3D Convolutional Neural Networks for Human Action Recognition

Shuiwang Ji, Wei Xu, Ming Yang, *Member*, *IEEE*, and Kai Yu, *Member*, *IEEE*

Abstract—We consider the automated recognition of human actions in surveillance videos. Most current methods build classifiers based on complex handcrafted features computed from the raw inputs. Convolutional neural networks (CNNs) are a type of deep model that can act directly on the raw inputs. However, such models are currently limited to handling 2D inputs. In this paper, we develop a novel 3D CNN model for action recognition. This model extracts features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. The developed model generates multiple channels of information from the input frames, and the final feature representation combines information from all channels. To further boost the performance, we propose regularizing the outputs with high-level features and combining the predictions of a variety of different models. We apply the developed models to recognize human actions in the real-world environment of airport surveillance videos, and they achieve superior performance in comparison to baseline methods.

Index Terms—Deep learning, convolutional neural networks, 3D convolution, model combination, action recognition

1 Introduction

Recognizing human actions in the real-world environment finds applications in a variety of domains including intelligent video surveillance, customer attributes, and shopping behavior analysis. However, accurate recognition of actions is a highly challenging task due to cluttered backgrounds, occlusions, and viewpoint variations, etc. [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Most of the current approaches [12], [13], [14], [15], [16] make certain assumptions (e.g., small scale and viewpoint changes) about the circumstances under which the video was taken. However, such assumptions seldom hold in the real-world environment. In addition, most of the methods follow a two-step approach in which the first step computes features from raw video frames and the second step learns classifiers based on the obtained features. In real-world scenarios, it is rarely known what features are important for the task at hand since the choice of features is highly problem-dependent. Especially for human action recognition, different action classes may appear dramatically different in terms of their appearances and motion patterns.

Deep learning models [17], [18], [19], [20], [21] are a class of machines that can learn a hierarchy of features by building high-level features from low-level ones. Such learning machines can be trained using either supervised or unsupervised approaches, and the resulting systems have been shown to yield competitive performance in visual object recognition [17], [19], [22], [23], [24], human action recognition [25], [26], [27], natural language processing [28], audio classification [29], brain-computer interaction [30], human tracking [31], image restoration [32], denoising [33], and segmentation tasks [34]. The convolutional neural networks (CNNs) [17] are a type of deep models in which trainable filters and local neighborhood pooling operations are applied alternatingly on the raw input images, resulting in a hierarchy of increasingly complex features. It has been shown that, when trained with appropriate regularization [35], [36], [37], CNNs can achieve superior performance on visual object recognition tasks. In addition, CNNs have been shown to be invariant to certain variations such as pose, lighting, and surrounding clutter [38].

As a class of deep models for feature construction, CNNs have been primarily applied on 2D images. In this paper, we explore the use of CNNs for human action recognition in videos. A simple approach in this direction is to treat video frames as still images and apply CNNs to recognize actions at the individual frame level. Indeed, this approach has been used to analyze the videos of developing embryos [39]. However, such an approach does not consider the motion information encoded in multiple contiguous frames. To effectively incorporate the motion information in video analysis, we propose to perform 3D convolution in the convolutional layers of CNNs so that discriminative features along both the spatial and the temporal dimensions are captured. We show that, by applying multiple distinct convolutional operations at the same location on the input, multiple types of features can be extracted. Based on the proposed 3D convolution, a variety of 3D CNN architectures can be devised to analyze video data. We develop a 3D CNN architecture that generates multiple channels of information from adjacent video frames and performs convolution and subsampling separately in each channel. The final feature

- *S. Ji is with the Department of Computer Science, Old Dominion University, Suite 3300, 4700 Elkhorn Avenue, Norfolk, VA 23529-0162. E-mail: sji@cs.odu.edu.*
- *W. Xu is with Facebook, Inc., 1601 Willow Road, Menlo Park, CA 94304. E-mail: emailweixu@fb.com.*
- *M. Yang is with NEC Laboratories America, Inc., 10080 North Wolfe Road, SW3-350, Cupertino, CA 95014. E-mail: myang@nec-labs.com.*
- *K. Yu is with Baidu Inc., Baidu Building, Shangdi 10th Street, Haidian District, Beijing 100085, China. E-mail: yukai@baidu.com.*

Manuscript received 13 Apr. 2011; revised 28 Oct. 2011; accepted 17 Feb. 2012; published online 28 Feb. 2012.
Recommended for acceptance by C. Bregler.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-04-0227.
Digital Object Identifier no. 10.1109/TPAMI.2012.59.

0162-8828/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

representation is obtained by combining information from all channels. To further boost the performance of 3D CNN models, we propose to augment the models with auxiliary outputs computed as high-level motion features and integrate the outputs of a variety of different architectures in making predictions.

We evaluated the developed 3D CNN model on the TREC Video Retrieval Evaluation (TRECVID) data, which consist of surveillance video data recorded at London Gatwick Airport. We constructed a multimodule event detection system, which includes the 3D CNN as a major module, and participated in three tasks of the TRECVID 2009 Evaluation for Surveillance Event Detection [25]. Our system achieved the best performance on all three participating action categories (i.e., CellToEar, ObjectPut, and Pointing). To provide an independent evaluation of the 3D CNN model, we report its performance on the TRECVID 2008 development set in this paper. We also present results on the KTH data as published performance for this data is available. Our experiments show that the developed 3D CNN model outperforms other baseline methods on the TRECVID data, and it achieves competitive performance on the KTH data, demonstrating that the 3D CNN model is more effective for real-world environments such as those captured in the TRECVID data. The experiments also validate that the 3D CNN model significantly outperforms the frame-based 2D CNN for most tasks.

The key contributions of this work can be summarized as follows:

- We propose to apply the 3D convolution operation to extract spatial and temporal features from video data for action recognition. These 3D feature extractors operate in both the spatial and the temporal dimensions, thus capturing motion information in video streams.
- We develop a 3D convolutional neural network architecture based on the 3D convolution feature extractors. This CNN architecture generates multiple channels of information from adjacent video frames and performs convolution and subsampling separately in each channel. The final feature representation is obtained by combining information from all channels.
- We propose to regularize the 3D CNN models by augmenting the models with auxiliary outputs computed as high-level motion features. We further propose to boost the performance of 3D CNN models by combining the outputs of a variety of different architectures.
- We evaluate the 3D CNN models on the TRECVID 2008 development set in comparison with baseline methods and alternative architectures. Experimental results show that the proposed models significantly outperforms 2D CNN architecture and other baseline methods.

The rest of this paper is organized as follows: We describe the 3D convolution operation and the 3D CNN architecture employed in our TRECVID action recognition system in Section 2. Some related work for action recognition is discussed in Section 3. The experimental results on the TRECVID and KTH data are reported in Section 4. We conclude in Section 5 with discussions.

2 3D Convolutional Neural Networks

In 2D CNNs, 2D convolution is performed at the convolutional layers to extract features from local neighborhood on feature maps in the previous layer. Then an additive bias is applied and the result is passed through a sigmoid function. Formally, the value of an unit at position (x, y) in the jth feature map in the ith layer, denoted as v_{ij}^{xy}, is given by

$$v_{ij}^{xy} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right), \quad (1)$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, b_{ij} is the bias for this feature map, m indexes over the set of feature maps in the $(i-1)$th layer connected to the current feature map, w_{ijk}^{pq} is the value at the position (p, q) of the kernel connected to the kth feature map, and P_i and Q_i are the height and width of the kernel, respectively. In the subsampling layers, the resolution of the feature maps is reduced by pooling over local neighborhood on the feature maps in the previous layer, thereby enhancing the invariance to distortions on the inputs. A CNN architecture can be constructed by stacking multiple layers of convolution and subsampling in an alternating fashion. The parameters of CNN, such as the bias b_{ij} and the kernel weight w_{ijk}^{pq}, are usually learned using either supervised or unsupervised approaches [17], [22].

2.1 3D Convolution

In 2D CNNs, convolutions are applied on the 2D feature maps to compute features from the spatial dimensions only. When applied to video analysis problems, it is desirable to capture the motion information encoded in multiple contiguous frames. To this end, we propose to perform 3D convolutions in the convolution stages of CNNs to compute features from both spatial and temporal dimensions. The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together. By this construction, the feature maps in the convolution layer are connected to multiple contiguous frames in the previous layer, thereby capturing motion information. Formally, the value at position (x, y, z) on the jth feature map in the ith layer is given by

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right), \quad (2)$$

where R_i is the size of the 3D kernel along the temporal dimension, w_{ijm}^{pqr} is the (p, q, r)th value of the kernel connected to the mth feature map in the previous layer. A comparison of 2D and 3D convolutions is given in Fig. 1.

Note that a 3D convolutional kernel can only extract one type of features from the frame cube since the kernel weights are replicated across the entire cube. A general design principle of CNNs is that the number of feature maps should be increased in late layers by generating multiple types of features from the same set of lower level feature maps. Similarly to the case of 2D convolution, this can be achieved

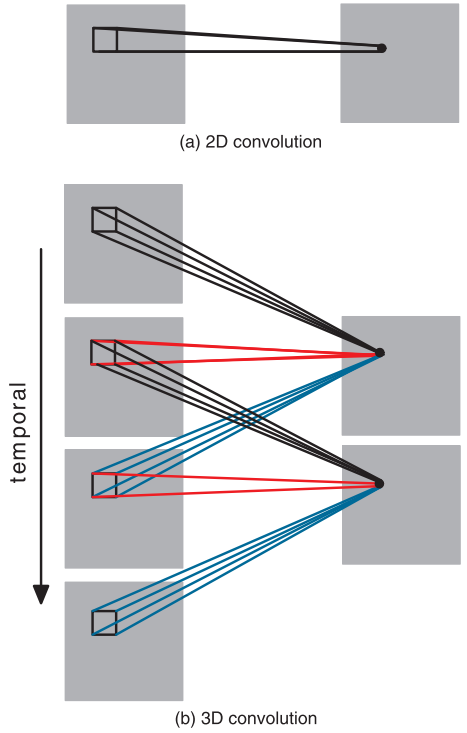


Fig. 1. Comparison of (a) 2D and (b) 3D convolutions. In (b) the size of the convolution kernel in the temporal dimension is 3 and the sets of connections are color-coded so that the shared weights are in the same color. In 3D convolution, the same 3D kernel is applied to overlapping 3D cubes in the input video to extract motion features.

by applying multiple 3D convolutions with distinct kernels to the same location in the previous layer (Fig. 2).

2.2 A 3D CNN Architecture

Based on the 3D convolution described above, a variety of CNN architectures can be devised. In the following, we describe a 3D CNN architecture that we have developed for human action recognition on the TRECVID data set. In this architecture, shown in Fig. 3, we consider seven frames of size 60×40 centered on the current frame as inputs to the 3D CNN model. We first apply a set of hardwired kernels to generate multiple channels of information from the input frames. This results in 33 feature maps in the second layer in

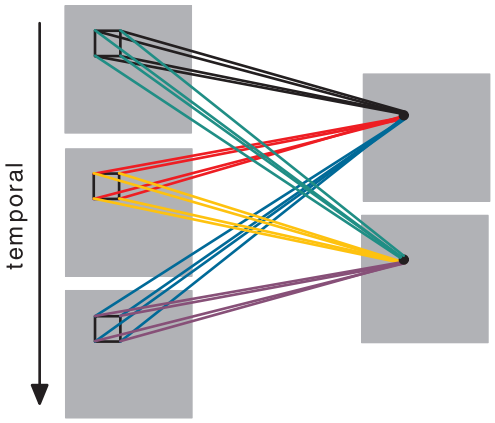


Fig. 2. Extraction of multiple features from contiguous frames. Multiple 3D convolutions can be applied to contiguous frames to extract multiple features. As in Fig. 1, the sets of connections are color-coded so that the shared weights are in the same color. Note that all six sets of connections do not share weights, resulting in two different feature maps on the right.

five different channels denoted by gray, gradient-x, gradient-y, optflow-x, and optflow-y. The gray channel contains the gray pixel values of the seven input frames. The feature maps in the gradient-x and gradient-y channels are obtained by computing gradients along the horizontal and vertical directions, respectively, on each of the seven input frames, and the optflow-x and optflow-y channels contain the optical flow fields along the horizontal and vertical directions, respectively, computed from adjacent input frames. This hardwired layer is employed to encode our prior knowledge on features, and this scheme usually leads to better performance as compared to the random initialization.

We then apply 3D convolutions with a kernel size of $7 \times 7 \times 3$ (7×7 in the spatial dimension and 3 in the temporal dimension) on each of the five channels separately. To increase the number of feature maps, two sets of different convolutions are applied at each location, resulting in two sets of feature maps in the C2 layer each consisting of 23 feature maps. In the subsequent subsampling layer S3, we apply 2×2 subsampling on each of the feature maps in the C2 layer, which leads to the same number of feature maps with a reduced spatial resolution.

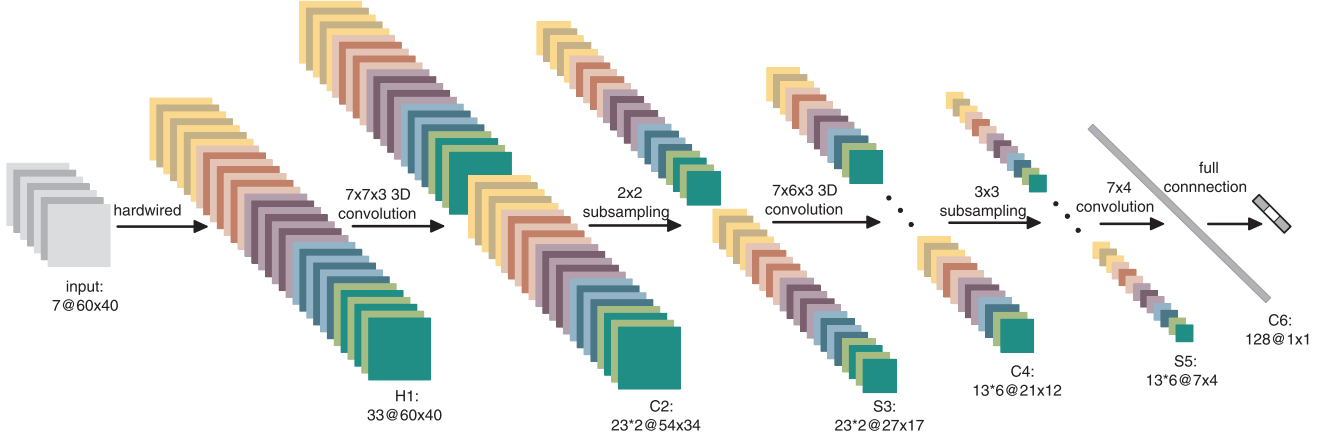


Fig. 3. A 3D CNN architecture for human action recognition. This architecture consists of one hardwired layer, three convolution layers, two subsampling layers, and one full connection layer. Detailed descriptions are given in the text.

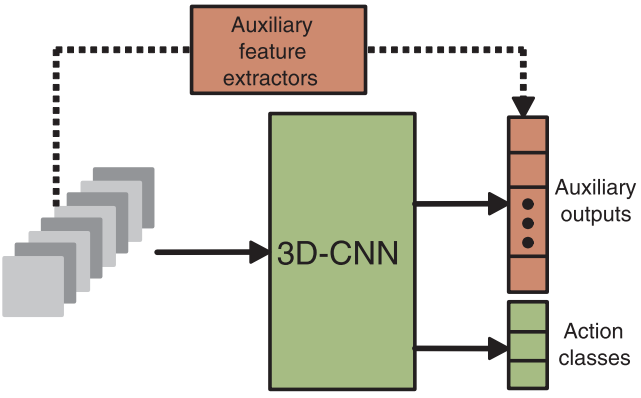


Fig. 4. The regularized 3D CNN architecture.

The next convolution layer C4 is obtained by applying 3D convolution with a kernel size of $7 \times 6 \times 3$ on each of the five channels in the two sets of feature maps separately. To increase the number of feature maps, we apply three convolutions with different kernels at each location, leading to six distinct sets of feature maps in the C4 layer, each containing 13 feature maps. The next layer S5 is obtained by applying 3×3 subsampling on each feature map in the C4 layer, which leads to the same number of feature maps with a reduced spatial resolution. At this stage, the size of the temporal dimension is already relatively small (3 for gray, gradient-x, gradient-y, and 2 for optflow-x and optflow-y), so we perform convolution only in the spatial dimension at this layer. The size of the convolution kernel used is 7×4 so that the sizes of the output feature maps are reduced to 1×1. The C6 layer consists of 128 feature maps of size 1×1, and each of them is connected to all 78 feature maps in the S5 layer.

After the multiple layers of convolution and subsampling, the seven input frames have been converted into a 128D feature vector capturing the motion information in the input frames. The output layer consists of the same number of units as the number of actions, and each unit is fully connected to each of the 128 units in the C6 layer. In this design, we essentially apply a linear classifier on the 128D feature vector for action classification. All the trainable parameters in this model are initialized randomly and trained by the online error back-propagation algorithm as described in [17]. We have designed and evaluated other 3D CNN architectures that combine multiple channels of information at different stages, and our results show that this architecture gives the best performance.

2.3 Model Regularization

The inputs to 3D CNN models are limited to a small number of contiguous video frames due to the increased number of trainable parameters as the size of input window increases. On the other hand, many human actions span a number of frames. Hence, it is desirable to encode high-level motion information into the 3D CNN models. To this end, we propose computing motion features from a large number of frames and regularizing the 3D CNN models by using these motion features as auxiliary outputs (Fig. 4). Similar ideas have been used in image classification tasks [35], [36], [37], but its performance in action recognition is not clear. In particular, for each training action we generate a feature vector encoding the long-term action information beyond the information contained in the input frame cube to the CNN. We then encourage the CNN to learn a feature vector close to this feature. This is achieved by connecting a number of auxiliary output units to the last hidden layer of CNN and clamping the computed feature vectors on the auxiliary units during training. This will encourage the hidden layer information to be close to the high-level motion feature. More details on this scheme can be found in [35], [36], and [37]. In the experiments, we use the bag-of-words features constructed from dense SIFT descriptors [40] computed on raw gray images and motion edge history images (MEHI) [41] as auxiliary features. Results show that such a regularization scheme leads to consistent performance improvements.

2.4 Model Combination

Based on the 3D convolution operations, a variety of 3D CNN architectures can be designed. Among the architectures considered in this paper, the one introduced in Section 2.2 yields the best performance on the TRECVID data. However, this may not be the case for other data sets. The selection of optimal architecture for a problem is challenging since this depends on the specific applications. An alternative approach is to construct multiple models and combine the outputs of these models for making predictions [42], [43], [44]. This scheme has also been used in combining traditional neural networks [45]. However, the effect of model combination in the context of convolutional neural networks for action recognition has not been investigated. In this paper, we propose constructing multiple 3D CNN models with different architectures, hence capturing potentially complementary information from the inputs. In the prediction phase, the input is given to each model and the outputs of these models are then combined. Experimental results demonstrate that this model combination scheme is very effective in boosting the performance of 3D CNN models on action recognition tasks.

2.5 Model Implementation

The 3D CNN models are implemented in C++ as part of NEC's human action recognition system [25]. The implementation details are based on those of the original CNN as described in [17] and [46]. All the subsampling layers apply max sampling as described in [47]. The overall loss function used to train the regularized models is a weighted summation of the loss functions induced by the true action classes and the auxiliary outputs. The weight for the true action classes is set to 1 and that for the auxiliary outputs is set to 0.005 empirically. All the model parameters are randomly initialized as in [17] and [46] and are trained using the stochastic diagonal Levenberg-Marquardt method [17], [46]. In this method, a learning rate is computed for each parameter using the diagonal terms of an estimate of the Gauss-Newton approximation to the Hessian matrix on 1,000 randomly sampled training instances.

3 Related Work

CNNs belong to the class of biologically inspired models for visual recognition, and some other variants have also been developed within this family. Motivated by the organization of visual cortex, a similar model, called HMAX [48], has been developed for visual object recognition. In the HMAX

TABLE 1
The Number of Samples on the Five Dates Extracted from the TRECVID 2008 Development Data Set

DATE\CLASS	CELLTOEAR	OBJECTPUT	POINTING	NEGATIVE	TOTAL
20071101	2692	1349	7845	20056	31942
20071106	1820	3075	8533	22095	35523
20071107	465	3621	8708	19604	32398
20071108	4162	3582	11561	35898	55203
20071112	4859	5728	18480	51428	80495
TOTAL	13998	17355	55127	149081	235561

model, a hierarchy of increasingly complex features is constructed by alternating applications of template matching and max pooling. In particular, at the S1 layer a still input image is first analyzed by an array of Gabor filters at multiple orientations and scales. The C1 layer is then obtained by pooling local neighborhoods on the S1 maps, leading to increased invariance to distortions on the input. The S2 maps are obtained by comparing C1 maps with an array of templates which were generated randomly from C1 maps in the training phase. The final feature representation in C2 is obtained by performing global max pooling over each of the S2 maps.

The original HMAX model is designed to analyze 2D images. In [16], this model has been extended to recognizing actions in video data. In particular, the Gabor filters in the S1 layer of the HMAX model have been replaced with some gradient and space-time modules to capture motion information. In addition, some modifications to HMAX, proposed in [49], have been incorporated into the model. A major difference between CNN and HMAX-based models is that CNNs are fully trainable systems in which all the parameters are adjusted based on training data, while all modules in HMAX consist of hard-coded parameters.

In speech and handwriting recognition, time-delay neural networks have been developed to extract temporal features [50]. In [51], a modified CNN architecture has been developed to extract features from video data. In addition to recognition tasks, CNNs have also been used in 3D image restoration problems [32].

4 EXPERIMENTS

We focus on the TRECVID 2008 data to evaluate the developed 3D CNN models for action recognition in surveillance videos. Meanwhile, we also perform experiments on the KTH data [13] to compare with previous methods.

4.1 Action Recognition on TRECVID Data

The TRECVID 2008 development data set consists of 49-hour videos captured at London Gatwick Airport using five different cameras with a resolution of 720×576 at 25 fps. The videos recorded by camera number 4 are excluded as few events occurred in this scene. In the current experiments, we focus on the recognition of three action classes (*CellToEar*, *ObjectPut*, and *Pointing*). Each action is classified in the one-against-rest manner, and a large number of negative samples were generated from actions that are not in these three classes. This data set was captured on five days (20071101, 20071106, 20071107, 20071108, and 20071112), and the statistics of the data used in our experiments are summarized in Table 1. Multiple 3D CNN models are evaluated in this experiment, including the one described in Fig. 3.

As the videos were recorded in real-world environments, and each frame contains multiple humans, we apply a human detector and a detection-driven tracker to locate human heads. The detailed procedure for tracking is described in [52], and some sample results are shown in Fig. 5. Based on the detection and tracking results, a bounding box for each human that performs an action was computed. The procedure to crop the bounding box from the head tracking results is illustrated in Fig. 6. The multiple frames required by the 3D CNN model are obtained by extracting bounding boxes at the same position from consecutive frames before and after the current frame, leading to a cube containing the action. The temporal dimension of the cube is set to 7 in our experiments as it has been shown that 5-7 frames are enough to achieve a performance similar to the one obtainable with the entire video sequence [53]. The frames were extracted with a step size of 2. That is, suppose the current frame is numbered 0; we extract a bounding box at the same position from frames numbered -6, -4, -2, 0, 2, 4, and 6. The patch inside the bounding box on each frame is scaled to 60×40 pixels.

Fig. 5. Sample human detection and tracking results from camera numbers 1, 2, 3, and 5 (left to right).

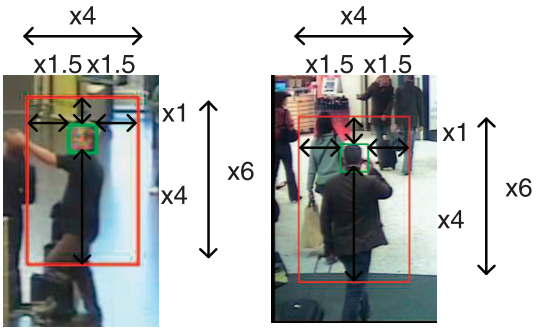


Fig. 6. Illustration of the procedure to crop the bounding box from the head tracking results.

To evaluate the effectiveness of the 3D CNN model, we report the results of the frame-based 2D CNN model. In addition, we compare the 3D CNN model with four other methods which build state-of-the-art spatial pyramid matching (SPM) features from local features computed on dense grid or spatiotemporal interest points (STIPs). For these methods, we construct SPM features based on local invariant features computed from each image cube as used in 3D CNN. Then a one-against-all linear SVM is learned for each action class. For dense features, we extract SIFT descriptors [40] from raw gray images or motion edge history images [41]. Local features on raw gray images preserve the appearance information, while MEHI is concerned with the shape and motion patterns. The dense SIFT descriptors are calculated every 6 pixels from 7×7 and 16×16 local image patches. For features based on STIPs, we employ the temporally integrated spatial response (TISR) method [54], which has shown promising performance on action recognition. The local features are softly quantized (each local feature can be assigned to multiple codebook words) using a 512-word codebook. To exploit the spatial layout information, we employ the spatial pyramid matching method [55] to partition the candidate region into 2×2 and 3×4 cells and concatenate their features. The dimensionality of the entire feature vector is $512 \times (2 \times 2 + 3 \times 4) = 8{,}192$. We denote the method based on gray images as $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{gray}}$, the one based on MEHI as $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{MEHI}}$, and the one based on TISR as $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{TISR}}$. We also concatenate $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{gray}}$ and $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{MEHI}}$ feature vectors into a single vector, leading to the 16,384D feature representation denoted as $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{gray+MEHI}}$.

In the first set of experiments, we report the performance of the 3D CNN architecture described in Fig. 3 as this model achieved the best performance. This architecture is denoted as $\mathrm{3D\text{-}CNN}^{\mathrm{s}}_{332}$ since the five channels are convolved separately (the superscript s) and the first two convolutional layers use 3D convolution and the last convolutional layer use 2D convolution (the subscript 332). We also report the performance of the regularized 3D CNN model based on $\mathrm{3D\text{-}CNN}^{\mathrm{s}}_{332}$. In this model, denoted as $\mathrm{3D\text{-}RCNN}^{\mathrm{s}}_{332}$, the auxiliary outputs are obtained by applying PCA to reduce the dimensionality of 8,192D $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{gray}}$ and $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{MEHI}}$ features to 150 dimensions and then concatenating them into a 300D feature vector.

We report the fivefold cross-validation results in which the data for a single day are used as a fold. The performance measures we used are precision, recall, and area under the ROC curve (ACU) at multiple values of false positive rates (FPR). The performance of the seven methods is summarized in Table 2, and the average performance over all action classes is plotted in Fig. 7. We can observe from these results that the 3D CNN models outperform the frame-based 2D CNN model, $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{gray}}$, and $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{MEHI}}$ significantly on the action classes *CellToEar* and *ObjectPut* in all cases. For the action class *Pointing*, the 3D CNN model achieves slightly worse performance than the other three methods. Concatenation of the $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{gray}}$ and $\mathrm{SPM}^{\mathrm{cube}}_{\mathrm{MEHI}}$ features yields improved performance over individual features, but the performance is still lower than that of the 3D CNN models. We can also observe that our models also outperform the method based on the spatiotemporal feature TISR. Overall, the 3D CNN models outperform other methods consistently, as can be seen from the average performance in Fig. 7. In addition, the regularized model yields higher performance than the one without regularization in all cases. Although the improvement by the regularized model is not significant, the following experiments show that significant performance improvements can be obtained by combining the two models.

To evaluate the effectiveness of model combination in the context of CNN for action recognition, we develop the three alternative 3D CNN architectures described below:

- $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{332}$ denotes the architecture in which the different channels are "mixed," and the first two convolutional layers use 3D convolution, and the last layer use 2D convolution. "Mixed" means that the channels of the same type (i.e., gradient-x and gradient-y, optflow-x, and optflow-y) are convolved separately, but they are connected to the same set of feature planes in the first convolutional layer. In the second convolutional layer, all five channels are connected to the same set of feature planes. In contrast, for models with superscript s, all five channels are connected to separate feature planes in all layers.
- $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{322}$ denotes a model similar to $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{332}$, but only the first convolutional layer uses 3D convolution and the other two layers use 2D convolution.
- $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{222}$ denotes a model similar to $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{332}$, but all three convolutional layers use 2D convolution.

The average performance of these three models, along with that of $\mathrm{3D\text{-}CNN}^{\mathrm{s}}_{332}$, is plotted in Fig. 8. We can observe that the performance of these three alternative architecture is lower than that of $\mathrm{3D\text{-}CNN}^{\mathrm{s}}_{332}$. However, we show in the following that combination of these models can lead to significant performance improvement.

To evaluate the effectiveness of model combination, we tuned each of the five models ($\mathrm{3D\text{-}RCNN}^{\mathrm{s}}_{332}$, $\mathrm{3D\text{-}CNN}^{\mathrm{s}}_{332}$, $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{222}$, $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{322}$, and $\mathrm{3D\text{-}CNN}^{\mathrm{m}}_{332}$) individually and then combined their outputs to make prediction. We combine models incrementally in order of decreasing individual performance. That is, the models are sorted in decreasing order of individual performance and they are combined incrementally from the first to the last. The reason for doing this is that we expect individual models with high performance will lead to more significant improvements when they are combined. We report the combined performance for each combination in Table 3 and

TABLE 2
Performance of the Seven Methods under Multiple False Positive Rates

Method	FPR	Measure	CellToEar	ObjectPut	Pointing	Average
3D-RCNN$^{s}_{332}$	0.1%	Precision	0.5717	**0.7348**	0.8380	**0.7148**
		Recall	0.0211	**0.0348**	0.0169	**0.0243**
		AUC($\times 10^3$)	0.0114	**0.0209**	0.0096	**0.0139**
	1%	Precision	0.3917	**0.5384**	0.7450	**0.5584**
		Recall	0.1019	**0.1466**	0.0956	**0.1147**
		AUC($\times 10^3$)	0.6272	**0.9044**	0.5665	**0.6993**
3D-CNN$^{s}_{332}$	0.1%	Precision	**0.6433**	0.6748	0.8230	0.7137
		Recall	**0.0282**	0.0256	0.0152	0.0230
		AUC($\times 10^3$)	**0.0173**	0.0139	0.0075	0.0129
	1%	Precision	**0.4091**	0.5154	0.7470	0.5572
		Recall	**0.1109**	0.1356	0.0931	0.1132
		AUC($\times 10^3$)	**0.6759**	0.7916	0.5581	0.6752
2D-CNN	0.1%	Precision	0.3842	0.5865	0.8547	0.6085
		Recall	0.0097	0.0176	0.0192	0.0155
		AUC($\times 10^3$)	0.0057	0.0109	0.0110	0.0092
	1%	Precision	0.3032	0.3937	0.7446	0.4805
		Recall	0.0505	0.0974	0.1020	0.0833
		AUC($\times 10^3$)	0.2725	0.5589	0.6218	0.4844
SPM$^{cube}_{gray}$	0.1%	Precision	0.3576	0.6051	0.8541	0.6056
		Recall	0.0088	0.0192	0.0191	0.0157
		AUC($\times 10^3$)	0.0044	0.0108	0.0110	0.0087
	1%	Precision	0.2607	0.4332	0.7511	0.4817
		Recall	0.0558	0.0961	0.0988	0.0836
		AUC($\times 10^3$)	0.3127	0.5523	0.5915	0.4855
SPM$^{cube}_{MEHI}$	0.1%	Precision	0.4848	0.5692	0.8268	0.6269
		Recall	0.0149	0.0166	0.0156	0.0157
		AUC($\times 10^3$)	0.0071	0.0087	0.0084	0.0081
	1%	Precision	0.3552	0.3961	0.7546	0.5020
		Recall	0.0872	0.0825	0.1006	0.0901
		AUC($\times 10^3$)	0.4955	0.4629	0.5712	0.5099
SPM$^{cube}_{gray+MEHI}$	0.1%	Precision	0.5279	0.6422	**0.8657**	0.6786
		Recall	0.0177	0.0226	**0.0211**	0.0205
		AUC($\times 10^3$)	0.0098	0.0129	**0.0122**	0.0116
	1%	Precision	0.3753	0.4668	**0.7671**	0.5364
		Recall	0.0951	0.1102	**0.1078**	0.1043
		AUC($\times 10^3$)	0.5822	0.6431	**0.6478**	0.6244
TISR	0.1%	Precision	0.3391	0.6201	0.8553	0.6048
		Recall	0.0081	0.0205	0.0193	0.0160
		AUC($\times 10^3$)	0.0041	0.0121	0.0112	0.0091
	1%	Precision	0.2831	0.4628	0.7502	0.4987
		Recall	0.0625	0.1083	0.0983	0.0879
		AUC($\times 10^3$)	0.3717	0.6134	0.5897	0.5249

The AUC scores are multiplied by 10^3 for ease of presentation. The highest performance in each case is shown in bold face.

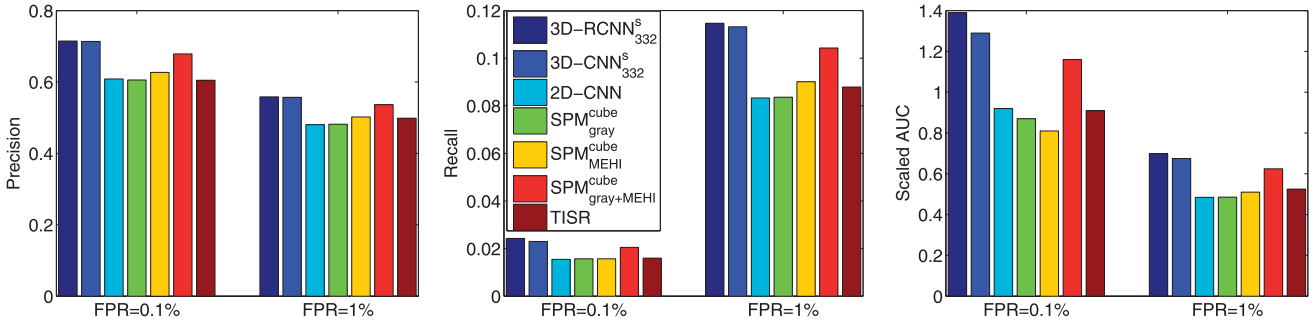


Fig. 7. Average performance comparison of the seven methods under different false positive rates. The AUC scores at $\mathrm{FPR} = 0.1$ and 1 percent are multiplied by 10^5 and 10^3, respectively, for better visualization.

plot the average performance in Fig. 9. We can observe that combination of models leads to significant performance improvements in almost all cases except the combination of 3D-CNN$^{m}_{332}$, which leads to slight performance degradation. This shows that the different architectures encode complementary information, and combination of these models

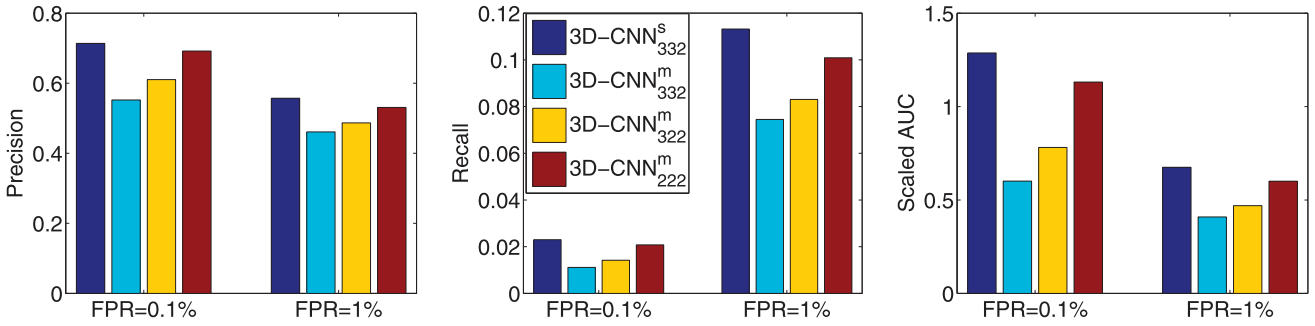


Fig. 8. Average performance comparison of the four different 3D CNN architectures under different false positive rates. The AUC scores at FPR = 0.1 percent and 1 percent are multiplied by 10^5 and 10^3, respectively, for better visualization.

effectively integrates this information though the performance of some of the individual models is low. Figs. 10, 11, and 12 show some sample actions in each of the three classes that are classified correctly and incorrectly by the combined model. It can be observed that most of the misclassified actions are hard to recognize even by human.

To highlight the performance improvements over our previous result in [26], we report the best performance achieved by the methods in [26] and that of the new methods proposed in this paper in Table 4. We can observe that our new methods in this paper improve over the previous results significantly in all cases.

4.2 Action Recognition on the KTH Data

We evaluate the 3D CNN model on the KTH data [13], which consist of six action classes performed by 25 subjects. To follow the setup in the HMAX model, we use a 9-frame cube as input and extract foreground as in [16]. To reduce the memory requirement, the resolutions of the input frames are reduced to 80×60 in our experiments as compared to 160×120 used in [16]. We use a similar 3D CNN architecture as in Fig. 3, with the sizes of kernels and the number of feature maps in each layer modified to consider the $80 \times 60 \times 9$ inputs. In particular, the three convolutional layers use kernels of sizes 9×7, 7×7, and

TABLE 3
Performance of Different Combinations of the 3D CNN Models

FPR	Measure	Method	CellToEar	ObjectPut	Pointing	Average
0.1%	Precision	1	0.5717	0.7348	0.8380	0.7148
		1+2	0.6547	0.7477	0.8710	0.7578
		1+2+3	0.6716	**0.7736**	0.8810	0.7754
		1+2+3+4	**0.7015**	0.7588	**0.8869**	**0.7824**
		1+2+3+4+5	0.6948	0.7477	0.8853	0.7759
	Recall	1	0.0211	0.0348	0.0169	0.0243
		1+2	0.0299	0.0372	0.0220	0.0297
		1+2+3	0.0323	**0.0429**	0.0242	0.0331
		1+2+3+4	**0.0369**	0.0395	**0.0256**	**0.0340**
		1+2+3+4+5	0.0364	0.0372	0.0252	0.0329
	AUC($\times 10^3$)	1	0.0114	0.0209	0.0096	0.0139
		1+2	0.0194	0.0213	0.0109	0.0172
		1+2+3	0.0206	0.0240	0.0117	0.0187
		1+2+3+4	0.0227	**0.0244**	**0.0132**	**0.0201**
		1+2+3+4+5	**0.0230**	0.0237	0.0123	0.0197
1%	Precision	1	0.3917	0.5384	0.7450	0.5584
		1+2	0.4301	0.5573	0.7795	0.5890
		1+2+3	0.4623	**0.5700**	0.7902	0.6075
		1+2+3+4	0.4737	0.5621	0.7942	0.6100
		1+2+3+4+5	**0.4816**	0.5657	**0.7973**	**0.6149**
	Recall	1	0.1019	0.1466	0.0956	0.1147
		1+2	0.1193	0.1583	0.1156	0.1311
		1+2+3	0.1357	**0.1666**	0.1232	0.1418
		1+2+3+4	0.1424	0.1612	**0.1261**	**0.1433**
		1+2+3+4+5	**0.1428**	0.1590	0.1256	0.1425
	AUC($\times 10^3$)	1	0.6272	0.9044	0.5665	0.6993
		1+2	0.7564	0.9758	0.6762	0.8028
		1+2+3	0.8428	**1.0240**	0.7490	0.8720
		1+2+3+4	**0.8990**	0.9708	**0.7862**	**0.8853**
		1+2+3+4+5	0.8966	0.9539	0.7522	0.8675

In this table, numbers 1 through 5 represent the models 3D-RCNN$^{s}_{332}$, 3D-CNN$^{s}_{332}$, 3D-CNN$^{m}_{222}$, 3D-CNN$^{m}_{322}$, *and* 3D-CNN$^{m}_{332}$, *respectively. The highest performance in each case is shown in bold face.*

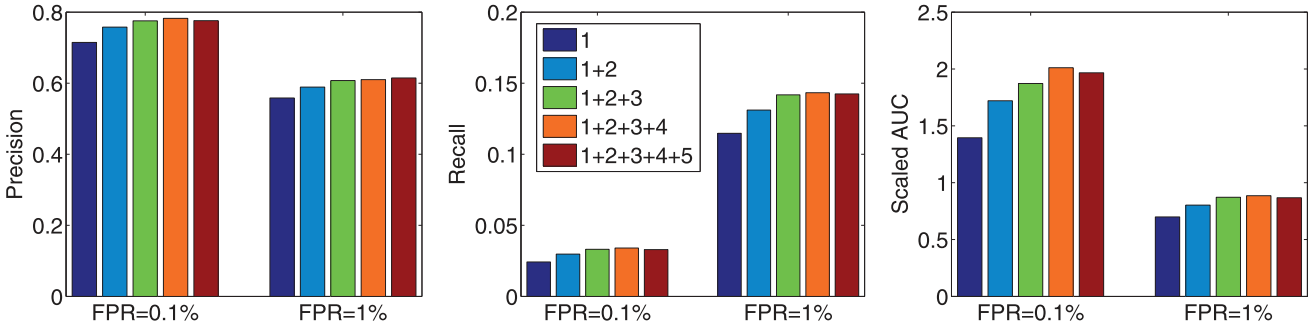


Fig. 9. Performance of different combinations of the 3D CNN architectures. The AUC scores at FPR = 0.1 and 1 percent are multiplied by 10^5 and 10^3, respectively, for better visualization. See the caption of Table 3 and the text for detailed explanations.

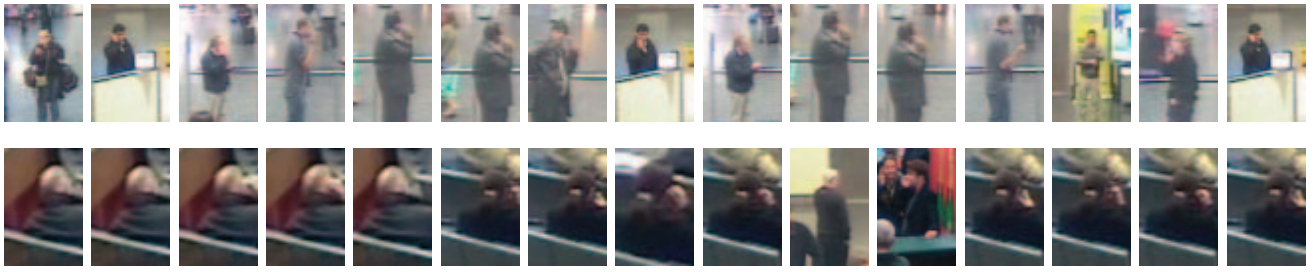


Fig. 10. Sample actions in the CellToEar class. The top row shows actions that are correctly recognized by the combined 3D CNN model, while the bottom row shows those that are misclassified by the model.

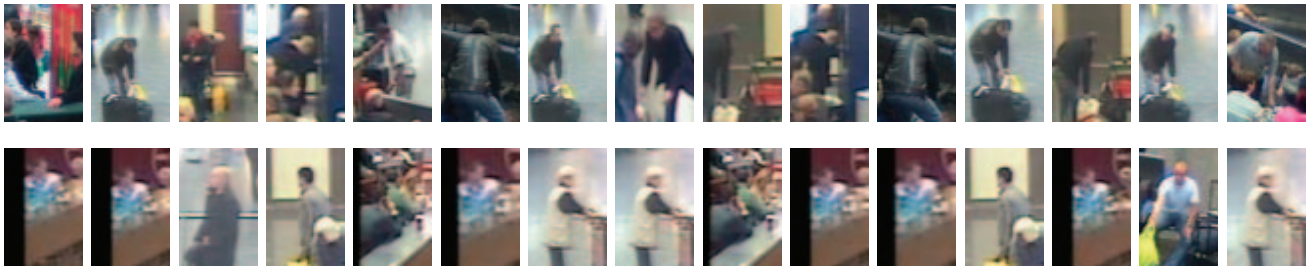


Fig. 11. Sample actions in the ObjectPut class. The top row shows actions that are correctly recognized by the combined 3D CNN model, while the bottom row shows those that are misclassified by the model.

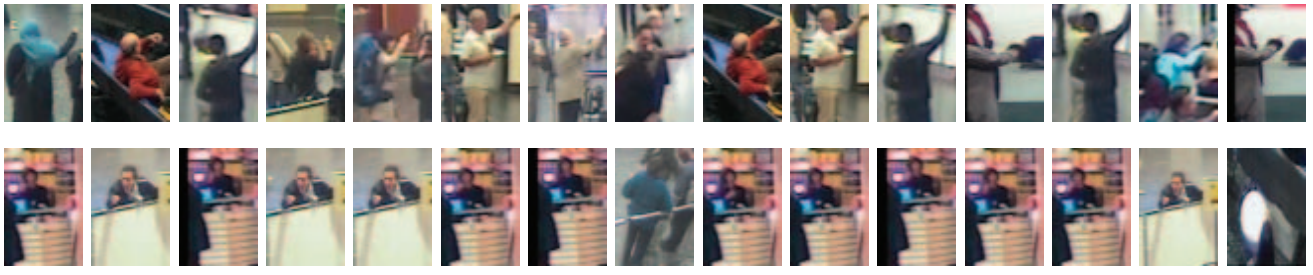


Fig. 12. Sample actions in the Pointing class. The top row shows actions that are correctly recognized by the combined 3D CNN model, while the bottom row shows those that are misclassified by the model.

6×4, respectively, and the two subsampling layers use kernels of size 3×3. By using this setting, the $80 \times 60 \times 9$ inputs are converted into 128D feature vectors. The final layer consists of 6 units corresponding to the six classes.

As in [16], we use the data for 16 randomly selected subjects for training and the data for the other nine subjects for testing. Majority voting is used to produce labels for a video sequence based on the predictions for individual frames. The recognition performance averaged across five random trials is reported in Table 5 along with published results in the literature. The 3D CNN model achieves an overall accuracy of 90.2 percent as compared with 91.7 percent achieved by the HMAX model. Note that the HMAX model uses handcrafted features computed from raw images with fourfold higher resolution. Also, some of the methods in Table 5 used different training/test splits of the data.

TABLE 4
Comparison of the Best Performance Achieved by Our Previous Methods in [26] (ICML Models) with the New Methods Proposed in This Paper (New Models)

	Precision		Recall		AUC	
FPR	0.1%	1%	0.1%	1%	0.1%	1%
New models	0.7824	0.6149	0.0340	0.1433	0.0201	0.8853
ICML models	0.7137	0.5572	0.0230	0.1132	0.0129	0.6752

TABLE 5
Action Recognition Accuracies in Percentage on the KTH Data

Method	Boxing	Handclapping	Handwaving	Jogging	Running	Walking	Average
3D CNN	90	94	97	84	79	97	90.2
Schüldt [13]	97.9	59.7	73.6	60.4	54.9	83.8	71.7
Dollár [14]	93	77	85	57	85	90	81.2
Niebles [56]	98	86	93	53	88	82	83.3
Jhuang [16]	92	98	92	85	87	96	91.7
Schindler [53]	–	–	–	–	–	–	92.7

Note that we use the same training/test split as [16] and other methods use different splits.

5 Conclusions and Discussions

We developed 3D CNN models for action recognition in this paper. These models construct features from both spatial and temporal dimensions by performing 3D convolutions. The developed deep architecture generates multiple channels of information from adjacent input frames and perform convolution and subsampling separately in each channel. The final feature representation is obtained by combining information from all channels. We developed model regularization and combination schemes to further boost the model performance. We evaluated the 3D CNN models on the TRECVID and the KTH data sets. Results show that the 3D CNN model outperforms compared methods on the TRECVID data, while it achieves competitive performance on the KTH data, demonstrating its superior performance in real-world environments.

In this paper, we considered the CNN model for action recognition. There are also other deep architectures, such as the deep belief networks [19], [23], which achieve promising performance on object recognition tasks. It would be interesting to extend such models for action recognition. The developed 3D CNN model was trained using a supervised algorithm in this paper, and it requires a large number of labeled samples. Prior studies show that the number of labeled samples can be significantly reduced when such a model is pretrained using unsupervised algorithms [22]. We will explore the unsupervised training of 3D CNN models in the future.

References

[1] I. Laptev and T. Lindeberg, "Space-Time Interest Points," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* pp. 432-439, 2003.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[3] J. Liu, J. Luo, and M. Shah, "Recognizing Realistic Actions from Videos," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1996-2003, 2009.

[4] Y. Wang and G. Mori, "Max-Margin Hidden Conditional Random Fields for Human Action Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 872-879, 2009.

[5] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic Annotation of Human Actions in Video," *Proc. 12th IEEE Int'l Conf. Computer Vision,* pp. 1491-1498, 2009.

[6] Y. Wang and G. Mori, "Hidden Part Models for Human Action Recognition: Probabilistic versus Max Margin," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 7, pp. 1310-1323, July 2011.

[7] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," *Proc. British Machine Vision Conf.,* p. 127, 2009.

[8] M. Marszalek, I. Laptev, and C. Schmid, "Actions in Context," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2929-2936, 2009.

[9] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "View-Independent Action Recognition from Temporal Self-Similarities," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 1, pp. 172-185, Jan. 2011.

[10] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing Human Actions in Still Images: A Study of Bag-of-Features and Part-Based Representations," *Proc. 21st British Machine Vision Conf.,* 2010.

[11] Q. Le, W. Zou, S. Yeung, and A. Ng, "Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 3361-3368, 2011.

[12] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* pp. 726-733, 2003.

[13] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Proc. 17th Int'l Conf. Pattern Recognition,* pp. 32-36, 2004.

[14] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," *Proc. IEEE Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* pp. 65-72, 2005.

[15] I. Laptev and P. Pérez, "Retrieving Actions in Movies," *Proc. 11th IEEE Int'l Conf. Computer Vision,* pp. 1-8, 2007.

[16] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A Biologically Inspired System for Action Recognition," *Proc. 11th IEEE Int'l Conf. Computer Vision,* pp. 1-8, 2007.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE,* vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[18] G.E. Hinton and R.R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science,* vol. 313, no. 5786, pp. 504-507, July 2006.

[19] G.E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation,* vol. 18, pp. 1527-1554, 2006.

[20] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning,* vol. 2, no. 1, pp. 1-127, 2009.

[21] Y. Bengio and Y. LeCun, "Scaling Learning Algorithms towards AI," *Large-Scale Kernel Machines,* L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds., MIT Press, 2007.

[22] M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[23] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng, "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations," *Proc. 26th Ann. Int'l Conf. Machine Learning,* pp. 609-616, 2009.

[24] M. Norouzi, M. Ranjbar, and G. Mori, "Stacks of Convolutional Restricted Boltzmann Machines for Shift-Invariant Feature Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.

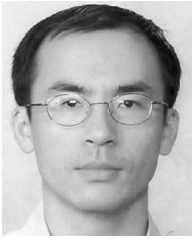
[25] M. Yang, S. Ji, W. Xu, J. Wang, F. Lv, K. Yu, Y. Gong, M. Dikmen, D.J. Lin, and T.S. Huang, "Detecting Human Actions in Surveillance Videos," *Proc. TREC Video Retrieval Evaluation Workshop,* 2009.

[26] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *Proc. 27th Int'l Conf. Machine Learning,* pp. 495-502, 2010.

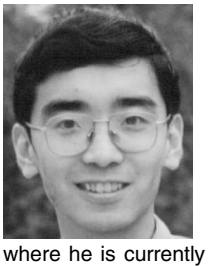
[27] G.W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional Learning of Spatio-Temporal Features," *Proc. 11th European Conf. Computer Vision,* pp. 140-153, 2010.

[28] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," *Proc. 25th Int'l Conf. Machine Learning,* pp. 160-167, 2008.
[29] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised Feature Learning for Audio Classification Using Convolutional Deep Belief Networks," *Proc. Advances in Neural Information Processing Systems 22,* pp. 1096-1104, 2009.
[30] H. Cecotti and A. Graser, "Convolutional Neural Networks for P300 Detection with Application to Brain-Computer Interfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 3, pp. 433-445, Mar. 2011.
[31] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human Tracking Using Convolutional Neural Networks," *IEEE Trans. Neural Networks,* vol. 21, no. 10, pp. 1610-1623, Oct. 2010.
[32] V. Jain, J.F. Murray, F. Roth, S. Turaga, V. Zhigulin, K.L. Briggman, M.N. Helmstaedter, W. Denk, and H.S. Seung, "Supervised Learning of Image Restoration with Convolutional Networks," *Proc. 11th IEEE Int'l Conf. Computer Vision,* 2007.
[33] V. Jain and S. Seung, "Natural Image Denoising with Convolutional Networks," *Proc. Advances in Neural Information Processing Systems 21,* pp. 769-776, 2009.
[34] S.C. Turaga, J.F. Murray, V. Jain, F. Roth, M. Helmstaedter, K. Briggman, W. Denk, and H.S. Seung, "Convolutional Networks Can Learn to Generate Affinity Graphs for Image Segmentation," *Neural Computation,* vol. 22, no. 2, pp. 511-538, 2010.
[35] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. Xing, "Training Hierarchical Feed-Forward Visual Recognition Models Using Transfer Learning from Pseudo-Tasks," *Proc. 10th European Conf. Computer Vision,* pp. 69-82, 2008.
[36] K. Yu, W. Xu, and Y. Gong, "Deep Learning with Kernel Regularization for Visual Recognition," *Proc. Advances in Neural Information Processing Systems 21,* pp. 1889-1896, 2009.
[37] H. Mobahi, R. Collobert, and J. Weston, "Deep Learning from Temporal Coherence in Video," *Proc. 26th Ann. Int'l Conf. Machine Learning,* pp. 737-744, 2009.
[38] Y. LeCun, F. Huang, and L. Bottou, "Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2004.
[39] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. Barbano, "Toward Automatic Phenotyping of Developing Embryos from Videos," *IEEE Trans. Image Processing,* vol. 14, no. 9, pp. 1360-1371, Sept. 2005.
[40] D.G. Lowe, "Distinctive Image Features from Scale Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
[41] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong, "Human Action Detection by Boosting Efficient Motion Features," *Proc. IEEE Workshop Video-Oriented Object and Event Classification,* 2009.
[42] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, pp. 123-140, 1996.
[43] Y. Freund and R.E. Schapire, "Experiments with a New Boosting Algorithm," *Proc. 13th Int'l Conf. Machine Learning,* pp. 148-156, 1996.
[44] J. Kittler, M. Hatef, R.P. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 3, pp. 226-239, Mar. 1998.
[45] L.K. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 10, pp. 993-1001, Oct. 1990.
[46] Y. LeCun, L. Bottou, G. Orr, and K. Muller, "Efficient Backprop," *Neural Networks: Tricks of the Trade,* G. Orr and M. Klaus-Robert, eds., Springer, 1998.
[47] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What Is the Best Multi-Stage Architecture for Object Recognition?" *Proc. 12thIEEE Int'l Conf. Computer Vision,* 2009.
[48] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object Recognition with Cortex-Like Mechanisms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 3, pp. 411-426, Mar. 2007.
[49] J. Mutch and D.G. Lowe, "Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields," *Int'l J. Computer Vision,* vol. 80, no. 1, pp. 45-57, Oct. 2008.
[50] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature Verification Using a Siamese Time Delay Neural Network," *Proc. Advances in Neural Information Processing Systems 6,* pp. 737-744, 1994.
[51] H.-J. Kim, J.S. Lee, and H.-S. Yang, "Human Action Recognition Using a Modified Convolutional Neural Network," *Proc. Fourth Int'l Symp. Neural Networks,* pp. 715-723, 2007.
[52] M. Yang, F. Lv, W. Xu, and Y. Gong, "Detection Driven Adaptive Multi-Cue Integration for Multiple Human Tracking," *Proc. 12th IEEE Int'l Conf. Computer Vision,* pp. 1554-1561, 2009.
[53] K. Schindler and L. Van Gool, "Action Snippets: How Many Frames Does Human Action Recognition Require?" *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[54] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong, "Detecting Video Events Based on Action Recognition in Complex Scenes Using Spatio-Temporal Descriptor," *Proc. 17th ACM Int'l Conf. Multimedia,* pp. 165-174, 2009.
[55] S. Lazebnik, C. Achmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2169-2178, 2006.
[56] J.C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int'l J. Computer Vision,* vol. 79, no. 3, pp. 299-318, 2008.

Shuiwang Ji received the PhD degree in computer science from Arizona State University, Tempe, Arizona, in 2010. Currently, he is working as an assistant professor in the Department of Computer Science at Old Dominion University (ODU), Norfolk, Virginia. His research interests include machine learning, data mining, and bioinformatics. He received the Outstanding PhD Student Award from Arizona State University in 2010 and the Early Career Distinguished Research Award from ODU's College of Sciences in 2012.



Wei Xu received the BS degree from Tsinghua University, Beijing, China, in 1998 and the MS degree from Carnegie Mellon University (CMU), Pittsburgh, in 2000. From 1998 to 2001, he was a research assistant in the Language Technology Institute at CMU. In 2001, he joined NEC Laboratories America working on intelligent video analysis. He has been a research scientist at Facebook since November 2009. His research interests include computer vision, image, and video understanding, machine learning, and data mining.



Ming Yang received the BE and ME degrees in electronic engineering from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree in electrical and computer engineering from Northwestern University, Evanston, Illinois, in June 2008. From 2004 to 2008, he was a research assistant in the computer vision group of Northwestern University. After his graduation, he joined NEC Laboratories America, Cupertino, California, where he is currently a research staff member. His research interests include computer vision, machine learning, video communication, large-scale image retrieval, and intelligent multimedia content analysis. He is a member of the IEEE.

Kai Yu received the PhD degree in computer science from the University of Munich in 2004. He is now the director of the Multimedia Department at Baidu. This work was done when he was the head of the Media Analytics Department at NEC Laboratories America, where he managed an R&D division working on image recognition, multimedia search, video surveillance, sensor mining, and human-computer interaction. He has published more than 70 papers in top-tier conferences and journals in the area of machine learning, data mining, and computer vision. He has served as area chair for top-tier machine learning conferences, e.g., ICML and NIPS, and taught an AI class at Stanford University as a visiting faculty member. Before joining NEC, he was a senior research scientist at Siemens. He is a member of the IEEE.