A Review on Visual Recognition of RGB Images and Videos by Learning from RGB-D Data

ALPHONSA JOSE
M.Tech Student, Electronics and Communication Engineering
St. Josephs College of Engineering And Technology, Palai, Kerala
E-mail: alphonsajose80@mail.com

SREEJITH V
Assistant Professor Dept. ECE
St. Josephs College of Engineering And Technology Palai, Kerala
E-mail: sreejith.vasu@sjcetpalai.ac.in

Abstract. Domain adaptation provides the possibility of research that allows changes in data distribution across training and testing datasets. Recognizing the RGB images by learning RGB-D data contains the additional depth information. The unsupervised domain adaptation (UDA) take advantage of the additional depth features. UDA deals with domain mismatch between the source and the target. The various adaptation techniques deals with the source and target domain. The domain mismatch is minimized by describing a projection matrix that is optimized by reducing the Maximum Mean Discrepancy (MMD) and aligning the source and target domains. To optimize the depth information the correlation between different types of features are to be maximized. In order to simultaneously cope with the domain mismatch issues, a unified framework called domain adaptation from multi-view to single-view (DAM2S) is learned. The effectiveness of the proposed methods for recognizing RGB images and videos by learning from RGB-D data is demonstrated by comprehensive experiments for object recognition, cross dataset and cross-view action recognition.

1. Introduction

Visual recognition gives the ability to see, hear or become aware of something through the senses and uses deep learning algorithms to analyse images that can give insights into ones visual content. One can organize images libraries, understand an individual image, recognize food, detect faces, and create custom classifiers for specific results that are tailored for the needs. Visual recognition quickly and accurately tag, classify and search visual content using machine learning. Visual recognition understands the contents of images, analyse images for scenes, objects, faces, colors, food, and other subjects that gives insights to the visual content. Humans have the remarkable ability to organize their perceptual input instead of a collection of values associated with
individual photoreceptors, to perceive a number of visual groups, usually associated with objects or well-defined paths of objects. Algorithms and representations is used to understand the methods to recognize objects, scenes, place and peoples. There is an increasing interest in developing new technologies using depth images and videos for various visual recognition tasks. It is obtained by capturing depth information with the RGB-D equipments namely Kinect sensors. The depth information cannot be applied to most ordinary visual recognition applications, in which images and videos are captured by conventional RGB cameras. Several labelled RGB-D datasets [1] are released for visual recognition tasks.

2. Domain Adaptation Methods

The domain adaptation is related to machine learning methods. The distribution of test data is different from that of training data [2]. The related works includes several adaptation methods. This includes the (a). Unsupervised Domain Adaptation (UDA) methods (b). Heterogeneous Domain Adaptation (HDA) methods (c). Multi-view Domain Adaptation method (e). Multi Domain Adaptation from Heterogeneous Sources (MDA-HS).

2.1. Unsupervised Domain adaptation Method

It contains no labelled data in the target domain [3]. Domain adaptation in deep architectures is adapted that can be trained on large amount of labeled data from the source domain and large amount of unlabeled data from the target domain.

2.2. Heterogeneous Domain Adaptation Method

The method does not require any labelled target domain samples [4]. In contrast this have both visual and depth features in the source domain, while the depth features are not available at the testing stage.

2.3. Multi-view Adaptation Methods

The samples in the source and the target domains have multiple type of features[5].

2.4. Multi-view Domain Adaptation from Heterogeneous Methods

The samples from the target domain have all types of features from all source domains. In contrast, it have single view features in the target domain [6].

3. Domain Adaptation With Additional Features

The approach to domain adaptation that is appropriate exactly in the case when one has enough target data to do slightly better than just using only source data is described.
In the framework, the visual features and depth features from the RGB images/videos and depth images/videos is extracted. To handle the distribution mismatch, we learn a common subspace that is parameterized by a matrix P for the visual features from two domains.

3.1. Learning Projection For Domain Adaptation

Two strategies for learning the projection matrix P is proposed. The strategies include reducing the Maximum Mean Discrepancy (MMD) and aligning two subspaces. The first strategy is to minimize the Maximum Mean Discrepancy (MMD) criterion, which is widely used to measure the distribution difference between the data sampled from two datasets. The second strategy seeks a domain invariant feature space by learning a mapping function which aligns the source subspace with the target one.

3.2. Incorporating Depth Information

In order to effectively incorporate the depth features to learn more robust classifiers it includes two different strategies, one is maximizing the feature correlation where Canonical Correlation Analysis (CCA) is one pioneering work in two-view learning, which aims to learn two projection matrices to map the training samples with different features into a common subspace, such that the feature correlation can be maximized. Kernel Canonical Correlation Analysis (KCCA) is an extension of CCA by applying the kernel trick. The two classifiers for visual and depth features based on the projected visual and depth samples in the common subspace is learned. The objective function is formulated as the DAM2S-A algorithm. The second is enforcing the classifier consistency done by using the MMD based regularizer with the objective function formulated as the DAM2S-B algorithm and the Kernel SA based regularizer with the objective function formulated as the DAM2S-C algorithm.

3.3. Summary of DAM2S Approaches

The domain adaptation with additional features proposed two regularizers to reduce the domain distribution mismatch and two forms of regularizers for incorporating the depth features, which leads to three DAM2S algorithms, DAM2S-A, DAM2S-B, and DAM2S-C, the MMD based regularizer in DAM2S-A and DAM2S-B, and the kernel SA based regularizer in DAM2S-C. For incorporating the depth information, we utilize the feature correlation maximization based regularizer in DAM2S-A, and the classifier consistency based regularizer in DAM2S-B and DAM2S-C. The parameter matrix P or parameter matrices (P, Q) can be written in a single matrix in the dual form.

4. Evaluation of Effectiveness

The effectiveness of our proposed three algorithms for different visual recognition tasks, including object recognition, cross-dataset human action recognition, and cross-view
human action recognition is evaluated.

4.1. Object Recognition

The proposed three DAM2S algorithms is evaluated for object recognition by using the RGB-D Object dataset as the source domain and the Caltech-256 dataset [7] as the target domain. The RGB-D Object dataset contains the color and depth images of different objects from 51 categories.

4.2. Cross-Dataset Human Action Recognition

For cross-dataset action recognition, the Hollywood 3D Dataset are used as the source domain, and the Hollywood2 dataset as the target domain. The Hollywood2 dataset is a widely used benchmark dataset for human action recognition, which contains 1,707 (823 in the training set and 884 in the test set) RGB videos from 12 human actions cropped from the Hollywood movies. Similarly, the Hollywood 3D dataset contains 650 RGB-D video clips from 14 human actions cropped from the Hollywood 3D movies.

4.3. Cross-View Human Action Recognition

A recently released multi-view RGB-D action dataset ACT42 is used for the experiments, which contains 2,648 RGB-D videos from 14 human actions captured by Kinect from four different view points. To evaluate the proposed algorithms, we use the RGB-D videos from the first two views as the source domain, and use only the RGB videos from the remaining two views as the target domain, which leads to 1,324 RGB-D videos in the source domain, and 1,324 RGB videos in the target domain.

5. Conclusions

A new framework is discussed for recognizing the RGB images and videos by learning from a set of labelled RGB-D data. A review of different adaptation techniques provides the advantage of understanding the domain mismatch between the source and the target. The framework for domain adaptation is formulated as a new Unsupervised Domain Adaptation (UDA) problem, in which both visual and depth features are considered in the source domain and the visual features are considered in the target domain. Three DAM2S algorithms are described by taking into consideration the various datasets. The algorithms take into account the depth features in the source domain and will reduce the domain distribution mismatch between the source and target domains. In order to evaluate the effectiveness of the three DAM2S algorithms, comprehensive experiments for object detection, cross-dataset human action recognition and cross-view human action recognition have been demonstrated. The comparison between the three algorithms are discussed. The algorithm play a crucial role in computer vision. It is actually a measurement that computes the properties of 3D world by identifying the
visual data and information. The algorithms are used to understand the methods to recognize objects, scenes, place and peoples is the method of visual recognition.

References

[1] Wen Li, Lin Chen, Dong Xu, Senior Member, IEEE, and Luc Van Gool Visual Recognition in RGB Images and Videos by Learning from RGB-D Data, IEEE Trans. on Pattern Analysis and Machine Intelligence, TPAMI. 2017.
[2] K. Lai, L. Bo, X. Ren, and D. Fox, A large-scale hierarchical multiview RGB-D object dataset, ICRA, 2011.
[3] L. Duan, I. W. Tsang, D. Xu, and T. Chua, Domain adaptation from multiple sources: A domain-dependent regularization approach, T-NNLS, vol. 23, no. 3, pp. 504518, 2012.
[4] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa Center for Automation Research, University of Maryland, College Park, MD 20742 USA, Domain Adaptation For Object Recognition.
[5] B. Kulis, K. Saenko, and T. Darrell, What you saw is not what you get: Domain adaptation using asymmetric kernel transforms, CVPR, 2011.
[6] D. Zhang, J. He, Y. Liu, L. Si, and R. D. Lawrence, Multi-view transfer learning with a large margin approach, KDD, 2011.
[7] G. Griffin, A. Holub, and P. Perona, Caltech-256 object category dataset California Institute of Technology, Tech. Rep., 2007.