Neural Networks 22 (2009) 425–449

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Invariant template matching in systems with spatiotemporal coding: A matter of instability

Ivan Tyukin [a,b,c], Tatiana Tyukina [a,b], Cees van Leeuwen [a,*]

[a] Laboratory for Perceptual Dynamics, RIKEN (Institute for Physical and Chemical Research) Brain Science Institute, 2-1, Hirosawa, Wako-shi, Saitama, 351-0198, Japan

[b] Department of Mathematics, University of Leicester, University Road, Licester, LE1 7RH, United Kingdom

[c] Department of Automation and Control Processes, St-Petersburg State University of Electrical Engineering, St-Petersburg, Prof. Popova street 5, 197376, Russia

ARTICLE INFO

Article history:
Received 2 May 2007
Received in revised form 4 December 2008
Accepted 26 January 2009

Keywords:
Template matching
Nonlinear parameterization
Weakly attracting sets
Adaptation
Convergence
Lyapunov-unstable

ABSTRACT

We consider the design principles of algorithms that match templates to images subject to spatiotemporal encoding. Both templates and images are encoded as temporal sequences of samplings from spatial patterns. Matching is required to be tolerant to various combinations of image perturbations. These include ones that can be modeled as parameterized uncertainties such as image blur, luminance, and, as special cases, invariant transformation groups such as translation and rotations, as well as unmodeled uncertainties (noise). For a system to deal with such perturbations in an efficient way, they are to be handled through a minimal number of channels and by simple adaptation mechanisms. These normative requirements can be met within the mathematical framework of weakly attracting sets. We discuss explicit implementation of this principle in neural systems and show that it naturally explains a range of phenomena in biological vision, such as mental rotation, visual search, and the presence of multiple time scales in adaptation. We illustrate our results with an application to a realistic pattern recognition problem.

1. Notational preliminaries

We define an image as a mapping $S_0(x, y)$ from a class of locally bounded mappings $\mathcal{S} \subseteq L_\infty(\Omega_x \times \Omega_y)$, where $\Omega_x \subseteq \mathbb{R}$, $\Omega_y \subseteq \mathbb{R}$, and $L_\infty(\Omega_x \times \Omega_y)$ is the space of all functions $f : \Omega_x \times \Omega_y \to \mathbb{R}$ such that $\|f\|_\infty = \operatorname{ess\,sup}\{\|f(x, y)\|, x \in \Omega_x, y \in \Omega_y\} < \infty$. Symbols x, y denote variables on different spatial axes. The value of $S_0(x, y)$ depends on the domain of interest (e.g. brightness, contrast, color, etc.). Our interpretation of images as functions from $L_\infty(\Omega_x \times \Omega_y)$ is motivated mostly by the fact that in the domain of vision the characteristic values are usually bounded. We will treat them as such unless information to the contrary is available.

We assume that within a system an image is represented as a set of pre-specified templates, $S_i(x, y) \in \mathcal{S}$, $i \in \mathcal{I} \subset \mathbb{N}$, where $\mathcal{I}$ is the set of indices of all templates associated with the image $S_0(x, y) \in \mathcal{S}$. Symbol $\mathcal{I}^+$ is reserved for $\mathcal{I}^+ = \mathcal{I} \cup 0$.

The solution of a system of differential equations $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t, \boldsymbol{\theta}, \mathbf{u}(t))$, $\mathbf{u} : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$, $\boldsymbol{\theta} \in \mathbb{R}^d$ passing through point $\mathbf{x}_0$ at $t = t_0$ will be denoted for $t \geq t_0$ as $\mathbf{x}(t, \mathbf{x}_0, t_0, \boldsymbol{\theta}, \mathbf{u})$, or simply as $\mathbf{x}(t)$ if it is clear from the context what the values of $\mathbf{x}_0$, $\boldsymbol{\theta}$ are and how the function $\mathbf{u}(t)$ is defined.

By $L^n_\infty[t_0, T]$, $t_0 \geq 0, T \geq t_0$ we denote the space of all functions $\mathbf{f} : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ such that $\|\mathbf{f}\|_{\infty,[t_0,T]} = \operatorname{ess\,sup}\{\|\mathbf{f}(t)\|, t \in [t_0, T]\} < \infty$; $\|\mathbf{f}\|_{\infty,[t_0,T]}$ stands for the $L^n_\infty[t_0, T]$ norm of $\mathbf{f}(t)$.

Let $\mathcal{A}$ be a set in $\mathbb{R}^n$ and $\|\cdot\|$ be the usual Euclidean norm in $\mathbb{R}^n$. By the symbol $\|\cdot\|_{\mathcal{A}}$ we denote the following induced norm:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf_{\mathbf{q} \in \mathcal{A}} \{\|\mathbf{x} - \mathbf{q}\|\}.$$

In case x is a scalar and $\Delta \in \mathbb{R}_{>0}$, notation $\|x\|_\Delta$ stands for the following

$$\|x\|_\Delta = \begin{cases} |x| - \Delta, & |x| > \Delta \\ 0, & |x| \leq \Delta. \end{cases}$$

2. Introduction

This article deals with the challenges and opportunities that spatiotemporal representation of visual information offers for visual pattern recognition. We will consider spatiotemporal pattern representation in the framework of template matching, the oldest and most common method for detecting an object in an image. According to this method the image is searched for items that match a template. A template consists of one or more local arrays of values representing the object, e.g. intensity, color, or

* Corresponding author.
E-mail addresses: tyukinivan@brain.riken.jp, I.Tyukin@le.ac.uk (I. Tyukin), tatianat@brain.riken.jp (T. Tyukina), ceesvl@brain.riken.jp (C. van Leeuwen).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.01.014

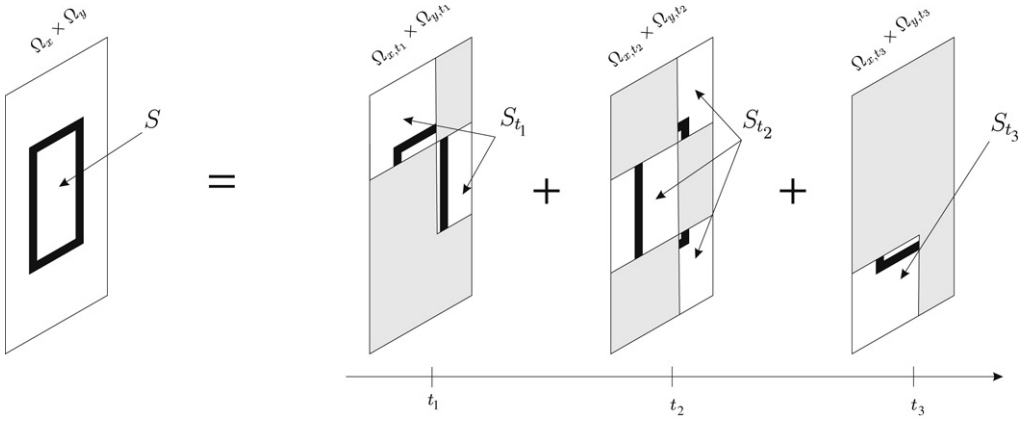

Fig. 1. Spatial sampling of image $S(x, y) : \Omega_x \times \Omega_y \to \mathbb{R}_+$ according to the factorization of $\Omega_x \times \Omega_y$ into subsets $\Omega_{x,t_1} \times \Omega_{y,t_1}, \Omega_{x,t_2} \times \Omega_{y,t_2}, \Omega_{x,t_3} \times \Omega_{y,t_3}$.

texture. A similarity value[1] is calculated between these templates and domains of the image, and a domain is associated with the template once their similarity exceeds a given threshold.

Despite the simple and straightforward character of this method, its implementation requires us to consider two fundamental problems. The first relates to *what* features should be compared between the image $S_0(x, y)$ and the template $S_i(x, y)$, $i \in \mathcal{I}$. The second problem is *how* this comparison should be done.

The normative answer to the question of *what* features should be compared invokes solving the issue of optimal image representation, ensuring most effective utilization of available resources and, at the same time, minimal vulnerability to uncertainties. Principled solutions to this problem are well known from the literature and can be characterized as *spatial sampling*. For example, when the resource is frequency bandwidth of a single measurement mechanism, the optimality of spatially sampled representations is proven in Gabor's seminal work (Gabor, 1946).[2] In classification problems, the advantage of spatially sampled image representations is demonstrated in Ullman, Vidal-Naquet, and Sali (2002). In general, these representations are obtained naturally when balancing the system resources and uncertainties in the measured signal. A simple argument supporting this claim is provided in Appendix A.

A variety of sophisticated spatial sampling methods exists (Blake, Curwen, & Zisserman, 1994; Bueso, Angulo, Quian, & Alonso, 1999; Gabor, 1946; Lee & Yuille, 2006). Here we limit ourselves to spatial sampling in its elementary form, which is achieved by factorizing both the domain $\Omega_x \times \Omega_y$ of the image S_0 and the templates S_i, $i \in \mathcal{I}$ into subsets:

$$\Omega_x \times \Omega_y = \bigcup_t \Omega_{x,t} \times \Omega_{y,t}, \quad t \in \Omega_t, \Omega_{x,t} \subseteq \Omega_x, \Omega_{y,t} \subseteq \Omega_y. \quad (1)$$

Factorization (1) induces sequences $\{S_{i,t}\}$, where $S_{i,t}$ are the restrictions of mappings S_i to the domains $\Omega_{x,t} \times \Omega_{y,t}$. These sequences constitute sampled representations of S_i, $i \in \mathcal{I}^+$ (see Fig. 1). Notice that the sampled image and template representations $\{S_{i,t}\}$ are, strictly speaking, sequences of functions. In order to compare them, scalar values $f(S_{i,t})$ are normally assigned to each $S_{i,t}$. Examples include various functional norms, correlation functions, spectral characterizations (average frequency or phase), or simply weighted sums of the values of $S_{i,t}$ over the entire domain $\Omega_{x,t} \times \Omega_{y,t}$. Formally, f could be defined as a functional, which maps restrictions $S_{i,t}$ into the field of real numbers:

$$f : L_\infty(\Omega_{x,t} \times \Omega_{y,t}) \to \mathbb{R}. \quad (2)$$

This formulation allows a simple representation of images and templates as sequences of scalar values $\{f(S_{i,t})\}$, $i \in \mathcal{I}^+$, $t \in \Omega_t$. We will therefore adopt this method here.

The answer to the second question, *how* the comparison is done, involves finding the best and simplest way possible to utilize the information that a given image representation provides, while at the same time ensuring invariance to basic distortions. Even though considerable attention has been given to this problem, a unified solution is not yet available. The primary goal of our current contribution is to present a unified framework to solve this problem for a class of systems of sufficiently broad theoretical and practical relevance.

We consider the class of systems in which spatially sampled image representations are encoded as temporal sequences. In other words, parameter t in the notation $f(S_{i,t})$ is the time variable. This type of representation is frequently encountered in neuronal networks (Gutig & Sompolinsky, 2006) (see also the references therein). Examples of similar representation schemes are widely reported in the neuroscience literature. For example Alonso, Usrey, and Reid (1996) show that patches of visual stimuli which are perceived as spatially close by the processing system (e.g. when the receptive fields of individual cells overlap) are encoded by similar firing spike patterns and vice versa. In our model spatially non-overlapping patches are represented by different sequences $\{f(S_{i,t})\}$, and identical images have identical temporal representations. Hence such systems have a claim to biological plausibility. In addition, they enable a simple solution to a well-known dilemma. This is about whether comparison between templates and image domains should be made on a large, i.e. global, or on a small, i.e. local scale. The solution to this dilemma consists in temporal integration. Let, for instance, $\Omega_t = [0, T]$, $T \in \mathbb{R}_{>0}$. Then an example of a temporally integral, yet spatially sampled, representation is:

$$f(S_{i,t}) \mapsto \phi_i(t) = \int_0^t f(S_{i,\tau}) \mathrm{d}\tau, \quad t \in [0, T], i \in \mathcal{I}^+. \quad (3)$$

The temporal integral $\phi_i(t)$ contains both spatially local and global image characterizations. Whereas its time derivative at t equals to $f(S_{i,t})$ and corresponds to spatially sampled, local representation $S_{i,t}$, the global representation $\phi_i(T)$ equals to the integral, cumulative characterization of S_i. An example illustrating

[1] Traditionally a correlation measure is commonly used for this purpose (Jain, Duin, & Mao, 2000).

[2] Consider, for instance, a system which measures image $S_i(x, y)$ using a set of sensors $\{m_1, \ldots, m_n\}$. Each sensor m_i is capable of measuring signals within the given frequency band Δ_i at the location x_i in corresponding spatial dimension x. Then according to Gabor (1946), sensor m_i can measure both the frequency content of a signal and its spatial location with minimal uncertainty only if the signal has a Gaussian envelope in x: $S_i(x, y) \sim e^{\sigma_i^{-2}(x-x_i)^2}$. In other words, the signal should be practically spatially bounded. This implies that the image must be spatially sampled.