Hierarchical Motion Evolution for Action Recognition

Hongsong Wang[1,2] Wei Wang[1,2] Liang Wang[1,2]
[1]Center for Research on Intelligent Perception and Computing, Institute of Automation
[2]National Laboratory of Pattern Recognition, Chinese Academy of Sciences
{hongsong.wang, wangwei, wangliang}@nlpr.ia.ac.cn

Abstract

Human action can be decomposed into a series of temporally correlated motions. Since the traditional bag-of-words framework based on local features cannot model global motion evolution of actions, models like Recurrent Neural Network (RNN) [15] and VideoDarwin [5] are accordingly explored to capture video-wise temporal information. Inspired by VideoDarwin, in this paper, we present a novel hierarchical scheme to learn better video representation, called HiVideoDarwin. Specifically, we first use different ranking machines to learn motion descriptors of local video clips. Then, in order to model motion evolution, we encode features obtained in previous layer again using a ranking machine. Compared with VideoDarwin, HiVideoDarwin captures the global and high-level video representation and is robust to large appearance changes. Compared with RNN, HiVideoDarwin can also abstract semantic information in a hierarchical way and is fast to compute and easy to interpret. We evaluate the proposed method on two datasets, namely MPII Cooking and Chalearn. Experimental results show that HiVideoDarwin has distinct advantages over the state-of-the-art models. Additional sensitivity analysis reveals that the overall results are hardly affected by parameter changes.

1. Introduction

Action recognition in videos has been an active research area due to its potential applications in video surveillance, video indexing, human computer interaction, etc. Significant progress has been made in recent years by designing local spatio-temporal features and adopting different encoding schemes [16, 24]. The very popular features are dense trajectories [23, 25] and its variants [10, 8], which track densely sampled local interest points in the short term (e.g., 15 frames) and compute hand-crafted features (e.g., HOG, HOF, MBHx, MBHy) based on the trajectories. These features are then encoded with Fisher vector (FV) or other super vector encoding methods (e.g., VLAD), which can achieve good performance for action classification. These methods can capture the discriminative local changing patterns around the interest points well, but fail to describe the global motion patterns or the long-term temporal evolution of features.

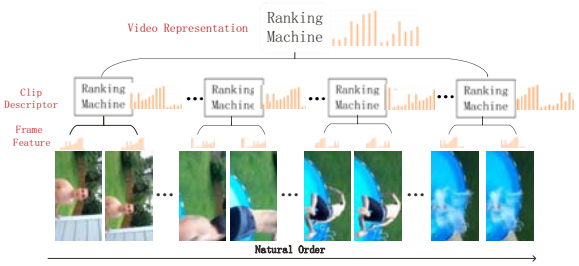

Figure 1. Pipeline of HiVideoDarwin. In this example, we construct global feature evolution with two layers. In the first layer, different ranking machines are performed for video clips to get local motion descriptors. In the second layer, these descriptors are combined to obtain final video representation.

Inspired by the great success of deep learning for image classification [13, 21, 22], object localization [6] and speech recognition [2], etc, there is a growing trend of learning video representations using deep neural networks. Ji et al. [9] and Karparthy et al. [12] extend the 2-D Convolutional Neural Networks (CNN) to the temporal dimension to learn 3-D video features. Simonyan et al. [20] propose a two-stream CNN framework by performing two CNNs on static frames and optical flows, respectively. These CNN based approaches only compute motion features in short time windows, so it is hard for them to capture global temporal motions. Considering that Recurrent Neural Network (RNN) has the generic powerful ability for sequence data [7], recently, Hei Ng et al. [15] connect Long Short-Term Memory (LSTM) cells to the output of the CNN and show performance improvement when compared with various convolutional temporal pooling architectures. Donahue et al. [3] develop a novel end-to-end trainable recurrent convolutional architecture suitable for a variety of visual tasks like action recognition, image description and retrieval. These

978-1-4799-6100-9/15/$31.00 ©2015 IEEE

RNN based approaches combine image information across a video over a longer time period, but are computationally expensive and difficult to train.

More recently, Fernando et al. [5] propose a method called VideoDarwin that can model video appearance evolution. They adopt a ranking machine to model the ordering of frames and use the parameters as video representation. Although this method shows promising results in a variety of action datasets, it has several drawbacks. Firstly, for long video sequences, a single ranking machine could hardly capture the global ordering. Secondly, the model is sensitive to the large appearance changes between adjacent frames due to the uncertain ordering relationship.

To address the above problems, in this paper, we propose Hierarchical VideoDarwin (HiVideoDarwin) for action recognition. Figure 1 shows the pipeline of our approach. Unlike VideoDarwin that uses one ranking machine to summarize all the frame descriptors, our HiVideoDarwin first performs separate ranking machines for the divided clips to get local motion representation. Then we treat them as inputs of a ranking machine in the next layer to model video motion evolution. Compared with original VideoDarwin, HiVideoDarwin models both local appearance evolution for clips and global motion evolution for video. Compared with deep learning based approaches, HiVideoDarwin can also abstract semantic information in a hierarchical way. Compared with traditional action recognition approaches, HiVideoDarwin can model global motion evolution for actions based on local features.

The main contribution is that we propose a hierarchical ranking based method to learn video representation, which considers both local order and global order of video frames to model local appearance evolution and global motion evolution. It is worth noting that our approach can also be applied to other large scale learning to rank problems. With the resulting new representation we obtain the state-of-the-art results in fine-grained action and gesture recognition tasks.

2. VideoDarwin

As Fernando et al. show in [5], a ranking machine can be used to model the appearance evolution of video frames. They use the parameters of the ranking machine as new video representation, named VideoDarwin.

Given a video with n frames $X = [x_1, x_2, \cdots, x_n]$, x_i is the representation of frame i, Fernando et al. define a simple vector valued function V over the time variable t, $V : t \rightarrow v_t$, where v_t combines the representations of all frames up to time t. In this way, the relative ordering constraints exist for v_t, which is, $v_1 \prec \cdots \prec v_t \prec \cdots \prec v_n$. Then a lot of learning to rank paradigms [14, 11] can be adopted to maintain the order and to abstract features. They use a pairwise linear ranking machine to illustrate the process. With parameter u, the ranking score of frame t is obtained as $\psi(v_t; u) = u^T v_t$, such that, $\forall t_i, t_j, t_i > t_j \Leftrightarrow u^T \cdot t_i > u^T \cdot t_j$. The objective can be optimized via a max-margin framework.

$$\begin{aligned} &\arg\min_u \tfrac{1}{2} \cdot \|u\|^2 + C \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} \varepsilon_{ij} \\ &s.t.\ u^T(v_{t_i} - v_{t_j}) \geq 1 - \varepsilon_{ij} \\ &\quad \varepsilon_{ij} \geq 0. \end{aligned} \tag{1}$$

In [5], the authors also analyze the generalization capacity to illustrate the learned parameter of the ranking machine shall be similar for similar actions. Since the parameter vector u captures the ordering of frames, it will be an ideal video representation.

3. Hierarchical VideoDarwin

Although VideoDarwin shows promising results for several action recognition tasks (generic action recognition, fine-grained actions recognition and gestures recognition), it has some shortcomings. Firstly, a single ranking machine could hardly capture the global ordering of long sequences. Secondly, if the sequence has large appearance changes in the frame level, such as shot transition, since there is no obvious appearance ordering constraints for video frames coming from different shots, the model would fail to capture the real evolution of appearance within the video.

The motivation of HiVideoDarwin is to model high-level and global motion evolution of long video with large appearance changes. The pipeline of HiVideoDarwin is shown in Figure 1. In this paper, we construct HiVideoDarwin with two layers. It can be easily generalized to more layers. The details of the two layer HiVideoDarwin are as follows.

3.1. Model Local Appearance Evolution

Before modeling video appearance evolution, we first extract and preprocess frame features in the same way as [5]. Let $V = [v_1, v_2, \cdots, v_n]$ be the frame descriptors, where the video is composed of n frames. We divide the long video sequence into s overlapping clips. Assuming each clip has the equal length of m frames, and overlapping frame number is h. Let define r as the overlap rate, which is $r = h/m$. The following equation satisfies:

$$m(s + r - s \cdot r) = n \tag{2}$$

The only two independent parameters for our model are r and s, and we conduct sensitivity analysis for them in the final part of experiment.

We apply different ranking machines for each video clip, and use the parameters of these models as clip motion descriptors. This layer can well capture the local appearance evolution for long video sequences.

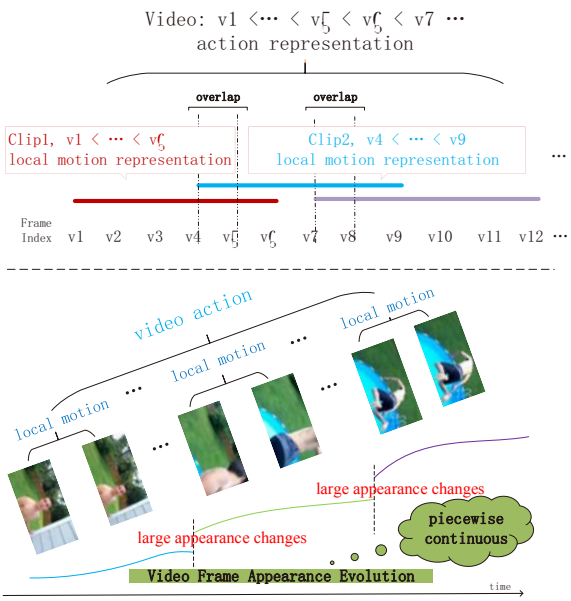

Figure 2. Top: A video is divided into overlapping clips. HiVideoDarwin first models frames ordering of clips to get local motion representation, then models clips ordering of the video to get action representation. Bottom: Action is composed of a series of ordered motions. For videos of large appearance changes, it will result in incorrect action representation if a single ranking machine is performed over all the frames since the order does not exist at appearance level.

3.2. Model Global Motion Evolution

As the parameters of the ranking machines in the first layer keep the ordering of frames for local clips, they can model the appearance evolution and serve as motion descriptors of these clips. Based on the observation that a certain action can be decomposed into a series of ordered motions, the relative ordering constraints exist for these subclips. In order to model global motion evolution of the video, we directly use these clip representations as inputs of the ranking machine in the next layer. This layer models the global motion evolution of action (see Figure 2). It can preserve semantic ordering of local clips even when the frame appearance changes greatly.

3.3. Computational Complexity

As illustrated in [1] that when a function can be compactly represented by a hierarchical architecture, it might need a very large architecture to be represented by a shallow one. Let us use the pairwise learning to rank scheme to discuss the computational complexity. VideoDarwin optimizes an objective defined over n^2 possible pairs for data with n examples (e.g., a video of n frames). HiVideoDarwin computes $s \cdot m^2 + s^2$ pairs. According to (2), this formula can be simplified as $s \cdot m^2 + s^2 \approx s \cdot m^2 \approx \frac{n^2}{s(1-r)^2}$, e.g., when $s = 0.2$, the computational burden is almost reduced to $\frac{2}{s}$ of the original.

4. Experiment

Now we evaluate the performance of HiVideoDarwin. The default parameters are $s = 5$ and $r = 0.2$. We use the non-linear forward and backward model (NL-RFDVD) in HiVideoDarwin, compare the best VideoDarwin model in [5] and other state-of-the-art methods. In addition, evaluations on the parameters are conducted.

4.1. Datasets

The MPII cooking activities dataset [18] was created for fine-grained action classification. It recorded 12 participants performing 65 different cooking activities with a total length of more than 8 hours. We use the bag-of-words histograms features (HOG, HOF, MBH and TRJ) provided by [18]. To compute HiVideoDarwin, we compute χ^2-kernel maps on histograms in the same way as [5]. Multi-class precision and recall and per class mean average precision (mAP) are computed using the same procedure as in [18].

The ChaLearn 2013 Gesture dataset [4] contains 23 hours of Kinect data of 27 persons performing 20 Italian gestures. The data is split into train, validation and test sets, with in total 955 videos each lasting 1-2min and containing 8-20 non-continuous gestures. For each frame we estimate the body joints using [19] to preprocess these data and extract frame descriptors in the same way as [5]. We report precision, recall, F1-score and mAP on the validation set, as done in [17, 28].

4.2. Experimental Results

The results of MPII cooking are shown in Table 1, from which we can see that HiVideoDarwin is the best. It surpasses VideoDarwin by $3.4\%, 6.7\%, 3.2\%$ in terms of multi-class precision and recall, per class mean average precision, respectively. The results verify our expectation that better action representation would be achieved by using a hierarchical structure to consider both local and global orders of video frames.

The results of ChaLearn are shown in Table 2. Although HiVideoDarwin outperforms the current state-of-the-art methods, the improvement is not significant when compared with original VideoDarwin. This can be explained that for this dataset, the number of frames for action clips is small, i.e., 40 frames in average, and VideoDarwin is capable of the ordering of such short sequence. In addition, the body joints that we use as frame descriptors are relatively stable temporally while our model is more suitable for features with sharp appearance changes.

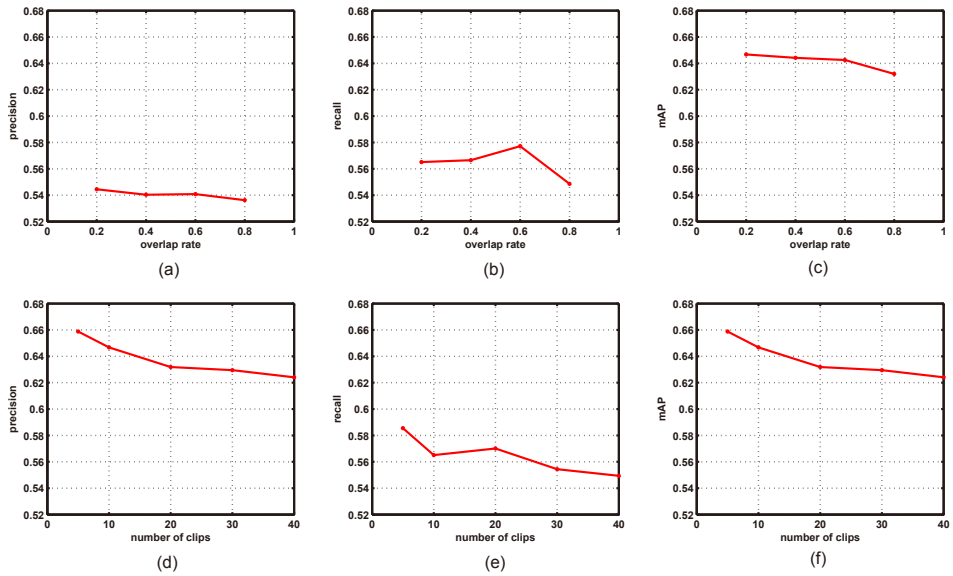

Figure 3. Evaluation of the number of clips and the overlap rate. For (a),(b),(c), the number of clips is fixed with 10, when we examine the influence of the overlap rate. For (d),(e),(f), the overlap rate is fixed with 0.2 when we examine the influences of the number of clips.

Approach	Multi-class		Per class
	Precision	Recall	mAP
Local Pooling [18]	49.4%	44.8%	59.2%
VideoDarwin [5]	50.8%	51.9%	62.7%
HiVideoDarwin	**54.2**%	**58.6**%	**65.9**%

Table 1. Comparison of HiVideoDarwin with VideoDarwin and Dense Trajectory on MPII cooking dataset.

Approach	Precision	Recall	F-score
Pfister et al. [17]	61.2%	62.3%	61.7%
Yao et al. [28]	-	-	56.0%
Wu et al. [26]	59.9%	59.3%	59.6%
VideoDarwin [5]	74.0%	73.8%	73.9%
HiVideoDarwin	**74.9**%	**75.6**%	**74.6**%

Table 2. Comparison of HiVideoDarwin with the state-of-the-art methods on ChaLearn gesture recognition dataset.

	Precision(%)		Recall(%)		mAP(%)	
	mean	std	mean	std	mean	std
$s = 10$	54.0	0.34	56.4	1.18	64.1	0.65
$r = 0.2$	53.9	0.42	56.5	1.42	63.8	1.42

Table 3. Statistical analysis for parameters. Note that std is short for standard deviation. In the first row we fix the number of clips with 10 and vary the overlap rate, in the second row we fix the overlap rate with 0.2 and vary the number of clips.

4.3. Parameter Evaluation

In this section, we evaluate the impact of the number of clips s and the overlap rate r on the performance by reporting multi-class precision and recall and per class mAP of different model parameters for the MPII cooking dataset. Evaluation is carried out for one parameter at a time. We evaluate the overlap rate from 0.2 to 0.8, and the number of clips from 5 to 40, the results are shown in Figure 3.

In Figure 3, we can see that generally, the performances increase with a lower value of both the overlap rate and the number of clips. For example, with $s = 5, r = 0.2$, the performance is the highest in our settings. The influence of the number of clips can be explained by some observed phenomenons, i.e., a simple action can be decomposed of several (e.g., four or five) motions. For the overlap rate, a small value of 0.2 is sufficient to establish connection between adjacent clips, and more overlap will cause redundancy thus reducing the performance.

Interestingly, recall curve has a local maxima with r $\approx$ 0.6. On one hand, the more overlap, the more similar subclips, which can lead to more similar video representations of the same action. On the other hand, as discussed above, the large overlap decreases accuracy, so there should be a balance between the two factors.

The mean and standard deviation of the evaluations are shown in Table 3, with average performance higher than current state-of-the-art methods and low standard deviation values, our model is insensitive to parameter changes, which reflects the robustness of HiVideoDarwin.

5. Conclusions

As action is a series of temporal motions, long sequence analysis is increasingly attracting researcher's attention. Long Short Term Memory (LSTM) networks show promising results for sequences of very long time lags of a unknown size [7], and are beginning to rise in action recognition [27, 15, 3]. However, these methods need a huge number of labeled videos and are computationally expensive and difficult to train. VideoDarwin regards the frames sequence as an ordered list and use a ranking machine to learn video representation. It shows good results on some datasets and is fast to compute, but it has limitations to process long sequences with large appearance changes. To address this problem, in this paper we have proposed a hierarchical VideoDarwin method, short for HVideoDarwin, which can model both local appearance evolution for clips and global motion evolution for video. Extensive experiments on two datasets indicate the effectiveness of HVideoDarwin. We will combine this method with LSTM networks in the future.

Acknowledgement

This work is jointly supported by National Natural Science Foundation of China (61420106015, 61175003, 61202328, 61572504) and National Basic Research Program of China (2012CB316300).

References

[1] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1), 2009. 3

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *TASLP*, 20(1), 2012. 1

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv*, 2014. 1, 5

[4] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ICMI*. ACM, 2013. 3

[5] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, volume 2, 2015. 1, 2, 3, 4

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*. IEEE, 2014. 1

[7] A. Graves et al. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, 2012. 1, 5

[8] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*. IEEE, 2013. 1

[9] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *TPAMI, IEEE Transactions on*, 35(1), 2013. 1

[10] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *ECCV*. Springer, 2012. 1

[11] T. Joachims. Training linear svms in linear time. In *ICKDD*. ACM, 2006. 2

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*. IEEE, 2014. 1

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[14] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3), 2009. 2

[15] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *arXiv*, 2015. 1, 5

[16] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *arXiv*, 2014. 1

[17] T. Pfister, J. Charles, and A. Zisserman. Domain-adaptive discriminative one-shot learning of gestures. In *ECCV*. Springer, 2014. 3, 4

[18] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*. IEEE, 2012. 3, 4

[19] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1), 2013. 3

[20] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 1

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv*, 2014. 1

[23] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*. IEEE, 2011. 1

[24] H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *arXiv*, 2015. 1

[25] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*. IEEE, 2013. 1

[26] J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In *ICMI*. ACM, 2013. 4

[27] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. *arXiv*, 2015. 5

[28] A. Yao, L. Van Gool, and P. Kohli. Gesture recognition portfolios for personalization. In *CVPR*. IEEE, 2014. 3, 4