2018 IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin).

Activity monitoring from RGB input for indoor action recognition systems

Gianluigi Ciocca, Alessio Elmi, Paolo Napoletano, and Raimondo Schettini
University of Milano-Bicocca
DISCo - Department of Informatics, Systems and Communication
Viale Sarca 336, 20126 - Milano, Italy
{ciocca,napoletano,schettini}@disco.unimib.it, alessio.elmi@gmail.com

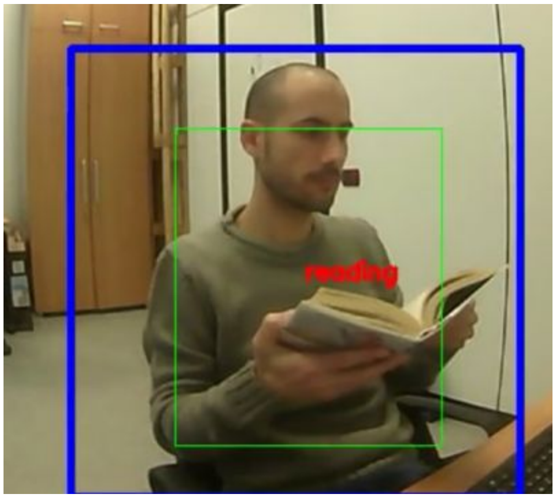

Fig. 1. Our system correctly classified reading action.

***Abstract*—In this work we present how some state-of-the-art action recognition techniques can be tailored to monitoring indoor human activities. At first, we analyze the most relevant algorithms and eventually we present an effective implementation based on the solely RGB input. Our preliminary results show that the proposed solution achieves competitive results with respect to other methods in the state of the art that take also advantage of multiple inputs.**

***Index Terms*—Action recognition, tracking.**

I. Introduction

Action recognition is one of the most challenging topic in computer vision. Its use-cases are numerous, ranging from security surveillance to elderly care, from augmented reality to sports analysis. Despite several applications, a general purpose solution is still considered far from being achieved. The main reason is that the definition of action itself embraces a very large number of attributes, that are hardly modeled in a single universal framework.

Different approaches for action recognition have been proposed, readers can refer to [1] for a complete overview. Recently, deep learning techniques used in the image domain have been adapted for action recognition. Simonyan et al. [2] showed the benefit of simultaneously training single frames and optical flow clips. Du Tran et al. [3] extended convolutional networks to the spatio-temporal domain with C3D. More recently, other newer image-based techniques have been translated to video classification, like inception modules [4] and residual blocks [5]. Apart from convolutional approaches, other authors have investigated the benefit of modelling the body pose and its evolution [6], [7].

Action recognition can be employed in domestic contexts, like smart home applications - where some specific actions or gestures trigger home automation technologies - or tracking elderly activities in order to spot anomalies in their behavior or monitor their activities.

Focusing on this last use case, we designed an activity monitor based exclusively on the RGB input, which looks for human subjects in the scene and tracks the activities for each of them (e.g. see Fig. 1). In the following sections we will illustrate the basic principles and we show that deep convolutional networks can be easily fine-tuned for finer-grained domain. To this end we conducted several experiments on the popular NTU RGB+D activity dataset [8].

II. Action classification pipeline

We adopted a classical top-down approach, where independent classifications are performed for every person spotted in the scene.

- **Frame packet**. Starting from a wide-angle single RGB camera, we continuously record a 32 frames circular buffer of a full resolution RGB tensor (32x640x480x3 with our equipment).
- **Pose estimation**. We select five frames per second and we individually feed them into a pose estimation network. This is a lighter version of [9], where a Mobilenet [10] is trained to detect joints and limbs, that are later grouped together to form human candidates.
- **Track management**. Every detected person will be assigned to some existing tracks or alternatively they will start a new one. For person re-identification in the video sequence we tested both associations based on euclidean distance between candidate centers and chi-squared histogram distance between person 2D bounding boxes. Assignment operations are computed based on the Hungarian algorithm [11]. Tracks that are no longer

978-1-5386-6095-9/18/$31.00 ©2018 IEEE

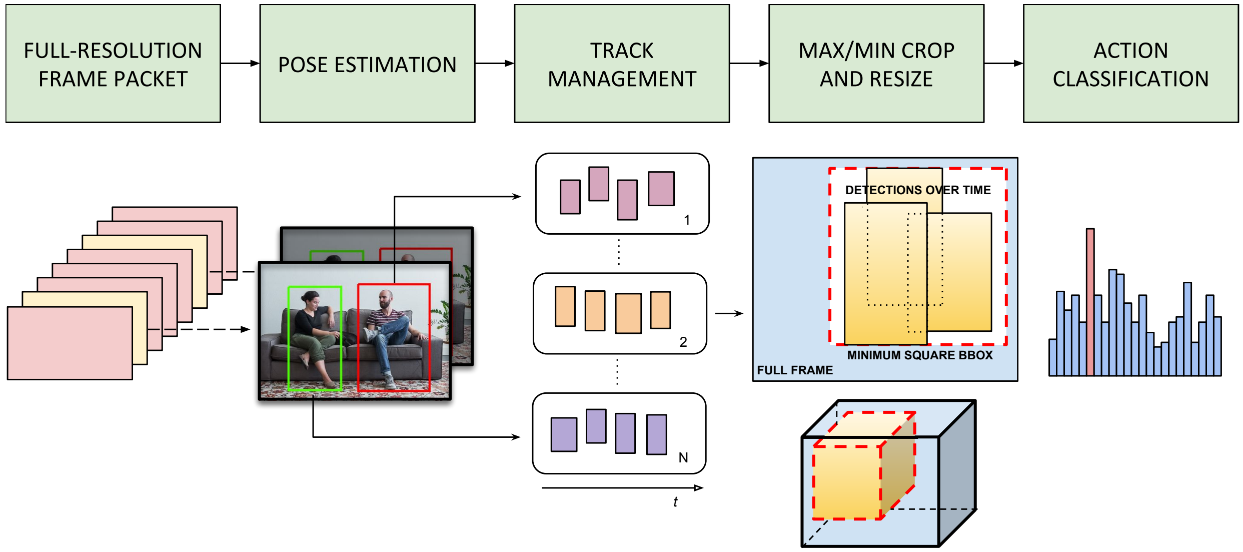

Fig. 2. Block diagram of the classification pipeline.

visible (i.e. they are not associated to any detection) for at least two detection intervals are discarded. At this point we have a finite number of tracks, each of them with a set of pose coordinates detected over time and these coordinates are continuously updated in a FIFO fashion.

- **3D bounding box**. For each track we compute its smallest bounding box, as the maximum of its coordinates over time. Bounding boxes are also adjusted based on some skeletal ratios, so that the subjects look equally big inside the box (otherwise seated subjects will have a tighter bounding box compared to those that are jumping). Every final bounding box will be used to slice the full-resolution buffered clip into different cropped 3D bounding box, that are finally scaled to a fixed dimension of 32x224x224x3.
- **Action classification**. Each 3D bounding box is then classified independently, through a deep spatio-temporal CNN. To accomplish this we fine-tuned the I3D network [4], which architecture is composed of two parallel - and independent - identical branches, one for RGB input, the other for the optical flow. We found out that the RGB branch alone can produce good results if properly trained, relieving us of the optical flow computation, which can be quite expensive from the computational point of view. Track predictions are averaged over time and if the dominant class is greater than a certain threshold the classification is considered reliable and therefore stored in the logging system.

A block-diagram of the entire pipeline is reported in Fig. 2.

III. Experiments

To verify if I3D architecture can be also applied in more compact and fine-grained domains, we conducted a set of fine-tuning experiments, starting from a pre-trained model trained on the Kinetics [12] dataset. A pose estimator network is placed before the I3D in order to get a crop of the people in the scene, in the same fashion presented in Section II.

We chose NTU RGB+D [8] as a workbench dataset, which comprises more than 50 thousands videos, covering 60 actions performed by 40 subjects. To the best of our knowledge it is the biggest dataset including a set of activities that better resemble our scenario. There are full-body actions (like *sitting down*, *jumping*, *falling*), some fine-grained gestures (*putting the palms together* that needs to be classified from *rubbing two hands*), human-object interaction (*reading* and *writing*) and people interaction (*stealing from someone's pocket* or *handshaking*). Even though skeleton coordinates, depth and infrared images are provided, only RGB sequences have been used here. We performed cross-subject experiments, adopting the usual train/test split provided by the original authors. In particular, we used the entire training video sequences, randomly and temporally segmented, during the training phase, while in the testing phase we only considered centered clips (potentially padded with the last frame in case of too few frames are available) and classified them accordingly. Some augmentation techniques have proved to be helpful, like horizontal flips or random shifting of the people bounding box.

IV. Result

Several experiments have been conducted, where we tried to variate the number of frames in the temporal dimension, going from 16 to 32 ($\sim$1 second), and the number of deepest layers to be fine-tuned (from *Logits* to *Mixed_5b*, using authors nomenclature). Table I reports the results of our most relevant experiments along with results of state-of-the-art algorithms on the same dataset. The highest accuracy on the test set has been obtained by increasing the temporal dimension and the number of trained layers ("Mixed5b Logits, 32 frames"), getting a value of 78.15%. These results are not far those obtained by the other works reported in the Table I, even

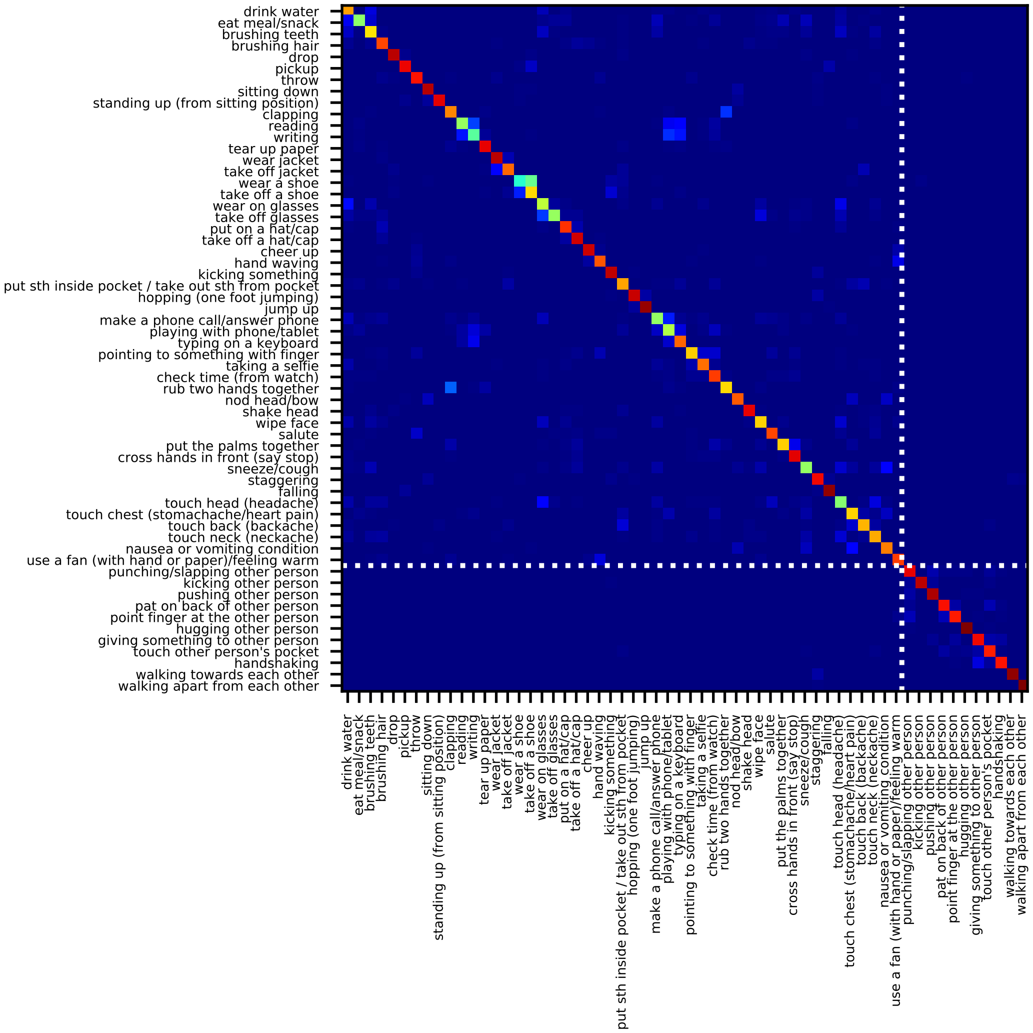

Fig. 3. Confusion matrix for the experiment with highest accuracy. It can be seen that action involving more than one person (last 11 classes) are grouped together.

though our approach exploits only RGB information. Training only the last network layer is not sufficient to get remarkable results as confirmed by the accuracy of 49.62%. Fig. 3 shows the confusion matrix of the “Mixed5b Logits, 32 frames” experiment. We can see that analyzing the classification errors we can identify two macro-groups of actions: one group corresponds to single-person actions, while the other comprises interaction between two subjects. The two macro-groups are practically never misclassified between each others. Lowest classification accuracy is obtained for the following action pairs: *reading/writing*, *wear/take off shoes*, *playing/calling with phone*. Finally, in Fig. 4 we report the accuracy curves on the test set for our three recognition approaches during training. The curves show how “Mixed5b Logits, 32 frames” outperforms the other two approaches since the first few iterations.

TABLE I

SOME REPRESENTATIVE STATE-OF-THE-ART ACCURACIES ON NTU RGB+D DATASET FOR DIFFERENT MODALITIES

Method	Modality	Cross-Sub
SSSCA-SSLM (Shahroudy et al. 2017) [13]	RGB+D	74.86%
Two-stream CNN (Li et al., 2017b) [14]	Skel	83.2%
DDMNIs (Wang et al. 2018) [15]	Depth	87.08%
Logits, 32 frames (*ours*)	RGB	49.62%
Mixed_5b Logits, 16 frames (*ours*)	RGB	66.96%
Mixed_5b Logits, 32 frames (*ours*)	RGB	78.15%

V. CONCLUSION

In this work we have shown how state-of-the-art action recognition techniques can be effectively and efficiently tailored to classify indoor human activities. Using our approach made possible to obtain results in the state of the art using only RGB modality.

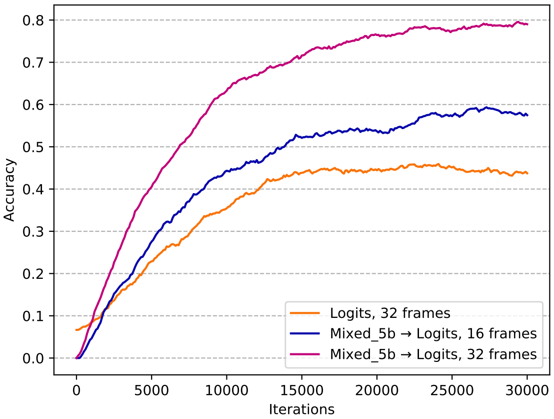

Fig. 4. Accuracy results for the most representative experiments. Accuracy is computed on clips randomly sampled over the entire test set. Final results are computed on the temporally centered clips.

Several improvements are in progress. First of all, it would be helpful to add a person re-identification system, which could assign detected track to a pool of known people (for example it could profile all family members), and therefore collect finer statistics for each of them. As a direct consequence of that, classification could be further improved by performing a per-subject fine-tuning. Another topic that would deserve a deeper attention is the anomaly detection. During our experiments we noticed that features extracted by our system can be easily modeled to detect unknown actions as outliers. Finally, we have already started testing this solution with low-end hardware, like the ones available in IoT contexts and smart home devices.

VI. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Jetson TX1 Kit and the Titan Xp GPU used for this research. This research leading to these results has received funding from The Home of Internet of Things (Home IoT), CUP: E47H16001380009 - Call "Linea R&S per Aggregazioni" cofunded by POR FESR 2014-2020.

References

[1] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

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

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[4] 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)*. IEEE, 2017, pp. 4724–4733.

[5] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[6] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[7] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data." in *AAAI*, vol. 1, no. 2, 2017, p. 7.

[8] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," *arXiv preprint arXiv:1604.02808*, 2016.

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[11] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 2, no. 1-2, pp. 83–97, 1955.

[12] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[13] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in rgb+ d videos," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[14] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 597–600.

[15] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based large-scale 3-d action recognition with convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1051–1061, 2018.