MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition

Yizhou Zhou*[1] Xiaoyan Sun[2] Zheng-Jun Zha[1] Wenjun Zeng[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia
zyz0205@mail.ustc.edu.cn, zhazj@ustc.edu.cn {xysun,wezeng}@microsoft.com

Abstract

Human actions in videos are three-dimensional (3D) signals. Recent attempts use 3D convolutional neural networks (CNNs) to explore spatio-temporal information for human action recognition. Though promising, 3D CNNs have not achieved high performanceon on this task with respect to their well-established two-dimensional (2D) counterparts for visual recognition in still images. We argue that the high training complexity of spatio-temporal fusion and the huge memory cost of 3D convolution hinder current 3D CNNs, which stack 3D convolutions layer by layer, by outputting deeper feature maps that are crucial for high-level tasks. We thus propose a Mixed Convolutional Tube (MiCT) that integrates 2D CNNs with the 3D convolution module to generate deeper and more informative feature maps, while reducing training complexity in each round of spatio-temporal fusion. A new end-to-end trainable deep 3D network, MiCT-Net, is also proposed based on the MiCT to better explore spatio-temporal information in human actions. Evaluations on three well-known benchmark datasets (UCF101, Sport-1M and HMDB-51) show that the proposed MiCT-Net significantly outperforms the original 3D CNNs. Compared with state-of-the-art approaches for action recognition on UCF101 and HMDB51, our MiCT-Net yields the best performance.

1. Introduction

Human action recognition is a fundamental yet challenging task with considerable efforts having been investigated for decades. Motivated by the notable-success of convolutional neural networks (CNNs) for visual recognition in still images, many recent works take advantage of deep models to train end-to-end networks for recognizing actions in videos [25, 9, 18, 35, 40, 20, 30], which significantly outperform hand-crafted representation learning methods [33, 23, 32, 17].

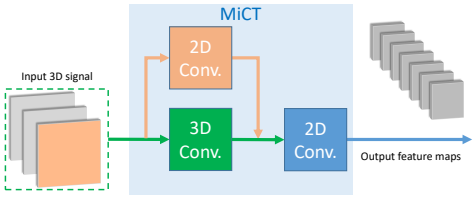

Figure 1. Illustration of our proposed MiCT that integrates 2D CNNs with the 3D convolution for the spatio-temporal feature learning.

Human actions in video sequences are three dimensional (3D) spatio-temporal signals. Jointly modeling spatio-temporal information via a 3D CNN in an end-to-end deep network provides a natural and efficient approach for action recognition. In spite of the large progress made by incorporating 3D CNN deep networks [30, 31, 11], the performance of action recognition in videos is still far from satisfactory compared with what has been achieved by 2D CNNs for visual recognition in images. Furthermore, state-of-the-art results for action recognition are obtained by a two-stream-like framework leveraging 2D CNNs pre-trained on huge image datasets [35, 7], though this approach does not provide systematic justification for its design choice [10].

Reconsidering current 3D CNN networks for action recognition, we notice that most of these methods share the same architecture that stacks 3D convolutions layer by layer, as proposed in C3D [30]. Since the spatial and temporal signals get coupled with each other through each 3D convolution, it becomes much more difficult to optimize the network with dozens of such 3D convolution layers because of the exponential growth of the solution space with respect to the case of 2D CNNs. Besides, the memory cost of 3D convolution is too high to be afforded in practice when constructing a deep 3D CNN, which makes the features of the current 3D CNNs usually not deep enough. For example, an 11-layer 3D CNN requires nearly 1.5 times as much memory as a 152-layer Residual Network. Based on the above observations, we believe it is very beneficial for a 3D CNN to limit the number of 3D convolution layers while increasing the depth of the feature maps. There may seem to be

* This work was performed while Yizhou Zhou was an intern with Microsoft Research Asia

a conflict between limiting the number of convolutions and increasing the depth of feature maps. The efficiency of the 2D convolution makes this a possibility.

In this paper, we present a new deep architecture to address this problem and improve the performance of 3D CNNs for action recognition with our proposed Mixed 2D/3D Convolutional Tube (MiCT). The MiCT enables the feature map at each spatio-temporal level to be much deeper prior to the next spatio-temporal fusion, which in turn makes it possible for the network to achieve better performance with fewer spatio-temporal fusions, while reducing the complexity of each round of spatio-temporal fusion by using the cross-domain residual connection. In contrast to the 3D CNNs that stack the 3D convolution layer by layer, the proposed MiCT, as shown in Fig.1, integrates 3D CNNs with 2D CNNs to enhance the feature learning with negligible increase in memory usage and complexity. Experiment results show that our proposed deep framework MiCT-Net with MiCT significantly enhances the performance of 3D CNNs for spatio-temporal feature learning and achieves state-of-the-art performance on three well-known benchmark datasets for action recognition.

2. Related Work

There exists an extensive body of literature on human action recognition. Here we outline work involving deep features and classify the related work into two categories, 2D CNN and 3D CNN based approaches, according to the convolutions used in feature learning.

2D CNN based. To explore the spatio-temporal information in human actions, the two-stream architecture is first proposed in [25] where two 2D CNNs are applied to the appearance (RGB frames) and motion (stacked optical flow) domains, respectively. Based on this architecture, several mechanisms are presented to fuse the two networks over the appearance and motion [15, 11, 9]. Li *et al*. extend the architecture via the multi-granular structure [18, 19]. A key volume mining deep framework is designed by Zhu *et al*. to identify key video clips and perform classification simultaneously [41]. Temporal segment networks is proposed which adopts a sparse temporal sampling strategy to enable long-range temporal observations [35].

On the other hand, early attempts to incorporate LSTM with traditional features have shown the potential of the LSTM-RNN network for modeling spatio-temporal information in action recognition [1, 2]. LSTM networks are employed to combine the frame-level features of 2D CNNs to explicitly model spatio-temporal relationships [40, 8]. Srivastava *et al*. [27] make use of LSTMs in an encoder-decoder framework for unsupervised video representation. Attention models are also presented based on the recurrent networks to weight the important frames [24] or highly relevant spatio-temporal locations as well [20].

Among these approaches, state-of-the-art performance is achieved in several large action databases. However, their success depends greatly on hand-crafted optical flow information, which is computationally expensive, and pre-trained 2D CNN models with huge datasets. The frame/optical flow based feature learning leaves the temporal evolution across consecutive frames not fully exploited [22]. In contrast, our scheme integrates 2D CNNs with the 3D CNNs for feature learning so that it is able to better exploit spatio-temporal information and benefit from pre-trained 2D CNNs, while requiring no additional complicated hand-crafted optical flows.

3D CNN based. The 3D CNN for action recognition was first presented in [14] to learn discriminative features along both spatial and temporal dimensions. Later, the C3D feature along with the corresponding 3D CNN architectures are presented in [30]. The out-of-the-box C3D feature has since been widely employed in many subsequent works on action recognition and detection [31, 21, 6, 39, 37, 4]. Varol *et al*. observe that utilizing a C3D network with longer temporal information can largely boost performance [31]. Wang *et al*. propose spatio-temporal pyramid pooling with LSTM to deal with arbitrary spatial and temporal sizing [37]. The performance of 3D CNNs is further improved by employing more complex spatio-temporal fusion strategies [7]. There are a few works that focus on ameliorating the downside of the 3D convolution based framework. Regarding 2D CNNs, 3D CNNs dramatically increase the number of parameters by extending spatial filters to spatio-temporal ones, thus greatly increasing both complexity and memory usage. To mitigate this drawback, Sun *et al*. factorize the 3D convolution kernel into a combination of a 2D spatial kernel and a 1D temporal kernel [28]. Similarly, Qiu *et al*. replace spatio-temporal 3D convolution with spatial and temporal convolutions in a residual connection style [22], which means these schemes are no longer 3D CNNs.

As mentioned before, all these 3D CNNs for action recognition follow the same structure - stacking the 3D convolutional module layer by layer. Frequent spatio-temporal fusions in the structure drastically increase the difficulty of optimizing the whole network and restrict the depth of feature maps in instances of limited memory resources. We address this problem by proposing a Mixed Convolution Tube (MiCT), which enables the 3D CNN to incorporate fewer 3D convolutions while also empowering feature learning by taking advantage of 2D CNNs to achieve better performance.

3. MiCT and Deep MiCT Network

In this section, we start with a brief introduction of the 3D convolution. We then give a detailed description of our proposed MiCT. Lastly, our simple yet efficient deep network, MiCT-Net, is presented for human action recognition.

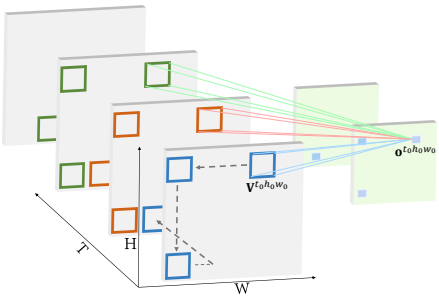

Figure 2. Illustration of a 3D convolution. The convolution kernels slide along both the spatial and temporal dimensions of the input 3D signal and generate the 3D spatio-temporal feature maps.

3.1. 3D Convolution

A 3D spatio-temporal signal, e.g. a video clip, can be represented as a tensor with a size of $T \times H \times W \times C$, where T, H, W, C denotes the temporal duration, height and width in the spatial domain, and number of channels, respectively. Kernels of a 3D convolution layer are then formulated as a 4D tensor $\mathcal{K} \in \mathbf{R}^{n_k \times t_k \times h_k \times w_k}$(we omit the channel dimension hereafter for simplicity), where l_k, h_k, w_k are the kernel size for the T, H, and w dimensions, and n_k denotes the number of kernels. As illustrated in Fig. 2, a 3D convolution layer takes the input 3D spatio-temporal features $\mathbf{V} = \{\mathbf{v}_{t,h,w}\}$ and outputs the 3D dimensional feature map $\mathbf{O} = \{\mathbf{o}_{t,h,w}\}$ by implementing convolution along both the spatial and temporal dimensions of the inputs (the stride size of the convolution is set to 1 for simplicity), which can be formulated as

$$\begin{aligned} \mathbf{O} &= \mathcal{K} \otimes \mathbf{V}, \quad where \\ \mathbf{o}_{t_0,h_0,w_0} &= [q^1_{t_0,h_0,w_0}, q^2_{t_0,h_0,w_0}, ..., q^{n_k}_{t_0,h_0,w_0}]^T, \\ q^n_{t_0,h_0,w_0} &= \sum_{t,w,h} \mathcal{K}_{n,t,w,h} \cdot \mathbf{V}^{t_0 h_0 w_0}_{t,w,h}. \end{aligned} \tag{1}$$

Here $\mathbf{V}^{t_0 h_0 w_0}$ is the sliced tensor that starts from the location (t_0, h_0, w_0) in $\mathbf{V}$ and has the same size as the kernel $\mathcal{K}^n$. $q^n_{t_0,h_0,w_0}$ denotes the value at (t_0, h_0, w_0) on the n^{th} feature map output by the n^{th} 3D convolution kernel.

3.2. MiCT

A 3D convolution couples spatio-temporal signals in an effort to effectively extract spatio-temporal features. However, when stacked together to form 3D CNNs, it also increases the difficulty of optimization, hinders 3D CNNs from generating deeper feature maps for high-level tasks due to unaffordable memory usage and high computational cost, and raises the demand on huge training sets. All these facts together limit the performance of current existing 3D CNNs on action recognition. In order to address these problems, we propose introducing 2D CNNs, which can be trained effectively, constructed deeply, and learned with huge datasets, to 3D convolution modules and form a new 3D convolution unit MiCT to empower feature learning, as illustrated in Fig.3. It integrates 2D convolutions with 3D convolutions to output much deeper feature maps at each round of spatio-temporal fusion. We propose mixing 3D and 2D convolutions in two ways, i.e. concatenating connections and cross-domain residual connections.

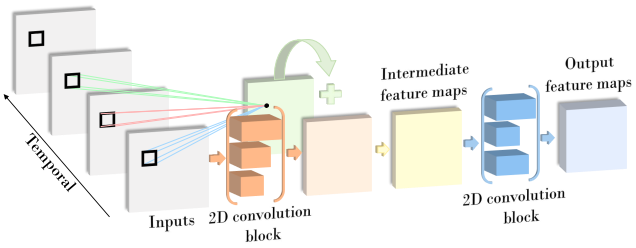

Figure 3. Illustration of MiCT that integrates 2D CNNs into 3D convolution for feature learning. In each MiCT, feature maps generated by the 3D convolutional module (green) are added to the ones produced by the residual 2D convolutional module (orange) on sampled 2D inputs. The combined feature maps are then fed into the concatenated 2D convolutional module (blue) to obtain the final feature maps.

3.2.1 Concatenating Connections

Fig. 4 illustrates the concatenated connection of 2D and 3D convolutions in the MiCT. We use MiCT$_{con}$ to represent the MiCT with only the concatenate connection hereafter. Denoting the feature map $\mathbf{O}$ at time t as $\mathbf{O}^t$, we have

$$\begin{aligned} \mathbf{O}^t &= \mathcal{M}(\mathbf{V}^t) \\ &= \mathcal{K} \otimes \mathbf{V}^t, \end{aligned} \tag{2}$$

where $\mathbf{V}^t \in \mathbf{R}^{l_k \times h \times w}$ is the sliced tensor from time t to time $t + l_k$. Since $\mathcal{M}(\cdot)$ only outputs linearly fused spatio-temporal feature maps based on Eq.(1) and (2), a 3D CNN has to stack enough of $\mathcal{M}(\cdot)$ for deep and high-level feature maps which requires dynamically increased memory usage, training samples, and training complexity. We thus propose enhancing $\mathcal{M}(\cdot)$ by a deeper and capable alternative $\mathcal{G}(\cdot)$ to extract much deeper features during every round of spatio-temporal fusion. $\mathcal{G}(\cdot)$ is supposed to meet three requirements. It should be computationally efficient, support end-to-end training, and be capable of feature learning for 2D and 3D signals. To meet these requirements, we design the function $\mathcal{G}(\cdot)$ by concatenating 2D CNNs after the 3D convolution to provide a very efficient deep feature extractor, denoted as

$$\mathcal{G}(\cdot) = \mathcal{H}(\mathcal{M}(\cdot)), \tag{3}$$

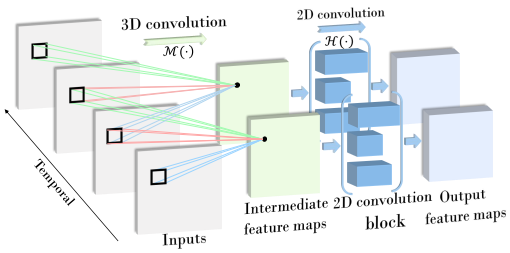

Figure 4. MiCT with concatenated connections. For an input 3D signal, the 3D convolution fuses spatio-temporal information and obtains intermediate feature maps which are fed into the 2D convolution block to generate the final feature maps.

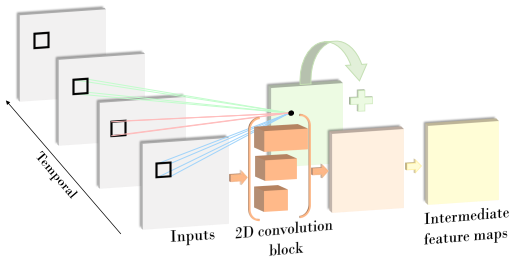

Figure 5. MiCT with a cross-domain residual connection. Spatio-temporal fusion is achieved by both the 2D convolution block to generate stationary features and 3D convolution to extract temporal residual information.

where $\mathcal{H}(\cdot)$ denotes the mapping function of a 2D convolution block. In other words, feature maps of a 3D input $\mathbf{V}$ using $\mathcal{G}(\cdot)$ are achieved by coupling the 3D convolution with the 2D convolution block serially in which the 3D convolution enables spatio-temporal information fusion while the 2D convolution block deepens feature learning for each 2D output of the 3D convolution.

3.2.2 Cross-Domain Residual Connections

The MiCT with only a cross-domain residual connection, denoted as MiCT$_{res}$, is illustrated in Fig. 5. It introduces a 2D convolution between the input and output of the 3D convolution to further reduce spatio-temporal fusion complexity and facilitate the optimization of the whole network. Following the notations in Eq. (1), we have

$$\begin{aligned} &\mathbf{o}^{'}_{t_0,h_0,w_0} = \mathbf{o}_{t_0,h_0,w_0} + \mathbf{S}^{t_0}_{h_0,w_0}, \\ &where\ \ \mathbf{S}^{t_0} = \mathcal{H}^{'}(\mathbf{V}^{t_0}). \end{aligned} \tag{4}$$

Here $\mathbf{V}^{t_0} \in \mathbf{R}^{h \times w}$ is the sliced tensor of input $\mathbf{V}$ at time t_0, $\mathbf{S}^{t_0}_{h_0,w_0}$ refers to the value at (h_0, w_0) on $\mathbf{S}^{t_0}$ obtained by $\mathcal{H}^{'}(\cdot)$, and $\mathcal{H}^{'}(\cdot)$ denotes a 2D convolution block. Unlike the residual connections in previous work [12, 9], the shortcut in our scheme is cross-domain, where spatio-temporal fusion is derived by both a 3D convolution mapping with respect to the full 3D inputs and a 2D convolution block mapping with respect to the sampled 2D inputs. We propose a cross-domain residual connection based on the observation that a video stream usually contains lots of redundant information among consecutive frames, resulting in redundant information in feature maps along the temporal dimension. By introducing a 2D convolution block to extract the very informative but static 2D features, the 3D convolution in MiCT$_{res}$ only needs to learn residual information along the temporal dimension. Thus the cross-domain residual connection largely reduces the complexity of MiCT in the learning for 3D convolution kernels.

Our proposed MiCT combines the two connections as shown in Fig. 3 and achieves the best performance among the three configurations MiCT$_{con}$, MiCT$_{res}$, and MiCT, which will be demonstrated in Section 4.

3.3. Deep MiCT Network

We propose a simple yet efficient deep MiCT Network (MiCT-Net in short) by stacking the MiCT together. The MiCT-Net takes the RGB video sequences as inputs and is end-to-end trainable. As shown in Fig. 6, it consists of four MiCTs, which means only four 3D convolutions are employed. For the 2D convolution blocks in each MiCT block, we partially follow the designs of BN-inception [29]. More details of the network architecture are provided in Table 3.3. The inception in the table refers to the architecture as shown at the top-right corner of Fig. 6. The batch normalization and ReLU layer after each convolution are omitted for simplicity. We also employ global pooling along the temporal dimension in the last layer of the network to enable the network to accept arbitrary-length videos as inputs.

Regarding the baseline C3D architecture [30, 31], the MiCT-Net contains fewer 3D convolutions for spatio-temporal fusion while it produceing deeper feature maps and limiting the complexity of the entire deep model. Moreover, unlike traditional 3D CNNs, our framework is able to take advantage of 2D models pre-trained on large image datasets. The pre-trained parameters on large image datasets potentially provide MiCT with more advanced initialization in 2D convolution blocks for feature learning.

4. Experiments

In this section, we first introduce the evaluation datasets, data augmentation, and training configuration used in our tests. We then evaluate the performance of our scheme by comparisons to both the baseline 3D CNN and state-of-the-art approaches. Regarding the MiCT-Net, we perform evaluations with three configurations, MiCT$_{con}$, MiCT$_{res}$, and MiCT, respectively, which also provide an ablation study of our MiCT for human action recognition.

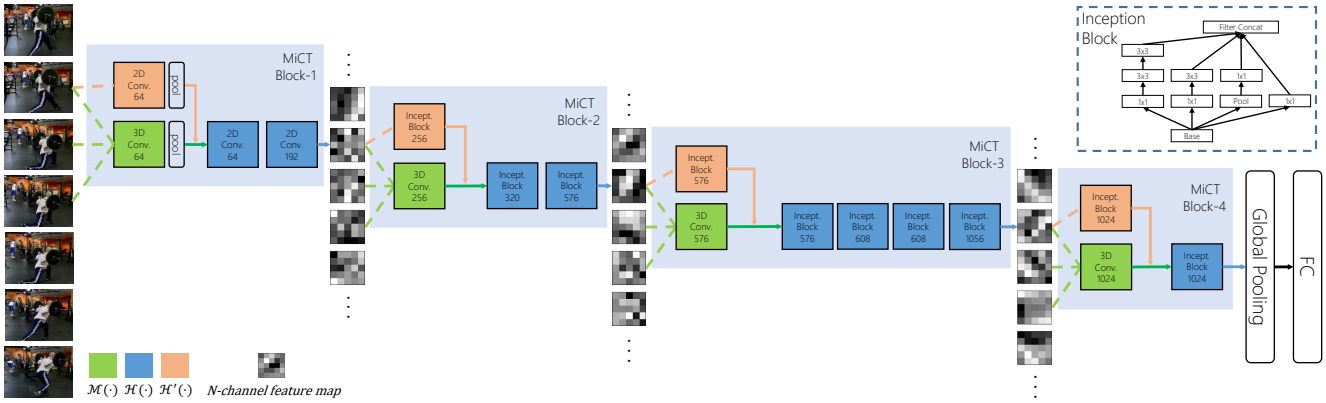

Figure 6. Illustration of the proposed MiCT-Net. Green blocks refer to 3D convolution. Orange blocks and blue blocks refer to 2D convolutions for cross-domain residual connections and concatenated connections, respectively. Each mosaic-like box denotes an n-channel feature map at time t (n=192, 576, 1056 in MiCT Block-1/2/3, respectively). The architecture details of each Incept. block are shown in the top-right area of the figure.

Type	$\mathcal{M}(\cdot)$ size / stride	$\mathcal{H}(\cdot)$ size / stride	$\mathcal{H}^{'}(\cdot)$ size / stride
block-1	3x7x7x64/(1,2) Maxpool/(1,2)	1x1x64/1 3x3x192/1	7x7x64/2 Maxpool/2
block-2	3x3x3x256/(2,1)	2xInception	1xInception
block-3	3x3x3x576/(2,1)	4xInception	1xInception
block-4	3x3x3x1024/(2,1)	1xInception	1xInception
pooling	global pooling on spatial dimension		
fc	$1024 \times num\ classes$		
pooling	global pooling		

Table 1. Architecture of MiCT-Net. The size and stride of $\mathcal{M}(\cdot)$ are denoted in $time \times height \times width \times number\ of\ kernels$ and $(stride_{temporal}, stride_{spatial})$, respectively. The sizes and strides of both $\mathcal{H}(\cdot)$ and $\mathcal{H}^{'}(\cdot)$ are shown in $height \times width \times number\ of\ kernels$ and $stride_{spatial}$, respectively.

4.1. Experiment Settings

Datasets. Three well-known benchmarks, UCF101[26], Sport1M[15], and HMDB-51[16], are used in the evaluations. UCF101 consists of 13,320 manually labeled videos from 101 action categories. It has three train/test splits. Each split has around 9,500 videos for training and 3,700 videos for testing. HMDB51 is collected from various sources, e.g. web videos and movies, which proves to be realistic and challenging. It consists of 6,766 manually labeled clips from 51 categories. Sport1M consists of 1.1 million automatically labeled sports videos from 487 categories.

Data augmentation. Our data augmentation includes random clipping, brightness and contrast adjustment, and temporal sampling. We resize each frame to 256×340 and crop 224×224 regions randomly. We also randomly flip frames horizontally and subtract the mean values from the R, G and B channels. Random brightness and contrast adjustments are also applied to each frame. Moreover, we adopt multiple down-sampling rates along the temporal dimension to generate training samples with multiple temporal resolutions to further diversify the pattern of an action. All random operations are consistent across all frames in one video sequence in training. During testing, we only use central cropping and a fixed down-sampling rate along the temporal dimension for simplicity.

Training configuration. We use the Adam Gradient Descent optimizer with an initial learning rate of $1e^{-4}$ to train the MiCT-related networks from scratch. The drop out ratio and weight decay rate are set to 0.5 and $5e^{-5}$, respectively, for all datasets (except HMDB51 for which the dropout ratio is 0.8). The gradient descent optimizer is adopted with a momentum of 0.9 to train our MiCT-Net initialized with the ImageNet pre-trained model. The initial learning rate is $1e^{-5}$. We employ the higher drop out ratio of 0.9 and the weight decay rate of $5e^{-4}$ to prevent over-fitting.

4.2. Comparison with the Baseline 3D CNN

We first evaluate the performance of our MiCT-Net in comparison with that of the baseline 3D CNN approach called C3D [30]. C3D is a typical and popular 3D CNN for action recognition which stacks the 3D convolutions layer by layer. We choose C3D as our baseline since it is the most direct way to show what has been improved by MiCT.

Table 2 exhibits the comparison results in terms of mean accuracy. The performance is evaluated at both clip and video levels. We also implement a C3D with batch normalization (BN) [13] for fairness comparison as our MiCT contains the BN module. Considering that long-term temporal inputs can significantly boost the performance of C3D [31], we evaluate the accuracy using inputs at the length of both 35 and 16 frames (except the test on Sports1M which only

Method	Aux-Data	BN	Test	UCF101(16f / 35f)	HMDB51(16f / 35f)	Sport1M(16f)	Model Size	Speed
C3D	-	N	Clip	44.0% / 47.9%	37.0% / 43.2%	44.9%	321MB	-
			Video	- / 50.2%	43.9% / 46.6%	60.0%		
C3D	-	Y	Clip	45.8% / 49.6%	38.4% / 44.9%	45.4%	-	-
			Video	49.3% / 51.7%	45.7% / 48.0%	60.8%		
MiCT$_{con}$-Net	-	Y	Clip	49.7% / 53.9%	41.2% / 48.6%	47.0%	-	-
			Video	53.6% / 56.1%	48.3% / 51.9%	63.4%		
MiCT$_{res}$-Net	-	Y	Clip	46.6% / 50.4%	38.9% / 45.5%	45.9%	-	-
			Video	50.1% / 52.8%	46.5% / 48.4%	61.2%		
MiCT-Net	-	Y	Clip	50.9% / 56.5%	43.9% / 51.1%	47.6%	221MB	-
			Video	54.6% / 58.7%	50.4% / 54.3%	64.1%		
MiCT-Net	ImageNet	Y	Clip	81.4% / 85.1%	48.1% / 55.3%	-	221MB	-
			Video	84.9% / 87.3%	55.2% / 58.0%	-		
C3D(1 Net)	Sport1M	N	Clip	- / -	- / -	-	321MB	323fps
	+I380K		Video	82.3% / -	- / -	-		
MiCT-Net	ImageNet	Y	Clip	84.3% / 87.8%	- / -	-	221MB	394fps
	+Sport1M		Video	88.6% / 89.1%	- / -	-		

Table 2. Comparison with C3D on UCF101, HMDB51 and Sport1M. The performance of C3D with batch normalization and C3D with 35 frames for training are reported based on our experimental results. It can be observed that the 3D CNN with either the concatenated connection (MiCT$_{con}$-Net) or the cross-domain residual connection (MiCT$_{res}$-Net) outperforms the traditional 3D CNN. Leveraging the ImageNet pre-trained model can further boost performance of the MiCT-Net.

involves 16-frame inputs). Experiment results shown in this table demonstrate that MiCT-Net significantly outperforms the baseline approach on all three benchmarks under different test conditions. It is also worth noting that C3D uses eight 3D convolutions while our model only uses four 3D convolutions but achieves better performance, which indicates that the MiCT can learn and represent spatio-temporal features much more efficiently and accurately than 3D convolution.

UCF101. We show the mean accuracy of different models on UCF101 in Table 2. The official data splits (3 splits) are used for both training and testing. We can observe that all three MiCT-based networks outperform the C3D network. Taking the test condition Video/35f/with BN as an example, the MiCT$_{con}$-Net increases accuracy by up to 6% and 4.4% compared with the C3D and C3D with BN, respectively, which demonstrates the importance of employing the concatenated 2D convolution blocks for enhanced spatio-temporal feature mapping. The MiCT-Net combining both MiCT$_{res}$ and MiCT$_{con}$ can further improve performance by 6.8%. The performance of our MiCT-Net can be boosted to 89.1% by leveraging the ImageNet pre-trained model, which leads to nearly 40% higher performance than that of C3D. This makes the proposed MiCT-Net much more practical when only a limited number of training samples is available. Moreover, given a larger training set, i.e. Sport1M to MiCT and I380K to C3D, the MiCT-Net still outperforms C3D with a 6.3% gain in accuracy.

HMDB51. Similar enhancements are achieved by the MiCT on HMDB51. Results show that the improvements brought by MiCT$_{con}$-Net vary from 4.2% to 5.4% and 2.8% to 3.9% compared to C3D with and without BN at the video level. The MiCT$_{res}$ contributes relatively less which may partially be due to the complicated temporal variations in this dataset. Putting these two together, the MiCT-Net obtains an additional 2.7% gain on average, and achieves 58% accuracy when using the ImageNet pre-trained model.

Sport1M. Sport1M is a very challenging dataset which contains long, weakly annotated videos. As some URLs of Sport1M are invalid, we can only access about 90% of the whole dataset. Experiments are conducted on these available data using the official data split. Results in this table show that our MiCT-Net can still outperform C3D with BN with a 3.3% gain in accuracy.

In addition to accuracy, we also evaluate the model sizes of both C3D and our MiCT. It is clear that the model size of MiCT is much smaller. All these faces show that the proposed MiCT-Net significantly outperforms the baseline approach C3D. It is more capable of handling spatio-temporal signals than 3D convolution and each component of the MiCT does help increase performance. It should be noted that the design of the MiCT makes it possible to use pre-trained models on very large image datasets, which brings considerable benefits for training and final performance.

4.3. Comparison with the State-of-the-Art Methods

We further demonstrate the advances of the proposed MiCT in comparison with state-of-the-art works for action recognition. Related results on UCF101 and HMDB51 are shown in Tables 3 and 4, respectively.

Our MiCT-Net is able to explore spatio-temporal information by requiring only RGB frames as inputs. In Ta-

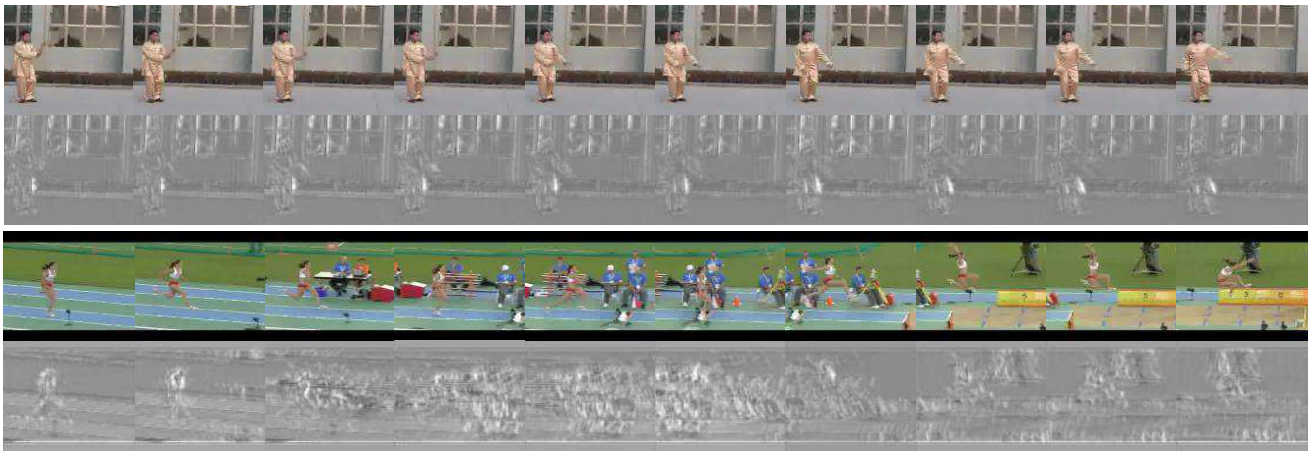

Figure 7. Visualization of the feature maps in the first MiCT block. It can be observed that those feature maps are very sensitive to object edges and regions containing large motions. Also, they are visually very similar to dynamic image [3] or motion blur, which records action changes along the temporal dimension.

Method	UCF101	HMDB51
Slow fusion [15]	65.4%	-
C3D [30]	44.0%[1]	43.9[2]%
LTC [31]	59.9%	-
Two-stream [25]	73.0%	40.5%
Two-stream fusion [11]	82.6%	47.1%
Two-stream+LSTM [40]	82.6%	47.1%
Transformations [36]	81.9%	44.1%
TSN [35]	85.7%	54.6%[3]
F_{ST}CN [28]	71.3%	42.0%
ST-ResNet [9]	82.2%	43.4%
Key-volume mining CNN [41]	84.5%	-
TLE(C3D CNN) [7]	86.3%	60.3%
TLE(BN-Inception) [7]	86.9%	63.2%
I3D [5]	84.5%	49.8%
P3D ResNet [22]	88.6%	-
MiCT-Net	**88.9**%	**63.8**%

Table 3. Performance comparison with the state-of-the-art results on UCF101 and HMDB51 with only RGB frames as inputs. The results with [1] are read from the figure in [30], [2] is the results reported in [31], and [3] is obtained from the released code of the paper.

Method	UCF101	HMDB51
C3D + IDT [30]	90.4%	-
TDD + IDT [34]	91.5%	65.9%
LTC [31]	91.7%	64.8%
LTC + IDT [31]	92.7%	67.2%
ST-ResNet + IDT [9]	94.6%	70.3%
P3D ResNet + IDT [22]	93.7%	-
Two-stream+LSTM [40]	88.6%	-
Two-stream(conv. fuse) [11]	92.5%	65.4%
Transformations [36]	92.4%	62.0%
TLE [7]	**95.6**%	**71.1**%
TSN (3 modalities) [35]	94.2%	69.4%
Spatio-temporal Network [38]	94.6%	68.9%
Two-stream MiCT-Net	94.7%	70.5%

Table 4. Performance comparison with state-of-the-art results on UCF101 and HMDB51.

ble 3, we show the performance comparison with state-of-the-art action recognition methods using only RGB inputs reported by those works for a fair comparison. The result of MiCT-Net is achieved by using inputs at the length of 75 frames. It can be observed that on both datasets, our MiCT-Net achieves the best performance, 88.9% on UCF101 and 63.8% on HMDB51, among all the compared methods. Even some of these referred works adopt advanced spatio-temporal fusion methods to the feature maps of the 2D CNNs [7, 41, 40], or learn very deep spatio-temporal features by either decomposing a 3D convolution into a 2D convolution along the spatial dimension and a 1D convolution along the temporal dimension to model spatio-temporal information [22] or directly inflating the state-of-the-art 2D CNN architecture into 3D CNN to take advantage of well-trained 2D models [5], the MiCT-Net still performs the best. Regarding 3D convolution based methods with the best accuracy up to 59.9% on UCF101 and 52.9% on HMDB51, our proposed MiCT-Net significantly enhances performance which indicates the large efficiency and accuracy brought by the proposed MiCT.

Additional motion information has proven helpful for action recognition in many previous works. Two-stream based proposals explicitly employ motion features, e.g. op-

tical flow, to boost performance. For example, the accuracy of using both RGB image and optical flow with LSTM increases to 88.1% in two-stream+LSTM [40]. Even so, our results are still comparable, which shows the efficiency and potential of the MiCT-Net. But the performance of the MiCT-Net with only RGB frames as inputs is still limited when considering the very sophisticated and well-designed frameworks that utilize hand-crafted motion features, such as TSN [35], a two-stream based framework that achieves 94.2% accuracy.

However, the MiCT is primarily proposed to enhance 3D CNNs. It can be applied to any framework that incorporates 3D convolutions. Some recent work use two-stream 3D CNNs for action recognition [37, 7]. Similar to the two-stream 2D CNNs, both optical flow and RGB frames can be employed as inputs. To better demonstrate the effectiveness of our MiCT, we present a simple two-stream MiCT-Net, where one stream takes as inputs the RGB frame and the other stream takes as inputs the optical flow. The architecture of the RGB stream is identical to the MiCT-Net, and the flow stream is almost the same as the MiCT-Net, except that we expand the channel dimension of the first convolution layer from 3 to 10. This is because for each RGB frame, we use 5 optical flow images around it and each flow image consists of two channels (x and y). We separately optimize the RGB stream and flow stream during training and simply average the inferences of the two streams as the final predictions during testing.

We compare the performance of the two-stream MiCT-Net with state-of-the-art works on both UCF101 and HMDB51. As shown in the Table 4, the performance of the MiCT-Net becomes further improved by incorporating the extra flow stream, which achieves state-of-the-art results of 94.7% on UCF101 and 70.5% on HMDB51. We claim that although 3D convolution is supposed to extract temporal information without requiring hand-crafted features, the optical flow still provides more detailed motion information that is very beneficial for action recognition. Please note that unlike many other two-stream solutions that employ lots of sophisticated designs, such as spatio-temporal residual connection [9], fusing the mid-layer features [11] or encoding the last layer features [7], we only employ a naive two-stream architecture with the proposed MiCT-Net to achieve state-of-the-art performance. Experiments and evaluations for different two-stream architectures to boost the performance are outside the scope of this paper.

4.4. Visualization

In order to better illustrate what MiCT has learned, we provide some feature maps output by the first MiCT block in Fig.7. Here we show two examples. The top one, Taichi, has slow motion. The feature maps learned by MiCT focus more on contextual information, such as body parts and background edges, which is very important for slow motion recognition, while the bottom one, LongJump, contains fast motion. Correspondingly, the feature maps learned with the MiCT focus more on the motion area and are visually very similar to motion blur, which indicates that the learned feature maps are trying to capture the time. Based on the above observation, our MiCT seems to be adaptive to the content, which is reasonable and consistent with human intuition.

5. Conclusion

In this paper, we propose the Mixed 2D/3D Convolutional Tube (MiCT) which enables 3D CNNs to extract deeper spatio-temporal features with fewer 3D spatio-temporal fusions and to reduce the complexity of the information that a 3D convolution needs to encode at each round of spatio-temporal fusion. We further present a deep network MiCT-Net based on the MiCT which is end-to-end trainable for human action recognition. Experiment results demonstrate that our MiCT-Net significantly outperforms traditional 3D CNNs for action recognition. Morever, the MiCT-Net achieves the best performance on both UCF101 and HMDB51 in comparison to state-of-the-art approaches with RGB input. We further show that the MiCT can be applied to other 3D CNN architectures, e.g. two-stream 3D CNNs, to achieve state-of-the-art performance, indicating that the proposed MiCT is general and efficient.

One explanation of the performance improvement of the MiCT-Net could be that a deep network generally benefits by deepening the network. Another is based on the observation that there is stronger temporal redundancy than a spatial one in videos. We therefore propose using additional 2D CNNs after each 3D convolution to further enhance the abstraction ability on the spatial domain. We also leverage the concept of ResNet to propose a cross-domain shortcut to facilitate 3D feature learning. However, the current MiCT-Net is very simple, without a comprehensive exploration of architecture, which indicates that the potential enhancement of MiCT-Net is very promising.

Acknowledgement

This work is partially supported by the Natural Science Foundation of China (NSFC, No.61622211, 61472392 and 61620106009), and the National Key Research and Development Plan of China (No.2017YFB1300201). We would like to thank all the reviewers for their insightful comments. We also appreciate the efforts of Dr. Steven Lin in the scientific manuscript editing of this paper.

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Action classification in soccer videos with long

short-term memory recurrent neural networks. *Artificial Neural Networks–ICANN 2010*, pages 154–159, 2010.

[2] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011.

[3] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3042, 2016.

[4] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 49–54. IEEE, 2016.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[6] A. Diba, A. M. Pazandeh, and L. Van Gool. Efficient two-stream motion and appearance 3d cnns for video classification. *arXiv preprint arXiv:1608.08851*, 2016.

[7] A. Diba, V. Sharma, and L. Van Gool. Deep temporal linear encoding networks. *arXiv preprint arXiv:1611.06678*, 2016.

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[9] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems*, pages 3468–3476, 2016.

[10] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7445–7454. IEEE, 2017.

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[14] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[16] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering 12*, pages 571–582. Springer, 2013.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[18] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo. Action recognition by learning deep multi-granular spatio-temporal video representation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 159–166. ACM, 2016.

[19] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo. Learning hierarchical video representation for action recognition. *International Journal of Multimedia Information Retrieval*, 6(1):85–98, 2017.

[20] Z. Li, E. Gavves, M. Jain, and C. G. Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 2017.

[21] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215, 2016.

[22] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5533–5541, 2017.

[23] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007.

[24] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[25] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[26] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[27] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, pages 843–852, 2015.

[28] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4597–4605, 2015.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[31] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[32] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[33] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[34] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015.

[35] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[36] X. Wang, A. Farhadi, and A. Gupta. Actions˜ transformations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2658–2667, 2016.

[37] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu. Two-stream 3d convnet fusion for action recognition in videos with arbitrary size and length. *IEEE Transactions on Multimedia*, 2017.

[38] Y. Wang, M. Long, J. Wang, and P. S. Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2017.

[39] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. *arXiv preprint arXiv:1703.07814*, 2017.

[40] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[41] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao. A key volume mining deep framework for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1991–1999, 2016.