A starting point for real-time human action detection

Yu Liu, Fan Yang, Dominique Ginhac

A starting point for real-time human action detection

Yu LIU, Fan YANG, Dominique GINHAC

Laboratoire ImViA, EA7535, Univ. Bourgogne Franche-Comté, Dijon, France

yu_liu@etu.u-bourgogne.fr, fanyang@u-bourgogne.fr
dominique.ginhac@ubfc.fr

Résumé – Analyser les actions humaines dans des séquences vidéo implique la compréhension du contexte spatial et temporel de la scène. Les réseaux de neurones convolutifs (CNNs) montrent des performances impressionnantes dans ce domaine. Cependant, la plupart des méthodes existantes fonctionnent hors ligne, non en temps réel et ne sont pas adaptées aux scénarios réalistes comme la conduite autonome et la surveillance publique. De plus, elles sont souvent trop gourmandes en consommation d'énergie pour être implémentées sur des systèmes embarqués. Dans ce papier, nous traçons d'abord un état de l'art des méthodes de détection des actions basées sur les CNNs. Puis nous proposons une chaîne de traitement rapide grâce à la propagation des caractéristiques d'apparence en utilisant les flux optiques. Notre approche est testée sur la base de données publique UCF-101-24. Les résultats expérimentaux obtenus valident son utilisation pour la détection d'actions en temps réel (40 fps).

Abstract – Analyzing videos of human actions involves understanding the spatial and temporal context of the scenes. State-of-the-art approaches have demonstrated impressive results using Convolution Neural Networks (CNNs). However, most of them operate in a non-real-time, offline fashion and are not well-equipped for many emerging real-world scenarios, such as autonomous driving and public surveillance. In addition, they are computationally demanding to be deployed on devices with limited power resources (e.g., embedded systems). This paper reviews state-of-the-art methods based on CNN for human action detection and related topics. Following that, we propose an initial framework to efficiently address action detection using flow-guided appearance features. We validate its performance on the UCF-101-24 dataset, and show that the method can achieve real-time action detection with a processing speed of 40 fps.

1 Introduction

Human action detection is a key element to video understanding. It has been an active research topic driven by many applications, such as assisted or autonomous driving, unmanned surveillance, and robot vision. These real-world scenarios often mandate not only on-site and real-time interpretation of scenes, but also robust recognition of events under restricted power budgets.

Moreover, applications such as surveillance in large environments and abnormal behavior detection in public, further demand having a network of cameras and exchange of information among local/central devices. The need to transmit and store redundant video streams imposes bottlenecks for effective analytic tasks. To manage such enormous data from multiple cameras without network overloads, efficient processing and extraction of relevant metadata at local devices become a fundamental system requirement. Instead of raw video streams, transmitting processed metadata between system components not only can minimize the content to be streamed, but also creates a smarter and cooperative framework.

With the recently rising Convolution Neural Network (CNN), object detection has progressed significantly with remarkable results. This motivates researchers to adopt CNN object detectors to action detection. To achieve spatio-temporal detection for action instances, existing approaches often link frame-level detections over time to create spatio-temporal tubes [1][2][3][4]. Handling every video frame independently is however non-optimal as the temporal continuity of videos is not fully exploited. On the one hand, distinguishing actions from a single frame can be ambiguous. On the other hand, neglecting the content similarity between successive frames imposes high processing cost and redundancy.

Our work focuses on computationally inexpensive human action detection potentially for embedded vision systems. In this paper, we first review state-of-the-art methods on action detection and related topics. We then describe our method which exploits video frames' continuity to save computation, and demonstrate its validity in the experiments.

2 Related work

Thanks to their remarkable results on object detection in images, CNN object detectors have been increasingly adopted for video action detection. This section briefly reviews recent works on both object and action detection.

Modern CNN object detectors can be grouped into two families. The first one uses a two-stage approach, first proposing object regions from images, and then performing classification and bounding box regression for each region [5]. Alternatively, YOLO [6] and SSD [7] directly classify and regress on a set of pre-defined boxes in a single pass. In exchange for minor drops in accuracy, these single-shot methods can achieve real-time

detection.

Extended from the image domain, video object detection has also been explored. Many existing methods link frame-level object boxes of consecutive frames into tubelets as a post processing operation [8][9]. Such an approach typically does not concern efficient processing. On the other hand, recently Zhu et al. [10] incorporate FlowNet [11] to propagate deep feature maps from sparse key frames to nearby non-key frames via flow fields. This accelerates video object detection as only a small number of key frames needs to go through the time-consuming deep feature extractor. In a similar spirit, Liu et al. [13] propagate frame-level information across frames using a recurrent convolutional architecture to enable near-real-time video object detection on low-powered mobile devices.

Concerning video action detection, many recent approaches rely on object detectors trained on action data. Furthermore, a popular way to capture actions' temporal information is the adoption of the two-stream framework [14], which performs detection on the appearance and motion stream separately followed by fusion and offline tubelet generation [1][4]. Others have also explored the use of multi-stream frameworks which take into account additional modalities such as human poses or semantics (e.g., objects) [15][16].

Targeting more realistic user scenarios, Singh et al. [4] achieve action detection in a real-time, online manner by combining the two-stream framework, SSD detectors, a fast optical flow estimator and their proposed online linking algorithm. Instead of making detection at the frame level, Kalogeiton et al. [2] propose an action tubelet detector, which learns to directly output sequences of action bounding boxes and scores.

3 Methodology

Typically in videos, image content varies slowly over consecutive frames. This phenomenon is also reflected in the corresponding CNN feature maps which encode high level semantics. This observation suggests that applying the complete CNN feature extraction for every video frame could be costly and redundant. As a starting point toward efficient action detection, we exploit neighboring frames' coherence to reduce computation. Different from the popular two-stream approach which explicitly uses motions as a separate stream, we make use of motions to efficiently guide the appearance features of key frames to their neighboring frames. This technique was applied by Zhu et al. [10] for video object detection, from which we hypothesize that action detection can also benefit.

Fig. 1 illustrates the proposed framework. During inference the deep and more expensive feature extraction network only runs on sparse key frames. Instead of being extracted from the feature network again, the feature maps of successive non-key frames are propagated from those of their preceding key frames. This is achieved by spatial warping for all locations and channels in the feature maps using optical flows. The flow fields are estimated by the corresponding pair of key and non-key frame.

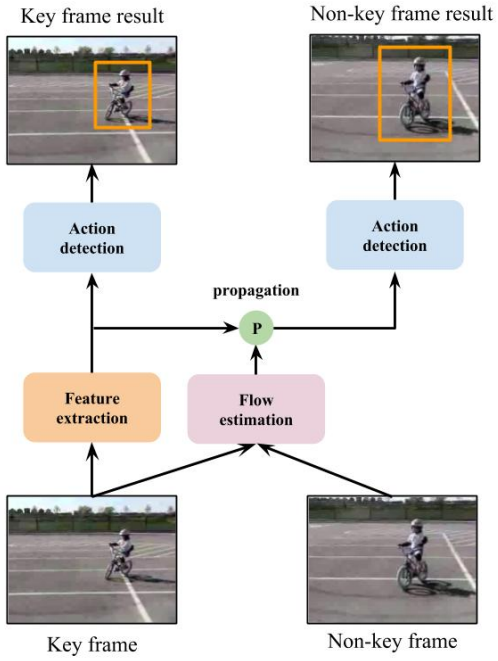


FIGURE 1 – Illustration of the flow-guided action detection framework based on [10].

Most existing two-stream approaches prepare motion streams pre-computed by traditional optical flow estimation methods. This incurs high consumption of time and storage, prohibiting online operations. We therefore adopt the framework of Zhu et al. [10] which integrates fine-tuning of flow estimation by a CNN network jointly with the feature extraction and detection networks. Computation reduction can be achieved as CNN flow estimation and feature map propagation are fast and inexpensive compared to CNN deep feature extraction, which in our framework is only used on a sparse set of key frames.

4 Experiments

We evaluate the proposed action detection framework on the UCF-101-24 [17] dataset. It is a subset of UCF-101 which is composed of realistic action videos across 101 action classes from YouTube. The UCF-101-24 consists of 24 classes in 3207 videos with frame-level localization annotations. We follow the work of Singh et al. [4], using 2290 of these videos for training and the remaining ones for testing.

We employ ResNet-101 [12], R-FCN [5] and FlowNet models for CNN feature extraction, action detection and flow estimation respectively. The entire system consisting of these sub-networks is trained end-to-end. During training, from each mini-batch a pair of nearby (a maximum offset of 9 frames) video frames, I_r and I_i, is randomly sampled, one being the reference frame. The appearance feature map f_r is first obtained from the reference frame, while both frames are fed to FlowNet to estimate the flow field. The estimated flow is then used to

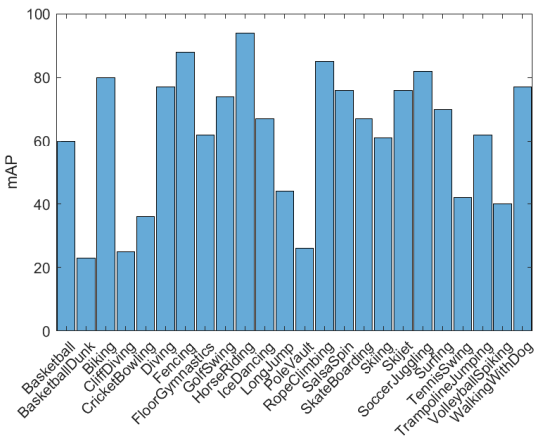


FIGURE 2 – Detection results (mAP) of individual action classes.

	Flow-guided	Baseline
Runtime (s) / frame	0.022	0.057
Accuracy (mAP)	62.2	67.1

TABLE 1 – Performance comparison between flow-guided and baseline method. The above runtime is reported as the mean speed over 10 frames.

propagate f_r to f_i, which will be the final feature map inputted in the detection network. The incurred localization and classification losses are then back-propagated to update all components of all sub-networks. Here, we use ResNet with ImageNet pre-training. FlowNet is pre-trained on the Flying Chair dataset [11]. The choice of individual components and training hyper-parameters are referred from the paper and released code of Zhu et al. [10].

In our experiment, we sample every 10 frame as a key frame during inference. In both training and inference, the size of an image is re-scaled to 600×800. The mean Average Precision (mAP) over Intersection over Union (IoU) at 0.5 is used as the evaluation metric. Predicted actions are considered correct only when the associated classes are correct and their bounding boxes reach the specified IoU with groudtruths.

The trained model achieves an average of 62.2% mAP over all action classes. Performances of individual classes are reported in Fig. 2. In Fig. 3 we display some action localization results at both key and non-key frames.

To assess the performance of the flow-guided approach, we compare it with a baseline method without guided features. In the baseline method, all frames are treated as key frames and go through feature extraction independently. Table 1 summarizes the comparison between the two in terms of accuracy and runtime. Overall, the flow-guided version demonstrates an average of 2.6 times speedup compared to the baseline with minor drops in accuracy. All experiments are conducted on an Nvidia GeForce GTX 1080 Ti GPU.

Finally, we conduct qualitative analysis on individual classes, especially over the ones which perform significantly worse than others. Poor performances may be associated with ambiguous context when learning from the appearance stream. For example, the action Basketball was often mis-classified as TennisSwing possibly due to the similarity of the court (Fig. 4a). Likewise, action RopeClimbing is initially classified as FloorGymnastics until the emergence of a clear rope (Fig. 4b). On the other hand, the action Fencing consistently performs well due to having unique appearances (i.e., white gears) that would not be confused with other action classes.

5 Conclusions and future works

In this paper we propose to adapt a recent work in video object detection to efficient video action detection. Our experiments demonstrate that this flow-guided method could enhance detection runtime by approximately three times, achieving real-time performance (40 fps) without losing significant accuracy. This validates our hypothesis that action detection could also benefit from exploiting the continuity between video frames, even though actions are conceptually more sensitive to temporal variations than objects.

As this work serves as our starting point for real-time action detection, we lay out rigorous research plans to follow. A key research direction is to incorporate the multi-stream framework into our existing work. We will explore using more modalities and feature aggregation techniques across frames to capture more temporal information. Our method, in terms of computational efficiency, may further benefit from smarter region proposal algorithms that attend to human presence in early frames. We believe incorporating the above tasks in hand will lead to a robust and efficient action detection solution suitable for embedded devices.

Acknowledgement

This work was supported by the H2020 Innovative Training Network (ITN) project ACHIEVE (H2020-MSCA-ITN-2017 : agreement no. 765866).

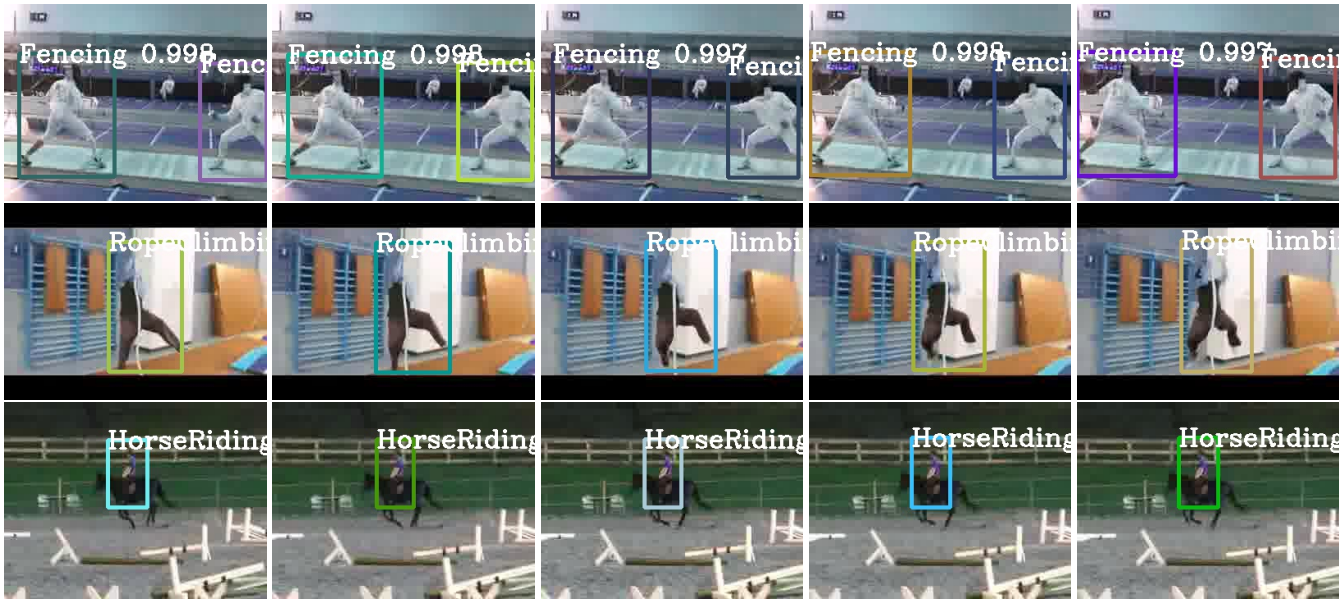
Références

[1] Saha, S., Singh, G., et al., *Deep learning for detecting multiple space-time action tubes in videos*, BMCV, 2016.

[2] Kalogeiton, V., Weinzaepfel, P., et al., *Action tubelet detector for spatio-temporal action localization*, IEEE ICCV, 2017.

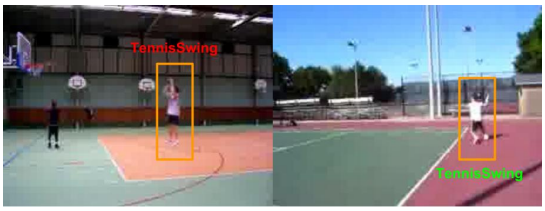
[3] Peng, X. and Schmid, C., *Multi-region two-stream R-CNN for action detection*, ECCV, 2016.

[4] Singh, G., Saha, S., et al., *Online real-time multiple spatiotemporal action localisation and prediction*, IEEE ICCV, 2017.

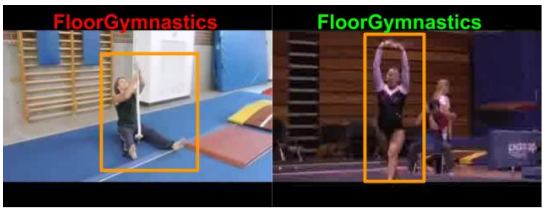


k k+2 k+4 k+6 k+8

FIGURE 3 – Flow-guided action detection results on UCF-101-24 dataset. The first column corresponds to detection results on key frames. The other four columns correspond to results of the following frames obtained with the propagated features.



(a) Incorrect detection



(b) Incorrect detection

(c) Correct detection

FIGURE 4 – Flow-guided action detection results on UCF-101-24 dataset. (a) both images exhibit similar context, but the correct classification of the left image is "Basketball". Likewise in (b), both images share similar context, but the correct action of the left image is "RopeClimbing".

[5] Dai, J., Li, Y., et al., *R-FCN : Object detection via region-based fully convolutional networks*, Journal NIPS, 2016.

[6] Redmon, J., Divvala, S., et al., *You only look once : Unified, real-time object detection*, IEEE CVPR, 2016.

[7] Liu, W., Anguelov, D., et al., *SSD : Single shot multibox detector*, ECCV, 2016.

[8] Han, W., Khorrami, P.,et al., *Seq-NMS for video object detection*, Journal CoRR, 2016.

[9] Kang, K., Li, H., et al., *T-CNN : Tubelets with convolutional neural networks for object detection from videos*, IEEE Trans. On TCSVT, 2018.

[10] Zhu, X., Xiong, Y., et al., *Deep feature flow for video recognition*, IEEE CVPR, 2017.

[11] Dosovitskiy, A., Fischer, P.,et al., *FlowNet : Learning optical flow with convolutional networks*, IEEE ICCV, 2015.

[12] He, K., Zhang, X., et al., *Deep residual learning for image recognition*, IEEE CVPR, 2016.

[13] Liu, M. and Zhu, M., *Mobile video object detection with temporally-aware feature maps*, IEEE CVPR, 2018.

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

[15] Zolfaghari, M., Oliveira, G.L., et al., *Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection*, IEEE ICCV, 2017.

[16] Wang, Y., Song, J., et al., *Two-stream SR-CNNs for action recognition in videos*, BMVC, 2016.

[17] Soomro, K., Zamir, A.R. and Shah, M., *UCF101 : A dataset of 101 human actions classes from videos in the wild*, Journal CoRR, 2012.