ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 45-50

Visual person tracking with a Supervised Conditioning-SOM

David Buldain, Elías Herrero, Rubén Cabrejas

Departamento de Electrónica y Comunicaciones, Universidad de Zaragoza,
María Luna 1, 50018 Zaragoza, Spain

Abstract. The classification problem of determining if a surveillance camera sees persons is tackled with two neural models: the Self-Organizing Map (SOM) with supervision as in a classical conditioning analogy and Multi Layer Perceptrons (MLP). The first model, that we call Conditioning-SOM (C-SOM) allowed a quick selection of input features with a good tradeoff between computational cost and classification performance. Finally, MLP classifiers were trained with the selected features. The classification performance of both neural models was very good with very simple features.

1. Introduction

Detection and tracking of persons with artificial vision systems is a useful task that is very dependent of the scene. It has been tackled with classical programing methods [9] and neural networks methods [1]. In this paper, this problem is revisited applying the Conditioning-SOM (C-SOM), a generalization of the supervised Self Organizing Map [4]. The C-SOM trains very quickly compared with the Multi-Layer Perceptrons (MLP) [6] or other supervised methods based in gradient descent of any cost function. The model also provides, as the original SOM, a visual inspection of the classification result. Both qualities were exploited by exhaustive simulations to permit a rapid selection of the more suited characteristics to be extracted from images for accomplishing the task, trying to minimize the time for developing MLP classifiers. Section 2 presents a brief description of C-SOM. Section 3 describe the problem. Section 4 and 5 present the selected characteristics and classification results with both neural models.

2. Conditioning-SOM

The Conditioning-SOM proposes to introduce a Classical Conditioning mechanism in the learning process of the SOM with multiple stimulus sources. The essence of Pavlov's Classical Conditioning is simple. You start with two things that are already connected or strongly related with each other (food and salivation). Then you add a third thing (bell) during the learning process and this third thing may become so strongly associated that it has the power to produce the behavior.

To introduce this conditioning method, lets divide data in a number of sub-patterns ‘D’ or stimulus, to be processed by each data-path ‘d’ (which number of components is $N^{(d}$). The path-weights, $\mathbf{w}_i^{(d}$, correspond to the fraction of weights of the neuron ‘i’ in the data-path ‘d’. Each data-path calculates the Euclidean distance between path-patterns and path-weights factorized by a gain coefficient, $g^{(d}$, to force different scales for the distances in each subspace. So the excitation of neuron 'i' becomes:

$$Exc_i = \sum_{d=1}^{D} g^{(d} dist(\mathbf{w}_i^{(d}, \mathbf{x}^{(d}) = \sum_{d=1}^{D} g^{(d} \sum_{j=1}^{N^d} (w_{ij}^{(d} - x_j^{(d})^2 \qquad \textbf{(1)}$$

Notice that, if all gains present value one, the neuron excitation is like in the original Supervised SOM model [4], where all the distances calculated in each data subspace are considered equivalent. This approach is implemented by J.Parhankangas with the function "som_supervised" in the SOMToolbox [7], but it only deals with class labels, however many other types of data sources and scaling of subspaces should be considered to exploit the capabilities of the model.

The Conditioning-SOM exploits this posibility by controling the gains coefficients in the learning and recall phases. During the learning phase, the path with a high gain acts as the conditional stimulus (food = supervisory labels). As the distance contribution of this path is more relevant than the rest of input stimulus, it decides the conditioned response (salivation = map activation). The stimulus with low gains provide a low contribution to the whole excitation, and does not decide which neuron is the winner, they are the conditioning stimulus (bell = data). In the learning process, the neuron prototypes associate the different data stimulus, and the data-paths with low gains are forced to be conducted by the organization of the map in the data-paths with high gains. Notice that this supervision is part of the competitive dynamic, instead of applying gradient descent methods as in [6][8].

During the recall phase, the gain of the supervisory-path is annulled (unconditional stimulus is not present) and the map gives its activation based only in the data (conditioning stimulus). If the new data contains enough information, it will be separated with the map structure forced by the labels.

Another difference between C-SOM and SOM model is their output response. SOM model outputs the winner-neuron index, but C-SOM outputs the winner-neuron weights in its supervision-path, as their distributio analysis gives light to the class confusion committed by the classifier with the selected features of the problem.

3. Problem description and Feature Extraction

The scene is a long corridor with sources of light in the ceiling, in the bottom and behind the camera that project shadows and reflects in the walls. Persons along the corridor have a different scale and can enter into the scene by lateral doors. We used 12 video sequences (382 seconds) with 5 frames per second of 240x320 pixels in 256 gray levels. The number of persons in the images varied from 0 to 6 and their moving blobs were generated with a motion detection algorithm.

Several methods of motion detection were tested [3], all of them gave blobs with broken contours and fragments, so the choice was the method with less computational cost, assuming that more imprecise blob-contours would be dealt by the neural classifier as a source of noise. Image foreground is obtained by subtracting the image with a stable background and filtering the gray-difference in two threshold steps that eliminate snow noise and rejects blobs with less than 100 pixels. Figure 1 shows one image of the scene and the corresponding detected blobs.

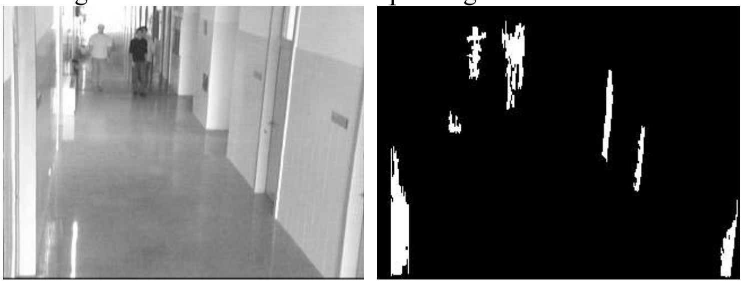


Figure 1: Scene frame with 3 persons in the corridor and the resulting blobs.

The blobs were scaled to a fixed size with 128x64 pixels. Features extracted from the blob-contour were: 1) Relative size and position of the bounding-box (BB) of the blob respect to the image frame (4 components), 2) Vertical histogram: counting pixels by columns (64), 3) Horizontal histogram: counting pixels by rows (128), 4) Left contour: relative position of each first row-pixel (128), 5) Right contour: relative position of each last blob row-pixel (128), 6) Upper contour: relative position of each upper blob column-pixel (64), 7) Bottom contour: relative position of each lower blob column-pixel (64), 8) Discrete Fourier Transform of 25x25 windows over image (625). Features extracted from the gray levels were: 9) Mean and variance of horizontal and vertical histograms of gray levels (4).

4. Selection of characteristics and network architectures

We used functions in SOMToolbox [7] for training maps in batch mode and for weight initialization. The map size along all experiments was 10x15 neurons, as this map size had a good tradeoff between classification performance and training time, although maps with more neurons provided better classification results.

The motion detection method generated three kinds of blobs: blobs from shadows and reflects ('No-Person'), blobs resulting from fragments of persons ('Fragment') and blobs containing complete persons ('Person'). The three classes were labeled with the values 0, 0.5 and 1 respectively, so the map has to separate the classes 'No-Person' and 'Person' with intermediate neurons dedicated to the class 'Fragments'. The figures 2 show histograms of the neuron activation represented over the U-matrix. Upper figures show the histograms when labels are present, while the bottom figures show the histograms when labels are absent. The left figures represent the neuron activation in the class 'Person', the middle figures show the class 'Fragment' and right figures show the class 'No-Person'. Notice that, if supervision is disabled, the main confusion appears between 'Fragment' class and the other two classes. The inspection of these topologically-structured confused patterns is a valuable tool for the refinement of the selected features. The map visualization also provides a distribution picture of the particular confussion sources found in the scene.

The histogram of supervision-weight values clarifies the distribution of neurons in the three classes (figure 3). It shows three groups of neurons associated to supervisory labels 0, 0.5 and 1 (classes: No-Person, Fragment and Person). In the recall phase, the pattern class is decided by comparing the supervision weight of the winner neuron with three class intervals that where chosen arbitrarily and could have been selected with a better criteria. Using the histogram of figure 3, we assigned class 'No-Person' to neurons with supervision-weights values in the interval [0, 0.2], class 'Fragment' to the interval (0.2, 0.75) and class 'Person' to [0.75, 1].

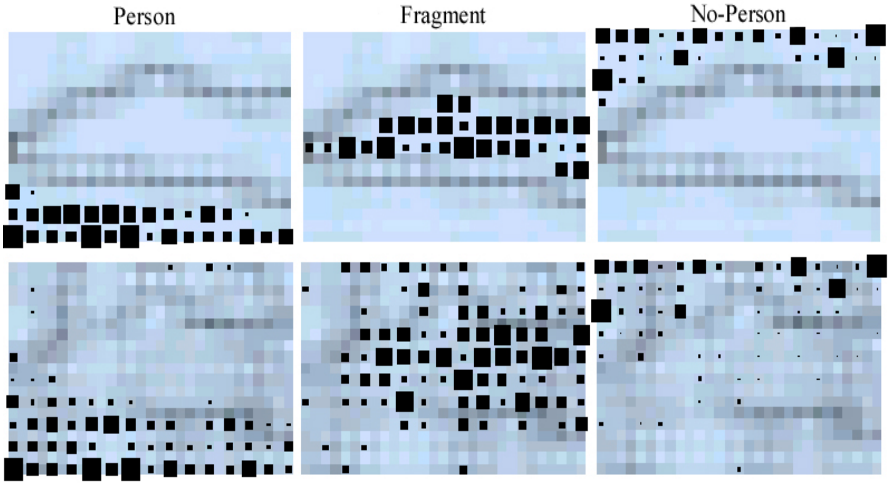


Figure 2: Histograms of neuron activation over the U-matrix representation. Upper figures represent class histograms when the map receives supervision. Bottom figures depict the map activation when supervision is disabled. Each column represents the activation for the classes: Person, Fragment and No-Person.

In a complete set of simulations with different characteristics and combinations of two of them, the best classification results appeared when maps processed the relative size and position of the bounding-box of the blob. The Means and deviations of the gray levels provided a better separation between classes 'Person' and 'No-Person', but introduced more confusion between classes 'Fragments' and 'Person'. This confusion is quite natural, as many fragments are person contours without head or legs, which characteristics are very similar to that of complete persons, so confusion between those classes is not as bad as confusion between classes 'Person' and 'No-Person'. A possible improvement of the system would introduce a stage to merge identified fragments into new blobs to be reprocessed by the classifier.

Once good characteristics were identified with C-SOM, the architecture selection for MLP was accomplished with the Bayesian regularization training method [5] following the next steps: a) to estimate a certain minimum number of cycles, T, needed for performance function (SSE, Sum Squared Error) to reach the lowest flat zone where intrinsic error of data is reached, b) to train T cycles different network architectures to select the one with the best compromise between classification error and network parameter utilization, c) to apply cross validation method and early stopping with 10 network exemplars with the selected architecture.

5. Classification results

The training set contained 1319 blobs: 478 with persons, 268 fragments of persons and 573 without persons. The test set contained 1394 blobs: 406 with persons, 329 fragments of persons and 659 without persons. Classification results in the test set for the networks processing both characteristics commented above are presented in confusion matrixes of figure 4 for C-SOM and for the average of the 10 MLPs. The best MLP architectures had two hidden layers: a) With relative size and position of the blob: (4:8:3:3), b) With relative size and position of the blob plus the statistics of gray content: (8:8:5:3). If we consider the confusion error with class ‘No-Person’ as the irrecoverable error of the classifier, C-SOM generate 10.1% error and MLP commits 8.46% of confusion error when only processing the relative size and position of the bounding-box (4 inputs). If statistic information of vertical and horizontal histograms of the gray levels is included (8 inputs), C-SOM produces 7.74% error and MLPs produce 7.17%.

The Cohen's Weighted Kappa-Index [2] is an improvement over using percentage agreement to evaluate inter-rater reliability when observing or otherwise coding categorical variables. Kappa has a range from 0-1, with larger values indicating better reliability. Kappa > 0.7 is considered satisfactory in a multi-class problem. The 4-input SOM has K = 0.736, the 8-input SOM has K = 0.775. The 4:8:3:3-MLP has K = 0.722 and the 8:8:5:3-MLP has K = 0.795.

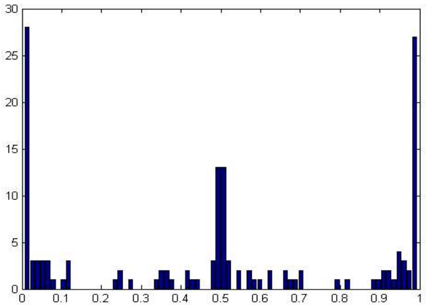
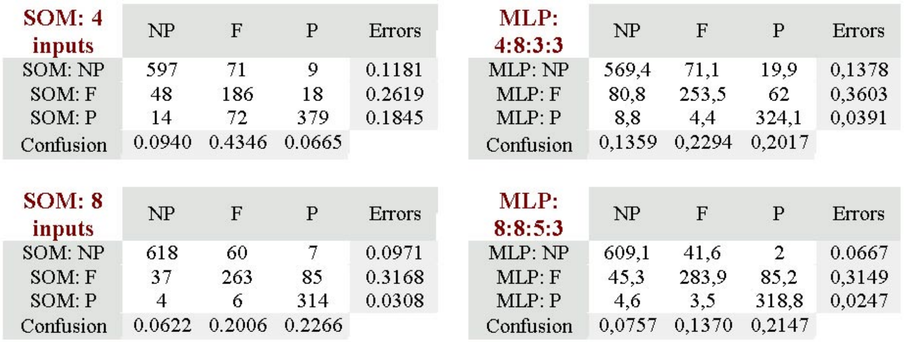


Figure 3: Histogram of values found in supervision weights of the map neurons. Three clusters of neurons are found around values 0, 0.5 and 1, that correspond to the respective classes: No-Person, Fragment and Person. Some interpolating neurons present supervision weights with inter-class values.

6. Conclusions

The Conditioning-SOM model extending the capabilities of the supervised SOM has been described with a classical conditioning methafor. It was tested in the classification problem of detecting persons with a camera. The study shows that, in this kind of problems with high amounts of noisy data, to use C-SOM for the initial development of the classifier is a recommended step, in order to select the input features to train classifiers with other supervised methods as Multi-Layer Perceptrons. The training time for MLP networks using the Levenberg-Marquardt method was greater in a factor 10-100 (depending on the size of the map) than the training time needed for developing the C-SOM classifier with similar classification results. The C-SOM also permits the visual analysis of the confussion nature in its map representation. As the confussion is topologically structured and depends strongly on the particular scene, this visualization aided considerably to select the best characteristics to refine the classification.



SOM: 4 inputs	NP	F	P	Errors
SOM: NP	597	71	9	0.1181
SOM: F	48	186	18	0.2619
SOM: P	14	72	379	0.1845
Confusion	0.0940	0.4346	0.0665	

MLP: 4:8:3:3	NP	F	P	Errors
MLP: NP	569,4	71,1	19,9	0,1378
MLP: F	80,8	253,5	62	0,3603
MLP: P	8,8	4,4	324,1	0,0391
Confusion	0,1359	0,2294	0,2017	

SOM: 8 inputs	NP	F	P	Errors
SOM: NP	618	60	7	0.0971
SOM: F	37	263	85	0.3168
SOM: P	4	6	314	0.0308
Confusion	0.0622	0.2006	0.2266	

MLP: 8:8:5:3	NP	F	P	Errors
MLP: NP	609,1	41,6	2	0.0667
MLP: F	45,3	283,9	85,2	0,3149
MLP: P	4,6	3,5	318,8	0,0247
Confusion	0,0757	0,1370	0,2147	

Figure 4: Confusion matrixes for two C-SOM architectures in the left column, and the mean confusion matrixes for 10 networks of two MLP architectures. The network classification results appear in rows and classes (NP = No-Person, F = Fragment, P = Person) in columns.

References

1. B. A. Boghossian, S. A. Velastin: Image processing system for pedestrian monitoring using neural classification of normal motion patterns. Measurement and Control, Vol.32, Issue 9, 261-264.
2. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological measurements, 20, 37-46, (1960).
3. E.Durucan, T.Ebrahimi: Change Detection and Background Extraction by Linear Algebra. Proc. IEEE, Vol. 89, (2001)
4. T.Kohonen: Self-Organizing Maps. 3rd Edition, Springer-Verlag Berlin Heidelberg New York (2001)
5. D.J.MacKay: Bayesian interpolation. Neural Computation, Vol 4, 415-447 (1992)
6. D.E. Rumelhart, J.L. McClelland (eds.): Parallel Distributed Processing. Vol 1: Foundations. MIT Press (1986).
7. SOM Toolbox: http://www.cis.hut.fi/projects/somtoolbox/
8. T. Villmann, B. Hammer, M. Strickert: Supervised Neural Gas for Learning Vector Quantization. Available in: http://citeseer.nj.nec.com/519513.html
9. C.Wren, A.Azarbayejani, T.Darrell, A. Pentland: Pfinder: Real-time tracking of the human body. IEEE Trans. Pattern Anal. and Machine Intelligence, 19 (7), 780-785, (1997).