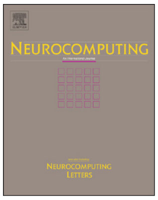
Neurocomputing 173 (2016) 93–101

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Adaptive incremental learning of image semantics with application to social robot

Hong Zhang [a,b,*], Ping Wu [a,b], Aryel Beck [c], Zhijun Zhang [c], Xingyu Gao [d]

[a] *College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China*
[b] *Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, China*
[c] *Institute for Media Innovation, Nanyang Technological University, Singapore*
[d] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form
2 June 2015
Accepted 2 July 2015
Available online 14 August 2015

Keywords:
Adaptive incremental learning
Image semantics
Social robot

ABSTRACT

Intelligent human–robot interaction is an interesting and challenging research topic both in areas of computational intelligence and robot control. In this paper we propose a novel adaptive and incremental image semantics learning framework based on the specific application platform of social robot. This endows the robot with the ability to learn to recognize new images based on previous human–robot interactions. In contrast with most of the intelligent image semantics learning works, which typically focus on how to recognize large scale of training data, this paper deals with how to learn image semantics from zero beginning and enrich the knowledge incrementally with human–robot interactions. In our framework, the user first presents unlabeled images to a humanoid robot for recognition; then the robot answers the user what's the image based on the semi-supervised incremental learning framework; thirdly user "teaches" the robot the right label of the image if the robot gives a wrong answer. Users can present more unlabeled images to the robot in the framework for teaching and learning. Extensive experiments and comparisons have validated the proposed methods with encouraging results. Our framework has a broad range of applications including education and rehabilitation.

1. Introduction

Image semantics understanding, which is a long-standing research topic in content-based multimedia analysis [1–3], plays an important role in different application areas, such as web data retrieval [4], biomedical diagnostics and recognition. Most of these researches focus on a large-scale training dataset and aim to achieve good performance on both training and testing datasets [5–7]. There are some other image semantics learning scenarios extensively, in which the image database increases from zero with the learning process. For example, when a child accesses to new knowledge he (she) usually learns bit by bit. Also in social robot research area how to teach a robot to learn knowledge from zero is an interesting and ambitious research topic [8,9]. However, there is a little image analysis work about such specific application scenarios. Moreover, due to the well-known semantic gap between low-level features and high-level semantics, image semantics understanding is still an open issue [1,10,11]. Most existing works on human–robot interaction ignore interactive learning of image semantics [12,13]. Therefore, in this paper we focus on two specific issues: how to understand image semantics with incremental learning and how to "teach" a social robot to recognize image semantics with the incremental learning framework.

It is difficult to teach a robot to recognize different images through human–robot interaction. For example, we first show robot images of "car" and the robot "remembers" what a car looks like; then we continue to show other images of "dog" and "bird", and the robot needs to find connections between visual features of images and high-level semantics. Besides, natural human–robot interaction is very important in order for the robot to react to the user's action or environment naturally. The ability of "learning image semantics" makes the user feel that the robot is more like an "intelligent life".

In this paper, we propose a novel adaptive framework of incremental image semantics learning on the social robot. The main idea of this paper is shown in Fig. 1. In the proposed framework, unlabeled images are first shown to the robot for recognition; then the robot makes decision of what the image is based on her learned knowledge; if the robot gives a wrong conclusion on the image label, the user could either let the robot guess again or directly "teaches" the robot the right label of the images. Our incremental learning algorithm is connected with the robot controller for decision making. Each time an image is shown

[*] Corresponding author at: College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China.
E-mail address: zhanghong_wust@163.com (H. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.07.104
0925-2312/© 2015 Elsevier B.V. All rights reserved.

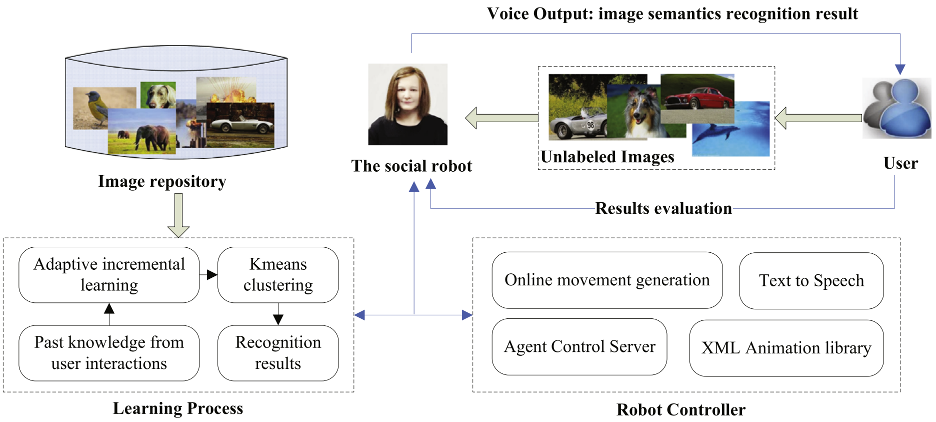


Fig. 1. Illustration of the adaptive incremental image semantics learning framework and its application on the social robot.

to the robot, it enriches the database increasing the robot's capacity to identify new images. During the interaction, the robot needs to understand the speech of the user and it should react appropriately expressions. Through periodic adaptive learning the robot's knowledge database is improved step by step. Therefore, it allows users to teach new skills to a humanoid robot in an innovative way. Moreover, our framework can be applied to many areas, such as the education and rehabilitation. Our framework is successfully integrated to a humanoid robot, allowing it to learn to recognize images from the beginning of zero knowledge. The robot (called the Nadine robot) used in our lab is a humanoid robot developed by Kokoro Company, Ltd. It has 27 Degrees of Freedom (DOF). In our application the robot communicates with users using body gestures and voice. For example, when users wave to the Nadine robot, she waves back.

This paper is organized as follows: Section 2 discusses related works on both image semantics recognition and social robots. Section 3 describes the incremental image semantics learning algorithm. Section 4 discusses the techniques used for our prototype on the Nadine robot. Section 5 presents experimental results and gives comparisons. We give concluding remarks in section 6.

2. Related works

This paper proposes an adaptive human–robot incremental image semantics learning framework, which involves adaptive image semantics recognition and robot activity control. Therefore, in this section, we discuss related works from the following two perspectives.

2.1. Adaptive image semantics recognition

It is a long-standing research topic to learn high-level image semantics from low-level visual content. In vector-based image analysis systems, we often concatenate different kinds of visual features into high-dimensional feature vectors for image representation. However, the existence of semantic gap makes it difficult to precisely recognize image semantics [1,14,15]. To bridge the semantic gap, many statistical methods and machine learning algorithms are used to build a robust image semantics learning system for different application scenarios. Meanwhile, nowadays gathering multimedia data is much easier than in the past. Multimedia repositories may lack in quality or may continuously change over time. Accordingly, adaptation methods are emerging topics in machine learning, computer vision and multimedia analysis to effectively process and utilize such kinds of multimedia data [6,16,17]. Transfer, domain adaptation and multi-task learning methods have been developed to better exploit the available data at training time [18,17].

In some multimedia applications, the target domain of interest contains very few labeled samples with limited knowledge [19], while an existing auxiliary domain is often available with a large number of labeled examples and useful knowledge. For example, in the area of social image analysis, since social multimedia data unevenly distribute on the Internet, paper [20] proposed a shared adaptive subspace learning framework to leverage a secondary text source to improve image retrieval performance from a primary dataset. The framework in paper [20] was validated on image and video retrieval tasks in which tags from the LabelMe dataset were used to improve image retrieval performance from a Flickr dataset and video retrieval performance from a YouTube dataset. Paper [6] proposed a framework for image attribute adaptation so as to automatically adapt the knowledge of attributes from a well-defined auxiliary image set to a target image set, thus assisting in predicting appropriate attributes for target images. Besides, for cross-media data analysis, Yi Yang et al. proposed a multi-feature fusion approach via nonlinear hierarchical regression to understand general semantics of multimodal documents which contained text, image and audio [21].

On the other hand, researchers have focused on semi-supervised methods in adaptive learning algorithms, which is also the main issue discussed in our paper. Since the robot's knowledge database increases step by step in human–robot interaction process, we propose semi-supervised learning algorithm to incrementally learn new unlabeled image semantics based on prior knowledge. Semi-supervised learning is widely used for different applications. For example, in video data analysis, paper [16] incorporated a semi-supervised process to utilize the unlabeled training videos in target domain, and by minimizing the difference of action prediction from still features and motion features, the still-to-motion adaptation was formulated into a joint optimization process. Zhang et al. mapped image and audio samples into an isomorphic feature subspace with kernel-based method, explored inherent multi-feature correlation with local linear regression and utilized external knowledge from relevance feedback [22]. In this way both the feature heterogeneity gap between image and audio and the semantic gap between low-level feature and high-level semantics were bridged, and flexible cross-media retrieval between image and audio was realized. Besides, Zhang et al. proposed a semi-supervised distance metric learning method based on local linear regression for image data clustering [5]. In paper [5], unlabeled samples were used to calculate the prediction error by means of local linear regression; labeled samples were used to learn discriminative ability. Then the knowledge learned from both labeled and unlabeled samples were fused into an overall objective function.