INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Multilayer and Multi Modal Fusion of Deep Neural Networks for Video Classification

BOGGARAPU SRINIVASULU
Research Scholar
Department of Computer Science and Enigineering
JS UNIVERSITY, Shikohabad ,Uttar Pradesh.

Dr. K. VIJAYA BHASKAR
Associate Professor
Department of Computer Science and Enigineering
JS UNIVERSITY, Shikohabad ,Uttar Pradesh.

Abstract: This paper presents a novel framework to combine multi- ple layers and modalities of deep neural networks for video classification. We first propose a multilayer strategy to simultaneously capture a variety of levels of abstraction and invariance in a network, where the convolutional and fully connected layers are effectively represented by the proposed feature aggregation methods. We further introduce a multi modal scheme that includes four highly complementary modalities to extract diverse static and dynamic cues at multiple temporal scales. In particular, for modeling the long-term temporal information, we propose a new structure, FC-RNN, to effectively transform pre-trained fully connected layers into recurrent layers. A robust boosting model is then introduced to optimize the fusion of multiple lay- ers and modalities in a unified way. In the extensive experiments, we achieve state-of-the-art results on two public benchmark datasets: UCF101 and HMDB51.

Keywords: Video Classification, Deep Neural Networks, Boosting, Fusion, CNN, RNN

I. INTRODUCTION

Content based video classification is fundamental to intelligent video analytics including automatic categorizing, searching, indexing, segmentation, and retrieval of videos. It has been applied to a wide range of real-word applications, for instance, surveillance event detection, semantic indexing, gesture control, etc. It is a challenging task to recognize unconstrained videos because 1) an appropriate video representation can be task-dependent, e.g., coarse ("swim" vs. "run") and fine-grained ("walk" vs. "run") categorizations; 2) there may be multiple streams of information that need to be taken into account, such as actions, objects, scenes, and so forth; 3) there are large intra-class variations, which arise from diverse viewpoints, occlusions and back- grounds. As the core information of videos, visual cues pro- vide the most significant information for video classification. Most traditional methods rely on the bag-of-visual-words (BOV) representation which consists of computing and ag- gregating visual features. A variety of local and global visual features have been proposed, for instance, GIST and SIFT can be used to capture static information in spatial frames, while STIP and improved dense trajectories (iDT) are widely employed to compute both appearance and motion cues in videos.

Recently there is a growing trend to learn robust feature representations with deep neural networks for various tasks such as image classification, object detection , natu- ral language processing, and speech recognition. As one of the most successful network architectures, the recent surge

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

of convolutional neural networks (CNN) has encour- aged a number of methods to employ CNN for video classification. Karparthy et al. made the first attempt to use a buffer of video frames as input to networks, however, the results were inferior to those of the best hand-engineered features. However, these methods focus on short or mid-term information as feature representations are learned in short-time windows. This is insufficient for video classification since complex events are better described by leveraging the temporal evolution of short-term contents. In order to capture long-term temporal clues in videos, re- current neural networks (RNN) were applied to explicitly model videos as an ordered sequence of frames.

CNN based video classification algorithms typically make predictions using the softmax scores or, alternatively, they use the last fully connected layer as a feature representation because CNN hierarchically computes abstract and invariant representations of the inputs. However, leveraging information across multiple levels in a network has proven beneficial for several tasks such as natural scene recognition, object segmentation and optical flow computation. This is somewhat expected since convolutional layers retain the spatial information compared to fully connected layers. For video classification, we argue that appropriate levels of abstraction and invariance in CNN for video representation are also task- and class-dependent.

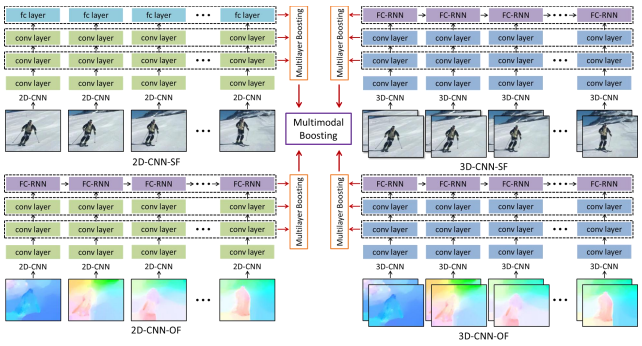


Figure 1: An overview of the proposed multilayer and multimodal fusion framework for video classification.

We use four modalities to extract highly complementary information across multiple temporal scales. For each single modality, discriminative representations are computed for convolutional and fully connected layers. We employ an effective boosting model to fuse the multiple layers and modalities. Box colors are encoded according to different networks: 2D-CNN and 3D-CNN with and without RNN. We propose FC-RNN to model long-term temporal information rather than using the standard RNN structure. Distinguishing "soccer game" and "basketball game" requires high-level representations to model global scene statistics. However, classification of "playing guitar" and "playing vio- lin" demands fine-scale features to capture subtle appearance and motion features. Therefore, leveraging the multilayer abstractions is able to simplify video classification.

Although a significant progress in recent years has been achieved in the development of feature learning by deep neural networks , it is clear that none of the features have the same discriminative capability over all classes. For example, videos of "wedding ceremony" are strongly as- sociated with static scenes and objects, while "kissing" is more related to dynamic motions. It is therefore widely accepted to adaptively combine a set of complementary features rather than using a single feature for all classes. Si- monyan et al. [36] proposed the two-stream networks based on 2D-CNN to explicitly incorporate motion information from optical flow to complement the static per-frame in- formation. A simple late fusion was adopted to combine the soft max scores of two networks by either averaging or with a linear classifier. This method has been widely utilized for video analysis [8, 46] thanks to the two complementary modalities and outstanding performance. Nevertheless, a question of which robust modalities to exploit and how to effectively perform multimodal fusion still remains open for video classification.

In this paper, we propose a multilayer and multimodal fusion framework of deep neural networks for video classification. The multilayer strategy can simultaneously capture a variety of levels of abstractions in a single network, which is able to adapt from coarse- to fine-grained categorizations. Instead of using only two modalities as in the two-stream networks [36], we

ISSN: 2057-5688

propose to use four highly complementary modalities in the multimodal scheme, i.e., 2D-CNN on a single spatial frame and optical flow image as well as 3D- CNN on a short clip of spatial frames and optical flow images. They not only effectively harness the static objects and dynamic motions in videos but also extensively exploit the multiple temporal clues. As for the fusion of multiple layers and modalities, we adopt a powerful boosting model to learn the optimal combination of them.

Fig. 1 illustrates the overview of our proposed multilayer and multimodal fusion framework. Given an input video, the four modalities are used to extract complementary in- formation at short and mid-term temporal scales. Instead of using the standard RNN structure, we propose FC-RNN to model the long-term temporal evolution across a whole video. FC-RNN takes advantage of pre-trained networks to transform the pre-trained fully connected (fc) layers into recurrent layers. In the following, we use 2D-CNN-SF, 2D- CNN-OF, 3D-CNN-SF, 3D-CNN-OF to indicate 2D-CNN and 3D-CNN on spatial frames and optical flow, respectively. For each individual network, an improved Fisher vector (iFV) is proposed to represent convolutional (conv) layers and an explicit feature map is used to represent fc lay- ers. We then employ a robust boosting model to learn the optimal combination of multiple layers and modalities. The main contributions of this paper are summarized as follows.

- We present a multilayer fusion strategy to capture mul- tiple levels of abstraction and invariance in a single network. We propose to use iFV and explicit feature map to represent features of conv and fc layers.
- We introduce a multimodal fusion scheme to incorpo- rate the four highly complementary modalities to ex- tract static and dynamic cues from multiple temporal scales. In particular for the long-term temporal model- ing, we propose FC-RNN to preserve the generalization properties of pre-trained networks.
- We adopt an effective boosting model for video classi- fication by fusing multiple layers and modalities in an optimal and unified way.
- In the extensive experiments, our method achieves su- perior results on the well-known UCF101 and HMDB51 benchmarks.

II. RELATED WORK

Videos have been studied by the multimedia community for decades. Over the years a variety of problems like mul- timedia event recounting, surveillance event detection, ac-tion search, and many more have been proposed. A large family of these studies is about video classification. Conventional video classification systems hinge on extraction of local features, which have been largely advanced in both detection and description. Local features can be densely sampled or selected by maximizing specific saliency functions. Laptev proposed STIP to detect sparse space-time interest points by extending the 2D Harris corner detector into 3D. Wang et al. introduced the improved dense trajectories (iDT) to densely sample and track interest points from multiple spatial scales, where each tracked interest point generates a set of descriptors to represent shape and motion. Many successful video classification systems use iDT with the motion boundary histogram (MBH) descriptor which is the gradient of horizontal and vertical components of optical flow. It is widely recognized as the state-of-the-art feature for video analysis.

After local feature extraction, a number of coding tech- niques have been proposed for feature quantization, e.g., sparse coding and locality-constrained linear coding . Then average pooling and max pooling are normally used to aggregate statistics from local features. Several more advanced coding methods, e.g., Fisher vector (FV) and vector of locally aggregated descriptors (VLAD) , have emerged to reserve high order statistics of local features and achieve noticeably better performance. However, these methods obviously incur the loss of spatio-temporal order of local features. Extensions to the completely orderless aggregation methods include spatiotemporal pyramid and super sparse coding vector . The graphical models, such as hidden Markov model (HMM) and conditional random

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

fields (CRF), are also popular methods to explore the long- term temporal information in videos.

The trajectory-pooled deep convolutional descriptor (TDD) was presented in to incorporate videos' temporal nature by using trajectory constrained sampling and pooling. This method shares the advantages of both hand-engineered fea- tures and deep-learned representations. While the improved networks using RNN can model long-term temporal order, our proposed multimodal method provides multi-temporal scales with short, mid, and long-term time contexts.

Figure 2: Illustration of multilayer representation and fusion.

The proposed feature aggregation methods are used to represent fully connected and convolutional layers over time. The introduced boosting algorithm is applied to combine the representations from multiple layers.

III. MULTI LAYER REPRESENTATIONS

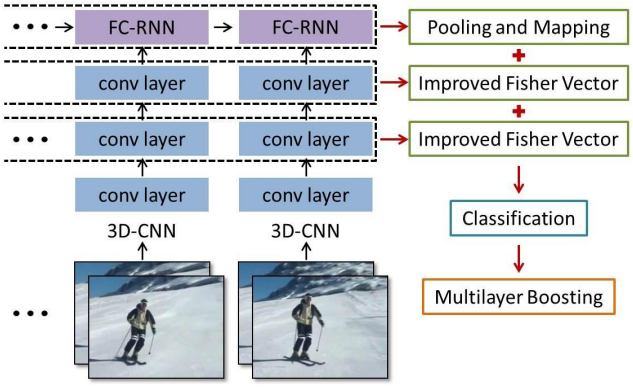
As a hierarchical feed-forward architecture, CNN progressively computes abstract and invariant representations of inputs. Recognition algorithms based on CNN often make predictions based on SoftMax scores or the last layer which is the most resistant to variables in the preceding layers. How ever, we argue that various abstractions such as poses, articulations, parts, objects, etc, learned in the intermediate lay- ers can provide multiple semantics from fine-scale to global descriptions for video classification. Moreover, we propose a concept of convlet to utilize the spatial information reserved in conv layers to refine the final feature representation. In this section, we describe the detailed procedures to compute multilayer representations as illustrated in Fig. 2.

A. Improved Fisher Vector by Convlet

Recent work on visualizing and understanding CNN re- veals that conv layers demonstrate many intuitively desir- able properties such as strong grouping within each feature map and exaggeration of discriminative parts of objects. Therefore, a set of appropriate levels of compositionality in conv layers are able to supply plenty of fine-scale information to the category-level semantics. Meanwhile, the features of multiple layers come for free because they are already ex- tracted during the forward pass. Furthermore, compared to fc layers, conv layers contain the spatial information. It can be applied to adaptive pooling and feature refinement because the discriminative information for video classifica- tion is often unevenly distributed in spatial domain.

IV. FC-RNN STRUCTURE

Most networks hinge on short or mid-term contents such as a single frame or a buffer of



frames where features are independently extracted for video classification. We believe that there are important connections between frames and the entire video is supposed to be processed as an ordered sequence. To address this intuition, we propose a simple and effective structure, FC-RNN, to transform a network pre-trained on separate frames or clips to deal with video as a whole sequence.

learn the specific order of videos in the training set, there- fore we randomly permute the order of training videos for each epoch. This operation slows down convergence but im- proves generalization. The regularization term which forces to learn weights with smaller l2-norm also helps generalization. With intention of preventing the gradients from exploding in recurrent layers, we employ soft gradient clip- ping in the following way. For each computed gradient g in stochastic gradient descent (SGD), we check if its l2-norm g is greater than a pre-defined threshold $\delta = 10$. If that is the case, we rescale the gradient to g $g\delta/\|g\|$. We find that without gradient clipping the

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

explosion of gradient val- ues is a critical barrier to successfully training the networks. To further improve generalization, we train networks with drop-out on the outputs of recurrent layers. During train- ing, we set the outputs of the recurrent layers to 0 with a probability of p = 0.5, and scale the activations of other neurons by a factor of p/(1 − p).

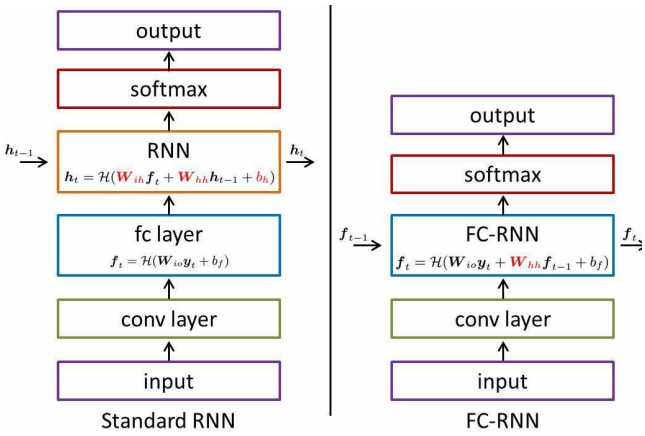


Figure 3: Comparison of standard RNN and FC- RNN. The variables in red correspond to the parameters that need to be trained from scratch.

V. MULTIMODAL REPRESENTATIONS

Since the visual information in videos is a juxtaposition of not only scenes and objects but also atomic actions evolving over the whole video sequence, it is favorable to capture and combine both static appearances and dynamic motions. To address this challenge we bring the multimodal approach to model a variety of semantic clues in multi-temporal scales. Fig. 1 demonstrates the proposed four modalities with mu- tually highly complementary information in short, mid, and long-term temporal contexts.

The two networks operating on spatial frames (single frame in 2D-CNN-SF and short clip of frames in 3D-CNN-SF) can capture objects and scenes that are strongly correlated to certain video categories, e.g., snow and mountains in Fig. 1 indicate skiing. 2D-CNN-SF is essentially an image classifi- cation network which can be built upon the recent advances in large-scale image recognition methods and datasets. 3D- CNN-SF selectively attends to both motion and appearance cues through spatio-temporal convolution and pooling oper- ations. It encapsulates the mid-term temporal information as the network's input is a short video clip (e.g., 16 spatial frames). The colorized optical flow maps enable us to reduce over-fitting and training time by leveraging on the pre-trained models from large-scale image datasets. Since the input is a single colorized image, 2D-CNN-OF captures the fine-scale and short-term temporal information between a pair of adjacent frames. 3D-CNN-OF models the high or- der motion cues such as spatial and temporal derivatives of optical flow which has been successfully applied to hand- engineered features [44]. This modality also encapsulates the mid-term temporal clues. Similar to 3D-CNN-SF, FC- RNN is also employed to learn the long-term temporal order of 2D-CNN-OF and 3D-CNN-OF.

Outputs of last four layers of each network are used to represent videos. Special attention is paid for the mini-batch assembling to deal with varied video length. We fill consequently all frames of a video into a mini-batch and fill with another video if there is still space in the mini-batch. When the limit of a mini-batch is reached and there are frames left then we fill them in the next. In the case there are no more frames to fill into a mini-batch, we fill them with zeros and these examples are not used in computation.

VI. EXPERIMENTS

In this section, we extensively evaluate the proposed mul- tilayer and multimodal fusion method on the two public benchmark datasets for video classification: UCF101 and HMDB51 . In all experiments, we use LIBLINEAR as the linear SVM solver. Experimental results show that our algorithm achieves the state-of-the-art results on the two benchmarks.

A. Experimental Setup

- **Datasets**

The UCF101 dataset contains 101 action classes with large variations in scale, viewpoint, illumination, camera motion, and cluttered background. It consists of 13,320 videos in total. We follow the standard experimental set- ting as

ISSN: 2057-5688

to use three training and testing splits. In each split, 20% of training data is used as validation set for boost- ing model selection. The first split of UCF101 (denoted as UCF101*) is also used to evaluate and understand the con- tribution of individual components. We report the average accuracy over the three splits as the overall measurement.

The HMDB51 dataset is collected from a wide range of sources from digitized movies to online videos. It contains 51 categories and 6,766 videos in total. This dataset includes original videos and stabilized ones. Our evaluations are based on the original version. There are 70 videos for training and 30 videos for testing in each class. We use 40% of training data as validation set to perform model selection for boosting. We follow the evaluation protocol defined in to use three training and testing splits and report the mean accuracy over the three splits.

- **Implementations**

We implement the networks of four modalities in Theano with cuDNN4 on NVIDIA DIGITS DevBox with four Titan X GPUs. 2D-CNN and 3D-CNN in the experiments are initialized by VGG16 pre-trained on ImageNet and C3D videos. For 2D-CNN we skip every second frame and oper- ate on a single frame . Training frames are gen- erated by random cropping and flipping video frames, while for testing, only a central crop with no flipping is evaluated. Since the two datasets are of quite different sizes, we apply different learning rate scheduling. For UCF101, we fine-tune 9 epochs with initial learning rate $\lambda = 3 \quad 10{-}4$ and divided by 10 after each 4 epochs. For HMDB51, we perform fine- tuning for 30 epochs with the same initial learning rate but divided by 10 after every 10 epochs. All network parameters that are not with pre-trained weights are initialized with random samples drawn from a zero-mean normal distribution with a standard deviation of 0.01. We use the frame-wise negative log-likelihood of a mini-batch as the cost function, which is optimized by SGD with a momentum of 0.9.

B. Experimental Results

- **Evaluation of Feature Aggregations**

We first evaluate the performance of iFV to represent conv layers in different modalities. Compared to the traditional aggregation methods, iFV retains high order statistics; in particular, it adaptively weights the features of a conv layer according to the associated spatial weights learned by the proposed convlet. We keep 300 out of 512 components in PCA. For the spatial weight normalization, we find sigmoid is more discriminative than softmax, e.g., iFV with sigmoid outperforms that with softmax by 0.6% for conv5 layer in 2D-CNN-SF. The sigmoid function is therefore used in the following experiments. We set K = 128 Gaussian compo- nents for both methods so the final feature dimension is 76.8K. We compare iFV with the conventional FV in Table 1 where iFV consistently outperforms FV for conv layers in all modalities with the improvements ranging from 0.6% to 2.5%. It is observed to be more improved for conv4 than conv5 probably because of the finer spatial information preserved in the lower layer. These improvements clearly show the advantages of utilizing the spatial discriminability learned by convlets in conv layers to enhance the feature representation.

We employ temporal max pooling to aggregate fc layers, which are further extended by the explicit feature map to approximate non-linear kernels. This representation is not only equipped with additional non-linearity but also bene- fits from the efficiency of learning and prediction in linear SVM. We demonstrate the results of fc layers in 3D-CNN- SF with approximated non- linearities in Table 2. Both fc6 and fc7 are transformed to recurrent layers by FC-RNN. We use l2-norm and z = 3 in explicit feature map so the extended feature dimension is 28,672. The baseline method is the linear representation by temporal max pooling with- out feature mapping. We evaluate three additive non-linear kernels: $\chi 2$, Jensen-Shannon and intersection kernels, which are widely used in machine learning and computer vision. All non-linear representations outperform the linear one, es- pecially the representation with intersection kernel achieves the best results. We thus use the intersection non-linearity

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

approximation to represent fc layers in the following experiments.

Modality	Layer	FV [34]	iFV
2D-CNN-SF	conv4	74.2%	**76.7%**
	conv5	79.6%	**80.6%**
2D-CNN-OF	conv4	75.6%	**78.1%**
	conv5	81.9%	**82.6%**
			%
			%
			%
			%
			V to
			es on
			ection
			.9%
	fc7	82.4%	82.9%	83.0%	**83.2%**

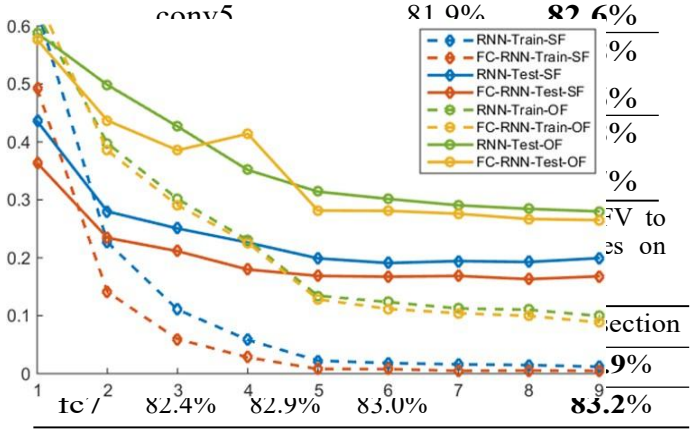


Table 2: Comparison of different non-linear approx- imations to represent fully connected layers in 3D- CNN-SF on UCF101*.

Figure 4: Comparison of the proposed FC-RNN and the standard RNN in training and testing of 3D- CNN-SF and 3D-CNN-OF on UCF101*.

➢ Evaluation of FC-RNN

Our method extensively extracts the static and dynamic information in multi-temporal scales. 2D-CNN and 3D- CNN on spatial frames and optical flow images compute features from short-term and mid-term temporal contexts. FC-RNN is then employed to model each video as an ordered sequence of frames or clips to capture the long-term temporal order. Since FC-RNN maintains the structure of a pre-trained network to the greatest extent, it is therefore effective to preserve important generalization properties of the network when fine-tuned on a smaller target dataset. Moreover, FC-RNN achieves higher accuracy and is faster to converge compared to the standard RNN. We compare the training and testing performances of our proposed FC-RNN and the standard RNN in Fig. 4. To avoid figure clutter, we demonstrate the comparison for 3D-CNN-SF and 3D-CNN- OF, similar phenomena is observed on 2D-CNN-OF as well. FC-RNN is generally able to alleviate over-fitting and con- verge faster, e.g., FC-RNN outperforms standard RNN and LSTM by 3.0% and 2.9% on 3D-CNN-SF. In comparison to the networks without recurrent connections, FC-RNN significantly improves the modalities of 2D-CNN-OF, 3D-CNN- SF and 3D-CNN-OF by 3.3%, 3.2% and 5.1%, respectively. This is evident to show the benefits of FC-RNN in modeling the long-term temporal clues.

➢ Evaluation of Multilayer Fusion

Here we evaluate the multilayer fusion on combining var- ious layers for individual modalities. Table 3 shows the per- formance of each single layer across different modalities and the fusion results on the two datasets. Although the last layer in a network is the most sensitive to category-level se- mantics, it is not unusual for lower layers to have on par or superior results, e.g., conv5 of 2D-CNN-OF on UCF101 and conv5 of 2D-CNN-SF on HMDB51. So it is of great potential to exploit the intermediate abstractions such as parts, objects, poses, articulations and so on for video clas- sification. It is also of interest to observe that most layers produce accuracies better than the baseline of softmax, i.e., the prediction outputs of a network. This again validates the merit of the proposed feature aggregation methods to represent conv and fc layers.

If we use the boosting algorithm to combine multiple lay- ers, the fusion result significantly outperforms the baseline for all modalities, especially for 3D-CNN-OF with 7.2% and 7.9%

ISSN: 2057-5688

gains on UCF101 and HMDB51. This demonstrates that various abstractions extracted in multiple layers are of rich complementarity. Although boost-c is more flexible to have class-specific mixing coefficients, its results are inferior to those of boost-u. This is because the model of boost-c tends to be over-fitting since the C M parameters to fit in boost-c requires more training data than the M parameters in boost-u. We thus use boost-u in the following fusion ex- periments. 3D-CNN-SF is the best modality before fusion as it jointly models appearance and motion information. Af- ter multilayer fusion the other two modalities involving dy- namic cues are enhanced to the similar performance level, which shows the boosting method is successful to maximize the capability of a network.

➢ Evaluation of Multimodal Fusion

We now demonstrate the multimodal fusion to combine the proposed four modalities. Since our networks are ini- tialized by the models pre-trained on large-scale image and video datasets, it is natural to fine-tune these networks for the two modalities of spatial frames. However for the other two modalities involving optical flow, they are distant from the source if we regard fine-tuning as a way of domain trans- formation. We introduce a simple but effective method to bridge the two domains—initialize optical flow networks by spatial frame models that have been fine-tuned on the tar- get domain. As shown in Table 4, compared to the networks directly fine-tuned on the source model (i.e., not initialized by 3D-CNN-SF), our initialization remarkably improves the results by 3.1% and 4.1% for 3D-CNN-OF trained with and without FC-RNN.

We finally compare our results with the most recent state- of-the-art methods in Table 3. Our method produces the best accuracy on UCF101 with a clear margin to other com- peting algorithms. It is more challenging to fine-tune net-works and train boost-u on HMDB51, where each train- ing split is 2.6 times smaller than UCF101.

UCF101	(%)	HMDB51	(%)
STIP + BOVW [21]	43.9	STIP + BOVW [21]	23.0
DT + MVSV [4]	83.5	DT + MVSV [4]	55.9
iDT + HSV [33]	87.9	iDT + HSV [33]	61.1
C3D [42]	85.2	iDT + FV [44]	57.2
LRCN [8]	82.9	Motionlets [3]	42.1
TDD [46]	90.3	TDD [46]	**63.2**
RNN-FV [25]	88.0	RNN-FV [25]	54.3
Two-Stream [36]	88.0	Two-Stream [36]	59.4
MultiSource CNN [32]	89.1	MultiSource CNN [32]	54.9
Composite LSTM [39]	84.3	Composite LSTM [39]	44.1
Ours	**91.6**	Ours	**61.8**

Table 3: Comparison of the multimodal fusion to the state-of-the-art results.

Our method still achieves superior performance on HMDB51, while other competitive results are based on the improved dense trajectories which require quite a few hand-crafted process such as dense point tracking, human detection, camera mo- tion estimation, etc. As shown on UCF101, large train- ing data is beneficial for training networks and boosting, so we are planing to explore the techniques such as multi-task learning and temporal elastic deformation to increase the effective training size of HMDB51.

VII. CONCLUSION

In this paper, we have presented a novel framework to fuse deep neural networks in multiple layers and modalities for video classification. A multilayer strategy is proposed to in- corporate various levels of semantics in each single network. We employ effective feature aggregation methods, i.e., iFV and explicit feature map to represent conv and fc layers. We further introduce a multimodal approach to capture diverse static and dynamic cues from four highly complementary modalities in multiple temporal scales. FC-RNN is then proposed to effectively model long-term temporal order by leveraging the generalization properties of pre-trained net- works. A powerful boosting model is in the end used for the optimal combination of multilayer and multimodal represen- tations. Our approach is

ISSN: 2057-5688

extensively evaluated on two public benchmark datasets and achieves superior results compared to a number of most recent methods.

REFERENCES

1) D. Borth, T. Chen, R. Ji, and S. Chang. Sentibank:

2) large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In ACM Multimedia, 2013.

3) T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In ECCV, 2004.

4) Z. Cai, L. Wang, X. Peng, and Y. Qiao. Motionlets: mid-level 3D parts for human motion recognition. In CVPR, 2013.

5) Z. Cai, L. Wang, X. Peng, and Y. Qiao. Multi-view super vector for action recognition. In CVPR, 2014.

6) S. Chen, J. Wang, Y. Liu, C. Xu, and H. Lu. Fast feature selection and training for AdaBoost-based concept detection with large scale datasets. In ACM Multimedia, 2010.

7) G. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. TASLP, 2012.

8) Demiriz, K. Bennett, and J. Taylor. Linear programming boosting via column generation. JMLR, 2000.

9) J. Donahue, L. Hendricks, S. Guadarrama, and

10) M. Rohrbach. Long-term recurrent convolutional networks for visual recognition and description. In CVPR, 2015.

11) R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. JMLR, 2008.

12) P. Fischer, A. Dosovitskiy, E. Ilg, P. Hausser, C. Hazirbas,

13) Golkov, P. Smagt, D. Cremers, and T. Brox. FlowNet: learning optical flow with convolutional networks. In ICCV, 2015.

14) J. Francoise, N. Schnell, and F. Bevilacqua. A multimodal probabilistic model for gesture based control of sound synthesis. In ACM Multimedia, 2011.

15) P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In ICCV, 2009.

16) J. Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. TPAMI, 2009.

17) Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In CVPR, 2015.

18) K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CoRR, 2015.

19) H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In CVPR, 2010.

20) S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. TPAMI, 2013.

21) W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. Loui. Short-term audio-visual atoms for generic video concept classification. In ACM Multimedia, 2009.

22) Karpathy, G. Toderici, S. Shetty, T. Leung,

23) R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.

24) Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

25) H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and

26) T. Serre. HMDB: a large video database for human motion recognition. In ICCV, 2011.

27) G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and

ISSN: 2057-5688

28) M. Jordan. Learning the kernel matrix with semidefinite programming. JMLR, 2004.

29) Laptev. On space-time interest points. IJCV, 2005.

30) Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In CVPR, 2008.

31) G. Lev, G. Sadeh, B. Klein, and L. Wolf. RNN Fisher vectors for action recognition and image annotation. In CoRR, 2015.