



# Learning Mid-Level Motion Features for the Recognition of Body Movements



Rodrigo Sigala\*, Thomas Serre\*, Tomaso Poggio\*, Martin Giese\*

ARL, UKT – Tübingen, Germany

Supported by National Institute of Health grant 2R01-EY-07861-11, Volkswagenstiftung, DFG and HESP

McGovern Institute, MIT, Cambridge, MA

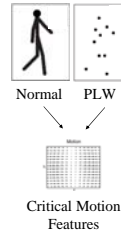
\*ARL, HCBR, Dept. of Cognitive Neurology, University Clinic Tübingen, Germany

\*CBCL, Mc Govern Institute, Brain & Cognitive Sciences Dept., M.I.T., Cambridge MA, USA

## Introduction

### Robust Recognition of Body Movements based on Critical Mid-Level Features

- Humans can recognize body movements (e.g. walking and running) accurately and robustly.
- This robustness is demonstrated by the ability of subjects to recognize body movements from strongly impoverished stimuli, like *Point-like walkers (PLW)*, which consist only of a small number of illuminated dots that move like the joints of a human actor [6]. Subjects can even recognize gender, emotions, and identity from these stimuli.
- A possible explanation for this robust generalization is that the brain extracts specific motion features (of intermediate complexity) that are shared by both stimuli classes (normal walker and PLW).
- The nature of these features is unknown, and it has been discussed whether they are based predominantly on motion or form information [7]. In a recent study, combining methods from image statistics and psychophysical experiments, it was shown that robust recognition can be accomplished based on mid-level motion features [2].



### A Neurally inspired Mechanism for Learning Mid-level Form Features from Natural Images

- Humans also recognize static objects (e.g. cars, houses or faces) very accurately and robustly.
- Robust object recognition in every-day life requires independence of background features, object size, view-point and light conditions.
- A physiologically plausible *Standard Model (SM)* of object recognition has been proposed in [8]. The SM accounts for different sets of physiological and psychophysical data on visual object recognition, and successfully recognizes idealized stimuli (e.g. without cluttered background). However, this model has failed so far to recognize objects in real-world conditions with background clutter and varying illumination.
- By optimizing the selectivity of the intermediate layers of the SM for an extraction of maximally informative mid-level features the recognition performance of the SM in real-world conditions could be substantially improved (see Serre & Poggio, VSS 05, poster # 744), reaching performance levels that are competitive with state-of-the-art computer vision systems [9].
- Mid-level feature learning can be realized with a simple neurally inspired *memory trace* algorithm, which requires only local neural operations that can be easily implemented by real neurons (Serre & Poggio, VSS 05).

### Performance comparison with benchmarks

OBJECT CATEGORY	SYSTEM	ROC (avg)	ZFA	AUSROC
CARS	SM-MeT algo + SVM	94.9%	99.9%	95.5%
	(Poggio et al. 2005)	N/A	N/A	90.3%
LEAFES	SM-MeT algo + SVM	84.9%	70.1%	84.9%
	(Walker et al. 2006)	84%	80%	96.9%
FACES	SM-MeT algo + SVM	86.3%	97.2%	96.9%
	(Poggio et al. 2005)	84.6%	N/A	N/A
MOTORBIKES	SM-MeT algo + SVM	84.9%	25.9%	84.9%
	(Poggio et al. 2005)	62.8%	N/A	N/A
FACES-CIRCLE	SM-MeT algo + SVM	82.6%	29.0%	87.6%
	(Poggio et al. 2005)	84.9%	61.8%	86.9%
CARS-CIRCLE	SM-MeT algo + SVM	81.5%	1%	97.4%
	(Liang 2004)	N/A	6%	97.4%



### Question: Is the Same Neural Mechanism Suitable for Learning Mid-level Features for Biological Motion Recognition?

- Adaptation of memory trace algorithm (MeT) for learning motion features.
- What is the form of optimized mid-level motion features? Are they congruent with experimental data?
- Does an optimization of mid-level features, like in object recognition, increase performance of biological motion detection under difficult conditions (e.g. in presence of background clutter)?

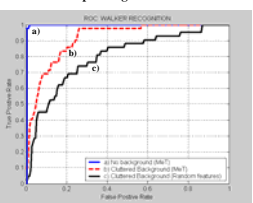
## Results

Performances (area under the ROC) are shown in the table below for our system using the MeT algorithm (MeT), without and with motion clutter in the background (Clutter). For comparison, we also show the results for stimuli in motion clutter when the mid-level features were defined by selecting randomly positioned regions from the stimuli (Rand). Results are compared for the four different tested classifiers.

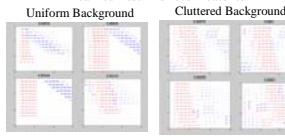
The best features for the walker-detection task are shown below for the simulations without and with motion clutter. An important observation is that many of these features are characterized by horizontal opponent motion.

The ROC curves for the three test conditions are shown in the figure on the right. Performance after training with the MeT rule without clutter is almost perfect. This is true for a robust classifier such as SVM, but also for simpler classifiers such as NN and MAU. In presence of clutter selection by the MeT rule is significantly better than random selection. (The 5-NN outputs SVM classifier, probably due to overfitting). This robust performance is consistent with results for object recognition in the shape pathway with the SM (Serre & Poggio, VSS 05). This suggests a key role for plasticity in intermediate and higher visual areas of cortex for the realization of robust recognition. (See [10] for details, and Jastorff et al., VSS 05, poster # 935.)

### Receiver Operating Characteristics



### Learned Best Mid-level Features



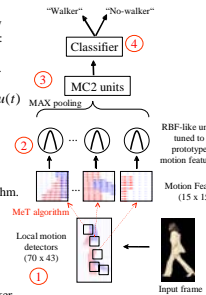
WALKER RECOGNITION PERFORMANCES (AUC)				
LEARNING	MAU	SVM	J-NN	k-NN (k=5)
MeT (No Bgd)	0.976190	0.999646	0.980952	0.961905
MeT (Bgd)	0.864823	0.911565	0.957143	0.971429
Random (Bgd)	0.721532	0.794714	0.876190	0.890476

## Methods

### Model for the Recognition of Biological Movements

The model corresponds to the motion pathway of the model in [5] and contains a hierarchy of neural detectors that are selective for motion features with different levels of complexity:

- Local motion energy-detectors whose responses are derived from optic-flow fields. Detectors for 70 x 43 different spatial positions and 4 different directions.
- Units temporally smooth assuming a simple linear low-pass filter:  $\tilde{u}(t) = r(t) - u(t)$  where  $r(t)$  is the motion energy signal,  $u(t)$  the detector output and  $\tau = 228$  ms.
- The MS2 units encode prototypical motion features of intermediate complexity. They combine the responses of detectors of the previous layer within a limited spatial region (15 x 15).
- Modeled by Gaussian RBF units whose centers are determined by the MeT algorithm. The responses of detectors increase with the similarity of the local motion energy patterns of the present stimulus with the RBF center.
- The MC2 units pool the responses of all mid-level motion detectors of the same type within a limited spatial receptive field using a MAXIMUM operation. The responses of these units are partially position-invariant.
- They define the input of a classifier that detects the presence or absence of a walker.

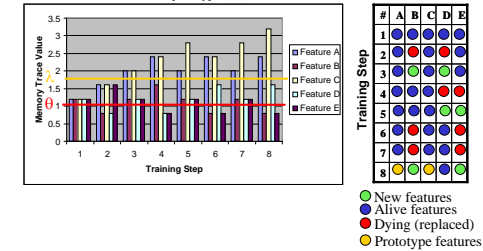


### Classification of the MC2 Units

- The "**Maximally Activated Unit**" (MAU) classifier is biologically plausible. It corresponds to a RBF unit whose centers are trained with the output signals from the MC2 level for the learned movements. If the activation of these units exceeds a given threshold the stimulus is classified as that particular action. Otherwise the classification result is negative.
- k-Nearest Neighbor (k-NN)**, is a classifier whose response is based on RBF units that are trained as for the MAU classifier. During classification the label of a test example is determined by the majority of the labels assigned by the units that correspond to the  $k$  nearest neighbors in the training set (we tested for  $k = 1$  and  $k = 5$ )
- Support Vector Machine (SVM)** classifiers [13], have been used in many recent machine vision systems. Although SVMs are not biologically plausible, they provide a typically well-performing classification back-end, which is useful to derive a measure for the quality of the learned features.

### The Memory Trace (MeT) Algorithm

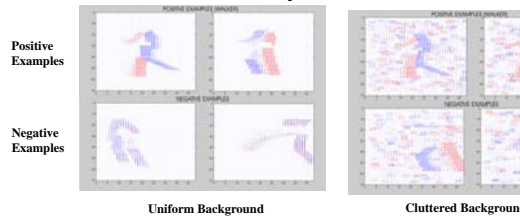
The MeT algorithm (see Serre & Poggio, VSS 05) is a new biologically inspired method for the unsupervised learning of frequently occurring features. The algorithm is based in a previous work of [3] who exploits a simple *memory trace rule* for the learning of shift invariance. Our trace rule assumes that the MS2 units keep record of their recent synaptic activity by an internal memory trace signal. The different features compete for the activations that a given stimulus produces. Successful activation of a feature results in an increase of its memory trace signal. Otherwise, the trace signal decays. Features whose memory trace falls below a fixed threshold  $\theta$  ("die") are eliminated (●) and replaced by new features. The other features remain in the competition ("alive") (●). New features (●) are generated by choosing a randomly positioned local region in the visual field, and by using the outputs of the motion energy detectors within this region for the present stimulus as feature vector. Finally, only features whose memory trace value exceeds a particular threshold  $\lambda$  are considered as prototypes (●).



### Walker Detection Task

The performance of our model was evaluated using a walker-detection task. We used stimuli with a uniform background, and with motion clutter. Stimuli were generated from five actors whose joint trajectories were tracked from videos (one gait cycle with 42 frames) [5]. The walking sequences of five actors were used as positive examples, and other human actions (e.g. running, boxing, jumping) as negative examples. We selected randomly different sets of these sequences for training and testing of the model. To introduce motion clutter for the same stimuli, we added 100 moving squares (3x3 pixels) at random positions in each stimulus frame, defining random motion with uniform distribution of motion energy over the different directions.

### Optic flow fields



## Conclusions

- We have presented a biologically-inspired local learning rule for the optimization of mid-level motion features of a hierarchical neural model for the recognition of biological movements.
- We found that learning of optimized mid-level features substantially improves the performance of the model, in particular in presence of motion clutter.
- Similar results have been obtained with a model for shape processing in the ventral pathway using the same learning rule (Serre & Poggio, VSS 05). This suggests a key role of visual experience and plasticity throughout the whole visual cortex.
- In addition, we found that for the detection of walkers, our algorithm learned optimized motion features that are characterized by horizontal opponent motion, for training with and without motion clutter. The same technique could be applied to optimize form features for the recognition of biological movements from body postures [5].
- The importance of opponent motion features seems to be supported by psychophysical and imaging results [2,12]: Opponent horizontal motion might be a critical feature for the recognition of walkers, and degraded point-light stimuli. Electrophysiological experiments indicate the existence of opponent motion-selective neurons, e.g. in monkey areas MT and MST [1, 11].

## References

- Born, R. T. (2000). Center-surround interactions in the middle temporal visual area of the owl monkey. *J. Neurophysiol.* 84, 2658-2669.
- Castile A, Giese M. (2005) Critical features for the recognition of biological motion. *J. of Vision*, 5(4), 348-360.
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, vol. 3, pp. 194-200, 1991.
- Fukushima, K. (1980) Neocognition: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193-202.
- Giese M A, Poggio T (2003) Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience* 4, 179-192.
- Johansson, G. (1973) Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201-211.
- Mather, G., Radford, K., & West, S. (1992). Low-level visual processing of biological motion. *Proc. R. Soc. Lon. B*, 249(1325), 149-155.
- Riesenhuber, M. & Poggio, T. (1999) Hierarchical models for object recognition in cortex. *Nat. Neuroscience* 2, 1019-1025.
- Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. To appear in CVPR 2005.
- Sigala, R., Serre, T., Poggio, T., Giese, M. (2005). Learning Features of Intermediate Complexity for the Recognition of Biological Motion. Proceedings of the International Conference on Artificial Neural Networks, ICANN 2005, accepted.
- Tanaka, K., Fukuda, Y. & Saito, H. (1989). Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey. *J. Neurophysiol.* 62, 626-641.
- Vaina et al. (2001). Functional neuroanatomy of biological motion perception in humans. *PNAS* 98, 11656-11661.
- Vapnik, V. (1998) *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

## Acknowledgements

We are grateful to A. Castile for his support with the simulations. ARL is supported by the Volkswagenstiftung, DFG SFB 550, and HESP. CBCL is sponsored by Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0209289, National Science Foundation-NIH (CRNS) Contract No. EIA-0218693, National Science Foundation (TR) Contract No. IIS-0209289, National Science Foundation-NIH (CRNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRNS) Contract No. EIA-0218506, and National Institutes of Health (Come) Contract No. 1 P20-MH66239-01A1.