Assessing similarity of emergent representations based on unsupervised learning

Juha Raitio, Ricardo Vigário, Jaakko Särelä and Timo Honkela
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT
E-mail: forename.surname@hut.fi

***Abstract*— According to a connectionist view, mental states consist of the activations of neural units in a connectionist network. We consider the similarity of representations that emerge in unsupervised, self-organization process of neural lattices when exposed to color spectrum stimuli. Self-Organizing Maps (SOM) are trained with color spectrum input, using various vectorial encodings for representation of the input. Further, the SOM is used for heteroassociative mapping to associate color spectrum with color names. Recall of association between the spectra and colors is assessed. It shows that the SOM learns representations for both stimuli and color names, and is able to associate them successfully. The resulting organization is compared through correlation of the activation patterns of the neural maps when responding to color spectrum stimuli. Experiments show that the emerged representations for stimuli are similar with respect to the partitioning-of-activation-space measure almost independently of the encoding used for input representation. This adds a new example in favour of the usability of the state space semantics.**

I. Connectionist networks and representation of content

The state of a connectionist network is the momentary activation levels of neurons [1]. A particular state may occur as a response to stimuli. These have a representation in the space spanned by the possible activations of the neurons in the network. Vice versa, any pattern of activation in the network may represent some, maybe latent, information. According to a connectionist view, mental states consist in these activations [1]. Networks with different neural architectures may reach comparable mental representations or states. Therefore connectionists have been puzzled with a criterion for determining when activations in two connectionist networks have similar content – or even, when they are representing the same mental state.

Fodor and Lepore [2] argue that connectionist theory of mind cannot give a satisfactory account of different individuals being in the same mental state, for the identity of content follows from the identity of networks, but this condition will never be satisfied in practice. Laakso and Cottrell [3] note the same problem in their statement: *If connectionism is to be an adequate theory of mind, we must have a theory of representation for neural networks that allows for individual differences in weighting and architecture while preserving sameness of content.*

In this article, we consider a method from Laakso and Cottrell [3] for comparing the similarity of representations in connectionist networks, and examine the possibilities of exploiting it for comparing emergent representations in unsupervised learning networks. We report the results based on applying this method as a similarity measure for representations emerging in the Self-Organizing Maps.

II. Measuring the similarity of state space representations

A straightforward way of measuring the similarity of the state space representations in a network, or between two networks having the same configuration, is to measure the distance between the activation levels of their corresponding neurons. In this *position-in-activation-space view* of similarity [1], the proximity of the state space representations are clearly dependent on the positions of activation. It is unclear, however, how two networks with different number of neurons could be compared according to this view, for common distance measures are only defined for vectors of equal lengths.

Identifying content with characteristic groupings of activation patterns was proposed by Churchland [4]. There, it is claimed that people react to the world in a similar way, because their activation spaces are similarly partitioned. Laakso and Cottrell acknowledge this as an evident solution, for it allows different individuals to represent the same latent information without needing to have identical networks.

Adopting this *partitioning-of-activation-space view* to similarity of representations, Laakso and Cottrell [3] propose that content is associated with relative positions in the partitioning of activation space. The momentary representations should then be compared by each representation's location relative to other possible activations in the same network.

Further, Laakso and Cottrell develop a method for assessing the similarity of representations in two networks by comparing their partitionings through correlating the distances between all pairs of activation patterns in each network:

1) Collect the activation patterns evoked by inputs and compute the distances between these representations.
2) Compute the correlation between the distances determined for all different state spaces.

The distances effectively capture the structure of representational space and eliminate the need to match the dimensions of the two spaces.

Laakso and Cottrell [3] test their measure with different architectures of MLP networks that learn to classify colors based on spectral stimuli. The reported results show that

1) networks with the same architecture that were given differently encoded spectral inputs learn similar internal representations,
2) networks receiving identical stimuli learn nearly identical representations, even when their architecture differs.

In computing the similarity of the distances between points in two representational spaces, Laakso and Cottrell [3] provide a *partitioning-of-activation-space criterion* for semantic similarity that answers the challenge Fodor and Lepore [2] place on state space semantics.

III. Emergent representations and association

Laakso and Cottrell experiments base on supervised learning to associate color names with color spectra. This, we think, is not a particularly plausible approach, though maybe intentionally simplified, in the context where they present it:

- differencies in encoding correspond to differencies in sensory organs across animal species,
- different numbers of neurons in the hidden layer of the MLP correspond to variation according to individual and cross-species differences in brain capacity and

The concept of color cannot be adequately studied only by considering the logico-semantic structure of color words. One has to take into account the color as a physical phenomenon. Color naming also requires consideration of the qualities of the human color perception system [5, 6].

We believe that color spectra stimuli as a physiological input to a connectionistic system provides basis for the emergence of latent representations in an unsupervised manner, irrespectively of whether the colors have known symbols or not. Adopting this approach, no color names or ready content are needed for the formation of meaningful representations of the stimuli. An association between a color symbol (name) and spectrum stimulus could grow by their simultaneous excitation of an unsupervised learning connectionist network.

The partitioning-of-activation-space criterion of Laakso and Cottrell can be generally applied to measure the similarity between any two neural representations [3]. In the following we introduce tools to study this criterion and to repeat their experiments in the unsupervised learning framework.

IV. Methodology

A. Self-Organizing Maps

The set of input samples to a connectionist network is described by a real vector $\mathbf{x}_j \in R^n$ where j is the index of the sample. Each node in the Self-Organizing Map (SOM) [7] contains a model vector $\mathbf{m}_i \in R^n$, which has the same number of elements as the input vectors. The nodes of the map form an array with a definite topology. The array is often a two dimensional rectangular grid.

The net outcome of the adaptation process (see [7] for details) is that ordered values for the $\mathbf{m}_i$ emerge over the array. Initial values of the $\mathbf{m}_i$ can be arbitrary. The basic properties of this ordering are that the distribution of the model vectors tends to approximate the density of input vectors, and that the organization of model vectors in the array is such that the mapping tends to preserve the topology of the input space.

The *output or the activation of the SOM*, as a response to stimulus $\mathbf{x}_j$, is the excitation of the best matching unit (BMU) and its neighboring neurons, whose values are determined by the application. This is referred to as *postsynaptic activation* in [8].

A detailed description about the selection of the parameters, variants of the map, and many other aspects have been covered in [7]. The SOM could be considered as an artificial neural network model of the brain [8] e.g. regarding the observed ordered "maps" in the cortex. It can also be viewed as a model of unsupervised statistical machine learning, as an adaptive knowledge representation scheme, as a statistical tool for multivariate analysis, or as a data mining and visualization tool.

B. Associative Mappings with the SOM

Assume two input patterns $\mathbf{x}^{(A)} \in R^{n_1}$ and $\mathbf{x}^{(B)} \in R^{n_2}$ are concatenated to form a single input vector $\mathbf{x}^{(AB)} \in R^{n_1+n_2}$. $\mathbf{x}^{(A)}$ and $\mathbf{x}^{(B)}$ may encode some information A and B presented simultaneously to the SOM. Now the model vectors $\mathbf{m}_i$ have components corresponding to A and B, respectively:

$$\mathbf{m}_i = \begin{bmatrix} \mathbf{m}_i^{(A)} \\ \mathbf{m}_i^{(B)} \end{bmatrix}. \tag{1}$$

During training, the SOM builds an association between A and B. To evoke this association, the BMU c is defined on the basis of $\mathbf{m}_i^{(A)}$ and $\mathbf{x}^{(A)}$ only. An estimate of $\mathbf{x}^{(B)}$, in the sense of the SOM mapping, is obtained as the vector $\mathbf{m}_c^{(B)}$. This recall of the $\mathbf{m}_c^{(B)}$ is referred to as *associative mapping* by Kohonen [7].

If the variance of the patterns $\mathbf{x}^{(A)}$ are large compared to the variance of the $\mathbf{x}^{(B)}$, then component $\mathbf{x}^{(B)}$ of the input has in general little significance in computation of the distance that determines the BMU. Consequently, the organization of the map is not affected by $\mathbf{x}^{(B)}$, but the SOM learns to approximate $\mathbf{x}^{(B)}$ in the SOM neighborhood of $\mathbf{x}^{(A)}$. The special case, where $\mathbf{x}^{(B)}$ is not used in finding the BMU at all is referred to as *heteroassociative mapping* [7].

V. Experiments

A. Data

In order to compare our results with those Laakso and Cottrell presented for the MLP, we prepared the spectrophotometer measurements [9] for the "Munsell book of color: matte finish collection" [10] in the same manner as in [3]. This resulted in 640 patterns of color spectrum $\mathbf{x}^{(S)}$ consisting of colors red, yellow, green, blue and purple with hue values 2.5, 5, 7.5 and 10. The pattern is a 12-dimensional vector, where each component represents the reflectance intensity of a color chip measured at 25 nm intervals from 400 nm to 700 nm ranging

from 0 to 4095. These spectrum patterns were further encoded as described in [3] into *binary*, *real*, *gaussian* and *sequential* representations $\mathbf{x}^{(S_b)}, \mathbf{x}^{(S_r)}, \mathbf{x}^{(S_g)}, \mathbf{x}^{(S_s)}$ having dimensions 96, 12, 60 and 3, respectively. The first two encodings represent the frequency in 8-bit binary and real basis. In the Gaussian, the frequency range is evenly partitioned by five Gaussians in a manner akin to the trichromacy of color vision. Each color is therefore represented by five real numbers. The sequential encoding was formed by numbering the patterns sequentially with three-digit decimal numbers resulting in three-dimensional input. [3] The color names where given symbols R, Y, G, B, P and encoding by the binary vectors

$$\mathbf{x}_j^{(C)} = \begin{cases} [1\,0\,0\,0\,0]^{\mathrm{T}}, & \text{if the symbol is} \quad \mathrm{R} \\ [0\,1\,0\,0\,0]^{\mathrm{T}}, & \text{"} \quad \mathrm{Y} \\ [0\,0\,1\,0\,0]^{\mathrm{T}}, & \text{"} \quad \mathrm{G} \\ [0\,0\,0\,1\,0]^{\mathrm{T}}, & \text{"} \quad \mathrm{B} \\ [0\,0\,0\,0\,1]^{\mathrm{T}}, & \text{"} \quad \mathrm{P} \end{cases} \tag{2}$$

Every sixth of the patterns was taken in the test set and the rest were used as the training set.

B. Testing procedure

To study the effects of encoding, a sample of five SOMs (Sec. IV-A), each with random initial values for model vectors, were trained for the four encodings of the color spectrum. Each SOM was configured to use 13x9 neurons in hexagonal lattice and the Gaussian neighborhood function. To find the effect of the map size, maps of sizes 3x2, 5x3, 8x5, 10x8, 13x9, 15x11, 18x14, 20x15, 22x18 and 25x20 were trained additionally for the binary and sequential encodings.

Following the partitioning-of-activation-space criterion (Sec. II), Euclidean distances of the activations on each map were computed for every pair of the test input patterns. Pearson correlations and their p-values were then computed for these distances. The activation of a neuron was computed with a Gaussian neighborhood function [7], where the radius was set to $1/10$ of the smaller of the dimensions of the map lattice.

For the emergence of an association between spectrum input $\mathbf{x}^{(S)}$ and color symbol input $\mathbf{x}^{(C)}$, these were concatenated to form a single input vector during training (Sec. IV-B). The color symbol part of the input was not used in finding the BMU. After training, the map units were labelled with the color symbol, whose component had the highest value in the color part $\mathbf{m}_i^{(C)}$ of the model vectors. This *strongest association* for each test pattern was compared with the respective color symbol of the test pattern. The performance of recall of the color symbol was recorded.

SOM Toolbox [11] has been utilized throughout this study when working with the SOM.

VI. Results

A. Similarity of the emergent representations

First we want to get an understanding on how similar the input representations resulting from the four encodings of the color spectra are. For this purpose, the correlations between the distances between every input pattern pair are computed as described in Sec. II. Correlations are strong except for the sequential encoding, where the pattern distances only correlate weakly with distances of the other encodings (Fig. 1). The strong correlations indicate that the respective encodings have preserved the relative distances between the patterns to large degree. In these encodings, spectrum samples that are alike are close to each other and between different samples there is relatively longer distance. The reason for the weak correlations with the sequentially encoded patterns is the peculiarity [3] of the sequential encoding itself that hardly reflects the distances between the original physical spectrum patterns.

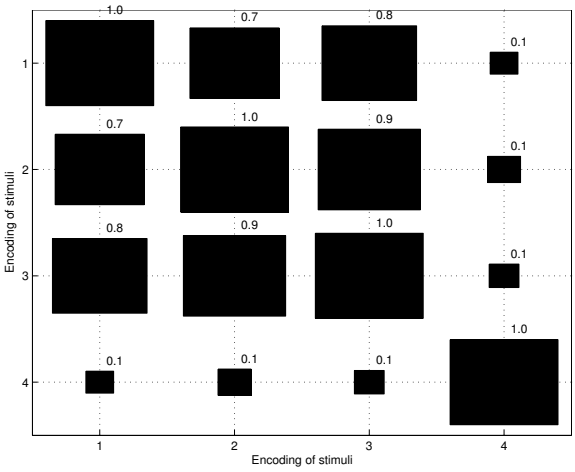

Fig. 1. Representations for the stimuli are similar to each other except in the case of sequential encoding. The Hinton correlations between pairwise distances of the input patterns of different encodings. The area of a box is proportional to the correlation. Black boxes indicate significant correlation (p-value < 0.05). Numbering of the encodings: 1 for binary, 2 for real, 3 for gaussian and 4 for sequential.

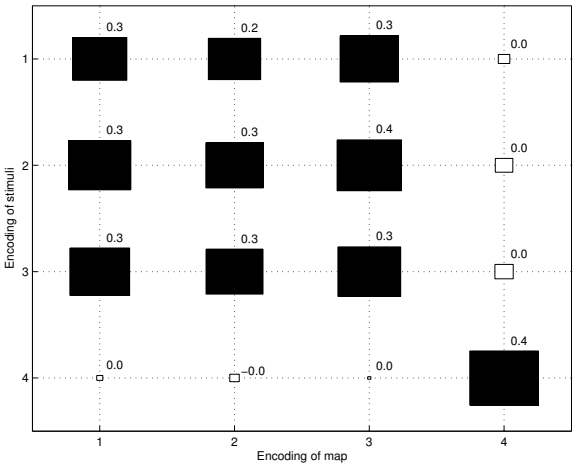

Fig. 2. Emerged representations for the stimuli in the maps are similar to the representations of the stimuli to some degree except for the sequential encoding. The Hinton diagram displays the mean correlations between distances between input patterns for each encoding and distances between activations of five networks trained on each encoding. Black boxes indicate mean p-value less than 0.05. Numbering of the encodings: 1 for binary, 2 for real, 3 for gaussian and 4 for sequential.

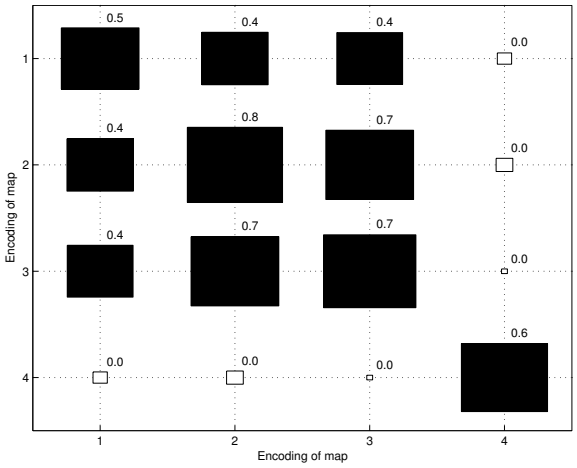

Fig. 3. The emergent representations in the maps are similar irrespectively of the encoding the map was trained with, except for the maps that were trained with the sequentially encoded stimuli. The Hinton diagram displays the mean correlation between activations of five networks trained on each encoding and five networks trained on each other encoding. The area of a box is proportional to the correlation. Black boxes indicate mean p-value less than 0.05. Numbering of the encodings: 1 for binary, 2 for real, 3 for gaussian and 4 for sequential.

Next we examine the similarity between the representation that has emerged in the map for a given input stimulus and the representation of the stimulus itself. For this purpose, the distances between the activations of the map evoked by each input pair are computed. These are found to be similar to some degree across encodings (Fig. 2). The correlations, though significant, are weaker than those between encodings (Fig. 1). Only the sequentially encoded stimulus is not similar to any representation, but its own. As the SOM forms a non-linear mapping that tends to preserve the topology of the input space, there is an expected similarity between the distances of the input patterns and the distances of their representations — in proportion to the similarity of the stimuli encodings.

Finally we compare the emerged organization of representations in the maps, with respect to the partitioning-of-activation-space view of similarity. The distances do correlate, irrespectively of the encoding that the map was trained with, except for the maps that were trained with the sequentially encoded stimulus (Fig. 3). They correlate only in the maps trained using sequential encoding.

The representations are similar, when they correlate, as similar color spectra activate approximately equal positions in the maps, relative to activations stimulated by other spectra samples. If one compares individual responses between two maps they may seem to have no relation at all. This is due to the degree of freedom available to the organization during training. The SOM may take different directions in the organization, depending on the initial values or other randomness in the training phase. It is worth noting that the correlations seem to be stronger between the activations than between the activations and stimuli (Fig. 2).

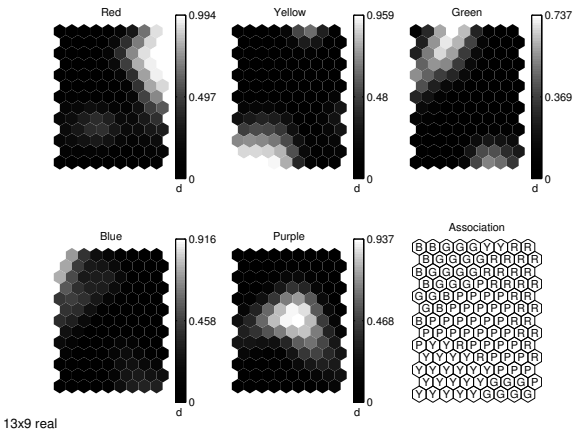

Fig. 4. The strength of the association for each color name in the 13x9 units of a map. This map was trained with the real encoded color spectrum and the encoded color names as input. Strong associations for each color are grouped together.

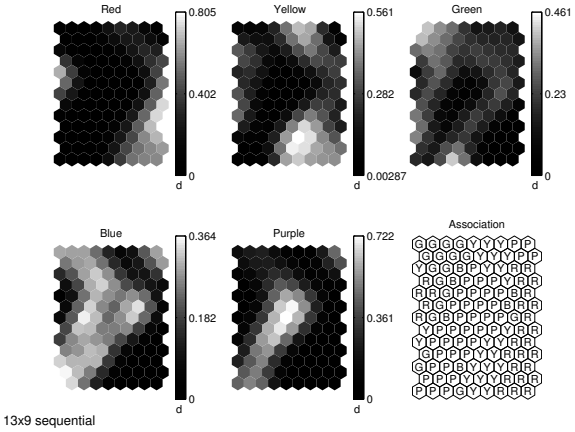

Fig. 5. The strength of the association for each color name in the 13x9 units of a map. This map was trained with the sequentially encoded color spectrum and the encoded color names as input. For most of the colors the association is not strong and is overlapping with associations for other colors.

B. Emergence of association between colors and spectra

The association between color symbols and spectra has emerged during training by their simultaneous input to the map. The result can be analysed by plotting, for every map unit, the values for the components that match the color symbols in the model vectors, $\mathbf{m}_i^{(C)}$. Values near 1 for such a component indicate that the unit has received little other input than 1 for the component in question. This can only happen with the chosen color symbol encoding (2), if spectra that map into this unit, have the same color. Hence high values for a component indicate strong association between the respective color and the stimulus patterns mapping into the unit. Values near 0 indicate that the spectra mapping into the unit is not associated to that color at all. The strength of the association increases as the values for a component increase from 0 to 1.

In the map receiving the real encoded stimuli, color components have high values in distinct regions of the map (Fig. 4). Units in these regions are images of spectra that have

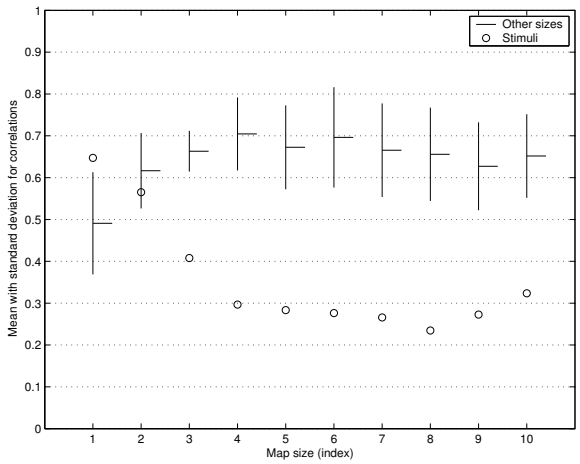

Fig. 6. Correlations between the representations in the maps trained on the real encoding versus map size, and their correlation to input patterns. Representations get more similar to each other as the map size grows, but less similar to input patterns. Map sizes are 3x2, 5x3, 8x5, 10x8, 13x9, 15x11, 18x14, 20x15, 22x18 and 25x20.

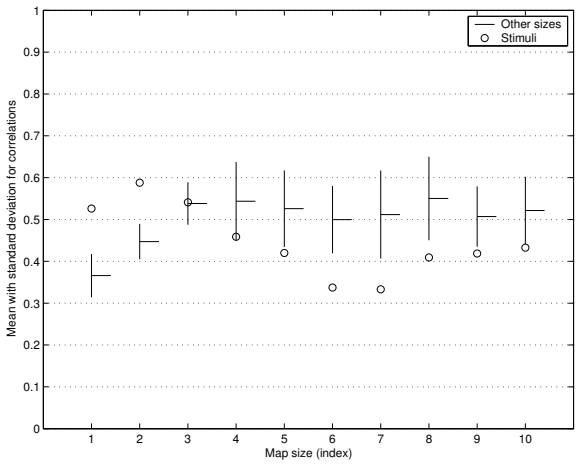

Fig. 7. Correlations between the representations in the maps trained on the sequential encoding versus map size, and their correlation to input patterns. Representations get slightly more similar to each other as the map size grows, but slightly less similar to input patterns. Map sizes are 3x2, 5x3, 8x5, 10x8, 13x9, 15x11, 18x14, 20x15, 22x18 and 25x20.

the respective color symbol as the component. For the real encoding, the same colors are very well grouped together in the input space.

In contrast to the real encoding, for the sequentially encoded input, the regions occupied by the color symbols are noticeably overlapping for some colors (Fig. 5). This happens because the input patterns have little information about the color. One of the tree dimensions of the input carries the information. Association can grow only along that dimension.

When we compare the results for the recalling of the color symbols of the input patterns, one has to remember that the maps were not specifically trained to recall the color symbol, but the color component having the highest value indicates the association. For the real and sequential encoded maps the association was correct for 68% and 40% of the test patterns, respectively. This further supports the idea that the encodings are qualitatively different.

C. Effects of scaling of the map capacity

The number of units in the SOM determines its capacity to differentiate between input patterns. In order to rule out the possibility that the results are remarkably dependent on the map capacity, we compared the results for maps of different sizes for the real and sequentially encoded inputs.

When we examine the similarity between the SOM representations and the real encoded stimuli as a function of map size (Fig. 6), the inputs are similar to their representation for the maps having less than 80 units. For the larger maps this similarity is not very strong. On the other hand, if we look at the similarity of the representations as the maps grow larger, we note that it increases up to sizes of 40 units and then saturate. Similarity between the representations is clearly greater than the similarity between representations and stimuli for map sizes larger than 40 units.

In contrast to the real encoding, the maps trained with the sequential encoding show little difference between the similarity across representations and the similarity between representations and stimuli (Fig. 7) — though the same trends for the effects of the size can be recognised. In this encoding there is little detailed structure that the SOM could reveal even with greater capacity.

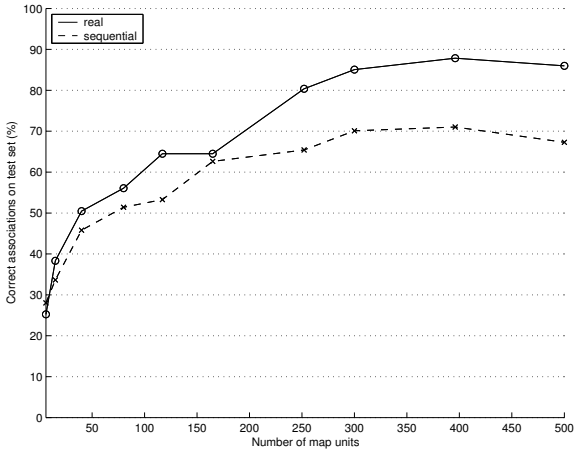

Fig. 8. Correct association for color symbol improves with increasing map size. Accuracy is better for the real encoded spectrum than for the sequentially encoded. The map sizes are 3x2, 5x3, 8x5, 10x8, 13x9, 15x11, 18x14, 20x15, 22x18 and 25x20.

As the capacity of a map grows, its resolution in discriminating between inputs improves. This is evident in Fig. 8 that shows the accuracy improving up to sizes of 300 units and then saturating at the level of 85% for the real encoding and 70% for the sequential encoding.

VII. Discussion

We have studied the relationship between continuous (perceptual) domain and discrete (symbolic, linguistic) domain in supervised learning framework (see also [12]). In particular,

we have considered how different encodings or representations of the input data influence concept formation process.

Figures 6 and 7 show that there is a much weaker relationship between the stimulus and the coded representations than across representations. This was also reported in [3]. More interestingly, when the map size is small, i.e., there are not enough degrees of freedom to account for the complexity of the data to be coded, the best it can do is to get close to reproducing the input. This is the reason for the poor results found for small sizes of the maps — metaphorically, it would correspond to being able to simply reproduce the inputs in a 'parrot-like' manner.

When the degrees of freedom increase, the map representation is able to reach 'meaningful' coding of the inputs, in such a way that formation of the internal semantics occurs, hence getting more distant from the inputs, but better structured. Using a similar analogy as in the above, one could say metaphorically that the map is capable of understanding the meaning of what it is producing. After reaching a certain degree of complexity, any increase of map size can only help refining the structuring.

Figure 7 shows that, if the input encoding is 'unnatural', it can be expected that without a clear external constrain to the representation, i.e., supervision, all maps can not reach the desired representation. The maps then stay in the level of simply reproducing as much as possible the input pattern. We could say that these maps have not found any significant internal structure, content, in the stimuli.

The measure of similarity presented in [3] is easily transposable to unsupervised mapping. We still find it to be a very useful one. Emergent representations follow a similar path as supervised codings, as different systems (e.g. varying sizes of maps) reach similar formation of the core content.

We have shown that supervision is not needed in order to gain meaningful representations regardless of the input encoding if the encoding can be considered 'natural'. Of course, raw input may not always be sufficient source for meaning conceptual organization but some external or secondary information is necessary. However, we claim that the statistical characteristics of the primary input data is a reasonable starting point for the formation of conceptual structures.

In the future, we plan, e.g., to study more in detail the notion of 'naturalness' of encoding schemes, and the sources of variation and ambiguity in the concept formation process. We also aim to take into account the apparent hierarchical nature of many conceptual structures.

VIII. Conclusions

The motivation behind the present paper was to examine Laakso and Cottrell findings regarding measures of similarity between representations [3], in emergent, i.e. unsupervised environments. We observed the following:

1) the SOM learns representations both for stimuli and color symbols and is able to associate them successfully,
2) application of the partitioning-of-action-space criterion for measuring the similarity of the latent representations for the stimuli show that the representation are alike almost independently of the encoding used for input.

The discovered usability of this criterion for the emergent representations, adds new support in favour of the state space semantic view of mind, and gives a counter example against the challenges Fodor and Lepore [2] have placed on the connectionist theory.

Acknowledgement

The authors would like to thank the participants of the seminar on Statistical and adaptive approaches to conceptual modeling [13] organized by the Laboratory of Computer and Information Science at Helsinki University of Technology.

References

[1] P. M. Churchland, "Some reductive strategies in cognitive neurobiology," in *A neurocomputational perspective: the nature of mind and the structure of science*. Cambridge, MA: MIT Press/Bradford Books, 1986, pp. 279–309.

[2] J. A. Fodor and E. Lepore, "Paul Churchland and state space semantics," in *The Churchlands and their critics*, R. N. McCauley, Ed. Blackwell, 1996.

[3] A. Laakso and G. Cottrell, "Content and cluster analysis: assessing representational similarity in neural systems," *Philosophical psychology*, vol. 13, no. 1, pp. 47–76, 2000.

[4] P. M. Churchland, "Learning and conceptual change," in *A neurocomputational perspective: the nature of mind and the structure of science*. Cambridge, MA: MIT Press/Bradford Books, 1989, pp. 231–253.

[5] C. Hardin, *Color for Philosophers*, extended ed. Indianapolis/Cambridge: Hackett Publishing Company, 1995.

[6] S. Zeki and L. Marini, "Three cortical stages of colour processing in the human brain," *Brain*, vol. 121, pp. 1669–1685, 1998.

[7] T. Kohonen, *Self-Organizing Maps*, 3rd ed., ser. Springer Series in Information Sciences. Springer-Verlag, 2001.

[8] T. Kohonen and R. Hari, "Where the abstract feature maps of the brain might come from," *Trends Neurosci.*, vol. 22, pp. 135–139, 1999.

[9] Anonymous, "Joensuu spectra databases," 2003, http://cs.joensuu.fi/spectral/databases/.

[10] ——, *Munsell book of color: matte finish collection*. Baltimore: Munsell Color, 1976.

[11] J. Vesanto, J. Himberg, and E. Alhoniemi, "SOM Toolbox for Matlab 5," Helsinki University of Technology, Espoo, Finland, Publications in Computer and Information Science A57, 2000.

[12] T. Honkela, "Self-organizing maps in symbol processing," in *Hybrid neural systems*. New York, NY, USA: Springer, 2000, pp. 348–362.

[13] T. Honkela, K. I. Hynnä, K. Lagus, and J. Särelä, Eds., *Adaptive and Statistical Approaches in Conceptual Modeling*, ser. Publications in Computer and Information Science, no. A75. Espoo, Finland: Helsinki University of Technology, 2004.