# MULTIMEDIA EVENT DETECTION VIA DEEP SPATIAL-TEMPORAL NEURAL NETWORKS

*Jingyi Hou, Xinxiao Wu, Feiwu Yu, and Yunde Jia*

Beijing Laboratory of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, Beijing 100081, P.R. China
{houjingyi; wuxinxiao; yufeiwu; jiayunde}@bit.edu.cn

## ABSTRACT

This paper proposes a novel method using deep spatial-temporal neural networks based on deep Convolutional Neural Network (CNN) for multimedia event detection. To sufficiently take advantage of the motion and appearance information of events from videos, our networks contain two branches: a temporal neural network and a spatial neural network. The temporal neural network captures motion information by Recurrent Neural Networks with the mutation of gated recurrent unit. The spatial neural network catches object information by using the deep CNN, to encode the CNN features as a bag of semantics with more discriminative representations. Both the temporal and spatial features are beneficial for event detection in a fully coupled way. Finally, we employ the generalized multiple kernel learning method to effectively fuse these two types of heterogeneous and complementary features for action recognition. Experiments on TRECVID MEDTest 14 dataset show that our method achieves better performance than the state of the art.

***Index Terms***— spatial-temporal networks, recurrent neural networks, multimedia event detection

## 1. INTRODUCTION

Automatically detecting events in videos attracts increasing research interests due to its wide applications in video surveillance, video content analysis, and video retrieval. However, it is difficult to design an efficient and robust feature for event detection due to the large intra-class variations in unconstrained events. Recent methods for event detection include combining multiple hand-crafted features [1, 2, 3] and applying Convolutional Neural Networks (CNNs) [4, 5, 6]. The latest works on CNNs [4, 5] show that the CNN-based method achieve better performance than other hand-craft based methods.

Encouraged by the success of CNN-based methods for event detection, we propose deep spatial-temporal neural networks for detecting events in videos. As shown in Fig. 1, the spatial-temporal nets exploit the spatial and temporal information from deep CNN features of event videos. The CNN is pre-trained on the ImageNet dataset [8] for better extracting frame-level features, followed by two-branches: the temporal net and the spatial net which capture the temporal motion information and the spatial related object information of events, respectively. Since both of the motion and object information are significantly important and beneficial for describing multimedia events, we additionally adopt the multiple kernel learning to effectively combine these two different types of features for recognition.

Specifically, the Recurrent Neural Networks (RNNs) with the MUTation of gated recurrent units (MUT) [9] are used as the temporal net to process the sequential information. To avoid the over-fitting problem caused by the lack of labeled data, the RNNs with single layer are practicable since they are trained by the discriminative frame-level video features extracted from the deep CNN. Different from previous methods [12, 13] which explore the motion information by using optical flow images, we use the RNNs to learn temporal relationship between frames from the original video with more effective video representations. Moreover, the event videos contain not only motion concepts but also object concepts, so the spatial net is needed to catch the static object information. The bag of semantics (BoS) [10] is used in our spatial net to encode object concepts. While it is difficult to find a robust and accurate semantic description, the deep CNN classifiers trained by the large-scale ImageNet dataset precisely settle this problem. The $1,000$ object categories of ImageNet are supposed as the semantic concepts, and each frame of the event video can be described by the semantic descriptors which can be represented as a multinomial distribution over the semantic concepts. Considering that not all concepts in a video are discriminative for detecting the event, we further use the vector of locally aggregated descriptors (VLAD) [14] to encode the frame-level semantic descriptors to eliminate the redundant information. Finally, Generalized Multiple Kernel Learning (GMKL) [11] is applied to fuse the spatial and temporal features for event detection.

The contributions of this work are as follows. The spatial-

Research supported by the National Science Foundation of China under Grant Nos. 61472038 and 61375044.

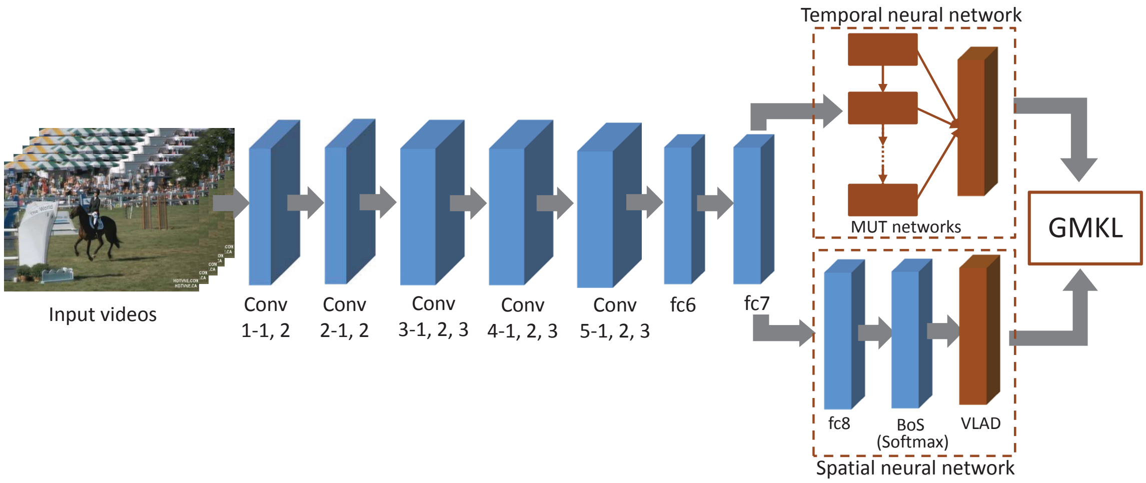

**Fig. 1**. An illustration of the proposed method. The network in blue is the entire VGG CNN [7] with 16 layers pre-trained on ImageNet [8]. "Conv" indicates convolutional layers and "fc" means the fully connected layer. Each "Conv" followed by a max-pooling operation contains 2 or 3 convolutional layers. From "fc7" the CNN splits into two branches. The upper branch is RNNs with MUTation of gated recurrent units (MUT) [9] mining motion information. The lower branch continues the original CNN to the end, and the Bag of Semantic (BoS) [10] encoding is applied to catch the object information. Descriptors extracted from the two branches are fused using the Generalized Multiple Kernel Learning (GMKL) [11] for multimedia event detection.

temporal networks are proposed to learn better CNN descriptions for multimedia event detection. RNNs with MUTation gated recurrent unit are first used to explore the motion information of event videos. BoS is first applied to the spatial net to get deeper and more discriminative representation which is complementary with representation of temporal net for event detection.

## 2. RELATED WORK

From the view of CNNs utilizing spatial and temporal information of videos, the most related works to our method are the two-steam convolutional networks for action recognition [12, 13]. Simonyan *et al.* [12] proposed the two-stream convolutional networks for action recognition. Their spatial network is a pre-trained CNN, and they trained a temporal network with optical flow sequential images as input. Wang *et al.* [13] proposed trajectory-pooled deep convolutional descriptor for action recognition. They integrated the two-stream CNNs proposed in [12] and a successful hand-craft feature for video classification called improved trajectories [15]. Both the applications of the two-stream CNNs achieved satisfactory results under auxiliary hand-crafted descriptors. However, their deep temporal net needs to be trained on a substantial number of optical flow image sequences. Due to the limited number of training data in multimedia event dataset, it is impractical to learn the parameters of the temporal net with such a small number of data for event detection. Different from the two-stream networks, we add two branches after the deep CNN to extract spatial and temporal features. Accordingly, our deep spatial-temporal networks do not rely on any hand-crafted features, and do not need large-scale training procedure either.

From the view of temporal neural network, our temporal net is most related to Recurrent Neural Networks (RNNs) with Long Short-Term Memorys (LSTM) [16] and Gated Recurrent Unit (GRU) [17]. Classical RNNs have the limitation of learning long-term representations of videos due to the vanishing and exploding gradient problems. Many specific RNNs are proposed to handle these problems. The Long Short-Term Memory (LSTM) neural networks with special gating schemes demonstrate the effectiveness in processing long-time sequential information [18, 19]. However, the purposes of many components of LSTM are obscuring, and the dimension of the hidden state in an LSTM is twice the number of memory cells due to its intrinsic structure with separate hidden states. The recently-introduced GRU which is an alternative to the LSTM has better performance and simpler structure without divided hidden states than LSTM. We use the RNNs with the MUTation of gated recurrent unit (MUT) [9] to process the sequential information as the temporal network in our networks. The MUT1 [9] applied in our temporal net is a kind of MUT, where the RNNs with MUT1 achieved the top one performances in many tasks among different RNNs including LSTM and GRU.

## 3. DEEP SPATIAL-TEMPORAL NEURAL NETWORKS

In this section, we describe the proposed deep spatial-temporal networks which explore the information of the motion and object concepts in event videos. We first introduce the pre-trained CNN (the blue part in Fig. 1). Secondly, the details of MUT based temporal net, and the BoS encoded spatial net are elaborated. Finally, the Generalized Multiple Kernel Learning (GMKL) [11] is introduced for event detection.

### 3.1. The pre-trained CNN

Our deep spatial-temporal networks start with an pre-trained deep CNN with 16 weight layers of VGG [7] for ILSVRC 2014 classification task. The dataset of this task includes 1.2M training images categorized into 1000 classes. As depicted in Fig. 1, the CNN contains 5 blocks of convolutional layers, and each block is followed by a max-pooling operation. There are 2 convolutional layers in each of the first 2 blocks, and the other 3 blocks have respective 3 layers inside and hence 13 convolutional layers altogether. In this paper, we use a kind of abbreviation to denote each convolutional layer (*i.e.*, Conv 2-1 refers to the first convolutional layer of the second block), whose notation is inconsistent with [7]. The fc means the fully connected layer, where fc6 and fc7 layer have 4096 channels each, and the fc8 which performs 1000 categories contains 1000 channels. We extract the frame-level CNN descriptors from the fc7 layer for temporal net and the softmax layer for spatial net, respectively. These two parts of works can be simultaneously implemented by using the Caffe toolkit [20] with the model shared by [7].

### 3.2. MUT based temporal neural network

The MUT-based neural networks are known as RNNs with MUTation of gated recurrent unit which are able to model longer sequential time-series data. Accordingly, we use the one of the MUT-based networks, MUT1, to learn the complex dynamics of the event video automatically. Different from the traditional RNNs, the MUT-based networks have gated units to replace the hidden nodes of the RNNs. Unlike the LSTM, the gated unit of an MUT is not separate. The MUT1 is defined by the following equations:

$$
\begin{aligned}
z_t &= \mathrm{sigm}(W_{xz}x_t + b_z) \\
r_t &= \mathrm{sigm}(W_{xr}x_t + W_{hr}h_t + b_r) \\
\tilde{h}_t &= \tanh(W_{hh}(r_t \odot h_t) + \tanh(x_t) + b_h) \\
h_{t+1} &= \tilde{h}_t \odot z_t + h_t \odot (1 - z_t),
\end{aligned}
$$

where $x_t, z_t, r_t, h_t, \tilde{h}_t$ respectively denote the input data, update gate, reset gate, activation and the candidate activation of time step $t$. The $W_*$ is the weight matrix and the $b_*$ is the bias. The $\odot$ is element-wise vector product.

In this work, we process the frame-level descriptors extracted from the fc7 layer of CNN by using single-layer MUT1 networks to obtain the temporal net features. To learn the model parameters of the MUT1-based networks, we use the cross entropy as loss function and minimize the loss function by Back Propagation Through Time (BPTT). BPTT is to expand an MUT1-based net over steps and uses the back propagation algorithm to train the nets. For the sake of none decaying or exploring error when back propagating through internal states of memory cells, the truncated BPTT [16] is used. With the truncated BPTT, errors arriving at a memory cell would not be propagated back anymore. Because of the variable number of the input video frames, we introduce masks to the input data. The temporal-net is implemented by Theano toolkit [21].

### 3.3. BoS encoded spatial neural network

We use the Bag of Semantics (BoS) to encode the frame-level video features extracted from the fc8 layer. Since the BoS is to map each frame into the classifier with multinomial distribution output, the softmax classifier of the pre-trained CNN can be applied directly. The theoretical effectiveness of the BoS encoding is as follow:

Assuming that a video $V = [\pi_1, \ldots, \pi_t]^{\mathrm{T}} \in \mathcal{R}^{t \times c}$ can be described as the multinomial distributions of its frames, where $t$ is the number of frames, $c$ is the number of concepts, $\pi_i = [\pi_{i1}, \ldots, \pi_{ic}]$ denotes the posterior probability distribution of the $i^{th}$ frame, and $\pi_{ij}$ is the probability of the $i^{th}$ frame belonging to the $j^{th}$ semantic concept. The expected log-likelihood of video $V$ can be expressed as

$$
\begin{aligned}
E(\mathcal{L}) &= E(\log \prod_{i=1}^{t} \prod_{j=1}^{c} \pi_{ij}^{I_{ij}}) = \sum_{i=1}^{t} \sum_{j=1}^{c} E(I_{ij}) \log \pi_{ij} \\
&= \sum_{i=1}^{t} \sum_{j=1}^{c} \pi_{ij} \log \pi_{ij}, \qquad (1)
\end{aligned}
$$

where $I_{ij}$ is an indicator function to identify whether the $i^{th}$ frame belongs to the $j^{th}$ class, and $E(I_{ij})$ represents the expectation of $I_{ij}$. As shown in Eq.(1), $E(\mathcal{L})$ only depends on $\pi$, therefore the video is available to be described by BoS. After getting the BoS features of a event video from the softmax classifier, we map the features into the natural parameter space BoS, $[s_1, \ldots, s_t]^{\mathrm{T}}$, using the logarithmic function $s_i = \log \pi_i + \sum_j e^{s_j}$ to describe the video since the posterior probability produced by the softmax classifier is non-Euclidean nature.

Videos encoded by BoS remain the frame-level features, therefore we leverage the VLAD-$k$ [22] which is a soft assignment version of the traditional VLAD to encode these features to video-level features. Firstly, we use PCA to reduce the dimension of BoS features to $S = [s'_1, \ldots, s'_t]^{\mathrm{T}} \in \mathcal{R}^{t\times d}$. Secondly, the K-means is applied to generate $K$ cluster centers $B = [b_1, \ldots, b_K]^{\mathrm{T}} \in \mathcal{R}^{K\times d}$. Then the VLAD-$k$ of the $i^{th}$ frame in a video is calculated by

$$d_i = [\omega_1(s'_i - b_1); \ldots; \omega_K(s'_i - b_K)], \tag{2}$$

where

$$\omega_n = \frac{I(s'_i, b_n)\exp(-\|s'_i - b_n\|_2^2)}{\sum_{m=1}^{K} I(s'_i, b_m)\exp(-\|s'_i - b_m\|_2^2)} \tag{3}$$

is the localized soft assignment weight with $I(s'_i, b_n)$ indicating that if the $b_n$ belongs to the $k$ nearest neighbors of $s'_i$, $I(s'_i, b_n) = 1$, otherwise $I(s'_i, b_n) = 0$. Finally we utilize the intra-normalization [23] and max pooling for VLAD to get the descriptors of the spatial net.

### 3.4. GMKL for event detection

Given training spatial net features $\{d_i | i = 1, 2, ..., N\}$ and temporal net features $\{t_i | i = 1, 2, ..., N\}$, the generalized multiple kernel learning is used to fuse these two kinds of features and train multiple binary classifiers for multimedia event detection. The decision function is defined as

$$f(s, t) = c_1 w_1^T \phi_1(d) + c_2 w_2^T \phi_2(t) + b, \tag{4}$$

where $c_1$ and $c_2$ are the combination coefficients of the two kinds of features with the constraints that $c_1 + c_2 = 1$ and $c_1, c_2 \geq 0$, $w$ and $b$ are parameters of the standard SVM, and $\phi(\cdot)$ is the a function mapping spatial or temporal net features to high dimensional space. $c$, $w$ and $b$ are learned by solving

$$\begin{aligned} \min_{w,b,c} \quad & \frac{1}{2}\sum_{t=1}^{2}\Big(c_t\|w_t\|^2 + c_t^2\Big) + C\sum_{i}^{N} l(y_i, f(d_i, t_i)) \\ \text{s.t.} \quad & c_1 + c_2 = 1, c_1, c_2 \geq 0, \end{aligned} \tag{5}$$

where $l(y_i, f(d_i, t_i)) = \max(0, 1 - y_i f(d_i, t_i))$ is the loss function, and $y_i = \{+1, -1\}$ is the label of the $i$-th training sample. Similar to [11], Eq.(5) can be reformulated by replacing the SVM with its dual form:

$$\begin{aligned} \min_{c} \quad & \frac{1}{2}(c_1^2 + c_2^2) + J(c_1, c_2) \\ \text{s.t.} \quad & c_1 + c_2 = 1, c_1, c_2 \geq 0, \end{aligned} \tag{6}$$

where

$$\begin{aligned} & J(c_1, c_2) = \\ & \max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \big(c_1 k_1(d_i, d_j) + c_2 k_2(t_i, t_j)\big) \\ & \text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, 2, .., N, \end{aligned} \tag{7}$$

$\alpha$ is the dual variable, $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$ are kernel functions for temporal and spatial net features, respectively. Here, the RBF kernel function $k_1(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ and the linear kernel function $k_2(x_i, x_j) = x_i^{\mathrm{T}} x_j$ are used, where $\gamma > 0$ is the kernel parameter. Following [11], Eq.(6) is solved by iteratively updating the linear combination coefficients $c$ and the dual variable $\alpha$.

## 4. EXPERIMENTS

### 4.1. TRECVID MEDTest 14 dataset

We conduct our experiments on the challenging NIST TRECVID MEDTest 14 dataset [1]. The dataset is known as the largest publicly available video corpora for event detection. There are 20 complex events in the dataset, namely "attempting a bike trick", "cleaning an appliance", "dog show", "giving directions to a location", "marriage proposal", "renovating a home", "rock climbing", "town hall meeting", "winning a race without a vehicle", "working on a metal crafts project", "beekeeping", "wedding shower", "non-motorized vehicle repair", "fixing musical instrument", "horse riding competition", "felling a tree", "parking a vehicle", "playing fetch", "tailgating" and "tuning musical instrument". These categories are identified as E21-E40. The total number of training videos is 8,030. Each event class has about 100 positive videos, and the remaining 4,983 videos are negative exemplars which do not belong to any event mentioned above. The testing dataset contains about 23,000 videos.

### 4.2. Experimental settings

In our experiments, we extract CNN features every 5 frames and pre-process each frame by following [4]. For temporal net, the entire MUT1 model which contains an MUT1 layer, a fully connected layer and a softmax layer is trained on the MED14 positive training videos and about 150 negative training videos selected randomly. The maximum timestep is set to 20, and we spilt long-time videos to several videos to expand the number of training data to about 8,000. We conduct the training procedure of MUT1-based networks on a computer without using GPU, and the convergence only takes a few hours. The number of MUT1 units is set to 200 which is smaller than the number of input units. The reason is that the temporal features can be embedded in low dimensional manifold of the input space. Once the MUT1-based nets are trained, for a new input video, we get its feature from the MUT1 layer. For spatial nets, we firstly use PCA with whitening to reduce the 1000-D BoS features to 256-D. Then we apply k-means to obtain 256 centers for VLAD-$k$, where the learning samples are all from training set and $k$ is set to be 5. When using the GMKL for event detection, we utilize 5-fold cross-validation to choose the parameter of the cost $C$

[1] http://www.nist.gov/itl/iad/mig/med14.cfm

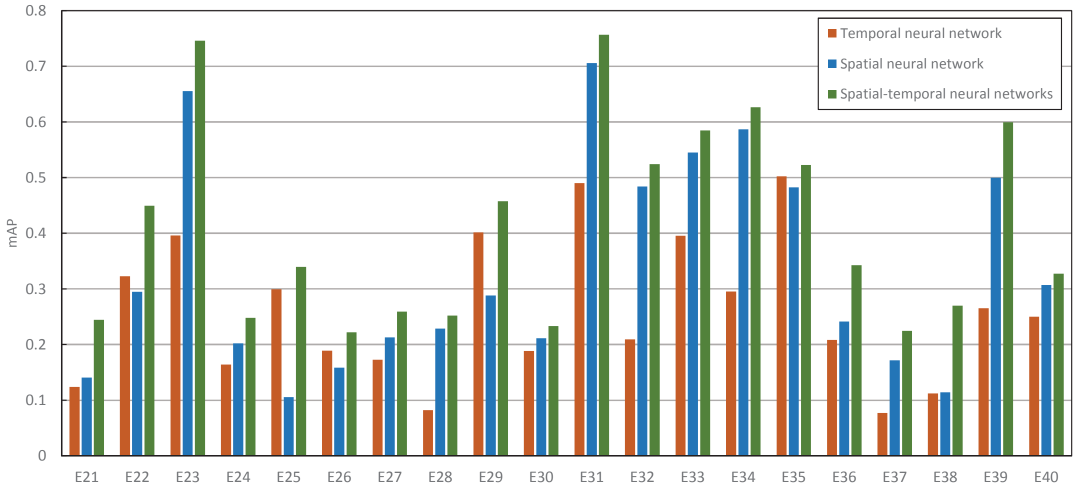

**Fig. 2**. Per event performance comparisons among our spatial neural network, temporal neural network and spatial-temporal neural networks on MEDTest 14 dataset.

in SVM, and the kernel parameter $\gamma$ is set to be the average of the RBF kernel to save the cost training time. When validating the effectiveness of the temporal net, we use 5-fold cross-validation to choose $C$ and $\gamma$.

There are 2 standard training conditions for TRECVID MED: 100Ex and 10Ex. In 100Ex, all the positive exemplars are used for training, and in 10Ex, each event class has only 10 positive exemplars for training. Training MUT1-based networks needs plenty of exemplars, hence, we only concentrate on the 100Ex condition. We use the mean Average Precision (mAP) for binary classification to evaluate the performance for multimedia event detection.

### 4.3. Experimental results

We compare our method with several state-of-the-art methods as follows: Improved Dense Trajectory [15] encoded by Fisher Vector with $\chi^2$ kernel SVM (IDT-FV) [17], average pooling of key-frame CNN features with $\chi^2$ kernel SVM (CNN-avg) [17], frame-level CNN features encoded by VLAD with linear SVM (CNN-VLAD) [4], CNN Latent Concept Descriptors (LCD) of Spatial Pyramid Pooling encoded by VLAD with linear SVM (LCD-SPP-VLAD) [4], Average Late Fusion of CNN-VLAD and LCD-SPP-VLAD (Avg-LF) [4], Spatial-Temporal pooling of frame-level CNN features (CNN-STpooling) [24], and combining the Spatial-Temporal pooling of frame-level CNN features with IDT-FV (CNN-STpooling+IDT-FV) [24]. All the deep learning based methods mentioned above use adaption of the ImageNet pre-trained CNNs to the task of event detection.

Table 1 shows that our method with spatial-temporal nets capturing the motion and object information outperforms methods using only temporal or spatial features, which demonstrates that our spatial-temporal features can preciously represent the event videos. The result of the proposed spatial-temporal net is higher than CNN-STpooling+IDT-FV method which also captures the temporal and spatial information, although the separate spatial and temporal nets are respectively similar to CNN-pooling and IDT-FV in mAPs. The probable reason is as follow: features of MUT are extracted on the CNN features, so the temporal information described by MUT features and spatial information characterized by CNN features are easy to be integrated, while CNN features and IDT features are in 2 totally different feature spaces and it is hard to be fused well and properly. The mAP improvements of feature fusions are 0.011 for Avg-LF, 0.055 for CNN-pooling + IDT-FV , and 0.07 for our spatial-temporal nets, which shows our fusion is efficient.

Consequently, we evaluate the effectiveness of fusing the motion and object information. To this end, we conduct experiments of event detection using the temporal net and spatial net separately. The classifier for detection of the two nets are SVMs with the same kind of kernels as the kernels used in multi kernel learning, namely RBF kernel and linear kernel. Fig. 2 is the per-event mAP comparison among our spatial net, temporal net and spatial-temporal nets. As can be seen in Fig. 2, the spatial-temporal nets outperform the two single nets on all event, which suggests complementarity of the temporal net and spatial net on feature expression.

## 5. CONCLUSION

In this paper, we propose an effective model called deep spatial-temporal networks for multimedia event detection. The temporal net implemented by Recurrent Neural Networks with the MUTation of gated recurrent unit captures motion information, and the spatial net catches the object information by encoding the CNN features with bag of semantics. These t-

**Table 1**. Performance comparisons on MEDTest 14 dataset.

| Method | mAP |
|---|---|
| IDT-FV [15] | 0.270 |
| CNN-avg [17] | 0.329 |
| CNN-VLAD [4] | 0.332 |
| LCD-SPP-VLAD [4] | 0.357 |
| Avg-LF [4] | 0.368 |
| CNN-STpooling [24] | 0.332 |
| CNN-STpooling + IDT-FV [24] | 0.387 |
| Temporal neural network | 0.257 |
| Spatial neural network | 0.332 |
| **Spatial-temporal neural networks** | **0.412** |

wo kinds of information are both demonstrated to be effective for event detection, and are fused by multiple kernel learning to represent event videos. Experiments on TRECVID MEDTest 14 dataset shows good performances of the proposed method.

## 6. REFERENCES

[1] Zhongwen Xu, Yi Yang, Ivor Tsang, Nicu Sebe, and Alexander G Hauptmann, "Feature weighting via optimal thresholding for video analysis," in *ICCV*. IEEE, 2013, pp. 3440–3447.

[2] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *CVPR*. IEEE, 2012, pp. 1298–1305.

[3] Amir Tamrakar, Saad Ali, Qian Yu, Jingen Liu, Omar Javed, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *CVPR*. IEEE, 2012, pp. 3681–3688.

[4] Zhongwen Xu, Yi Yang, and Alexander G Hauptmann, "A discriminative cnn video representation for event detection," in *CVPR*, 2015, pp. 1798–1807.

[5] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alexander G Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *CVPR*, 2015, pp. 2568–2577.

[6] Vignesh Ramanathan, Kevin Tang, Greg Mori, and Li Fei-Fei, "Learning temporal embeddings for complex video analysis," *arXiv preprint arXiv:1505.00315*, 2015.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[9] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever, "An empirical exploration of recurrent network architectures," in *ICML*, 2015, pp. 2342–2350.

[10] Yu Su and Frédéric Jurie, "Improving image classification using semantic attributes," *IJCV*, vol. 100, no. 1, pp. 59–77, 2012.

[11] Manik Varma and Bodla Rakesh Babu, "More generality in efficient multiple kernel learning," in *ICML*. ACM, 2009, pp. 1065–1072.

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

[13] Limin Wang, Yu Qiao, and Xiaoou Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015, pp. 4305–4314.

[14] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*. IEEE, 2010, pp. 3304–3311.

[15] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*. IEEE, 2013, pp. 3551–3558.

[16] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

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

[19] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[21] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[22] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *arXiv preprint arXiv:1405.4506*, 2014.

[23] Relja Arandjelovic and Andrew Zisserman, "All about vlad," in *CVPR*. IEEE, 2013, pp. 1578–1585.

[24] Shengxin Zha, Florian Luisier, Walter Andrews, Nitish Srivastava, and Ruslan Salakhutdinov, "Exploiting image-trained cnn architectures for unconstrained video classification," *arXiv preprint arXiv:1503.04144*, 2015.