A Feature Selection Approach for Emulating the Structure of Mental Representations

Marko Tscherepanow[1,4], Marco Kortkamp[1], Sina Kühnel[2,4], Jonathan Helbach[1], Christoph Schütz[3,4], Thomas Schack[3,4]

[1]Applied Informatics, Faculty of Technology
[2]Physiological Psychology, Faculty of Psychology and Sport Sciences
[3]Neurocognition and Action, Faculty of Psychology and Sport Sciences
[4]CITEC, Cognitive Interaction Technology, Center of Excellence
Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany

{marko,mkortkam}@techfak.uni-bielefeld.de
skuehnel@cit-ec.uni-bielefeld.de

Abstract. In order to develop artificial agents operating in complex ever-changing environments, advanced technical memory systems are required. At this juncture, two central questions are which information needs to be stored and how it is represented. On the other hand, cognitive psychology provides methods to measure the structure of mental representations in humans. But the nature and the characteristics of the underlying representations are largely unknown. We propose to use feature selection methods to determine adequate technical features for approximating the structure of mental representations found in humans. Although this approach does not allow for drawing conclusions transferable to humans, it constitutes an excellent basis for creating technical equivalents of mental representations.

Keywords: Feature selection, Mental representations, Memory.

1 Introduction

One of the biggest challenges today is the endeavour to copy or emulate memory as it is found in humans and animals. In principle, memory constitutes the basis for any kind of learning to be performed. Therefore, a multitude of approaches related to the topic of memory in artificial systems have been proposed. They adopt single properties of natural memory, in particular, its structure [11], its processes [1], or mental representations [3].

A crucial problem with developing artificial agents using memory systems is the formation of appropriate technical representations of perceptual data. Similar to natural agents possessing cognitive capabilities, technical memory systems have to obey the principle of cognitive economy [7]; i.e., the amount of data needs to be diminished before it is stored. Otherwise, the deluge of incoming sensory information would quickly consume the entire memory. Nevertheless,

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part III, LNCS 7064, pp. 639–648, 2011.

the formed representations need to contain the relevant information. Two important methods for achieving this goal are the formation of categories [7] and dimensionality reduction [4].

The goal of our work consists of the emulation of the structure of human mental representations by means of features that can be computed from visual stimuli (images). In order to comply with the principle of cognitive economy, the resulting feature sets should be as small as possible. Therefore, several feature selection methods are compared. As the selected features contain the information to replicate the results obtained from humans, we assume that they are good candidates for representing the corresponding images in artificial systems.

In Section 2, we introduce different methods for analysing mental representations in humans. Afterwards, popular feature reduction methods are discussed in Section 3. Our complete approach is described in Section 4 and evaluated in Section 5. Finally, Section 6 summarises the most important outcomes.

2 Psychological Background

One way to obtain knowledge about human mental representations consists of conducting experiments in which subjects assign labels to perceived stimuli (e.g., [10]). From these, conclusions about the internal concepts and features used for classification can be drawn. But degrees of class membership are usually not reflected. In [14], a method explicitly avoiding semantic groups was applied: The subjects successively split presented images into two groups. Here, images of one group should share a common global aspect, structure, or certain elements. Afterwards the subjects were asked to verbally describe the splitting criteria used. Hence, the features utilised for splitting were associated with a label, e.g., naturalness, which itself represents a concept.

Structural Dimensional Analysis (SDA) [13] constitutes an alternative approach to the analysis of mental representations. In contrast to the methods introduced above, it does not require labels provided by subjects. In cognitive psychology, SDA is a well-established method for psychometrically investigating the representational structure of concepts in long-term memory. The concepts under analysis are verbally defined by the experimenter, e.g., 'wood', 'brush', and 'hat' [13]. This original SDA method was extended to the analysis of the representational structure of motor skills, which is called Structural Dimensional Analysis-Motoric (SDA-M) [2,16]. The extension from verbally defined concepts to movements was achieved by introducing so-called basic action concepts (BACs), which represent components of complex movements that are characterised by perceivable features. In this context, SDA was shown to work with visual stimuli as an alternative to verbal descriptions.

3 Relevant Feature Reduction Methods

For numerous machine learning techniques, a feature reduction step is required in order to avoid problems arising from the *curse of dimensionality* [8]. Fea-

ture reduction methods aim to decrease the dimensionality of the feature space while minimising the information loss. This is achieved by removing irrelevant and redundant information. In general, feature reduction can be divided into two principal methods: feature extraction (e.g., principal component analysis [9] and independent component analysis [9]) and feature selection (e.g., minimum-redundancy-maximum-relevance [15] and genetic algorithms [19]). Feature extraction computes a functional mapping for which the underlying meaning of the features is lost. In contrast, feature selection chooses salient features from the original feature set and thereby preserves the features' semantics. Moreover, unselected features do not need to be computed. For these reasons, we decided to focus on feature selection rather than feature extraction approaches.

Feature selection methods can be divided into filters and wrappers. While filters operate independently of the utilised machine learning technique and optimise pre-selected criteria [8,15], wrappers optimise the actual learning results provided by an induction algorithm [19]. As a consequence, wrappers often lead to better results while filters are less computationally expensive [8]. Additionally, feature selection approaches can be distinguished depending on the way they determine sets of relevant features. Some methods first measure the quality of individual features and rank them [12]. Then, the top-ranked features are selected. Since the actual number of required features is hard to choose, other approaches directly select feature subsets [8,15]. These feature subset selection methods usually provide better results than ranking-based techniques, since they account for redundancies and complex interdependencies of the considered features [15].

Structural Dimensional Analysis itself can identify common features in different representational units (items) [13,2,16]. This is achieved by means of factor analysis [9], which is frequently applied in psychology. It provides meaningful factors explaining observed results and is closely related to the feature extraction methods mentioned above. However, the factors themselves cannot be computed from the stimuli. They rather are unobservable variables describing the experimental results obtained from human subjects.

4 Our Approach

An overview of our approach is given in Fig. 1. First, SDA is performed in order to measure the representational distances of a set of images in different human subjects (cf. Section 4.1). Then, feature-based representations of these images are determined. In principle, any kind of real-valued features which can be computed from individual images could be applied here. From this initial set, subsets of features are selected. An initial set or a selected subset are considered valid, if they enable an accurate mapping to the measured representational distances; i.e., if such a mapping exists, we assume that the respective features contain all information required for the reproduction of the representational distances found in humans. But the mapping itself may be very complex and non-linear.

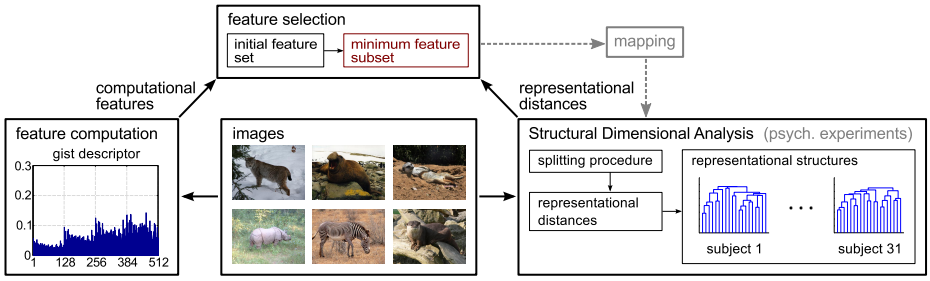


Fig. 1. Principal approach. A feature selection method chooses minimum sets of computable features in order to determine an efficient technical representation of natural images. The representational distances determined by SDA serve as ground truth, which is to be approximated using the respective feature subsets.[1]

In order to test for the existence of a mapping from the feature-based representations of the images to the measured representational distances, we attempt to compute an adequate regression model. Provided that such a model has been found, it is concluded that the applied feature set suffices to represent the images under consideration. In contrast to traditional machine learning approaches, the application of distinct test and training datasets is neither possible nor necessary for the training of the regression models. Firstly, such training sets would not be representative for the complete input distribution, as the underlying human information processing is too complex and results in unpredictable representational distances between untrained stimuli. Secondly, we aim at explaining observed data, similar to SDA, and do not require good generalisation properties of the regression models. But unlike the regression models, the determined feature subsets are validated.

As the maximum number of images and, therefore, the amount of available samples is very limited due to the algorithmic properties of SDA (see Section 4.1), we decided to apply Support Vector Regression[2] (SVR) [17]. Regarding the task of feature selection, several methods are compared (see Section 4.2).

4.1 Generating Ground Truth Data – SDA

In a first step, SDA seeks to gain information about the distance between representational units corresponding to a set of n_s selected stimuli. Since the structure of mental representations can only be explicated by subjects to a limited extent, this is achieved by a special splitting technique: one stimulus is chosen as an anchor and the remaining stimuli are compared to it (in random order) and manually classified as 'similar' or 'dissimilar'. This is repeated for the resulting subsets until they become too small to be split or the subject decides that

[1] lynx, sea-elephant, meerkat, and otter: CC-by-SA 3.0 Unported; rhinoceros and zebra: CC-by-A 2.0 Generic

[2] We used the ν-SVR implementation of LIBSVM, version 3.0.

further splitting is not reasonable. Thus, a decision tree is constructed. The splitting procedure is repeated in such a way that each stimulus serves as an anchor. Therefore, the number of constructed decision trees equals the number of stimuli.

In order to obtain a distance measure, the algebraic sums along all branches are computed for each decision tree. Here, stimuli classified as 'dissimilar' obtain a negative sign and elements classified as 'similar' a positive sign. From the resulting values, a matrix is constructed, with its elements s_{ik} denoting the sum for stimulus k with respect to anchor i. These sums are z-transformed:

$$z_{ik} = \frac{s_{ik} - \mu_i}{\sigma_i} \quad \text{, with} \quad \mu_i = \frac{1}{n_s}\sum_{k=1}^{n_s} s_{ik} \quad \text{and} \quad \sigma_i = \sqrt{\frac{1}{n_s}\sum_{k=1}^{n_s}(s_{ik} - \mu_i)^2}. \quad (1)$$

Then, a correlation matrix is computed. The individual correlation r_{ij} of two stimuli i and j is further transformed into the Euclidean distance measure d_{ij}:

$$d_{ij} = \sqrt{2n_s}\sqrt{1 - r_{ij}} \quad \text{, with} \quad r_{ij} = \frac{1}{n_s}\sum_{k=1}^{n_s} z_{ik} z_{jk}. \quad (2)$$

The computed distances d_{ij} are subjected to a hierarchical cluster analysis which reveals the representational structure of the stimuli and constitutes the second step of SDA. As the mental representations differ between the individuals of a population, the measured structures exhibit differences as well. The third step comprises a cluster-dependent factor analysis revealing underlying dimensions in the structured set of representations and the final step consists in testing for invariance within and between groups of subjects.

Our work focusses on the first step, as the distance values provided therein completely define the representational structure revealed by cluster analysis. Since d_{ii} always equals 0 independent of the representational structure and the underlying representations, we decided to omit these values. Furthermore, the number n_d of available distance values is reduced due to symmetry (d_{ij}=d_{ji}). It amounts to $\frac{1}{2}n_s(n_s-1)$. Due to the high number of comparisons which have to be performed by the subjects (up to $O(n_s^3)$), the number of obtainable distance values is very limited. In particular, n_s should not be chosen higher than 20. Otherwise, the decisions made regarding the similarity of stimuli may become inconsistent.

4.2 Feature Selection

In order to select adequate features, it must be considered that each subject has individual mental representations of the images and, therefore, the mappings from the stimuli to their representations and the resulting representational structures may vary considerably. Nevertheless, it would be beneficial if the data from different subjects could contribute to a common feature subset, as the amount of available data is considerably increased this way. Furthermore, we assume that

the principal way of information processing does not differ considerably between different healthy human subjects. Thus, our approach aims at aligning the feature subsets found for all subjects, in addition to pursuing the traditional goal of minimising the number of selected features.

The nature of the task at hand implies the usage of a wrapper approach, as the quality criterion consists of the accurate approximation of the representational distances d_{ij} provided by SDA. Hence, we decided to apply a genetic algorithm [6], due to the flexibility of this method. The developed algorithm is specifically tailored to the problem at hand. Nevertheless, it would be advantageous if standard methods could be applied as well. Therefore, we analysed two further feature selection approaches, namely Correlation-based Feature Selection (CFS) [8] and ReliefF [12].[3] As these methods are filters, we expected them to be less computationally expensive than the genetic algorithm. But they are not able to process the subjects individually while simultaneously aligning their results. Hence, we applied these methods to the collective data of all considered subjects, in order to find a single feature subset.

Genetic Algorithm In our genetic algorithm (GA), a candidate solution, also called an individual, constitutes a combination of a feature subset and an associated regression model approximating the representational distances of a specific subject. As a result, the genome of each individual comprises two components: (i) the feature genome $\underline{g}^f$ defining the selected features (and the dimensionality of the feature space) and (ii) the parameter genome $\underline{g}^p$ defining the parameters for the SVR. Here, three possible kernels – linear, radial basis function (RBF), and sigmoid – are considered depending on the parameter *type*.

While feature subsets are defined by binary genes denoting whether a specific feature is selected or not, the SVR parameters are encoded as numerical values from the interval [0,1]. For the regularisation constant C and the kernel parameters γ, κ, and ϑ, these numerical genes are mapped to the interval [0.00001, 10000]. The feature genome is adapted by bit mutation [6] with the probability p_m and uniform cross-over [6]. For the parameter genome, a mutation operator for real-valued genes[4] [5] and arithmetic cross-over [5] are utilised. p_c denotes the cross-over probability for both operators. In the initial generation, features are randomly selected with the mutation probability p_m.

In order to align the feature sets selected for different subjects, each feature i is assessed by a weight

$$w(i) = \frac{\sum_{A \in \mathcal{E}} g_A^f(i)}{\sum_{j=1}^{n_f} \sum_{A \in \mathcal{E}} g_A^f(j)} \tag{3}$$

reflecting the frequency of its occurrence in the set $\mathcal{E}$, which summarises the elite individuals of the current generation for all subjects. Here, n_f denotes the number of features.

[3] For CFS and ReliefF, the implementations of WEKA, version 3.6.3, were used.
[4] Changes are sampled from the Gaussian $\mathcal{N}(0, 0.025^2)$.

The three goals explained above are reflected by the fitness function which is used for evaluating the performance of each individual A:

$$F(A) = 1 - (1 - c_f - c_h) \underbrace{E(A)}_{\text{(i)}} - c_f \underbrace{\frac{1}{n_f} \sum_{i=1}^{n_f} g_A^f(i)}_{\text{(ii)}} - c_h \underbrace{\left(1 - \sum_{i=1}^{n_f} g_A^f(i) w(i) \right)}_{\text{(iii)}} . \quad (4)$$

Component (i) minimises the regression error $E(A)$, component (ii) minimises the size of the chosen feature subset, and component (iii) assures the alignment of selected features across all subjects.

The constants should be chosen as follows: $1 \gg c_f \gg c_h$. By this, the regression error obtains the highest priority, followed by the feature set size and the alignment of feature sets between different subjects. As the influence of the components (ii) and (iii) is very small compared to the regression error $E(A)$, we applied rank-based selection [5].

The final feature subset consists of those features which were applied by all elite individuals of the final generation.

5 Results

The suitability of the considered feature selection methods was analysed using 20 images (400×300 pixels) showing different, complete and centred animals in their natural environment (cf. Fig. 1). The representational distances between these images were measured for s=31 subjects (16 male, 15 female; age: 21–46) resulting in n_d=190 samples per subject and 5,890 samples in total. In order to alleviate the evaluation, the distances of each subject were normalised to the interval $[0, 1]$. As an example for feature-based image representations, we employed the well-established gist descriptor (n_f=512) [14]. For the evaluation, ten different splittings of the set of subjects into subsets of 20 training subjects and 11 test subjects each were randomly created. The data of the respective training subjects is applied so as to select salient features. These suitability of these features is tested with respect to the test subjects.

For each training subject, the genetic algorithm optimised 100 individuals, the 10 fittest of which were regarded as elite individuals.[5] Figure 2 depicts the development of the mean regression error $\mu(\overline{E})$ for the respective training subjects and the mean number of selected features $\mu(\overline{n}_{sf})$ during the optimisation. In addition, the mean normalised histogram intersection $\mu(\overline{H})$ [18] is plotted.[6] It measures the similarity between the feature usage histograms of the elite individuals assigned to different subjects. Figure 2 demonstrates that the genetic algorithm achieves the goals stated in Section 4.2.

[5] The remaining parameters were set to the following values: p_c=0.25, c_f=0.01, and c_h=0.0005.

[6] Over-lined symbols denote average values over all elite individuals and/or training subjects, while μ and σ denote the mean and the standard deviation over the different splittings, respectively.

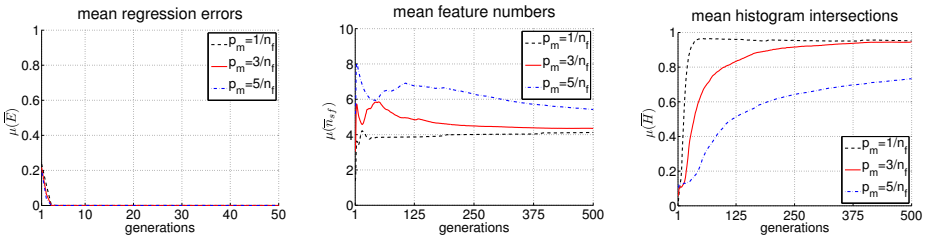
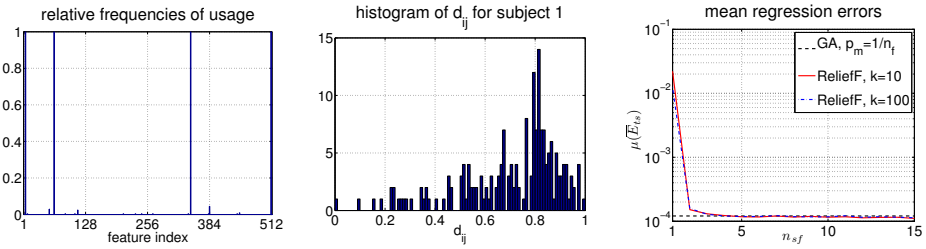


Fig. 2. Results of the evolutionary optimisation performed by the genetic algorithm. The regression error decreases rapidly during the first generations (left). The feature number (centre) and the alignment of the features subsets of different subjects (right) require more generations to converge. Here, larger values of the mutation probability p_m retard the optimisation process.



(a) (b) (c)

Fig. 3. Relative frequencies of feature usage by the elite individuals after a single run of the genetic algorithm (a), distribution of the normalised representational distances d_{ij} for a single subject (b), and mean regression errors of ReliefF for the test subjects depending on the number of selected features n_{sf} (c).

In order to determine the final feature subset, the relative frequencies of the usage of features by the elite individuals of the final generation are analysed. An exemplary result is shown in Fig. 3(a). Those four features, which were used by all elite individuals, constitute the resulting feature subset.[7]

The validity of the chosen feature subsets was tested using the data of the respective test subjects (see Table 1). The parameters ν, C, and γ for the SVR (RBF kernel) were determined by grid search (11 values per parameter) individually minimising the regression error E_{ts} for each test subject. The genetic algorithm was compared to CFS and ReliefF. In case, the computation of the regression models did not terminate using the default criterion (ϵ=0.001), the respective splitting was omitted.[8] The results for CFS using the default parameters and different search directions are given in Table 1, as well. Here, it must be

[7] Due to the random nature of the genetic algorithm and redundancies in the initial feature set, the actually selected features varied across different trials. But their number was approximately constant.

[8] CFS, forward: splitting 4; ReliefF, k=100: splittings 6 and 8

Table 1. Means μ and standard deviations σ of the regression errors E_{ts} for the test subjects and the corresponding sizes n_{sf} of the chosen feature subsets.

feature selection approach	$\mu(\overline{E}_{ts})$	$\sigma(\overline{E}_{ts})$	$\mu(\overline{n}_{sf})$	$\sigma(\overline{n}_{sf})$
genetic algorithm, p_m=1/n_f	$1.21{\cdot}10^{-4}$	$6.07{\cdot}10^{-6}$	3.9	0.7
genetic algorithm, p_m=3/n_f	$1.25{\cdot}10^{-4}$	$8.92{\cdot}10^{-6}$	4.1	0.7
genetic algorithm, p_m=5/n_f	$1.23{\cdot}10^{-4}$	$9.57{\cdot}10^{-6}$	3.9	0.7
CFS, forward	$1.17{\cdot}10^{-4}$	$6.52{\cdot}10^{-6}$	11.0	2.67
CFS, backward	$1.17{\cdot}10^{-4}$	$5.34{\cdot}10^{-6}$	13.5	2.06
CFS, bi-directional	$1.15{\cdot}10^{-4}$	$5.71{\cdot}10^{-6}$	10.2	1.99

considered that the majority of the distances d_{ij} is centred around a single peak (see Fig. 3(b)). Therefore, very small errors are required in order to preserve the representational structure.

Both the genetic algorithm and CFS enable the approximation of the representational distances with high accuracy. But the feature subsets determined by CFS are larger. This is likely to be a result of the collective processing for all training subjects.

In contrast to our approach and CFS, ReliefF does not directly select feature subsets but provides quality assessments and a ranking. Figure 3(c) depicts the mean regression error depending on the number of selected features using two different neighbourhood sizes k. If the 4 top-ranked features are selected, the regression errors are comparable to the genetic algorithm and CFS. A further increase of the feature set size does not lead to significant improvements, although ReliefF collectively processed the data of all training subjects like CFS.

6 Conclusion

We compared several feature selection methods regarding their ability to select subsets of computable features enabling the emulation of the structure of mental representations found in humans. Standard feature selection methods, in particular CFS and ReliefF, achieved results comparable to a genetic algorithm that was specifically tailored to this problem. Using such methods, the results of SDA can be explained in terms of small sets of salient features which are directly computable from the stimuli. In the future, the resulting feature sets could be exploited to learn human-like representational structures in technical agents. For example, adequate feature subsets could be determined off-line. As they preserve the relevant information, their usage instead of the original stimuli would not reduce the potential learning capabilities of the agent during interaction with its environment. However, the amount of data to be stored would be considerably reduced.

Acknowledgements. This work was partially funded by the German Research Foundation (DFG), Excellence Cluster 277 “Cognitive Interaction Technology”.

References

1. Amor, H.B., Ikemoto, S., Minato, T., Jung, B., Ishiguro, H.: A neural framework for robot motor learning based on memory consolidation. In: Proceedings of the International Conference on Adaptive and Natural Computing Algorithms. LNCS, vol. 4432, pp. 641–648. Springer, Berlin (2007)
2. Bläsing, B., Tenenbaum, G., Schack, T.: The cognitive structure of movements in classical dance. Psychology of Sport and Exercise 10, 350–360 (2009)
3. Chartier, S., Giguère, G., Langlois, D.: A new bidirectional heteroassociative memory encompassing correlational, competitive and topological properties. Neural Networks 22(5–6), 568–578 (2009)
4. Edelman, S., Intrator, N.: Learning as extraction of low-dimensional representations. In: Goldstone, R.L., Schyns, P.G., Medin, D.L. (eds.) Perceptual Learning, pp. 353–380. Academic Press, San Diego (1997)
5. Engelbrecht, A.P.: Computational Intelligence. John Wiley & Sons, Hoboken, 2nd edn. (2007)
6. Fogel, D.B.: Evolutionary Computation. IEEE Press, Piscataway, 3rd edn. (2006)
7. Goldstone, R.L., Kersten, A.: Concepts and categorization. In: Healy, A.F., Proctor, R.W. (eds.) Handbook of Psychology, vol. 4: Experimental Psychology, pp. 599–621. John Wiley & Sons, Hoboken (2003)
8. Hall, M.A.: Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, Department of Computer Science, The University of Waikato, Hamilton, New Zealand (1999)
9. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, New York (2001)
10. Johansen, M.K., Kruschke, J.K.: Category representation for classification and feature inference. Journal of Experimental Psychology 31(6), 1433–1458 (2005)
11. Kawamura, K., Gordon, S.M., Ratanaswasd, P., Erdemir, E., Hall, J.F.: Implementation of cognitive control for a humanoid robot. International Journal of Humanoid Robotics 5(4), 547–586 (2008)
12. Kononenko, I., Šikonja, M.R.: Non-myopic feature quality evaluation with (R)ReliefF. In: Liu, H., Motoda, H. (eds.) Computational Methods of Feature Selection, pp. 169–191. Chapman & Hall/CRC, Boca Raton (2008)
13. Lander, H.J., Lange, K.: Untersuchungen zur Struktur- und Dimensionsanalyse begrifflich repräsentierten Wissens. Zeitschrift für Psychologie 204, 55–74 (1996)
14. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
15. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
16. Schack, T., Mechsner, F.: Representation of motor skills in human long-term memory. Neuroscience Letters 391, 77–81 (2006)
17. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Computation 12, 1207–1245 (2000)
18. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision 7(1), 11–32 (1991)
19. Yu, L., Chen, H., Wang, S., Lai, K.K.: Evolving least squares support vector machines for stock market trend mining. IEEE Transactions on Evolutionary Computation 13(1), 87–102 (2009)