Multi-channel spatio-temporal topographic processing for visual search and navigation

István Szatmári, Dávid Bálya, Gergely Tímár, Csaba Rekeczky, and Tamás Roska
AnaLogical and Neural Computing Systems Laboratory
Computer and Automation Institute of the Hungarian Academy of Sciences
1111 Budapest, Kende u. 13-17, Hungary

ABSTRACT
In this paper a biologically motivated image flow processing mechanism is presented for visual exploration systems. The intention of this multi-channel topographic approach was to produce decision maps for salient feature localization and identification. As a unique biological study has recently confirmed, mammalian visual systems process the world through a set of separate parallel channels and these representations are embodied in a stack of strata in the retina. Beyond reflecting the biological motivations, our main goal was to create an efficient algorithmic framework for real-life visual search and navigation experiments. In the course of this design the retinotopic processing scheme is embedded in an analogic Cellular Neural Network (CNN) algorithm where image flow is analyzed by temporal, spatial and spatio-temporal filters. The output of these sub-channels is then combined in a programmable configuration to form the new channel responses. In the core of the algorithm crisp or fuzzy logic strategies define the global channel interaction and result in a unique binary image flow. This processing mechanism of the algorithmic framework and the hardware architecture of the system is presented along with experimental ACE4k CNN chip results for several video flows recorded in flying vehicles.

Keywords: visual terrain exploration, spatio-temporal filters, topographic processing, optical navigation, Cellular Neural Network

1. INTRODUCTION
The aim of this bio-inspired development of visual exploration system is to accomplish some specific “crucial functions” found in successful, nature-tested mechanisms effectively used by biological organisms. These special functions are rather difficult to realize by conventional methods and intent is to learn salient principles from a variety of diverse species for a desired “crucial function”. Translating of these natural visual strategies and founding way to apply these results in areas such visual terrain exploration, classification, or even navigation will make it possible to achieve previously impossible projects in many fields of science. For instance, integrating these strategies into a compact device makes it possible to offer stable flight control and terrain following mechanism for small biomorphic fliers capable to solve such difficult tasks that are still considered unfeasible.

It is well known that the representation of the visual world around us is formed in parallel channels coding different characteristics (shape, movement, color, textures, etc.) and these are embodied in a stack of “strata” in the retina confirmed again by recent biological studies (see e.g. [1]). Each of these representations can be efficiently modeled in Cellular Nonlinear Networks (CNN, [2-6]). Different VLSI implementations of the CNNs have proven to be useful in many image processing applications. They have unparalleled speed for image processing functions (up to 10000 frames/sec) and provide unique capabilities (like wave type processing, see e.g. [7], [8]) to accomplish difficult image processing tasks.

Our work demonstrates that incorporating these successful biological strategies into engineering solutions in the field of vision applications constitute novel and useful algorithmic cornerstones and induce specific hardware and software architecture design. Here, we focus primarily to translate these biologically motivated image processing strategies to develop a small and robust visual search, recognition, and navigation system. This makes the first step toward a CNN based system-on-a-chip (SoC) terrain search and navigation system. During this work we have been developing stored program cellular nonlinear processing strategies for terrain exploration and classification; automatic adjustment of focus and scale of attention; and navigation support. The experiments have been carried out on aerial video-flows showing diverse terrain sites from a navigating plane (see sample images in Figure 1).
The paper is organized as follows: first, the biological motivations are considered, second, the algorithmic framework and the system architecture are described, third the multi-channel topographic processing is analyzed thoroughly, finally some experimental results are presented.

Figure 1: Samples from aerial image flows used during the terrain exploration and classification experiments.

2. BIOLOGICAL MOTIVATIONS

Building a multi-channel adaptive algorithmic framework for visual search and navigation relies strongly on biological motivations. It has been long known that visual system of mammalian species processes the world through a set of separate spatio-temporal channels. A recent study has confirmed that the organization of these channels begins in the retina where a vertical interaction across ten parallel stack representations can also be identified ([1]). We have decomposed the above model into a multi-channel adaptive (single-layer) CNN-UM analogic algorithm. Within this algorithmic framework the enhanced image flow is analyzed by temporal, spatial and spatio-temporal filters shown in Figure 2. After some enhancement the spatio-temporal filters extract various global and local features of the image flow. The output of these sub-channels is combined in a programmable configuration to form the new channel responses. Crisp or fuzzy logic strategies define the global channel interaction and result in a unique binary image flow for further processing. This is also combined with the output of a single/multiple step prediction and forms the final format of this binary image flow.

Figure 2: The framework of the multi-channel adaptive CNN-UM algorithm. Nonlinear spatio-temporal feature and change detectors are implemented in each channel and these are combined to obtain a unique binary output for further processing.
When building up the computing blocks of the multi-channel algorithmic framework the following key processing strategies learned from retina modeling experiments ([10], [11]) have been used (the associated image processing arguments are given in italic):

- signal flow normalization: *dynamic range optimization*
- attention and saccade mechanism: *region of interest selection*
- spatial, temporal and spatio-temporal decomposition of the input flow: *an efficient geometric distortion analysis requires a sparse signal representation*
- parallel on-off channel processing: *DC-component compensation*
- narrow and wide-field wave-type interaction: *efficient binary patch shaping with noise suppression*
- “vertical” interaction of the decomposed channels: *forming a unique detection output through optimized “cross-talk” of the individual channels*

3. SYSTEM DESCRIPTION

Based on biologically motivated assumption the following algorithmic framework and system architecture were proposed for visual terrain exploration and navigation system.

3.1 ALGORITHMIC FRAMEWORK

The general algorithmic framework designed for bioinspired visual search and navigation is shown in Figure 3. It contains the following key elements:

- Attention and selection mechanism (focus and scale)
- Topographic multi-channel preprocessing (CNN processing)
- Multi-target tracking
- Feature based classification
- Navigation parameter estimation

Assuming a large resolution array sensor input, only a certain part of the image flow is analyzed by an automatically adjusted focus and scale mechanism. This attention selection is connected in a feedback loop depending on further feature processing results. The selected sensor input undergoes a parallel multi-channel CNN processing, which provides a topographic (binary) output for the multi-target tracking (MTT, [9]) framework and produces various feature descriptors of scene for object classification. The output of the MTT sub-system consists of static and dynamic target attributes. Static target attributes include target feature descriptors such as centroid locations, contour and skeleton structure, orientation, size and others. Dynamic target attributes describe all targets in motion with their kinematic properties (this includes the optical flow: the estimate of the 2D motion field).

Figure 3: Main processing blocks and signal flow of the visual microprocessor architecture designed for bio-inspired visual search/navigation
Terrain classification is mainly built on static target attributes. There are several classifiers that could have been used for terrain feature analysis and global classification. Two types of classification approaches were considered: category based and model based with the on-line learning capability. For the first type, we have applied an adaptive resonance theory (ART, [12]) implemented and running also on the CNN. The idea behind this type of classification is to select the input vectors into different categories based on the distance between the input and the category prototypes. The model based approach incorporates the class knowledge from the beginning. The input vectors are collected and labeled with the correct class or class probabilities. This database is examined and used during the supervised learning. Bayesian networks, decision trees and fuzzy classifiers can be established from the statistical description from previous measurements. These two types of classification approaches can be combined and have influence on preprocessing strategies as we will see later. The last flight control module is mainly built on dynamic target attributes coming from the MTT module. In addition to several tracked features, the navigation parameter calculation requires an optical flow estimation also incorporated in this module. This motion field is calculated for each frame over a fixed uniform and/or adaptive non-uniform grid. Finally, the navigation estimator calculates the unit translation direction along with the rotation parameter estimates.

3.2 THE COMPACT CVM ARCHITECTURE

This algorithmic framework description envisions a platform with multi-task processing capability. This architecture provides a framework toward a fully integrated cellular sensor-computer and can be attributed as follows:

- Fault-tolerant visual computer
- Compact and low-power system
- Multi-task architecture
  - Attention and selection
  - Feature classification
  - Navigation parameter estimation
  - Multiple target tracking
- Defines a biologically inspired single chip sensor-computer

Such architecture have been designed and is under testing; and is referred to as the Compact Cellular Visual Microprocessor (COMPACT CVM). The COMPACT CVM architecture (Figure 4) builds on state-of-the-art CVM type (ACE16k [14]) and DSP type (TEXAS 6x) microprocessors and its algorithmic framework contains several feedback and automatic control mechanisms in between different processing stages.

![Figure 4: Main building blocks of the COMPACT CVM architecture](image)

The architecture is standalone and with the interfacing communication processor it is capable of 100 Mbit/sec information exchange with the environment (over TCP/IP). The COMPACT CVM is also re-configurable, i.e. it can...
be used as a monocular or a binocular device with a proper selection of a large resolution CMOS sensor (IBIS 5) and a low resolution CNN/CVM sensor-processor (ACE16k). This architecture provides a framework for a SoC design: toward a fully integrated cellular sensor-computer. In this paper the multi-channel topographic processing is analyzed in details other parts of the algorithm are discussed in [15-17].

4. TOPOGRAPHIC MULTI-CHANNEL PROCESSING WITH COMPRESSED FEATURE CHARACTERIZATION

This section discusses thoroughly the multi-channel topographic processing of the visual terrain search and classification system. As stated before, two types of classification approaches were considered: category based and model based. These methods can be considered as bottom-up and top-down approaches and have influence on strategy chosen for the preprocessing stage. In the final implementation, these approaches will be combined to each other to produce more robust feature descriptions for the classification.

4.1 BOTTOM-UP APPROACH

In Figure 5 examples are shown for calculating the topographic feature maps (the first parallel stage of the analogic algorithm described in Figure 2) for terrain video-flows. As illustrated (see the right most column of the figure) these maps describe the edginess, irregularity, rough/fine structures, connected structures etc. of the input. It is also obvious that the description is non-orthonormal in the feature space thus providing certain robustness when some of these specific estimation strategies are sensitive and not reliable enough for further processing.

Figure 5: Illustration for topographic feature maps (FM) calculated by CNN processing.

The choice among the possible channels was based on the statistical analysis of the extracted feature vectors. Factor analysis was carried out to identify the important features and diminish the unnecessary redundancies of the system. It is important to mention that this classical statistical method was used only to set up the channels and their interactions in advance and not for operational parameter tuning. Figure 6 and Figure 7 illustrate a statistical analysis of different feature maps. Spatio-temporal channels extract both global and local features. Global feature can be for instance, the pixel number of objects of a given scene, while a local feature describe a specific characteristic of a single object (e.g. orientation). It is shown that a global characterization of image feature descriptors and an object oriented feature extraction can be used jointly in further processing stages.
Figure 6: Image feature analysis with global statistical characterization of a test video-flow. Normalized image feature strength for 10 different channels are magnified. It can be seen that a global statistical analysis results in several similar and complementary channels. However, even this simplified characterization has a significant discriminative power which can be mainly seen in the first part of the flow.

Figure 7: Object oriented feature analysis for the same test video-flow. For all objects several local descriptors are calculated (area, orientation, eccentricity etc.) and their overall statistics are displaced in the individual figures (solid black: maximum, dashed black: minimum, solid blue: mean, solid red: median; horizontal axis: time; vertical axis: amplitude).

The factor analysis has showed that around six to ten major components exist among more than forty analyzed global and local features. The Figure 8 shows the derived flowchart of the implemented topographic preprocessing with
chosen spatio-temporal channels based on statistical analysis. It includes the programmable sub-channel interaction and mixture. This is also combined with the output of a single/multiple step prediction and forms the final format of this binary image flow towards the MTT part. Each image of the input flow is first enhanced (histogram modified), then the enhanced image is processed with three different channels: one channel extracts spatial information (Sobel like operation, vertical and horizontal lines), a second channel the temporal changes (background separation), while the third channel filters both spatial and temporal changes.

Figure 8: Algorithmic flowchart of the multi-channel topographic preprocessing of monocular data flow. It includes three major spatio-temporal filtering channels with programmable sub-channel interaction and prediction step. Further structural analysis extracts different modalities of the scene and objects.

Each channel is analyzed as a binary image and the above mentioned local features are extracted based on the assumption that the black patches are objects. These object features are computed such as area, minor axis, eccentricity, etc. The descriptive statistics is used to aggregate the same feature of the different objects such as minimum, average, variance, etc. The channels are analyzed also as scenes, and the same low-level spatial-temporal features are extracted. An edge-based separation is performed to distinguish between the stick-like and the box-like objects. A skeleton-based separation is computed to divide the objects as circle-like or tree-like items. The number of black pixels in these channels are used as features. All of these computations are made on the CNN-UM chip.

4.2 TOP-DOWN APPROACH

The model based approach requires some kind of description of possible terrain type characteristics in advance and this class knowledge is incorporated during the supervised learning. This database is examined and revised if necessary during the evaluation. Figure 9 shows some hand-drawn patterns as possible terrain types or classes. These patterns might serve as initial base classes and their some kind of statistical description as initial knowledge base at the beginning.
Figure 9: Hand-drawn examples of possible terrain classes. Some kind of statistical description of these prototypes might establish an initial knowledge base that during the supervised learning can be examined and revised.

Table 1 shows some possible basic patterns and their fuzzy type description as well. This description is much more compact than the category description and because it results in a human readable system makes it to be improved easily using domain knowledge.

<table>
<thead>
<tr>
<th>Class</th>
<th>Pattern</th>
<th>Parallel</th>
<th>Fork</th>
<th>Orientation</th>
<th>Fragmented</th>
<th>Grain</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-1: River bank or terrace like rock wall</td>
<td></td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>C-2: River, from higher attitude</td>
<td></td>
<td>M</td>
<td>H</td>
<td>M</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>C-3: River, from lower attitude</td>
<td></td>
<td>M</td>
<td>M</td>
<td>H</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>C-4: Rocky, cliff or Dust devil tracks</td>
<td></td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>C-5: Ground</td>
<td></td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td>H</td>
</tr>
</tbody>
</table>

Table 1: Terrain classes and their fuzzy description

Combining model based description with the category one might help to choose proper topographic filters and feature characteristics in order to reduce the number of features to be extracted still preserving the robustness through some redundancy.

5. EXPERIMENTAL RESULTS

The algorithmic framework and the architecture were implemented and emulated on the ACE-BOX visual computer embedded within the Aladdin Pro software framework ([18], [19]). The ACE-BOX contains a previous generation CNN-UM type microprocessor (ACE4K, [13]) and a Texas Instruments TMSC6202 250Mhz DSP. This system is capable of processing at a video frame rate. The initial measurement on the COMPACT CVM with the ACE16K [14] has been started and sub-parts of the algorithmic framework are initialized. Figure 10 show screenshot of the development environment running on a test video flow.
Figure 10: This screenshot shows the operation of the terrain exploration environment based on multiple feature extraction and CNN-A RT classification. The selected view point (focus) is the input of the multi-channel topographic analysis and extracted feature maps are displayed on the right part of the image. Feature intensities are shown at the bottom-right while category history is displayed at the bottom-left.

6. SUMMARY

We have proposed a bio-inspired multi-channel topographic processing based on sensor-level CNN microprocessor architecture with a dedicated algorithmic framework for efficient terrain exploration, classification, and navigation. Experimental results were shown on aerial video data using a real-time CNN microprocessor based visual computer. Hopefully, these bio-inspired engineering experiments will lead to efficient and realizable system-on-a-chip architectures in several application areas including intelligent sensory system missions.

This work was partially funded by the ONR (TeraOps Ltd CA USA, No: 0295-R, sub-contract) and by the Hungarian National Research and Development Program, Grant No. NKFP 035/02/2001.

REFERENCES