Concept-based lexical-semantic unsupervised learning system

V I Filatov[1]
[1]Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, 49 Kronverkskiy pr., Saint Petersburg 197101, Russia

E-mail: sigal89@mail.ru

Abstract. Concept learning is essential for automatic knowledge acquisition and consists in linking semantic and linguistic data together. In this work, a system of concept learning, which mimics to some extent human infant learning, is presented. This system performs visual and audial feature extraction and construction of concept database by maximizing of their mutual information. Experiments show 90% recognition rate of learnt concepts.

1. Introduction

Within the history of artificial intelligence system engineering, scientists have proposed various ways of machine learning. At first, the best-known kinds of systems operating with knowledge were logic-based. Expert systems were based on binding existing knowledge in order to assist humans in decision-making. The next big breakthrough in intelligent systems consisted in development of automatic knowledge acquisition methods.

Most promising type of machine learning in knowledge acquisition is unsupervised learning since it doesn't require manually labeled datasets. Unsupervised learning systems are now very popular, and novel approaches are appearing every day. Some of them imitate animal reflexes [1].

This work describes novel algorithms and approaches to concept-based learning. An experimental computer system called Automatic Lexical-Semantic Artificial Intelligence System (ALSAIS) was designed. Underlying ideas for this system originate from different existing approaches to concept-based learning [2, 3] and also from the attempt to mimic human infant learning in machine knowledge acquisition methods [4].

2. ALSAIS description

The main ALSAIS idea is based on emulation of infant mechanisms of knowledge acquisition (namely, concept-learning). Concept-learning requires splitting information into semantic and linguistic data channels. Both these channels are necessary for constructing concepts. Semantic data contains visual features and can be represented by information collected with a video camera. Linguistic data is information which includes speech features; it can be represented by an audio channel or a text input.

By combining simultaneous extracted audio and video features, dictionary database is constructed. This database contains pairs of lexical-semantic features. One such pair in this article is referred to as a lexical-semantic concept.

At first, new information is transmitted to the system's video and audio inputs. Semantic data is represented by video information of a scene; linguistic data is represented by audio information. Each grabbed video frame contains video description of the scene with some object, which visual features

can be extracted from the frame. There is also a predefined condition that the objects of interest should be moving. Thus, the first-priority task is to find objects by performing motion segmentation. After that, the object is represented by a feature vector extracted from its image.

Although similar actions are performed with the audio channel, there is some difference. It is well-known that speech directed to infants is often redundant. The construction of an audio feature vector representing speech lexemes in this work is based on this assumption.

After construction of visual and audio feature vectors, their simultaneous analysis should be performed. Thus, the next task is to find correlations between audial and visual feature vectors. To do this, characteristics of the speech directed to infants should be taken into account. Here, we use mutual information as the learning criterion, which is quite popular in the domain of learnable computer vision systems [5], but first of all, feature vectors must be prepared for using. Our system performs automatic clustering in feature space and separate visual feature vectors between them. This gives us an opportunity to get particular features which describe parts of the object; we will refer to these features as visual cluster's characteristics.

By using the maximization of mutual information between the audial lexemes and the visual cluster's characteristics, the system constructs a dictionary of their pairs and calculates their estimated probabilities. The system only adds the pairs with maximum mutual information to our database.

ALSAIS can be described as a set of modules solving the following tasks: segmentation; constructing the visual lexemes feature vectors; constructing the audial lexemes feature vectors; clustering the feature vectors; maximization of mutual information.

All of these steps are repeated yielding an ability to add new concepts and their features to the knowledge base. Thus, the constructed knowledge base will contain all the descriptions of the objects in the scene. This knowledge allows the computer system to collect information about the objects being learnt.

2.1. Segmentation. As it was mentioned above, the process of segmentation consists in separating background pixels and object pixels in the grabbed image. In our system, two algorithms are combined together: optical flow and SURF descriptor.

2.2. Visual lexemes building. Each grabbed frame can be described by a visual feature vector. There are particular visual events which correspond to considerable changes in the current visual scene. The absence of the perceptible optical flow changes in the grabbed frame is used as a separator between the visual events. This idea helps us to group feature vectors belonging to the same object presented in different frames.

2.3. Audial lexemes building. The feature vectors of audial lexemes are extracted from the input audio channel. Lexemes can be obtained by the assumption of redundancy of speech directed to infants. For this purpose the algorithm of finding longest high-frequency substrings is used.

2.4. Data clustering and the maximization of mutual information. Clustering is used to group particular object features in clusters.

Mutual information shows the link between audial and visual lexemes. By calculating mutual information between sequences of visual and audial features, audial-visual pairs are constructed. Pairs with largest amount of mutual information surpassing a threshold are added into the concept database.

3. System learning results

ALSAIS unsupervised learning is evaluated by additional module – Control video evaluation. This module contains several types of experimental tests of learning. Three types of objects are examined in testing video data:

1. Objects, for which the system was trained, for example: comb.

2. Other exemplars of the same class, for which the system was trained, for example: other exemplars of comb.
3. Objects which are unknown to the system.

The particular results of the control video evaluation module are represented in table 1. Each category is marked by the number in brackets. The experiment consists in applying the final knowledge base to the test video, which contains different objects. In this experiment, only feature extraction and recognition were used. The extracted features were matched with the existing concepts in the database.

Table 1. Particular learning results for several objects.

Lexemes	Object							
	Staple (1)	Comb (1)	Camera (1)	Toy (1)	Timer (1)	Remote control (3)	Toy (2)	Comb (2)
Comb	335	2473	393	336	212	357	936	1628
Jpen	112	116	20	163	23	167	125	202
Pen	190	118	19	252	51	260	179	254
Photoapp	516	714	1624	414	273	231	547	504
Stepler	1859	1279	921	529	481	435	1552	1076
Tea	397	487	148	486	163	371	813	592
Timer	752	878	625	702	560	370	735	741
Toy	635	756	411	2388	279	1097	1001	1062

To simplify the result information formed by the control video evaluation module, the digits in table 1 are represented by a number of similar features between the lexemes and the object. Evaluations show that the system unsupervised learning process is successful in more than 90% of the presented objects.

Acknowledgements
This work was financially supported by the Ministry of Education and Science of the Russian Federation, and by Government of Russian Federation, Grant 074-U01.

References

[1] Potapov A S and Rozhkov A S 2013 Cognitive robotic system for learning of complex visual stimuli *Proc. AIP Conf.* **1537** 54–59
[2] Liu J and Bhanu B 2002 Learning Semantic Visual Concepts from Video *Proceedings International Conference on Pattern Recognition* **2** 1061–1064
[3] Jia Y, Abbot J, Austerweil J, Griffiths T and Darell T 2013 Visual Concept Learning: Combining Machine Vision and Bayesian Generalization on Concept Hierarchies *Advances in Neural Information Processing Systems* **26** 1842–1850
[4] Roy D and Pentland A 2002 Learning words from sights and sounds: a computational model *Cognitive Science* **26** 113–146
[5] Potapov A S, Malyshev I A, Puysha A E, Averkin A N 2010 New paradigm of learnable computer vision algorithms based on the representational MDL principle *Proc. SPIE* **7696** 769606