INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Could early visual processes be sufficient to label motions?

Ivan Dimov, Pierre Kornprobst, Thierry Viéville

N° 5240

Juin 2004

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--5240--FR+ENG

Could early visual processes be sufficient to label motions?

Ivan Dimov, Pierre Kornprobst, Thierry Viéville

Thème BIO — Systèmes biologiques
Projet Odyssee

Rapport de recherche n° 5240 — Juin 2004 — 35 pages

Abstract: Biological motion recognition refers to our ability to recognize a scene (motion or movement) based on the evolution of a limited number of points acquired for instance with a motion capture tool. Much work has been done in this direction showing how it is possible to recognize actions based on these points. Following the reference work of Giese and Poggio [27], we propose an approach to extract such points from a video based on spiking neural networks with rank order coding. Using this estimated set of points, we verify that correct biological motion classification can be perfomed. We use some recent results of Thorpe et al. [51, 58, 16] who claim that the neural information is coded by the relative order in which these neurons fire. This allows to select a limited set of relevant points to be used in the motion classification. Several experiments and comparisons with previous neurological work and models are proposed. The result of these simulations show that information from early visual processes appears to be sufficient to classify biological motion.

Key-words: Motion classification, biological motion, spiking neurons, rank order coding schemes, support vector machine

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Des mécanismes de vision précoce suffiraient-ils pour classifier des mouvements?

Résumé : La reconnaissance de mouvements biologiques fait référence à notre capacité à reconnaître une scène (un geste ou un mouvement) à partir de l'évolution d'un nombre limité de points acquis par exemple avec un système de capture basé sur des amers collés sur le corps. Beaucoup de travail a été fait dans cette direction, montrant qu'il est tout à fait possible de reconnaître une action à partir de telles données. A partir du travail de référence de Giese et Poggio [27], nous expérimentons ce paradigme, à partir d'une séquence d'images en entrées, en extrayant ces amers, considérant un réseau de neurone à spike, avec un mécanisme de codage par rang. Nous utilisons les résultats récents de l'équipe de Thorpe [51, 58, 16] dont l'hypothèse est que l'information neuronale, à ce niveau, est codée par l'ordre relatif dans lequel les neurones émettent leur premier spike. Cela permet de sélectionner un nombre limité de point à soumettre au classificateur. A partir de ces données, on vérifie qu'il est possible de faire une classification correcte de mouvements biologiques. Plusieurs simulations et comparaisons avec des travaux en neurophysiologie sont proposés ici. Le résultat de ces simulations montre que l'information issue des mécanismes visuels précoces semble suffisante pour la classification de mouvements biologiques.

Mots-clés : Classification du mouvement, mouvements biologiques, neurones à spike, codage neuronal par rang, machine à vecteurs support

Contents

1 Introduction

1.1 Biological visual classification

Biological visual classification is a well-known and very common, but still intriguing fact. In the present work, data classification simply means being able to put a *unique label* on a given data input (e.g. “oh, there is a dog”). This differs from *categorization* (e.g. [3]) where not only a *label* but a more complex “semantic structure” is extracted from a given data input.

Recent series of experiments have enlightened this biological mechanism: data classification can be realized in the human visual cortex with latencies of about 150 ms [57] and even faster, [55] which, considering the visual pathway latencies [42], may only be compatible with a very specific processing architecture and mechanism [58]. Even "high level" visual data classification such as face recognition [17] can be realized at such a very fast rate.

It has been hypothesized that the underlying neural mechanism is based on a rank order coding scheme [21]: the neural information is coded by the relative order in which these neurons fire. The connexionist "Delorme and Thorpe" classification model presented in [56] is a biologically plausible model of this mechanism. It is based on spiking networks of neurons (quite different from usual neural networks, see e.g. [24] for a discussion).

Surprisingly enough, this experimental evidence is in coherence with algorithms derived from the statistical learning theory, following the work of Vapnik [61, 60]. More precisely, there is a double link: on the one hand the statistical learning theory offers tools to evaluate and analyze such biological models, and on the other hand the Delorme and Thorpe model is an interesting front-end for algorithms derived from the statistical learning theory.

This piece of theory has however never been experimented, regarding motion classification.

1.2 Biological motion classification

Biological motion recognition refers to our ability to recognize a scene (motion or movement) based on the evolution of a limited number of points acquired

for instance with a motion capture tool. M. Giese and T. Poggio [27] propose a biologically plausible neural model (not based on spiking neurons) for the recognition of biological motion and action, based on the availability of neurophysiological and imaging studies and experimental results. This section reminds the main ideas of their work.

The model is based on the key assumption that action recognition is based on learned prototypical patterns and exploits information from the ventral and the dorsal pathway. According to the model, the two pathways process form (ventral) and optic flow (dorsal) information (see Figure 1). Each pathway consists of a hierarchy of neural feature detectors with receptive field sizes that increase with the hierarchy level. The hierarchy levels of the form pathway are formed by simple cells, modeled by Gabor filters, more complex cells that respond maximally for oriented bars independent of their exact spatial position, view-tuned neurons selective for body poses, and motion pattern-selective neurons that are selective for the whole movement sequence. The hierarchy levels of the motion pathway are motion (energy) detectors, detectors for local optic flow field patterns (translation, expansion, and contracting flow), neurons selective for complex instantaneous optic flow patterns, and motion pattern-selective neurons. Each level in the hierarchy can be associated with areas in the macaque and human brain that contain neurons with similar properties.

The model was programmed and tested with stick figures performing different types of motions such as walking, limping, running and intermediate morphs. The motion was captured by manually tracking the joints of actors performing the three main motions and artificially generating the intermediate ones.

The model shows that several principles that are central for the recognition of stationary objects might be important also for the recognition of complex motion patterns. The first principle is a representation in terms of learned prototypical patterns. The second principle is a neural architecture that consists of hierarchies of neural detectors with gradually increasing feature specificity and invariance.

The model's architecture seems to be adequate to account for the invariance properties with respect to stimulus position, scaling, and speed that are

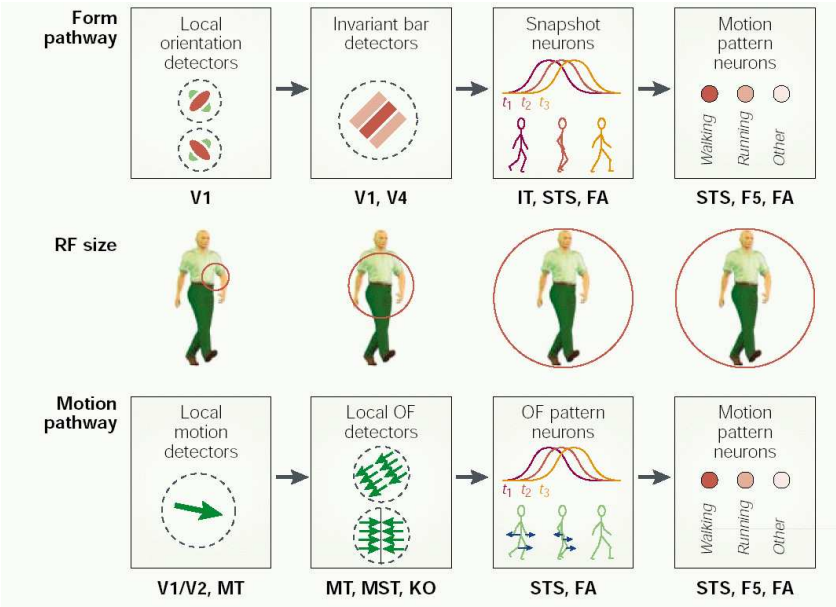

Figure 1: Block diagram of the model proposed by M. Giese and T. Poggio [27] that shows the two pathways for the processing of form and motion (optic flow). The approximate size of the receptive fields compared to typical stimuli is indicated in the middle row. Abbreviations: IT, inferotemporal cortex; KO, kinetic occipital cortex; OF, optic flow; RF, receptive field; STS, superior temporal sulcus; V1, primary visual cortex. Other abbreviations indicate corresponding areas in monkey and human visual cortex.

characteristic for recognition of biological motion. An important additional assumption in the model is the existence of recurrent neural network structures that associate sequential information over time. This assumption leads to predictions that can be physiologically tested, such as the existence of asymmetric lateral connections between motion pattern selective neurons. The prediction of the model that the recognition of biological movements is possible with the information from each pathway alone is consistent with clinical results showing that patients with lesions that include either only the human equivalent of IT, or the MT/V5 complex are still able to recognize complex biological move-

ments when the STS is spared. Only bilateral lesions of the STS have been reported to lead to severe deficits in the perception of biological movements.

Despite these results, the way biological motion is recognized is still an open problem. On the one hand [4] suggest that biological motion can be derived from dynamic *form* information without local image motion, on the other hand [10, 9] propose a new type of point-light stimulus which suggests that the detection of specific spatial arrangements of *opponent-motion features* can explain our ability to recognize actions.

1.3 Computer science related work

This section briefly reviews some related work from the computer science field.

In Computer Graphics

Using motion capture systems is commonly used in the movie making industry for special effects. They allow to have some real time acquisition of the joints positions. Based on the latters it is possible to animate some avatars. This data can also be used to analyze, recognize and generate motions. Many research has been carried out in this direction to synthetize new smooth motions from motion capture database (see for example [36, 28, 26]).

In Computer Vision

Instead of considering marked points positions across time, research in computer vision tries to handle directly the image sequence. There exists a wide litterature on event-based analysis of videos and our aim here is simply to remind some ideas. As far as human action recognition is concerned, many approaches have been proposed (see [2] for a review). Many approaches have been proposed, based on generic human model recovery [29, 31, 49], motion body parts tracking [53, 22], on motion periodicity analysis [14, 15, 48, 52], or based on new representations. This is a difficult task to find a right representation which allows to classify correctly novel data with a stored one. Many representations have been proposed such as Temporal Templates in [5], marginal histograms of spatio-temporal gradients at several temporal scales

[63], or motion motion descriptors [54, 39, 20].

This brief overview shows that the two communities have developped methods and applications based on different kinds of input, either stick representations (i.e. points) or image sequences. To bridge the gap, one possible question is how to extract from image sequences some points which could help for example for motion recognition. Such idea has been recently proposed in [37] where the authors show how to automatically extract the corners of the 2D+t volume (formed by the sequence) and use them for video interpretation. In this paper another proposition is discussed.

1.4 The present contribution

The Giese and Pioggio work is indeed a reference in this domain. As mentioned by the authors, in their experimental set they only consider stick figures (avoiding figure/background segmentation and eliminating some uncertainty in the figure detection), yielding these very promising results. In this work, the aim os to demonstrate that their framework is robust enough to deal with "true" images instead of stick figures. To show it, we will use the same data set as in [26], kindly given by the authors.

A step further, Giese and Pioggio consider general biological models where the brain activity is represented by a continuous scalar variable (e.g. related to the neural spike frequency) which is a valid assumption at this level of modelisation (see e.g. [13] for a discussion) but does not strictly corresponds to the true neural encoding (which is to be related to the spike train itself). Instead, we will consider networks of integrate-and-fire neurons using rank order coding schemes [16], which seem to be much more related to what is really encoded in the brain, at this small latency scale. Following this track, we would like to revisit the Giese and Poggio model (in fact a subset of it considering processing in V1 and MT) but using integrate-and-fire neural models. This is the second reason of this work.

A final reason of redoing Giese and Pioggo experimentation was to compare their results with "low level cues" in the following sense: in order to discriminate motions, do we need to consider long term features (i.e. trajectories of every joints which corresponds to global motion) or is it sufficient to work with

short term trajectories (i.e. local motion operators). In the original study, global displacements are integrated along the simulated dorsal/ventral visual pathways. In this comparative study we consider local motion cues only. The seminal idea of this choice is related to the Rubin work [50, 11] on segmentation where it is shown that early visual processes could be sufficient to perform the task: could also early visual processes be sufficient to label motions ? We simulate this situation here to help understanding this fact.

2 Classifying Motion with Points of Interest

2.1 System Overview

The computational problem that is addressed by this work is the recognition of biological motion in image sequences. Here we would like to focus on a biologically plausible mechanism considering the architecture of the brain [8].

The general problem is sub-divided into two main stages (see Figure 4). the first is the feature extraction and the second stage is the classification problem.

The feature vector extraction block

It takes as input a raw sequence of images as the ones shown in Figure 3 and extracts features from the sequence. Features could be edges, local motion computation, etc... In the brain, this typically corresponds to the V1 output. It is a huge map of values. In a computer system it is delivered in the form of a feature vector to the classifier block. The objective is to select the best features that will lead to the more robust classification. In particular we will investigate if rank order coding can be useful.

The classification block

It takes as input the feature vector and classifies it. It tells which class each feature vector belongs to. In our case it will be the type of action that the character is doing. Two classifiers are considered: the nearest neighbor (called RAW) [19] and the SVM [59, 30, 12] classifiers. The RAW classifier is a simple classification method based on minimum distances. High dimensional

training data (feature vector) together with the class it belongs to, is fed into the classifier. At least one training sample from each class is shown to the classifier. Later the distances [19], usually in a high dimensional space, between the sample to be classified and all the training samples previously fed to the classifier during the training stage, are calculated. The classification of the new sample is done by assigning it the same class as the training sample found with the smallest distance to it.

Experimental setup

The motion recognition simulation will be tested on a set of 40 biological motion image sequences (video samples) from two classes, walking and marching (see Figure 3).

The performance analysis consisted in starting the learning phase with one randomly choosen feature vector from each class and repeatedly incrementing the data sample (feature vector) in the walking and marching classes. When the training is completed the resting data samples are used as the testing set to quantify the classifier's error rate. The number of errors that the classifier commits are recorded and the error percentage is calculated over the total testing set.

The simulation process involved training the classifier with one sample from each category (class) and then using the rest of the data samples to test the error rate of the classifier. After completing the test a new sample of the test data from each category is added to the training set and the resting data is used to test the classifier. These steps are repeated successively until almost all the data is used as training data. Therefore as the training set grows the testing set is reduced and a smaller error rate is expected. This is illustrated in Figure 2 where after twelve training samples the classes are well defined.

This training and testing process is repeated until the testing set size is minimal (when the set size is equal to the number of classes). In this manner a more complete idea of the classification performace (the rate of correct classifications) is obtained for different training/testing set size ratios.

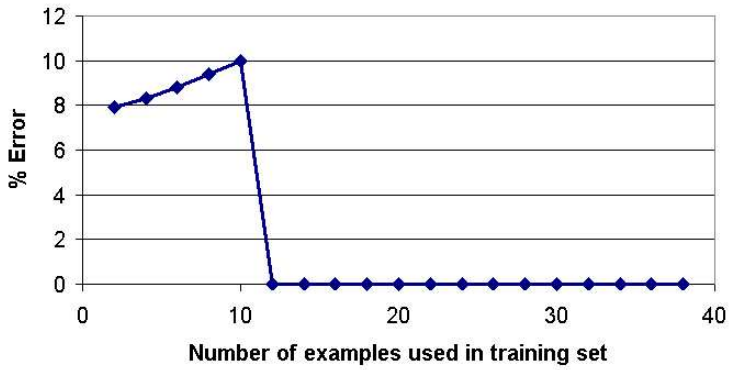

Figure 2: RAW classifier error rate obtained on the 40 Giese trajectories discriminating between walk and march motions. On the vertical axis is plotted the percentage error rate versus the amount of samples used to train the classifier.

2.2 Classifying Motion with Trajectories

2.2.1 Feature Vector Description

As dealt in [27, 25] the first technique is to extract manually the joints positions from the video samples of a subject repeatedly performing the two types of motions. Each feature vector consists of 12 spatio-temporal trajectories of the joints of a subject performing a motion, in Figure 5 the 12 joints are shown. There were roughly 20 different samples of the same subject walking and another 20 of marching.

The trajectories of each motion sample (video) are stored in a $12\times2\times20$ matrix, where each element in the matrix corresponds to an x or y coordinate of a joint in a specific time instant as shown in Figure 5. Figure 6 shows the smooth evolution of the joints position with time.

The coordinates are all relative to the hip joint, in other words the hip joint is considered as the origin. To generate the classifier feature vectors for the training and testing (classification) the three dimensional trajectory matrices are reshaped in a consistent manner to form a one column vector.

Figure 3: An example of each class of the Giese [25] image sequences database that were used to test the motion recognition. These classes of motions were chosen since walk and march are quite similar and thus more challenging to classify.

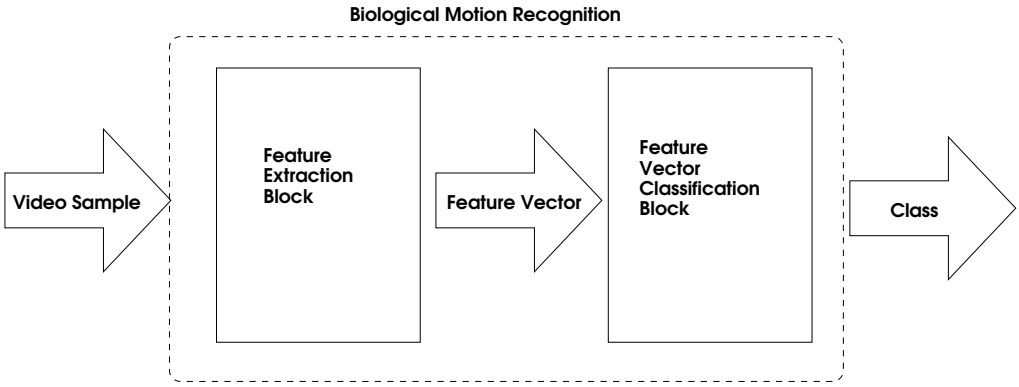

Figure 4: Block diagram of the Motion Recognition problem.

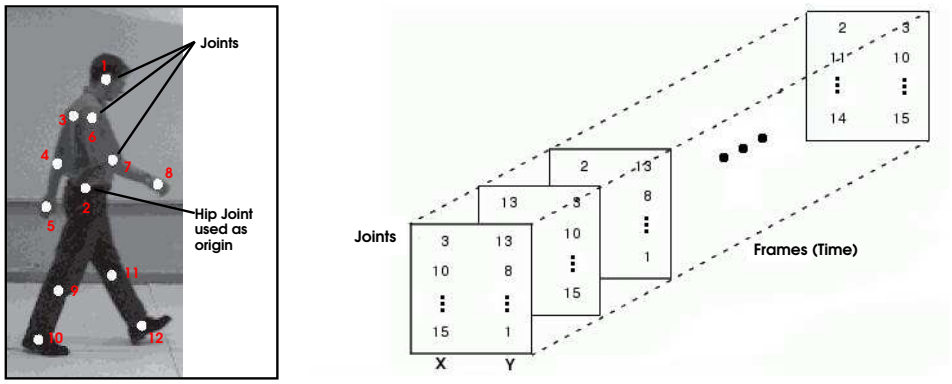

Figure 5: Giese approach input. Left: The joints positions are manually labelled. Right: Every position is then stored in a 3D matrix which can then be coded as a 1D feature vector.

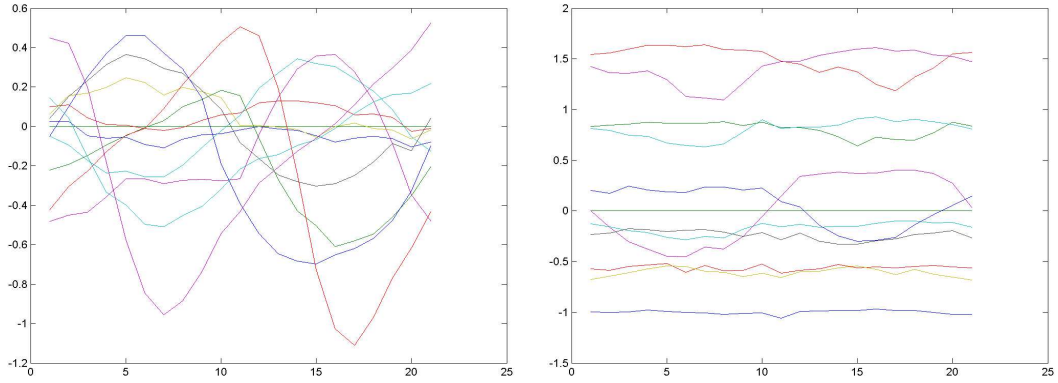

Figure 6: Evolution of the joints positions which will be stored at the same positions in the feature vector description. The left and right hand side graph are x and y coordinates respectively.

The reshaping order is not so important, the critical thing being to perform the exact same reshaping on every three dimensional trajectory matrices so that the elements in the training vector are consistant between eachother: the same row of two feature vectors should correspond to the same joint and time.

2.2.2 Classifiers Performance

Figure 7 shows the results obtained the RAW (black dots) and SVM classifiers according to he procedure described in section 2.1. In the case of SVM there were also a number of parameters that could to be adjusted. The main parameters are the Kernel Type (Linear, Polynomial, Radial Basis, Sigmoid), the Kernel Degree (for the polynomial classifier) and the error tolerance parameter. The manner in which these paremeters infered on the classification performance has been explored.

In this particular case, the RAW classifier seems to have a better performance than SVM. Such a fact is dicussed in [61]. Obviously, for the largest polynomial degree, with a large number of degrees of freedom, the estimation is unstable. Other estimations have very similar performances. As a consequence we are going to consider only one SVM estimator: the polynomial kernel degree 2 estimator.

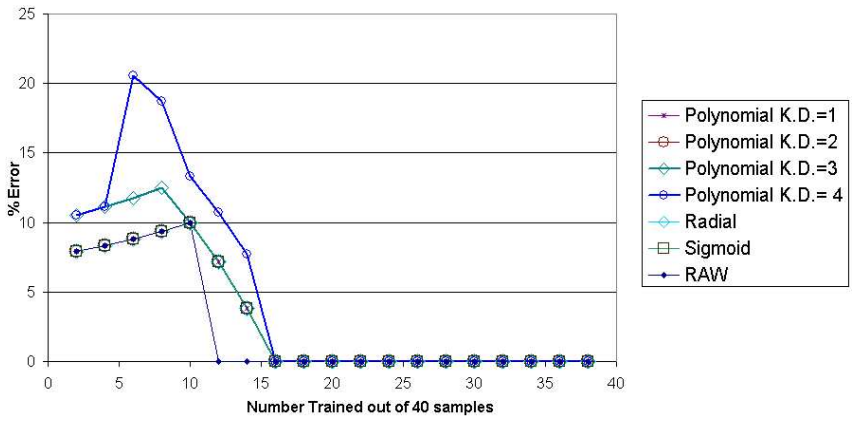

Figure 7: Motion classification performances using different classifiers

2.3 Classifying Motion with Spike Responses

2.3.1 Feature Vector Description (overview)

It is known that biological systems use spike coding [16, 51] to transmit first the most relevant information in an image. Based on this idea it has been shown previously that using such coding it is possible to extract the most important information from a static image in order to generate ultra fast classification [58]. The open question is weather such coding is also good for classifying biological motion. In other words weather spike coding is also able to extract rich information from every frame in a sequence of images and be able to classify with a good precision the types of motion?

The main difference in these experiments is the way the classification feature vectors are defined from the raw video samples of the two motions types. The different steps are summarized in the Figure 8) and are detailed in the next section. Figure 10 shows qualitatively the temporal relation between vector features elements. Contrary to joint position tracking which provides a smooth global information on body segment displacements (see Figure 6), spiking neurons output is a noisy version of this signal with only local temporal relations (small piece of joint position trajectory) separated by random jumps of the signal. As such, we claim that information in the spiking neurons output is mainly related to local motion information.

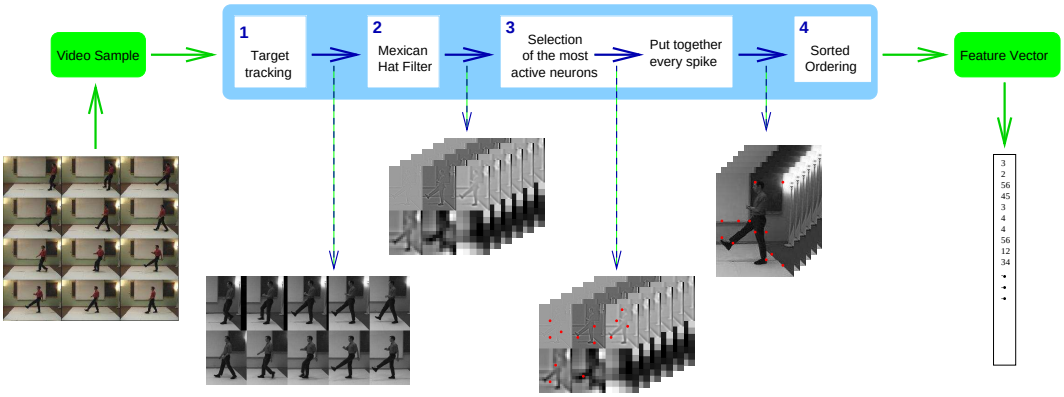

Figure 8: Overview of the feature vector extraction block.

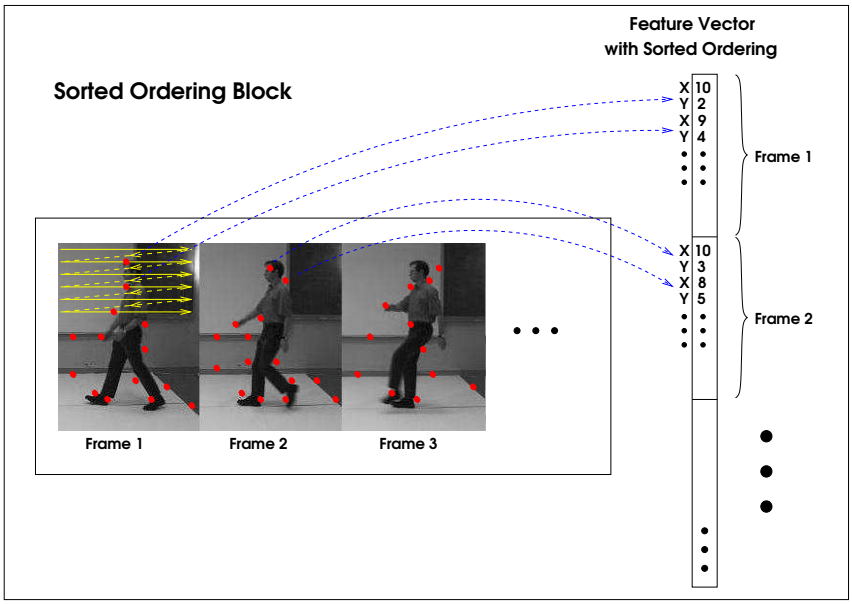

Figure 9: This figure shows the action that is performed by the sorted odering block which finally generates the feature vector. In frame 1 the scanning pattern that generates the sorted element ordering in the output feature vector, is shown.

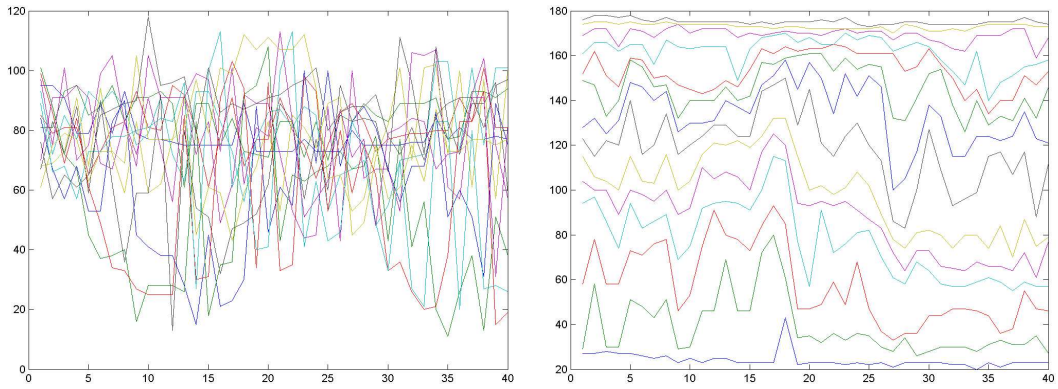

Figure 10: Examples of feature vectors given to the classifier using relevant spikes detection, automatically calculated from the spiking neural net . The left and right hand side graphs are x and y coordinates respectively.

2.3.2 Feature Vector Description (detailed steps)

Step 1: Target tracking

The goal is to obtain a body-centered sequence of the action. This was done in order to resemble the sequences that are analyzed by the human visual system while its tracking an object. As the eye moves, tracking a target while the target is displacing itself, the target is approximately centered onto the retina, this same effect was generated by cropping the videos to center the target in the frame. This also creates a sort of relative positioning, to the target coordinate system as the center of the target is approximately always in the center of the image. This was also used in [25] since every point coordinates are relative to the hip positions. From an implementation point of view, a simple moving object detector can be obtained using a thresholding technique over the *inter-frame difference* between a so-called *reference image* and the image being observed. Decisions can be taken independently point by point [62]. More complex approaches can also be used [46, 45, 47, 1, 32, 38, 6, 33, 34, 23, 18, 41]. In our case we have used the variational approach developed in [35] which allows to obtain a robust segmentation for noisy image sequences. The motion segmentation (separation foreground versus background) and the construction of a restored background are done in a coupled way, allowing the motion segmentation part to positively influence the restoration part and vice-versa. This approach is fully automatic and can be implemented for video-streams (i.e. with causality). From the binary masks obtained (every pixel is labelled as foreground or background), the most important foreground connected component is centered in a fixed size bounding box which is used to crop the video.

Step 2: Mexican hat filters

Each frame in the cropped videos was transformed using the Mexican hat transfer function in six different scales using the *matlabPyrTools* toolbox from Simoncelli[1] and the resulting coefficients from all the scales were normalized

[1]This Matlab source code for multi-scale image processing is available at http://www.cns.nyu.edu/~eero/software.html. It includes tools for building and manipulating Laplacian pyramids, QMF/Wavelets, and steerable pyramids.

(as explained [16]). The highest resulting normalized coefficients (which are proportional to the neural spike frequencies) and their coordinates within the frame were saved.

Step 3: Selection of the most active neurons using local inhibition

For each frame the procedure used to extract the most relevant spikes is the same as the one used in [51, 58, 16] (see also [40, 43]).

Step 4: Sorted ordering

As mentioned previously in order to generate the classification vector the spikes were sorted with respect to their position (coordinates x and y) so that the spikes which were on the top left hand corner of a frame are placed first and the ones that are in the bottom right hand corner are placed last (see Figure 9), this form of feature vector element ordering is called sorted ordering, due to the fact that the spikes are sorted according to their position in the image/frame. In that way there is a correspondence between the position of the spike's coordinate within the feature vector and the area of motion that generated the spike. Feature vectors were generated for the 20 Giese walking motion samples the other 20 marching samples, based on these vectors the classification performance was tested.

In Figure 11, an extraction of the top 40 spike positions over a few frames of the input motion film are shown, in order to illustrate what is a typical feature set.

2.3.3 Classifiers Performance

Feature vectors were generated for the 20 Giese walking motion samples and the other 20 marching samples. For each category feature vectors taking into account the top 10, 20, 30, 40 and 50 spikes were generated. This was done in order to test how sensitive the classification error is to the amount of spikes taken into account. Based on these vectors, classification performance was tested and the configurations with the lowest average classification error rate are shown together with the RAW Giese curve in Figure 12.

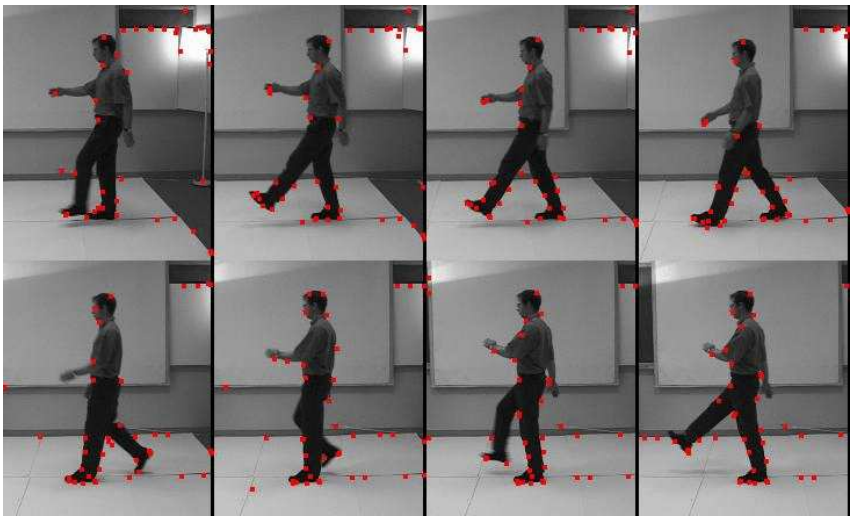

Figure 11: Top 40 spike positions over a few frames of the input motion film.

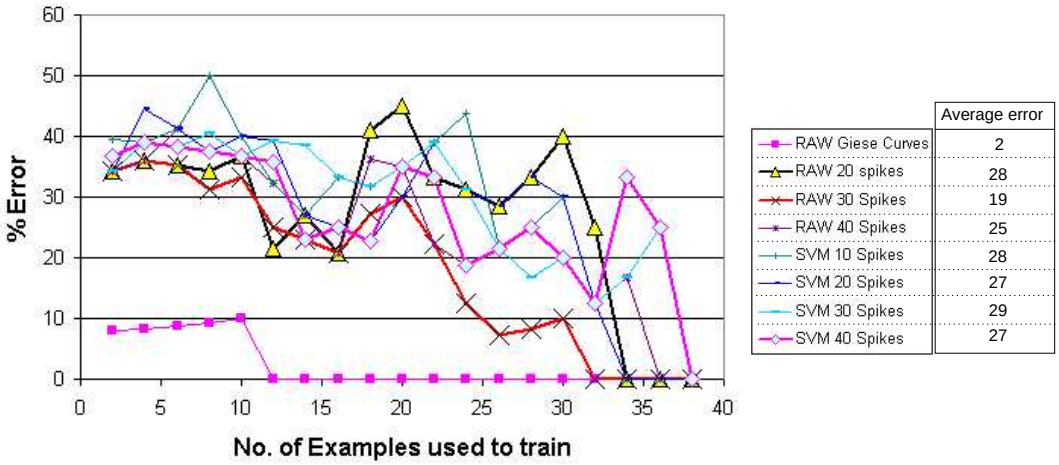

Figure 12: Motion classification performance based on the most important spike positions.

In Figure 12, the average classification errors of the different methods and features vector combinations are shown. From the figure, it is clear that the lowest average classification error of all the methods is the RAW classifier with the 30 spike feature vectors (RAW 30 spike).

Filtering Spikes from the Background Improve Classification

In order to improve classification performance, it has been tested to remove spikes coming from the background. This segmentation of the visual flux into layers (foreground versus background) is something which is performed in the brain by the MST area [8, 44] which "justifies" this idea. The results obtained from the segmentation step (see Figure 14) can be used to discard spikes from the background. The resulting relevant spikes ar shown in Figure 15 and its impact on classification performance is presented in Figure 13.

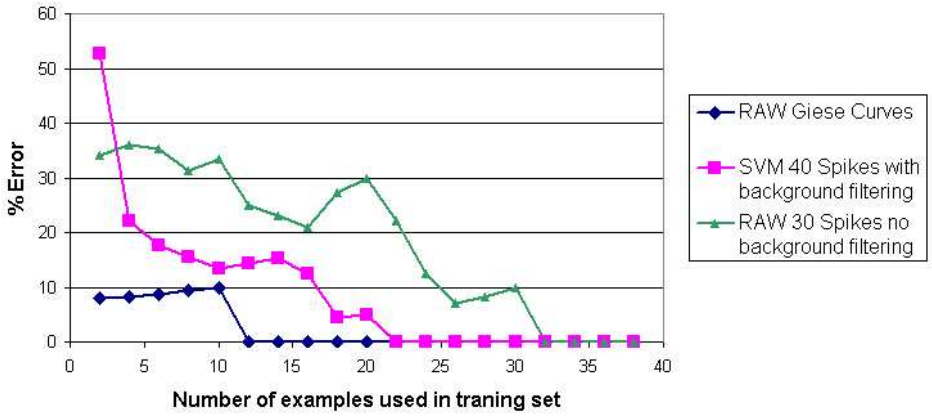

Figure 13: Filtering spikes from the background to improve classification. Note that choosing 30 or 40 spikes isn't significative as shown in Figure 12.

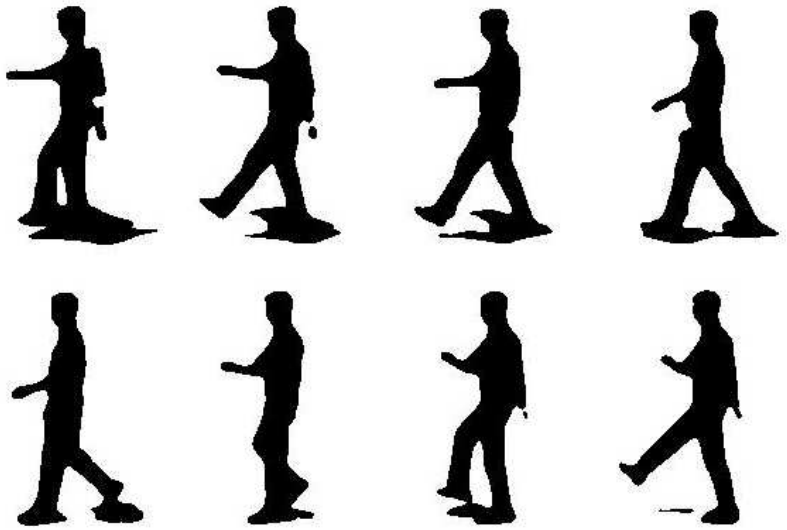

Figure 14: A sample of the binary masks obtained to discriminate between background and foreground spikes in the background filtering.

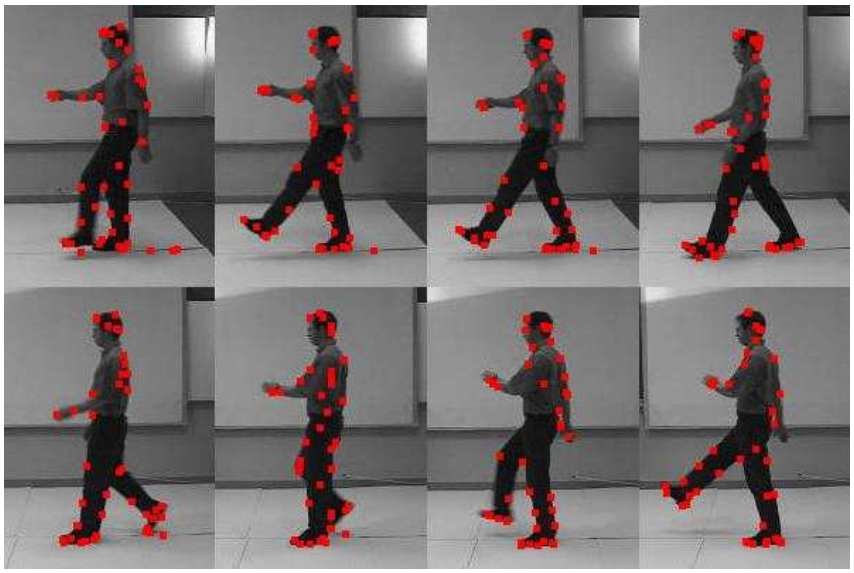

Figure 15: Top 40 spike positions over a few frames of the input motion film with background filtering.

Some Sequences Bring More Informations To The Training Phase

Although this does not correspond to a standard statistical paradigm, we have rerun the experiment, by choosing the "best" samples for training. In other words, this corresponds to a sample selection by an "expert" not a random sampling. If the richest, in information, samples or vectors are obtained and fed first as training vectors to the classifier much better results are obtained as shown in Figure 16, very similar, when not better to what is obtained with tragectories.

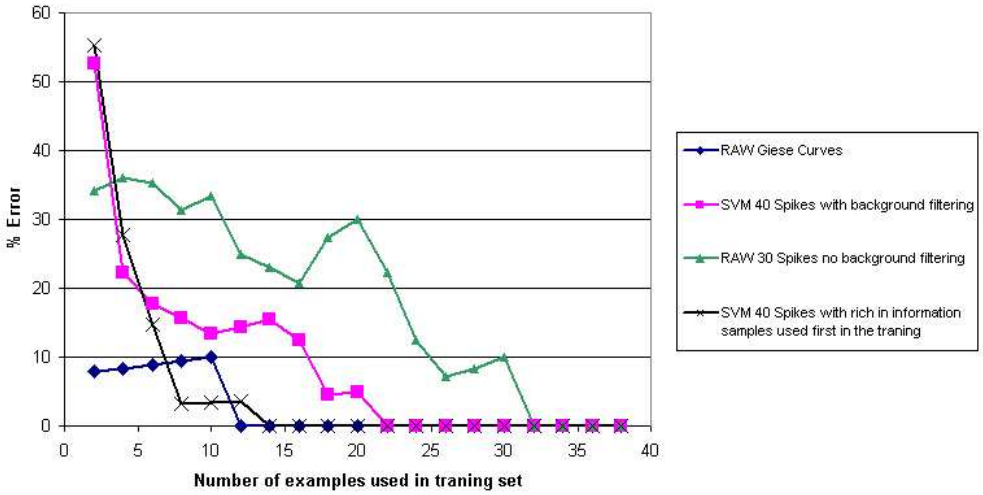

Figure 16: The figure shows that much better results can be obtained if the richest in information samples are selected as the first training samples.

Intervector Element Ordering on Classification isn't Crucial

A very legitimate question is, "How important is the impact of the previously mentioned feature vector element ordering system (see Figure 9) on the classification performance with the Giese data?" or "Is it crucial that the spike coordinates generating meaningful trajectories be sorted and grouped in the classification vector?". To answer this question we performed the following experiment on the Giese trajectories. Sorted ordering which is done by sorting the elements within the feature vector based on their location within the frame as described in Experimental Setup in subsection Classifying Motion with Spike Responses was compared to position-trajectory ordering were every position in the vector belongs to a specific trajectory (as was done in subsection Classifying Motion with Trajectories). The effects of the sorted ordering with respect to the position-trajectory ordering are shown in Figure 17. The Giese trajectories were used to compare the two types of ordering because it is complicated to do trajectory tracking on noisy data such as the top spike position data. With the Giese motion trajectories formed by the coordinates of the junctions over time, it is simple to generate position-trajectory ordering within the classification vectors.

As can be seen in Figure 17 the feature vector sorted ordering does not significantly increase the classification error rate with respect to position-trajectory ordering. This suggests that trajectory tracking is not a very crucial part of motion classification.

The main difference in these two conditions was the way the classification vectors (or the feature vectors) are generated from the Giese motion trajectories. On one hand, vectors were generated with position-trajectory ordering in other words they are the exact same vectors that were used to measure the classifier performance in subsection Classifying Motion with Trajectories. In these classification vectors each position within the vector belong to a certain trajectory of a certain joint and that is consistent throughout all the 40 samples. To generate the classification vectors with sorted ordering the same Giese trajectory data was used but within the feature vector the elements from each frame were sorted according to their position within the frame so that the coordinates of the different trajectories which were on the top left hand

corner of a frame are placed first and the ones that were in the bottom right hand corner are placed last (see Figure 9).

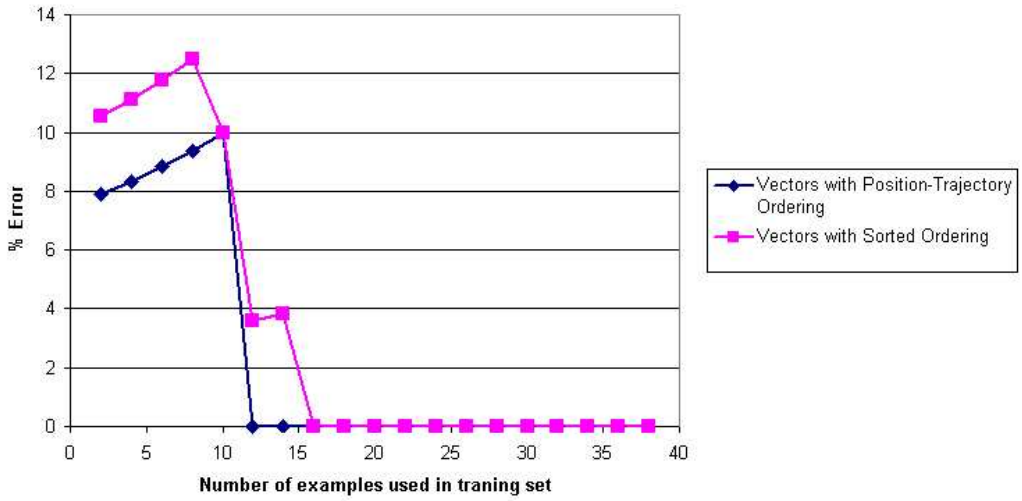

Figure 17: The impact of the feature vector element ordering on classification performance.

3 Conclusion

Following the reference work of Giese and Poggio [27], we have experimented considering a video sequence as the input and how to use spiking neural networks with rank order coding in order to extract relevant tokens. Using this data, we have verified that correct biological motion classification can be perfomed. This was not obvious since the data was "noisy" and points do not correspond to fixed joint's location but are quite noisy temporaly. This however, is in coherence with some recent studies [4, 10, 9] which are showing that motion recognition is still possible with perturbations on the marker's positions.

More precisely, we have selected the most relevant spike for every frame independently of their temporal organisation. This allows to select a limited set of relevant points to be used in the motion classification. A spatial matching pursuit like algorithm was implemented for that purpose. This corresponds to the recent theory of Thorpe et al. [51, 58, 16] who claim that the neural information is coded by the relative order in which these neurons fire.

A step further, these points are not "tracked" all along the sequence but simply related to nearby points in the previous frame, in other words only *local motion* cues have been taken into account. This is a key issue, because this result is coherent with the fact that information from early visual processes appears to be sufficient to classify biological motion. This does not means that "all is done" at this early stage but it suggests that such "cognitive" task can be realized in the so called "fast brain". The key idea behind this is that the complete perception uses feedbacks from early vision processes in order to drive the latter perception as discussed in [7].

As far as computer vision is concerned, this means that we can label such motion automatically and using feed-forward process (local motion cues feeding a SVM) useful in "real-time" systems. As dicussed in [60] the use of statistical learning theory is a key issue to obtain such functionnality.

Finally this experimental work of simulation, clearly confirm what Giese and Poggio [27] pointed out when proposing a model of cortical motion perception, and this work has been able to verify that the theory directly apply on raw image sequences, using spiking neurons and local motion detector. This experimental work also confirm that Thorpe et al. [58, 16] model is not only valid for static images, but should be valid for image sequences. Therefore, future work will consider the coding of the sequence using temporal causality, which means adapting the 2D coding done here to a 2D+t case. It is expected that the selected points will then be more relevant in terms of spatio-temporal events and then may be more suitable to classify motions.

Acknowledgments

The Giese dataset has been kindly provided by Dr. Giese and the present work has been realized thanks to this data set.

We are especially thankfull to Simon Thorpe and his group for powerfull ideas which are at the origin of this work. This work has been realized thanks to our common participation in the http://www-sop.inria.fr/odyssee/contracts/rivage ACI project.

The *matlabPyrTools* toolbox from Dr. Simoncelli (available online) was used to build and manipulate Laplacian pyramids. We would like also to thank Laurent Perrinet for providing us an example of matlab code using this library in order to decompose an image into spikes.

A Completing the [60] dataset

As an extra safeguard of the validity of the results that were obtained, two extra experiments were realized. Results are reported here, making profit of the experimental setup, although not directly related to the previous development.

Showing what the method can do: a simple example

The first was using the same method as in subsection "Classifying Motion from Spike Responses" to classify samples from two categories rotated and non rotated (see Figure 18). The samples in class "true" a logo with a different size, orientation on added noise. The class "false" is a non-logo, where the logo is mirrored or warpped. The classification results using the two types of methods RAW and SVM can be seen in Figure 19, there SVM classification based on spiking neurons is very powerfull, close to 0% after a traning with 100 samples.

As expected, classification performance is good because it is based on the location of the top spikes that are usually along edges. Here the logo caracteristic is indeed related to the edges. This result is going to be compared with Thorpe et al. spikenet results in a near future.

Showing what the method does not do: the animal detection example

The second experiment is just to remind us that although very efficient, we are far from what the brain can do. Let us see how the same algorithm as the one used in the previous experiment performs on classifying the presence of an animal in natural images (see Figure 20). The results can be seen in figure Figure 21.

Clearly it does NOT work, where as a primate brain performs this task with a succes better than 90% and in 100-150 msec [58]. Visual clasification is still an intriguing fact.

Figure 18: An example of each class of the logo images.

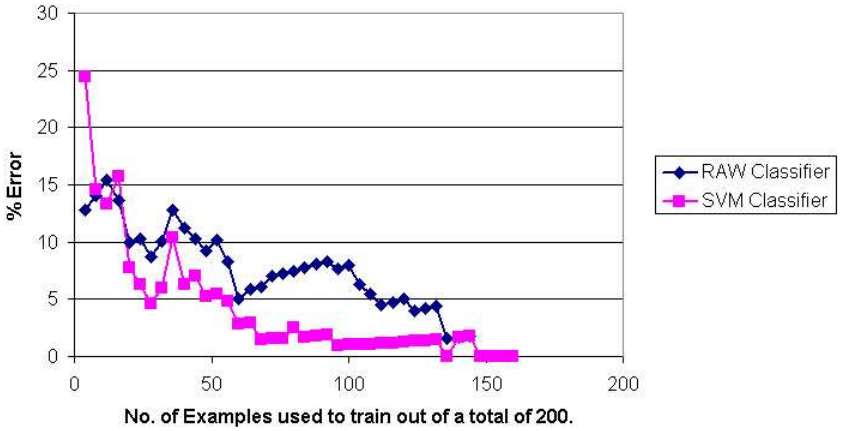

Figure 19: Classifying static logo images using the same algorithm as in subsection "Classifying Motion from Spike Responses" .

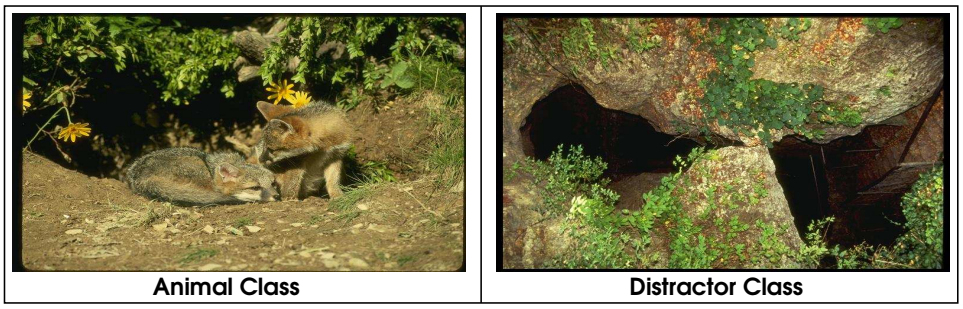

Figure 20: An example of each class of the animal images.

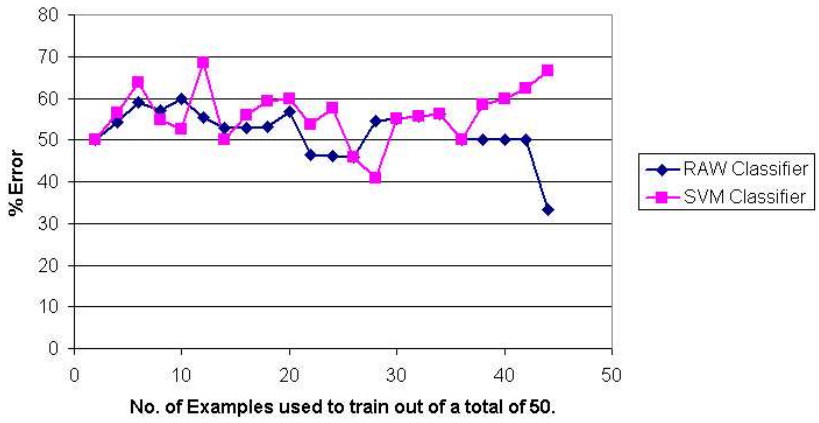

Figure 21: Classifying static animal images using the same algorithm as in subsection "Classifying Motion from Spike Responses" .

References

[1] T. Aach and A. Kaup. Bayesian algorithms for adaptive change detection in image sequences using markov random fields. *Signal Processing: Image Communication*, 7:147–160, 1995.

[2] J. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[3] Ruzena Bajcsy and Franc Solina. Three dimensional object representation revisited. In *Proceedings of the 1st International Conference on Computer Vision*, London, England, June 1987. IEEE Computer Society Press.

[4] J.A. Beintema and M. Lappe. Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences of the USA*, 99(8):5661–5663, 2002.

[5] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.

[6] J. Boyce. Noise reduction of image sequences using adaptative motion compensated frame averaging. In *IEEE ICASSP*, volume 3, pages 461–464, 1992.

[7] J. Bullier. Integrated model of visual processing. *Brain Res. Reviews*, 36:96–107, 2001.

[8] Y. Burnod. *An adaptive neural network: the cerebral cortex.* Masson, Paris, 1993. 2nd edition.

[9] A. Casile and M. Giese. Is form information necessary for the recognition of point-light walkers? In *ECVP*, 2003.

[10] A. Casile and M. Giese. Roles of motion and form in biological motion recognition. *Artifical Networks and Neural Information Processing, Lecture Notes in Computer Science 2714*, pages 854–862, 2003.

[11] C. Caudek and N. Rubin. Segmentation in structure from motion: modeling and psychophysics. *Vision Research*, 41:2715–2732, 2001.

[12] Chih-Chung Chang and Chih-Jen Lin. Training nu-support vector classifiers: Theory and algorithms. *Neural Computation*, 13(9):2119–214, 2001.

[13] J. Chey, S. Grossberg, and E. Mingolla. Neural dynamics of motion processing and speed discrimination. *Vision Res.*, 38:2769–2786, 1997.

[14] R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *5th Intl. Conf. on Automatic Face and Gesture Recognition*, 2002.

[15] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), August 2000.

[16] A. Delorme, L. Perrinet, and S. Thorpe. Network of integrate-and-fire neurons using rank order coding b: spike timing dependant plasticity and emergence of orientation selectivity. *Neurocomputing*, 38:539–545, 2001.

[17] A. Delorme and S. Thorpe. Face processing using one spike per neuron: resistance to image degradation. *Neural Networks*, 14:795–804, 2001.

[18] E. Dubois and S. Sabri. Noise reduction in image sequences using motion-compensated temporal filtering. *IEEE Transactions on Communications*, 32(7):826–831, July 1984.

[19] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification, 2nd edition.* Wiley Interscience, 2000.

[20] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the 9th International Conference on Computer Vision*, volume 2, pages 726–734, Nice, France, October 2003. IEEE Computer Society, IEEE Computer Society Press.

[21] J. Gautrais and S. Thorpe. Rate coding vs temporal order coding : a theorical approach. *Biosystems*, 48:57–65, 1998.

[22] D.M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[23] Stuart Geman, Donald E. McClure, and Donald Geman. A nonlinear filter for film restoration and other problems in image processing. *CVGIP: Graphics Models and Image Processing*, 54(4):281–289, July 1992.

[24] W. Gerstner and W. M. Kistler. Mathematical formulations of hebbian learning. *Biol Cybern*, 87:404–415, 2002.

[25] M.A. Giese and M. Lappe. Measurement of generalization fields for the recognition of biological motion. *Vision Research*, 38:1847–1858, 2002.

[26] M.A. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion patterns. *International Journal of Computer Vision*, 38(1):59–73, 2000.

[27] M.A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and actions. *Nature Reviews Neuroscience*, 4:179–192, 2003.

[28] M. Gleicher, H.J. Shin, L. Kovar, and A. Jepsen. Snap together motion: Assembling run-time animation. In *Symposium on Interactive 3D Graphics*, 2003.

[29] L. Goncalves, E. DiBernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3D. In *Proceedings of the 5th International Conference on Computer Vision*, pages 764–770, Boston, MA, June 1995. IEEE Computer Society Press.

[30] Y. Guermeur and H. Paugam-Moisy. Théorie de l'apprentissage de vapnik et svm, support vector machines. *READ*, 12(5):517–571, 1999.

[31] D. Hogg. Model-based vision: a paradigm to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[32] K. Karmann, A. Brandt, and R. Gerl. Moving object segmentation based on adaptive reference images. *Signal Processing: Theories and Applications*, V:951–954, 1990.

[33] A. Kokaram. Reconstruction of severely degraded image sequences. In *International Conference on Image Applications and Processing*, Florence, Italy, 1997.

[34] A. Kokaram and S.J. Godsill. A system for reconstruction of missing data in image sequences using sampled 3d ar models and mrf motion priors. In Bernard Buxton, editor, *Proceedings of the 4th European Conference on Computer Vision*, pages 613–624, Cambridge, UK, April 1996.

[35] P. Kornprobst, R. Deriche, and G. Aubert. Image sequence analysis via partial differential equations. *Journal of Mathematical Imaging and Vision*, 11(1):5–26, October 1999.

[36] L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. In Kurt Akeley, editor, *Proceedings of the SIGGRAPH*, volume 21. ACM Press, ACM SIGGRAPH, Addison Wesley Longman, 2002.

[37] I. Laptev and T. Lindeberg. Space-time interest points. In *Proceedings of the 9th International Conference on Computer Vision*, pages 432–439, Nice, France, 2003. IEEE Computer Society, IEEE Computer Society Press.

[38] S. Liou and R. Jain. Motion detection in spatio-temporal space. *Computer Vision, Graphics and Image Understanding*, (45):227–250, 1989.

[39] J. Little and J. Boyd. Describing motion for recognition. In *Proc. Int. Symp. Computer Vision*, pages 235–240, 1995.

[40] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12):3397–3414, 1993.

[41] R.D. Morris. *Image Sequence Restoration using Gibbs Distributions*. PhD thesis, Cambridge University, England, 1995.

[42] L.G. Novak and J. Bullier. *The Timing of Information Transfer in the Visual System*, volume 12 of *Cerebral Cortex*, chapter 5, pages 205–241. Plenum Press, New York, 1997.

[43] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1998.

[44] G. Orban. Visual processing in macaque area mt/v5 and its satellites (mstd and mstv). In Rockland et al., editor, *Cerebral Cortex*, volume 12, chapter 9, pages 359–434. Plenum Press, New York, 1997.

[45] N. Paragios and R. Deriche. Detecting multiple moving targets using deformable contours. In *Proceedings of the International Conference on Image Processing*, volume II of III, pages 183–186, Santa Barbara, California, October 1997.

[46] N. Paragios and R. Deriche. A PDE-based level set approach for detection and tracking of moving objects. In *Proceedings of the 6th International Conference on Computer Vision*, pages 1139–1145, Bombay, India, January 1998. IEEE Computer Society, IEEE Computer Society Press.

[47] N. Paragios and G. Tziritas. Detection and localization of moving objects in image sequences. *ICS/FORTH Technical Report, Accepted for publication in Signal Processing: Image Communication*, October 1996.

[48] R. Polana and R.C. Nelson. Detecttion and recognition of periodic, nonrigid motion. *The International Journal of Computer Vision*, 23(3):261–282, 1997.

[49] K. Rohr. Toward model-based recognition of human movements in image sequences. *CVGIP, Image Understanding*, 1:94–115, 1994.

[50] Nava Rubin. Figure and ground in the brain. *Nature Neuroscience*, 4:857–858, 2001.

[51] R. Van Rullen and S. Thorpe. Rate coding versus temporal order coding: What the retina ganglion cells tell the visual cortex. *Neural Computing*, 13(6):1255–1283, 2001.

[52] S.M. Seitz and C.R. Dyer. View-invariant analysis of cyclic motion. *The International Journal of Computer Vision*, 25(3), 1997.

[53] M. Shah and R. Jain. *Motion-based recognition*. Computational Imaging and Vision Series. Kluwer Academic Publisher, 1997.

[54] E. Shavit and A. Jepson. Motion understanding using phase portraits. In *Proc. IJCAI Workshop: Looking at People*, 1993.

[55] S. Thorpe. Ultra-rapid scene categorization with a wave of spikes. In *Biologically Motivated Computer Vision*, volume 2525 of *Lecture Notes in Computer Science*, pages 1–15. Springer-Verlag Heidelberg, 2002.

[56] S. Thorpe, A. Delorme, and R. VanRullen. Spike based strategies for rapid processing. *Neural Networks*, 14:715–726, 2001.

[57] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

[58] S.J. Thorpe and M. Fabre-Thorpe. Seeking categories in the brain. *Science*, 291:260–263, 2001.

[59] V. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.

[60] T. Viéville and S. Thorpe. A deterministic biologically plausible classifier. In *8th ICCNS*. Boston University, 2004.

[61] Thierry Viéville and Sylvie Crahay. A deterministic biologically plausible classifier. In *Computational Neuroscience Meeting*, volume 58-60C, pages 923–928. Elsevier, July 2003.

[62] O. Wenstop. Motion detection from image information. *Proceedings in Scandianvian Conference on Image Analysis*, pages 381–386, 1983.

[63] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proceedings of CVPR'01*, volume 2, pages 123–128, 2001.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399