Discrimination of locomotion direction at different speeds: A comparison between macaque monkeys and algorithms

Fabian Nater*[1], Joris Vangeneugden*[2], Helmut Grabner[1], Luc Van Gool[1,3], Rufin Vogels[2]

[1]Computer Vision Laboratory, ETH Zurich, Switzerland
[2]Laboratorium voor Neuro- en Psychofysiologie, K.U. Leuven, Belgium
[3]ESAT - PSI / IBBT, K.U. Leuven, Belgium

Abstract. Models for visual motion perception exist since some time in neurophysiology as well as computer vision. In this paper, we present a comparison between a behavioral study performed with macaque monkeys and the output of a computational model. The tasks include the discrimination between left and right walking directions and forward *vs.* backward walking. The goal is to measure generalization performance over different walking and running speeds. We show in which cases the results match, and discuss and interpret differences.

1 Introduction

Correctly recognizing biological motion is of utmost importance for the survival of all animals. The computer vision field has a long tradition in modeling human motion, especially of walking and running. Of those models, some recent ones have been inspired by neurophysiology. Human locomotion consists of motion patterns that involve different movements of all limbs and are therefore widely used to study visual motion perception on a psychophysical level but also for modeling with computational algorithms. Studies on action recognition in both fields suggest that human actions can be described using appearance/form or motion cues (*e.g.*, [1], [2]).

In this work, we set out to specifically investigate the perception of locomotion direction. While discrimination between right- and leftward walking is possible based on shape cues only (e.g. comparing momentary body poses), discriminating between forward and backward walking at least requires motion for a successful distinction [3]. A recent behavioral study in macaque monkeys investigated the perception of walking direction and how well these animals generalized from a trained categorization of walking to other walking speeds and running [4]. The question now arises how state-of-the-art methods in computer vision relate to these findings. More specifically, we focus on a recently proposed approach which encodes typical appearance and motion patterns in a hierarchical

*equally contributing first authors.

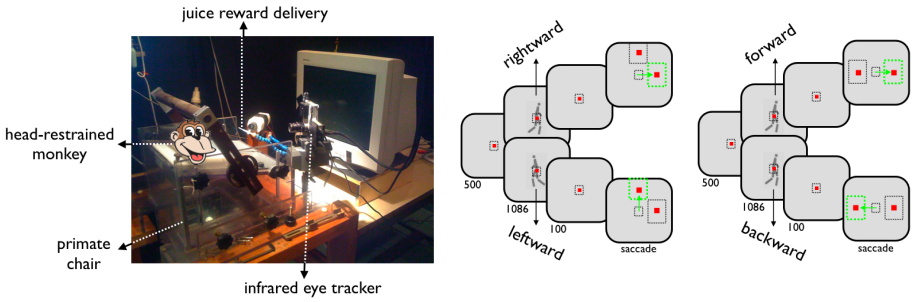
Work is funded by the EU Integrated Project DIRAC (IST-027787).

framework [5]. The two hierarchies of this model are intended to mirror the subdivision into "snapshot" and "motion" sensitive neurons that have been found in the brain of the macaque monkey, more specifically in the superior temporal sulcus within the temporal lobe [6]. While "snapshot" neurons encode for static body poses, "motion" neurons are driven by movement (i.e. kinematics). In the present paper we aim to compare the performances of this algorithm and of behavioral macaque responses regarding the visual coding of human locomotion at different walking and running speeds.

2 Behavioral study

2.1 Subjects and apparatus

Three rhesus monkeys (*Macaca mulatta*) served as subjects in this study. The heads of the monkeys were kept immobilized during the sessions (approx. 3h/day) in order to capture the position of one eye via an infrared tracking device (EyeLink II, SR Research, sampling rate 1000 Hz). Eye positions were sampled to assure that the subjects fixated the stimuli. In order to obtain a juice reward (operant conditioning), successful fixation, within a predefined window measuring $1.3° - 1.7°$, and a correct direct saccade towards one of the response targets were required, see Fig. 1.



(a) Setup (b) LR/RL task (c) FWD/BWD task

Fig. 1. Experimental setup for the behavioral study (a), illustration of the performed tasks (b,c). The gray fields indicate what was presented to the monkeys on the screen, with the respective durations indicated below each screenshot (times expressed in ms). The dotted rectangles around the red fixation point and the red response targets represent the windows in which eye positions had to remain prior to, during and after stimulus presentation (former case) or eye movements had to land, indicating the monkeys' decision (the latter case). Highlighted in green are the correct targets the monkeys had to saccade to in order to obtain a juice reward.

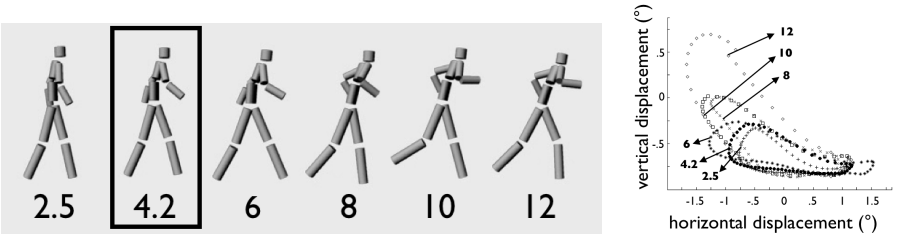
Each trial started with the presentation of a small red square at the center of the screen ($0.12°$ by $0.12°$). The subjects had to fixate this square for 500 ms, followed by the presentation of the stimulus (duration = 1086 ms; 65 frames at a 60 Hz frame rate). Before making a direct eye movement saccade to one of the two response targets, the monkeys had to fixate the small red square for another 100 ms. During the complete trial duration, monkeys had to maintain their eye position within the predefined window. Failure to do so resulted in a trial abort.

Response targets were located at 8.4° eccentricity, either on the right, left of upper part of the screen. The stimuli, described below, measured approximately 6° by 2.8° degrees (height/ width at the maximal lateral extension respectively).

All animal care, experimental and surgical protocols complied with national and European guidelines and were approved by the K.U. Leuven Ethical Committee for animal experiments.

2.2 Stimuli

Stimuli were generated by motion-capturing a male human adult of average physical constitution walking at 2.5, 4.2 or 6 or running at 8, 10 or 12 km/h. Specifications of the procedure can be found in [4]. We rendered enriched stimulus versions by connecting the joints by cylinder-like primitives, yielding *humanoid* renderings. Importantly, all stimuli were displayed resembling treadmill locomotion, *i.e.*, devoid of any extrinsic/translatory motion component, see Fig. 2.



(a) stimuli for the 6 speeds (b) ankle trajectories

Fig. 2. Stimuli presented in the behavioral and the computational experiments. In (a), snapshots/body poses of the 6 walking speeds are depicted with the training speed framed. (b) Ankle trajectories for the same speeds: with increasing speed, step size increases as well as vertical displacements grow.

2.3 Tasks and training

The three monkeys were extensively trained in discriminating between different locomotion categories. In a first task, they were instructed to discriminate between different facing directions (LR/RL task) when observing the stimulus (video) that shows a person that is either walking towards the right (LR_fwd) or towards the left (RL_fwd). The second task was designed to distinguish forward from backward locomotion (FWD/BWD task). In that case, the stimulus shows a person walking towards the right, but either forward (LR_fwd) or backward (LR_bwd). The LR_bwd condition was generated by playing the LR_fwd video in reverse. The start frames of the movie stimuli were randomized across trials to avoid that the animals responded to a particular pose occurring at a particular time in the movie. Training was done only at the 4.2 km/h walking speed.

Substantial training was needed for our monkeys to learn FWD/BWD discriminations, while LR/RL discrimination was made more easily (*cf.* [4]). *E.g.*, the number of trials required to reach 75% correct in a session for the LR/RL task was 1323 trials, while the same monkey required 37,238 trials to achieve a similar performance level in the FWD/BWD task (similar trends were observed in the other two subjects). Nevertheless, all three monkeys reached behavioral proficiency at the end of the training sessions.

2.4 Generalization Test

Trained at one speed only, the monkeys were tested for generalization to other speeds in the two described tasks. This was realized by interleaving trials of the trained speed with trials from the other speeds in a 90:10 ratio. Moreover, in order to avoid associative learning on these new stimuli, we always rewarded the monkeys on these other speed stimuli (still correct responses on the trained speed were required to obtain a juice reward).

3 Computational model

We recently developed a technique for the analysis of human behavior from unsupervised silhouette data [5]. In contrast to action recognition systems (*e.g.*, [7]) where specific actions are trained, our approach models the training data in an unsupervised and generative manner in order to redetect familiar patterns at runtime. Unknown queries are rejected in the same spirit as in [8]. Our approach was initially developed to monitor the behavior of (elderly) people in their homes, in this work we however apply it to the same data as used for the monkey behavioral study. An overview is depicted in Fig. 3 and described in the following.

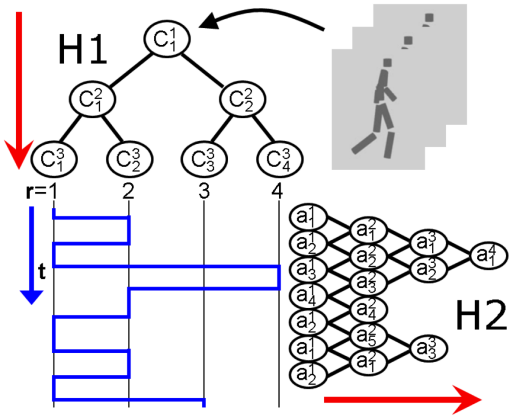


Fig. 3. Schematic overview of the developed computational model. From a set of data, *i.e.*, an image stream, our approach builds two hierarchies that encode the per frame appearance (snapshot, *H1*) and the actions (sequence, *H2*) in an unsupervised manner. *H1* is established from clustering the data in a hierarchical manner, *H2* analyzes how the cluster membership (the blue curve) evolves over time.

3.1 Appearance hierarchy (*H1*)

In a first hierarchical analysis stage, the per-frame appearance (snapshot) of the observed action is modeled. The training data are presented as normal video sequences of the *humanoid* stimuli. As input features we take binary segmented stimuli. These silhouettes are clustered (k-means) recursively in a top-down procedure, such that clusters are more specific on higher layers. This results in the structure *H1* as sketched in Fig. 3. The number of layers required, depends on the variability of the data. In this model, the root node cluster has to describe

all training features, whereas leaf node clusters only contain similar data and are more precise. Data points that arrive at one leaf node cluster are given the corresponding symbol.

3.2 Action hierarchy (*H2*)

In the action hierarchy, the sequence of symbols for subsequent frames in the training video data is encoded in a bottom-up process. Thus the evolution of the appearance over time is explored in *H2* as outlined in Fig. 3 (blue curve). If a change in the sequence of symbols appears, these symbols are combined to basic-level micro-actions, which are the low-level building blocks of *H2*. Inspired by [9], the combination of low-level micro-actions constitute higher-level micro-actions if they co-occur often enough. In this hierarchy a longer micro-action implies a more common action, as, again, such longer sequence is found to be common enough in the training data.

3.3 Analysis of unseen data

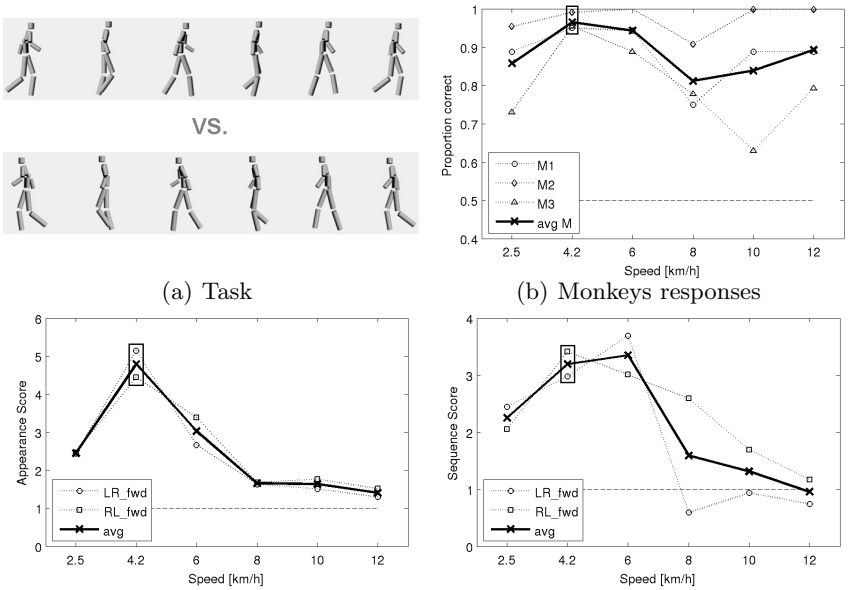
H1 and *H2* together provide a model of normal human behavior to which new data is compared at runtime. In *H1* a new frame's appearance is propagated as far as possible, seeking for the most specific cluster that still describes it well. Similarly, the observed history of appearances is matched in *H2* in order to evaluate how normal the observed action is. In [5] we show that this model can be used to track the (human) target, and that the model can be updated during runtime.

3.4 Generalization test

We use the same stimuli for training and tested the same tasks with the computational model as for the monkeys (*cf.* Sec. 2). After training with 4.2 km/h walking stimuli, new stimuli are presented to the model with different walking speeds in order to test the generalization capacity.

Since the model is designed to cope with larger amounts of data, in which recurring patterns are detected, 6 repetitions of the same stimuli were used for training. The appearance hierarchy (*H1*) consists of 4 layers resulting in 8 leaf node clusters. Separate models were trained for LR_fwd, RL_fwd and LR_bwd. During testing, we use the LR_fwd and the RL_fwd models for the LR/RL task, whereas in the FWD/BWD task we apply the LR_fwd and the LR_bwd models.

In the test phase, each applied model delivers two output values how well each test frame matches *H1* and *H2* (assuming *H1* has validated the stimulus), respectively. The value for *H1* captures the appearance only by measuring the similarity to one of the leaf node cluster centers. Additionally, *H2* requires the correct motion and searches for a corresponding micro-action with maximal length. To finally achieve the output score (appearance score from *H1*, sequence score from *H2*) we combine the two models, each trained for one of the two conditions relevant for the task. They are evaluated at each frame and a likelihood ratio is calculated and averaged across the whole stimulus. If for example a stimulus walking from left to right is described well in LR_fwd, but not in RL_fwd, the resulting score is high. On the other hand, if both models perform equally well, no clear decision can be drawn and the score is close to 1 (chance level in the case of the computational model).



(a) Task (b) Monkeys responses

(c) *H1* output (d) *H2* output

Fig. 4. LR/RL task: Monkeys performances and model scores for 6 tested speeds.

4 Results and discussion

In this section, the generalization performance for the different walking speeds is presented and compared between monkey behavior and computational model.

4.1 LR/RL task

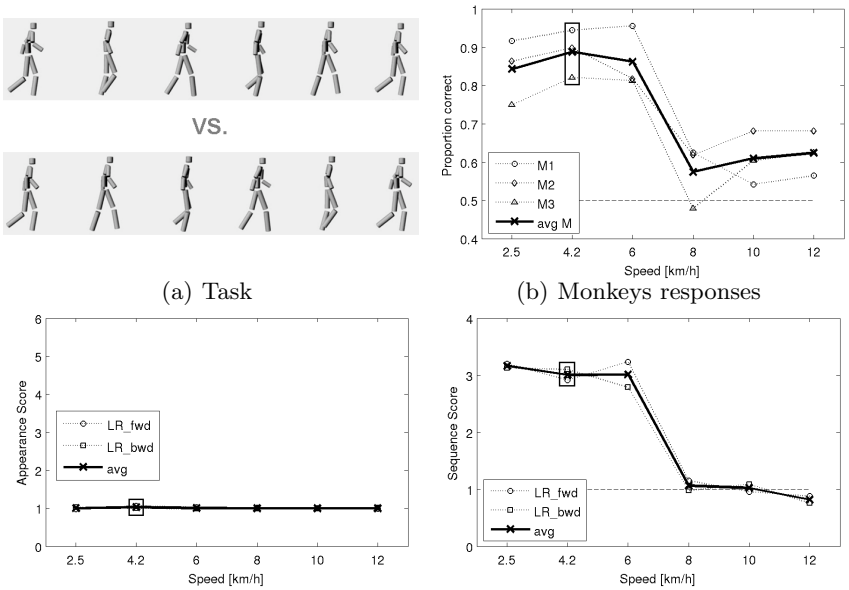
The results are depicted in Fig. 4, the monkeys responses are shown in panel (b), the appearance score (*H1*) in panel (c) and the sequence score (*H2*) in (d). Bold lines indicate the average results, dotted lines display individual performances for monkeys or different stimuli. Black boxes at 4.2 km/h point out the training speed. Chance level is marked with the dashed horizontal line.

In the behavioral study (Fig. 4(b)), categorization generalizes relatively well across the different walking and running speeds (binomial tests; $p < 0.05$ for 14 out of 15 generalization points). This suggests that the discrimination is based on spatial or motion cues that are common to the different speeds.

For the computational part, the results for *H2* (Fig. 4(d)) indicate a similar interpretation for slower walking speeds (2.5-6 km/h). In a more detailed analysis, we observe that for these speeds, the task can be solved already by only incorporating *H1*. Apparently, the appearances are distinctive enough. For the running stimuli on the other hand, silhouettes are different, thus the performance in $H1$ drops, which in turn drags down the *H2* scores.

4.2 FWD/BWD task

The results for the FWD/BWD task are visualized in Fig. 5 in the same manner as for the previous task. The behavioral data from the speed-generalization tests show that the categorization is specific to walking: in each monkey, generalization is significant (binomial test: $p < 0.05$) for the walking, but not the running



(a) Task (b) Monkeys responses

(c) *H1* output (d) *H2* output

Fig. 5. FWD/BWD task: Monkeys performances and model scores for 6 tested speeds.

patterns. In fact, in each monkey there is an abrupt drop of the performance when the locomotion changes from walking to running.

Computational findings show that the evaluation of the appearance (body poses) only is not sufficient for solving this task (Fig. 5(c)). This is not surprising, since the appearances are the same for both stimuli. However, if their ordering is considered (Fig. 5(d)), the task is solvable for walking speeds, and the scores resemble those of the monkeys. At higher speeds, due to wrong appearance classification in $H1$, the sequence is not reliable in $H2$ anymore.

The lack of significant transfer from the trained walking to running suggests that the animals learned a particular motion trajectory "template". Indeed, examination of the ankle trajectories (*cf.* Fig. 2(b)) reveals a relatively high similarity between those trajectories for the three walking speeds, which are in turn rather distinct from those of the three running patterns. This might also be a reason for the performance drop of the computational model.

4.3 General discussion

At the behavioral level in our monkeys we noticed a clear qualitative difference in generalization performances across tasks. Whereas the monkeys were quite apt at discriminating other speeds not seen before in the LR/RL task, a clear step-wise function was observed in the FWD/BWD task. In the LR/RL, when confronted with other walking speeds, *i.e.*, 2.5 or 6 km/h, all three monkeys could correctly categorize these locomotions significantly higher than chance level. However this was not the case when confronted with locomotions at running speeds, again a trend present in all three monkeys.

The broader generalization observed in the LR/RL task compared to the FWD/BWD task shows that such motion cues are less specific. Alternatively, the

monkeys might have used spatial features that are common to the walking and running *humanoids* that face in a particular direction. The fact that one could solve the LR/RL task quite simply by basing decisions on the presentation of just one frame could explain the observed (almost) perfect generalization. This is analogous to the first hierarchical analysis stage (*H1*), which works on the per-frame appearances of actions. However, at this stage, the model shows a slightly different pattern, performing quite robustly for the trained locomotion, with clear drop-offs already for the neighboring speeds. This is clearly due to overfitting of the model to the trained action. Monkeys have been exposed to other locomotion patterns before, in contrast to the computational model. When implementing the second hierarchical stage of the computational model (*H2*), which incorporates the evolution of the per-frame appearances over time, the model's performance resembles the monkey's performances more closely, especially for the FWD/BWD task. In summary, we see that monkeys have the capability to generalize well for simple tasks where snapshot information is sufficient. This might be due to prior knowledge based on different functional features, which is so far not included in the computational model at all.

5 Conclusions

In this work, we compared findings from behavioral studies to a particular, biologically inspired computer vision algorithm. The most important outcome is that a two stage computational system can, to some extent, reproduce monkey responses. The algorithm however has not the same generalization capacities which suggests that monkeys integrate the training in a broader manner than the computer does.

References

1. Giese, M.A., Poggio, T.: Neural mechanisms for the recognition of biological movements. Nature Reviews, Neuroscience **4** (2003) 179–192
2. Schindler, K., van Gool, L.: Action snippets: How many frames does human action recognition require? In: Proc. CVPR. (2008)
3. Lange, J., Lappe, M.: A model of biological motion perception from configural form cues. Journal of Neurosciences **26** (2006) 2894–2906
4. Vangeneugden, J., Vancleef, K., Jaeggli, T., Gool, L.V., Vogels, R.: Discrimination of locomotion direction in impoverished displays of walkers by macaque monkeys. Journal of Vision **10** (2010) 22.1–19
5. Nater, F., Grabner, H., Gool, L.V.: Exploiting simple hierarchies for unsupervised human behavior analysis. In: Proc. CVPR. (2010)
6. Vangeneugden, J., Pollick, F., Vogels, R.: Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. Cerebral Cortex **19** (2009) 593–611
7. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: Proc. CVPR. (2007)
8. Boiman, O., Irani, M.: Detecting irregularities in images and in video. In: Proc. ICCV. (2005)
9. Fidler, S., Berginc, G., Leonardis, A.: Hierarchical statistical learning of generic parts of object structure. In: Proc. CVPR. (2006)