



Bio-inspired visual sequences classification

Mauricio Cerda, Bernard Girau

► To cite this version:

Mauricio Cerda, Bernard Girau. Bio-inspired visual sequences classification. Brain Inspired Cognitive Systems 2010 - BICS 2010, Jul 2010, Madrid, Spain. pp. 20. hal-00504090

HAL Id: hal-00504090

<https://hal.science/hal-00504090>

Submitted on 19 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Bio-inspired visual sequences classification

2 Mauricio Cerda and Bernard Girau

3 **Abstract** The capacity to perceive and interpret highly complex visual patterns such
4 as body movements and face gestures, is remarkably efficient in humans and many
5 other species. Among others tasks, the classification of visual sequences without
6 context is one key problem to understand both the coding and the retrieval of spatial-
7 temporal patterns in the human brain. In this work we present a model able to per-
8 form classification of synthetic. Our model takes into account current knowledge
9 in experimental psychophysics and physiology. The presented model shows that
10 sparse spatial coding of spatial-temporal sequences could be sufficient to explain
11 both: classification with partial information and tolerance to time-warping. We are
12 also able to code temporal sequences with single populations of units, without the
13 need of explicit “snapshots” at each time instant.

14 1 Introduction

15 The understanding of the different principles and mechanism that exist in the brain
16 to perform perceptive and cognitive tasks, are since long time being studied by bi-
17 ologist. Yet, only in the last decades there is an increasing interest in the application
18 of these ideas in fields such as computer vision and robotics, the “bio-inspired”
19 methods.

20 There is a wide variety of questions in vision starting from what information to
21 process?, then how to analyze this data? and how to operate in natural conditions?
22 just to give a few examples. In this work, we are interested in the problem of rec-

Mauricio Cerda

INRIA-Loria Nancy Grand Est, Equipe Cortex - Bat. C040, 54506 Vandoeuvre-les-Nancy, France
e-mail: cerdavim@loria.fr

Bernard Girau

INRIA-Loria Nancy Grand Est, Equipe Cortex - B.P. 239, 54506 Vandoeuvre-les-Nancy, France
e-mail: bernard.girau@loria.fr

23 ognize spatial-temporal sequences, such as walking, jumping and running persons,
 24 see Figure 3. The applications are not necessary to human activities, but it is the
 25 problem motivates this work.

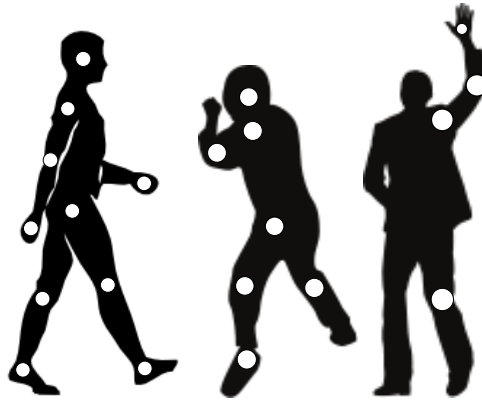
26 The work we present deals with the problem of how to code and differentiate
 27 spatial-temporal sequences, taking into account know properties of the human brain
 28 that we describe in 2.1. Some of the key ideas we consider are that classification can
 29 be performed using only a few points (see Figure 1) that can be extracted from the
 30 movement in the sequence, and a $2D$ coding of the sequence is the most likely, even
 31 if the movement is $3D$. Here, we propose a single neural mechanism to model these
 32 ideas.

33 To test our model, we perform several simulations over the case of single tra-
 34 jectories that could then describe other more complex spatial-temporal sequences.
 35 We show that we can retrieve other properties such as speed-invariance (time warp-
 36 ing ¹) and partial responses in time (to answer before the sequence is completely
 37 presented) [1].

38 We also compare our work to artificial vision techniques, to gain understand in
 39 the difference that could possibly have neural mechanisms and the current state-
 40 of-the-art techniques. The main difference is the extensive use of body models in
 41 computer vision (despite the fact that this is still in discussion among biologist)
 42 and the use of complete sequence to classify. Our model could explain both things:
 43 classification of sequences can be performed without an explicit model and it is
 44 possible to give classification answers since very early in the sequence.

45 The next section 2, presents an overview of experiments in biology and tech-
 46 niques in computer vision, to locate this work in both fields. Section 3 describes
 47 our model and the Results & Discussion presents the results of our simulations, and
 48 comparison against other model. Finally we presents the conclusion of this work
 49 in 5.

Fig. 1 Some example human movements (walking, fighting and waving). In these sequences there are locations (in white) that are more relevant in terms of the information they can contribute to be differentiated from other sequences.



¹ Commonly associated to temporal sequences, when the same sequence is delayed or compressed/dilated in time. Perceptive phenomena such as speed recognition can tolerate this kind of variation.

50 2 Overview

51 In this section, we present some experimental evidence in primates (humans) and
 52 available techniques in computer vision, to characterize the classification of visual
 53 patterns.

54 2.1 Biological overview

55 There is abundant experimental evidence related to the classification of spatial-
 56 temporal patterns in humans and primates. Most of these experiences come from
 57 experimental psychology as the classification task is associated to higher areas of
 58 the brain. More recently, different works have used medical imaging techniques to
 59 identify different zones of activation/inactivation. Despite these efforts questions
 60 such as: what is exactly the input to perform classification? and how exactly the
 61 coding and retrieval is performed?, are still in discussion [2]. To overview some rel-
 62 evant works, we summarize observed properties and the protocol used to support
 63 each one.

- 64 • **Robustness.** Even though visual signals can be severely diminish, stimuli as sim-
 65 ple as PL [3]² or even random PL [4] are sufficient to allow good pattern classifi-
 66 cation. Hence, a few points are sufficient to distinguish between several stimuli.
- 67 • **View dependent.** Recognition of visual patterns depends in the angle of view of
 68 the observer. Evidence show that the same subject decrease performance when
 69 the presented pattern is rotated, but experience can improve this performance.
 70 The normal tolerance is about 20 degrees [5, 6].
- 71 • **2D coding.** Experiments by [7] indicates that at least for the PL stimuli, a 2D
 72 representation is sufficient to explain brain coding schemes for body actions.
 73 This remain in discussion, because 3D representation could still exist with an
 74 intermediate 2D projection (top-down).
- 75 • **Foveal processing.** Several works show that peripheral areas of the visual [2], are
 76 significantly less sensitive to human action. We interpret this as an other possible
 77 simplification, one single area of interest can be process at the same time within
 78 a visual scene.
- 79 • **Feature extraction.** Different works [8, 4] indicate that the most relevant feature
 80 to perform classification of human sequences is the local motion (probably pro-
 81 cessed in areas V1/V5/MST, see Figure 2). This was tested using variations of
 82 the PL stimuli with occlusions. However, other work [9] show that static features
 83 could be also be used, movement information seems then to be the more relevant
 84 to classify, not the only one.
- 85 • **Temporal sensitivity.** Despite the robustness of the feature extraction, pattern
 86 matching in the brain seems to be extremely sensitive to temporal correlation

² Point-Light stimuli. Experiment proposed originally by G. Johansson in the 70', where only the joints of an actor were enlighten.

- 87 [2, 6]. Taking the PL stimuli as an example, it is possible to remove or even to
 88 change a few points but not to change the relative speed between these points.
- 89 • Code reading. Evidence exist [10] that single neurons in areas as EBA or FBA
 90 (see Figure 2) are sensitive to human actions such as walking, running, etc. Also
 91 evidence exist about areas sensitive to static features such as body posture, face
 92 expression, hands in the ventral pathway. However, body posture activation based
 93 in motion information only (dorsal pathway), is yet to be proved [6].

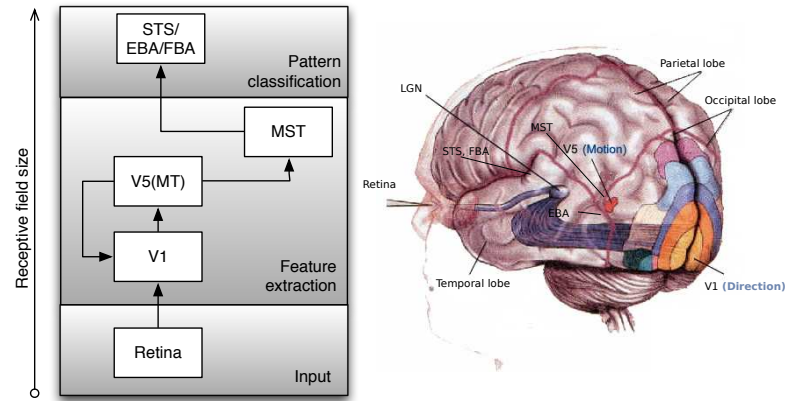


Fig. 2 Schematic view of some of the different visual areas involved in the recognition of human motion perception.

94 2.2 Computer vision

95 In the field of Computer Vision, a wide variety of algorithms [11] exists and have
 96 been applied to process video, and to perform pattern classification for this kind of
 97 signals. Although this large diversity of algorithms exist, the process can be divided
 98 in stages, to point out the different sub-tasks related to the problem. The stages
 99 we considered are feature extraction and pattern classification (pose estimation and
 100 recognition in [11]). Other stages such as initialization and tracking are in practical
 101 implementations absolutely necessary, but in this work there is no context to inter-
 102 pret or distractors to avoid; the target is already located, and there are no distractors
 103 in the scene.

104 **Feature Extraction** Feature extraction is about what information do we use to
 105 classify. One of the simplest features is pixel intensity, but others such as edges,
 106 silhouettes, color or combination of all of them can be used. More elaborated fea-
 107 tures also exist, such as PCA, ICA, SOM, VQ [13], that take into account statistical

information, to define the space where is more relevant to perform pattern classification. Even tough it is difficult to generalize due to the large number of techniques, silhouettes of the body are largely used [11].

Pattern Classification Pattern classification have been performed with techniques such as distances in some features space, HMM building states for each configuration, RBF with pattern prototypes, etc.. The available techniques are again quite large, but it is important to notice that most of the techniques use a a-priori model of the body and require a full movement to classify. However, there are “model free” techniques, and systems capable to answers in a few frames [1], but it is not a large percentage of the techniques as pointed out in [11].

3 Model description

To summarize, there is evidence in biology that local movement information is sufficient to perform classification, that the coding is more likely “2D”, but still with partial rotation-invariance. Also, highly robust to speed variations and capable to give answers before the full stimuli is presented, yet extremely sensitive to temporal variations (see Figure 3).

Since the features could be considered as several relevant trajectories in time (taking the idea of the PL stimuli), we start considering we are able to know the position of these points in time, and we want to differentiate trajectories in time. For that we use Continuum Neural Field Theory (CNFT) [14], where the visual visual is mapped to a populations of units or neurons (2D).

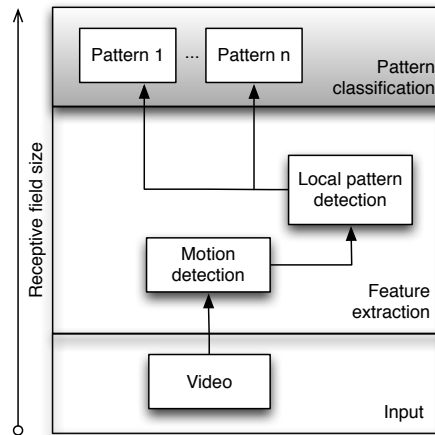


Fig. 3 Schematic view of the model we present.

3.1 Pattern classification (Asymmetric CNFT or ACNFT)

We build a classification system with the Eq. 1 for the activity m of each unit, where we use one m for each pattern we want to classify (see Figure 3), extending the work of [14]:

$$\frac{\partial m(\mathbf{x}, t)}{\partial t} + \tau m(\mathbf{x}, t) = \left[\int_0^{\mathbf{x}_f} w(\mathbf{x}', \mathbf{x}) m(\mathbf{x}', t) d\mathbf{x}' + I(\mathbf{x}, t) \right]^+ \quad (1)$$

here w determines the selectivity of the system and $[]^+$ is the maximum with 0. For the simple trajectory (line) we are considering, we use a periodic function and one Gaussian function along the trajectory axis.

$$w(\mathbf{x}, \mathbf{p}) = \alpha \exp\left(-\frac{(y - p_y)^2}{2\sigma}\right) (J_0 + J_1 \cos(2\pi(p_x - x)/l - \beta)) \quad (2)$$

The main parameters are the asymmetry β , the spatial size of the kernel σ and the total length of the path l . This function of the current unit position \mathbf{p} giving a weigh w for each position \mathbf{x} in the trajectory, being zero elsewhere.

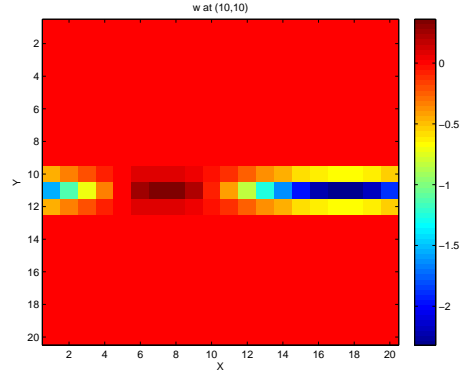


Fig. 4 Kernel function w at one position $\mathbf{p} = (10, 10)$ for all possible locations \mathbf{x} .

Looking at the Figure 4, the values for the function w far from the actual pattern trajectory are very close to zero. To give the final score for each input, we perform a temporally smoothed average as in [6]:

$$\frac{\partial s(t)}{\partial t} + \tau s(t)/2 = \int m(\mathbf{x}', t) d\mathbf{x}' \quad (3)$$

the decay term is written as $\tau/2$ to show that this equations dynamics should be slower than m , i.e. smaller than τ .

144 4 Results & Discussion

145 The simulations we performed were all for the single straight line trajectory, because
 146 is the most simple pattern we can consider, still useful to decompose more complex
 147 sequences. The objective in this simulation is to make the difference between the
 148 same trajectory in difference directions, controlling varying other variables.

149 4.1 Synthetic data

150 The data we generate to classify within two categories is left-to-right and right-
 151 to-left motion. The justification for this choose other than the simplicity is the use
 152 of this paradigm in experimental psychology [4], where one commonly used task
 153 is to difference left or right walking using the point-light-stimuli [3] in different
 154 conditions. The input we are using is defined as:

$$I(\mathbf{x}, t) = \exp\left(-\frac{(x - vt)^2 + (y - y_0)^2}{2\sigma^2}\right) + \mathcal{N}(0, \Sigma) \quad (4)$$

155 where $\mathbf{x} = (x, y)$, v is the input speed, y_0 is the location in the y axis, σ the input
 156 size. Finally we use additive Gaussian noise of mean 0 and variance Σ to modify
 157 the noise level.

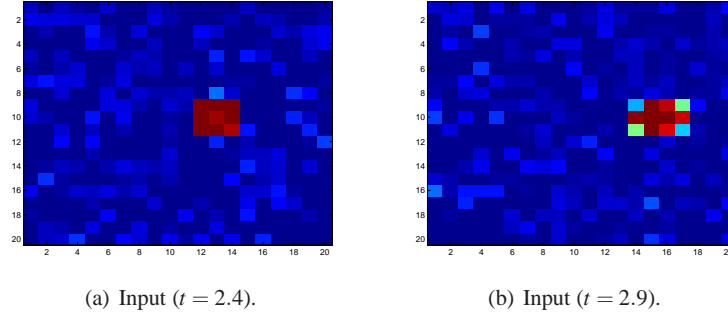
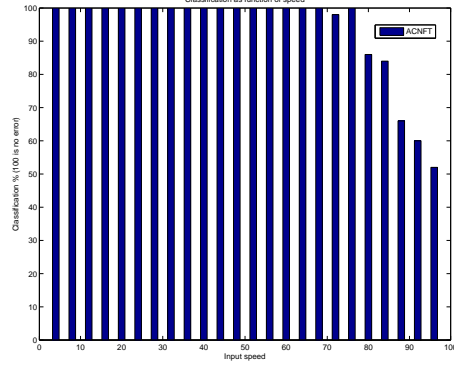


Fig. 5 Input sequence at two difference times, $\Sigma = 0.005$

158 Using directly this input (no feature extraction), we perform variations in three
 159 parameters: Σ (noise level), y_0 and v (input speed), considering 50 trials for each
 160 case. Using $v = 5$, $\Sigma = 0.005$, $y_0 = l/2$, $y_0 = l/2$, $tau = .15$, $J_0 = -9.8$, $J_1 = -13.5$,
 161 $\beta = 2$, $\sigma = .001$, $l = 20$ if not otherwise indicated. The method we use to simulate
 162 Eq. 1 is Runge-Kutta (4th order) with $dt = .1$. Extensive analysis of the parameters
 163 for the sinusoidal function can be found in [14].

164 4.2 ACNFT Simulation

165 **Test A, time-warping.** In this experiment, the input speed v in Eq. 4 was varied.
 166 The ACNFT was configured to recognize a given speed V . The input moves at speed
 167 v or $-v$ with v in $[V - \varepsilon, V + \varepsilon]$. The model it says to correctly classify if it can make
 168 the difference between this two inputs. Several trials (50) were performed to average
 169 the effect of noise.



170 **Fig. 6** Classification perfor-
 171 mance for different speed's
 172 with 50 trials. Level noise is
 173 $\Sigma = 0.005$.

170 The ACNFT could tolerate larger variations in speed using one trajectory as input.
 171 It is important to remember that the model was configured for speeds around
 172 $v = 5$, and as speed increased the absolute difference between v and $-v$ also in-
 173 creases.

174 **Test B, temporal response.** In this experiment, there is no variation of noise, show-
 175 ing the temporal evolution of the classification. The ACNFT was configured to rec-
 176 ognize a given speed V . The input moves at precisely speed V or $-V$. The model it
 177 says to correctly classify if it can make the difference between this two inputs. Sev-
 178 eral trials (50) were performed to average the effect of noise (Σ), that in that case
 179 takes three values 0, 0.05, .1, where we know from the Test C, the ACNFT drops
 180 performance as function of noise.

181 The ACNFT starts with a very poor performance, but very quickly it reaches a
 182 stable classification performance. it is important to notice that the full cycle happens
 183 at $t = 40$ and $t = 80$, but even before the performance reach the peak, temporally
 184 dropping at the transition point ($t = 40, 80$).

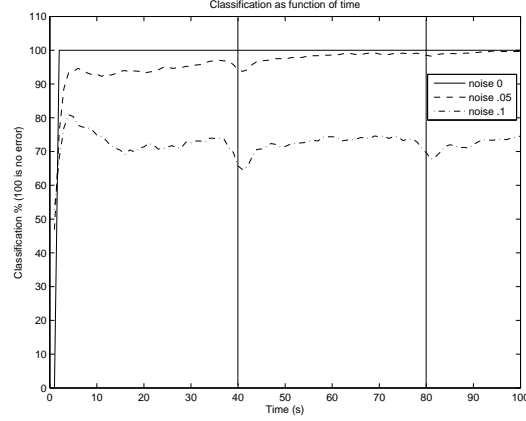


Fig. 7 Classification performance as function of time with 50 trials. Level noise (Σ) is 0, 0.05 and 0.1.

4.3 Comparison between ACNFT & STC

To gain further understanding we introduce a simpler spatial correlation mechanism. This mechanism keeps record of full snapshots for each time t .

4.3.1 Spatio-temporal correlation (STC)

To compare we choose a simple and more direct model, where at each time t we have a complete template of the input. The temporal sequence is build using also Eq. 3. This is a very naive approach to perform spatio-temporal classification, but it has the minimal required properties, to know: higher answer for a spatially well located input, higher answer for the right temporal order.

We can resume this system as:

$$C(t) = \sum I(\mathbf{x}, t) T(\mathbf{x}, t) \quad (5)$$

$$\frac{\partial S(t)}{\partial t} + \tau S(t)/2 = C(t) \quad (6)$$

T is the template. The Eq. 6, use the same kind of mechanism for sequentiality as the ACNFT model, smoothing out the spatial correlation over time.

Test C, noise tolerance. In this experiment, the noise level Σ in Eq. 4 was varied. Both classification systems: ACNFT and STC were configured to recognize a given input at speed V . The input moves at speed V or $-V$. One model it says to correctly classify if it can make the difference between this two inputs. Several trials (50) were performed to average the effect of noise.

At low noise level, STC and ACNFT give identical performance, as the noise level increase the ACNFT decrease its performance, until reaching the 50% (best than chance probability) around $\Sigma = .15$, see Figure 8. These results are function of

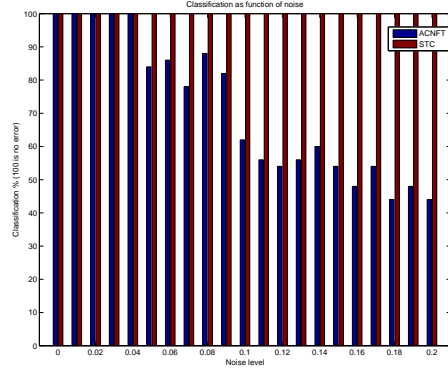


Fig. 8 Classification performance for different levels of noise using 50 trials. Input speed in all trials is $v = 5$.

205 w , if σ increases or if the kernel is not zero close the trajectory (wider trajectories),
 206 the tolerance to this kind of noise can be modified.

207 **Test D, position-invariance.** In this experiment, the location in the axe perpendicular to the trajectory y_0 in Eq. 4 was varied. Both classification systems: ACNFT
 208 and STC were configured to recognize a given speed V at one particular y_0 . The
 209 input moves at speed V or $-V$ but this time at different y_0 . One model it says to correctly
 210 classify if it can make the difference between this two inputs. Several trials
 211 (50) were performed to average the effect of noise.
 212

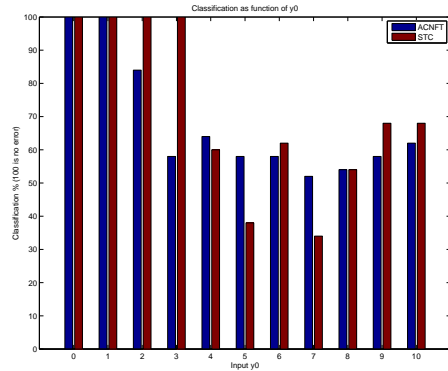


Fig. 9 Classification performance for different y_0 with 50 trials. Input speed in all trials is $v = 5$ and the level noise is $\Sigma = 0.005$.

213 The STC mechanism is very sensitive to this kind of variation by construction,
 214 the correlation is not invariant to spatial variation, dropping performance very fast.
 215 The ACNFT show similar properties, also quickly dropping performance. This can
 216 be explained by the definition of w , where the input does not requires to be exactly
 217 in the template position to activate the mechanism, but is limited by the size of the
 218 kernel σ , see Figure 4.

5 Conclusions

In this set of experiments we have show that the ACNFT model could perform classification of spatiotemporal sequences under different variations of the input. The presented model is also capable to answer with partial data, classifying even before the full temporal sequence is presented and to maintain performance for large variation of the speed for the same spatial pattern. We have also compare against the naive STC scheme, showing that the ACNFT model has basically similar spatial properties, dropping performance as function of noise and showing small spatial invariance.

These results show that the ACNFT exhibit several properties similar to how the human brain performs the classification of visual patterns: speed invariance (partial) and “on-line” classification. We also propose that experiences such as variations of the relative distance between PL stimuli and measurements of the temporal evolution of the response, could give further insides about the mechanism behind the brain processing of human motion sequences.

It still remains to show how to code more complex consequences, where multiple trajectories are necessary and the input is obtained by processing a real signal.

References

1. K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.
2. R. Blake and M. Shiffrar. Perception of human motion. *Annu. Rev. Psychol.*, 58:47–73, 2007.
3. G. Johansson. Visual perception of biological motion and model for its analysis. *Percept. Psychophys.*, 14:195–204, 1973.
4. Antonino Casile and Martin A. Giese. Critical features for the recognition of biological motion. *J. Vis.*, 5(4):348–360, 4 2005.
5. A. Puce and D. Perrett. Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London, Series B.*, 358:435–445, 2003.
6. Martin A. Giese and Tomaso Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4:179–192, 2003.
7. I. Bühlhoff, H. Bühlhoff, and P. Sinha. Top-down influences on stereoscopic depth-perception. *Nature neuroscience*, 1(3):254–257, July 1998.
8. Steven M. Thurman and Emily D. Grossman. Temporal “Bubbles” reveal key features for point-light biological motion perception. *J. Vis.*, 8(3):1–11, 3 2008.
9. R.J. Reid, A. Brooks, D. Blair, and R. Van der Zwan. Snap! recognising implicit actions in static point-light displays. *Perception*, 38(4):613–616, 2009.
10. Marius V. Peelen and Paul E. Downing. The neural basis of visual body perception. *Nature Reviews Neuroscience*, 8(8):636–648, 2007.
11. Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Comput. Vis. Image Underst.*, 81(3):231–268, 2001.
12. N. Rougier and J. Vitay. Emergence of Attention within a Neural Population. *Neural Networks*, 2005.
13. Vladimir S. Cherkassky and Filip Mulier. *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons, Inc., New York, NY, USA, 1998.

- 263 14. Xiaohui Xie and Martin A. Giese. Nonlinear dynamics of direction-selective recurrent neural
264 media. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65:051904, May 2002.