Anticipation of Human Actions with Pose-based Fine-grained Representations

Sebastian Agethen Hu-Cheng Lee Winston H. Hsu
National Taiwan University

Abstract

Anticipating an action that is about to happen allows us to be more efficient in interacting with our environment. However, prediction is a challenging task in computer vision, because videos are only partially available when a decision is to be made. Complicating the issue is that it is not always clear which of the visible activities in the scene are relevant to the action, and which ones are not. We suggest that the key to recognizing an action lies with the human actors, and that it is therefore necessary for the prediction process to attend to persons in a scene. In our work, we extract fine-grained features on visible human actors and predict the future via an L2-regression in feature space. This allows the regressed future feature to focus on the actor. Using this, the future action is classified. More specifically, the fine-grained extraction is guided by a pose prediction system that models current and future human poses in the scene. We run qualitative and quantitative experiments on the Charades dataset, and initial results show that our system improves action prediction.

1. Introduction

Action recognition is a popular field of research in computer vision, where upon inspection of data showing an activity, typically a video, a judgement in form of a classification is made. Oftentimes, it is however necessary to make a decision before the data relating to the action has become available for observation. In this scenario, we then need to solve an *action anticipation* [5, 8, 13] task. A plethora of applications rely on prediction systems. For example, an autonomously driving car needs to predict that a pedestrian is attempting to cross the road in order to be able to slow down in time. A different application is a household robot that anticipates its owner's next action in order to support him, or, at the very least, not hinder him.

The key challenge is that we are not able to observe the action itself. Fortunately, there are often common patterns that can be observed ahead of time. Previous work [13] attempted to find such patterns by unsupervised learning on large video datasets. However, without guidance as to what to look for, this approach has limited effectiveness, as large portions of the scenery are irrelevant to the activity. For the case of *human action anticipation*, which we tackle in this work, we suggest that such patterns can be found in the pose of visible persons, as well as in objects that a person interacts with. Hence, we can guide the prediction process by emphasizing on the people visible in a scene.

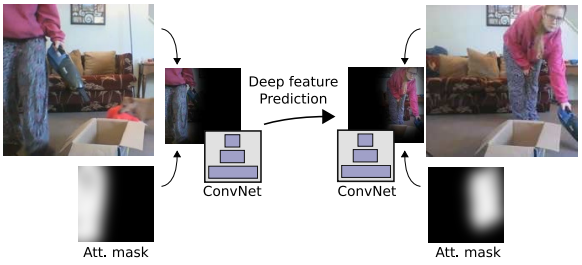


Figure 1: A person is about to vacuum the floor. Her pet dog is not relevant to the action. To focus the prediction process on the human actor, a pose-based attention mask is used to guide extraction of fine-grained representations from the image. The future representation is then learned by regression.

Similar to [13], we attempt to perform the prediction process inside the deep feature space in our work. This avoids unnecessary reconstruction of the future scenery pixel by pixel. Unlike the previous work however, we utilize the human pose as a means to learn an attention scheme. The attention scheme acts as a mask that removes information not relevant to any activity, both for the given past observation, as well as for the potential future. An input reduced in such way enables a more effective prediction process. As a side product, it also allows our proposed architecture to generate human poses for the as-of-yet unseen images in the future.

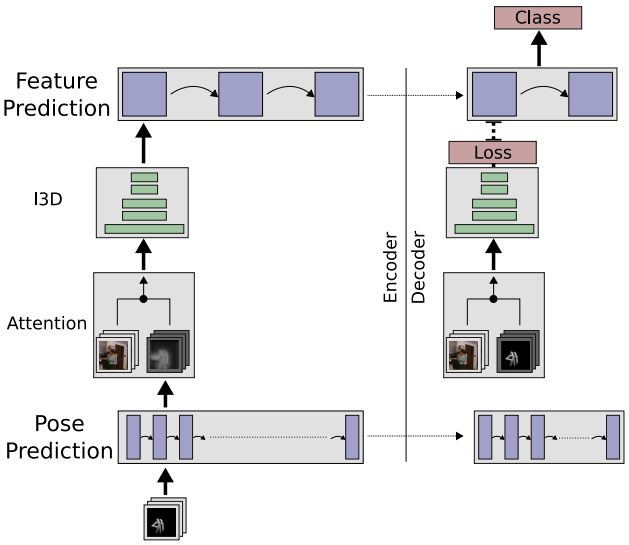


Figure 2: Our proposed system architecture. A *pose prediction* stream encodes past pose representations and predicts future representation with the help of a decoder. Reconstructed poses are used as masks that attend the input of a *feature prediction* stream.

2. Related Work

A number of methods [12, 15, 7, 14] attempt to pixelwise predict a future frame given an input of RGB frames. Srivastava *et al*. [12] made an initial attempt using an encoder-decoder LSTM architecture to predict a small number of frames. The work in [15] leverages human poses and then predicts a future frame in an adversarial fashion, showing some promising results.

Human pose has also successfully been used for Action Recognition [3, 4, 10]. In [3], pose-based features are created by extracting patches around human body parts, and aggregating them to learn a video descriptor. The work in [4] models human body parts via a more flexible pose attention mechanism. However, none attempt an action prediction scenario.

The work in [6] learns to anticipate human activities with the help of object affordances. However, such annotations are not available for most datasets.

Our work bears greatest similarity with that by Vondrick *et al*. in [13]. Instead of first predicting a high-dimensional RGB frame at the desired point in time in the future, they apply a deep regression network and predict the future in a much lower dimensional space, in their case the 4096-dimensional FC7 layers of AlexNet. The network is pretrained in an unsupervised fashion, and then annotated on a small-scale supervised dataset. However, we identify several flaws in their work: First, the deep feature extracted from AlexNet includes largely irrelevant information on the background scenery. Second, the features are extracted from a single past frame as well as a single future (groundtruth) frame. Any motion information that could help describe the action-to-be-predicted is lost.

3. Proposed Method

Our system, see Figure 2, is formed by two main components: A *pose prediction* stream and a *feature prediction* stream. We begin by describing the pose prediction stream.

3.1. Pose-based attention generation

The task of the pose prediction stream is two-fold: First, we can use it to predict the following poses of a person in the video. Based on these predicted poses, a classification may be made as well. Second, more importantly, we use the poses to generate meaningful attention masks, which will then be used to extract fine-grained features on the actor, see Sec. 3.2.

While one may attempt to learn the pose directly from RGB input, for our initial experiments, we make use of an existing pose detector to conclude where possible actors are located in the scene. In particular, we use the work by Cao *et al*. [1] to extract Part Affinity Fields (PAF). A PAF is a set of two-dimensional vectors that describe orientation of a body part at each location in an image. Given such a vector $\mathbf{v}_{x,y} = (v_1, v_2)$ of a PAF, we can consider $\|\mathbf{v}\|^2 = v_1^2 + v_2^2$ as a measure of whether a pixel (x, y) belongs to a body part or not. We use this body part map, visualized in the middle row of Figure 3, as input to our pose stream.

The first goal of the pose stream is to predict the future poses. To accomplish this, we implemented a ConvLSTM-based [9] encoder-decoder. An encoding ConvLSTM reads in the P-dimensional body part maps extracted on a video clip, where $P = 18$ in our case, and generates an intermediate state $(\mathbf{C}, \mathbf{H})$. The state is then used to initialize the decoding ConvLSTM, which will predict future poses. To approach our second goal, we can use the poses as a natural *attention mask* to remove background information from the observed (past) RGB video sequence.

3.2. Actor feature extraction

Our work proposes to improve the prediction step by attending to the main actors of an action. At the same time, changes in the background that do not contribute to the prediction process should be disregarded. It should be noted that background features can help describe the scene, and while no prediction is performed, a single (past) feature is extracted for classification.

In order to be able to attend to certain regions of a frame, we earlier described how we generate atten-

tion masks from the pose stream. These masks can be used in two different fashions: One can apply the mask on the RGB image itself, and then use a deep convolutional network to extract high-level features on that frame. Alternatively, we could also apply the attention mask at a later stage, for example on the last spatial layer of the convolutional stack.

In our ongoing work, we choose the 3D-convolutional network I3D [2] as a feature extractor. I3D is considered the state-of-the-art for action recognition on benchmarks such as UCF-101. We decide to apply the attention masks on the RGB image directly. This allows us to finetune the convolutional network in the hope that fine-grained, human-specific features can be learned. As the attention masks recovered from the pose stream have lower spatial resolution (in our case 28×28), we resize the masks in several steps up to a resolution of 224×224 using bilinear interpolation layers, each followed by a 1×1 convolution. The resizing operation supports gradient backpropagation.

3.3. Decoding future representation

In the previous section, we extracted fine-grained features of persons visible in the scene. These features describe the actors in the time ahead of the action-to-be-predicted, and we refer to them as past features. These past features are used as input to the *feature prediction* stream.

The prediction process in the feature prediction stream is again implemented by an encoder-decoder framework. An encoding recurrent network reads in the past features extracted previously on the observed video sequence of length T_{in}. The final state of the recurrent network is then copied, and used to initialize a decoding recurrent network, which does not take any input (or an all-zero input) and generates T_{out} output features. As prediction loss, we choose the L2-regression loss. As the outputs of recurrent networks often apply the hyperbolic tangens $\tanh(\cdot)$ as activation function, we add an additional 1×1 convolutional layer with ReLU activation in order to match the activation function used in I3D.

4. Evaluation

In the following, we first describe the used dataset, Charades [11], and then report our initial results.

4.1. Dataset: Charades

The Charades dataset consists of 9848 videos showing people acting out daily indoor activities. It is especially suited to our work, as it centers on human actions, which ensures that only a small fraction of frames lacks a visible pose. The are 157 activities, which are described by a verb-object structure, where the number of verbs and objects is restricted. There is a strong class imbalance in the dataset, and for our initial experiments we use a subset of seven classes[1] with large number of examples.

Each action in the dataset is annotated with two timestamps, the starting and end time of that action. For our experiments, we read in T_{in} frames ahead of the annotated starting time. In case that not enough frames exist, the last valid frame is repeated. For the future reference representation, we load T_{out} frames from the temporal center of the annotated action.

4.2. Quantitative results

For initial results, we run our system on a subset of seven Charades classes, and compare with the simpler feature regression approach in [13], marked as (i). We test three configurations: First, in the configurations marked as (ii) and (iv), we only input the body part maps generated from PAFs into the pose stream. In the second case, marked (iii), we additionally feed segmentation maps of a set of objects of interest. We choose 20 of the 80 annotated object classes in the CoCo dataset. Note that the objects only serve as input, i.e., we do not regress their future instances.

Table 1: Our initial results. The baseline configuration is based on [13], but now reads in a sequence of T_{in} frames instead of a single one. Our fine-grained prediction improves accuracy by 2% over the baseline.

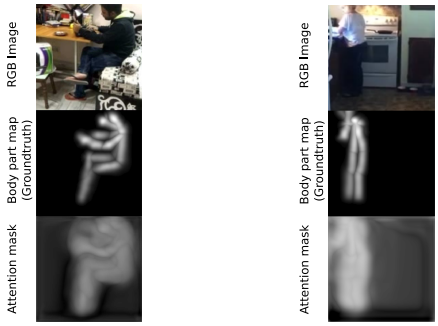
	Configuration		Accuracy
(i)	Vondrick *et al.* [13] +LSTM	Baseline	31.26%
(ii)	Fine-grained Pose +ConvLSTM	**Ours**	31.86%
(iii)	Fine-grained Pose+Obj +ConvLSTM	**Ours**	31.36%
(iv)	Fine-grained Pose +LSTM	**Ours**	**33.27%**

The results in Table 1 suggest that we can indeed gain some performance when attending to the main actors. Adding objects as additional input surprisingly had a detrimental effect on prediction accuracy. Our qualitative analysis reveals one reason for this.

4.3. Qualitative results

The pose stream reconstructs past poses, and predicts future poses. An exemplary reconstruction can be

[1] Putting something on a table, Drinking from a cup, Someone is going from standing to sitting, Someone is smiling, Someone is sneezing, Someone is standing up from somewhere, Someone is eating something



(a) (b)

Figure 3: (a) Attention mask generated from pose stream. (b) Objects fed to pose stream leave a trace in mask (contrast adjusted).

seen in Figure 3 (a). Figure 3 (b) reveals one cause for the reduced accuracy of configuration (iii): Although only the pose should be reconstructed, the mask also faintly shows objects in the scene, here a stove and a refrigerator. We conclude that relevant object information needs to be processed in a more elaborate manner.

5. Conclusion & Future Work

Human actors play an important role for human action anticipation. Our proposed system attends to visible persons in videos and extracts fine-grained features. We have shown that the use of such features can improve the prediction process. As a side product, our work also generates dense pose predictions.

Several open questions remain. Most importantly, the attention system is currently only using pose knowledge available from the pose prediction stream. However, some of the objects present in the scene, as well as the actors intention, expressed in particular by the gaze, may be useful information to generate a richer attention mask. Second, while the pose stream reconstructs past poses faithfully, we have not addressed future poses yet, which may be predicted by providing additional scene information.

Acknowledgement This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2634-F-002-004. We also benefit from NVIDIA grants and the DGX-1 AI Supercomputer. We are grateful to the National Center for High-performance Computing.

References

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

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. pages 4724–4733, 07 2017.

[3] G. Ch'eron, I. Laptev, and C. Schmid. P-CNN: Pose-based CNN Features for Action Recognition. In *ICCV*, 2015.

[4] W. Du, Y. Wang, and Y. Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3745–3754, 2017.

[5] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, Jan 2016.

[6] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, Jan 2016.

[7] M. Mathieu, C. Couprie, and Y. Lecun. Deep multi-scale video prediction beyond mean square error. 11 2015.

[8] M. Sadegh Aliakbarian, F. Sadat Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson. Encouraging lstms to anticipate actions very early. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 802–810, Cambridge, MA, USA, 2015. MIT Press.

[10] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. *CoRR*, abs/1805.02335, 2018.

[11] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016.

[12] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 843–852. JMLR.org, 2015.

[13] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106, 2016.

[14] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 613–621. Curran Associates, Inc., 2016.

[15] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision*, 2017.