] Limin Wang, Yu Qiao, and Xiaoou Tang. MoFAP: A multi-level representation for action recognition. *International Journal of Computer Vision*, 119(3):254–271, 2016.

[29] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[30] Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.