Handcrafted vs. Learned Representations for Human Action Recognition

1. Introduction

Human action recognition as one of the most active topics in computer vision has long been in the last few decades, and its potential applications can be found in many important areas such as surveillance, video annotation, multimedia retrieval and human-computer interaction. Action representations including both local and holistic representations play a fundamental role in action recognition, and good action representations can largely improve recognition performance. Conventional methods were mainly developed based on low-level handcrafted features by feature engineering, e.g., histogram of three-dimensional oriented gradients (HOG3D), histogram of oriented gradients and histogram of optical flow (HOGHOF) and spatial-temporal oriented energies. Recently, feature learning has started to draw increasing interest in visual recognition, especially in the image domain for scene classification and digit recognition. Nevertheless, both feature learning and feature engineering have their advantages in visual representation which, however, remains less explored in the video domain for action recognition.

Feature engineering is able to incorporate human ingenuity and prior knowledge and remain widely used in both image and video domains for visual representations. Compared to feature learning, feature engineering enjoys the flexibility and computational efficiency, and does not rely on large sets of samples for training. Feature learning, e.g., deep learning, can directly learn data representations from raw training samples and detect data-driven features for specific tasks. In contrast to feature engineering, feature learning can extract and organize the discriminative information from the data, and novel applications can be constructed faster. In addition, feature learning algorithms, e.g., descriptor learning, can also construct objective functions by parameterization of handcrafted features to learn task-specific descriptors. Up until now, feature representation for action recognition is still a chal-

lenging task far from being solved. It is highly desirable to design efficient and effective action representations by both feature engineering and feature learning, which will boost the progress of human action recognition towards practical applications.

This special issue targets researchers of broad fields from diverse communities, including machine learning, computer vision and pattern recognition and will encourage novel theories and advanced techniques of feature learning/engineering for human action recognition. The aim of this special issue is to collect publications of high quality papers on technical developments and practical applications around feature learning/engineering for both local and holistic representations of human actions.

2. Overview of accepted papers

Accepted papers cover a broad range of algorithms for spatial-temporal feature learning in human action recognition, which includes mid-level feature representation, deep learning, cross-domain transfer learning and metric learning, etc. We have received 21 submissions from all over the world and 12 of them have been finally accepted based on a strict and comprehensive review process. Each manuscript has been assigned to at least two peer reviewers and gone through at least two rounds of review.

2.1. A survey

The article "From handcrafted to learned representations for human action recognition: A survey" provides a timely and comprehensive survey of feature representation algorithms in up-to-date human action recognition systems. As the first survey that covers both handcrafted and learning-based action representations, this survey paper explicitly discusses the superiorities and limitations of exiting techniques from both kinds. This paper provides comprehensive analysis and comparisons between learning-based and handcrafted action representations respectively, which will serve as a guideline to inspire action recognition researchers towards the study of both kinds of representation techniques.

2.2. Mid-level feature representation

Mid-level feature representations, e.g., the bag-of-word (BoW) model, and vector of locally aggregated descriptors (VLAD), have been playing an

important role in human action recognition. A lot of improvement has been developed to enhance the mid-level representation models.

The article "Towards optimal VLAD for human action recognition from still images " proposes a optimized vector of locally aggregated descriptors (VLAD) for human action recognition from still images. In this work, VLAD has been improved by tackling three important issues including empty cavity, ambiguity and pooling strategies, which have been less investigated. The empty cavity limits the performance of VLAD and has long been overlooked. Moreover, the generalized max pooling (GMP) is for the first time incorporated to replace sum pooling in VLAD, which is more reliable for the final representation with improved performance. Experiments on four widely-used benchmarks to validate effectiveness of the optimized VLAD for human action recognition from still images.

The article "Robust geometric ℓ_p-norm feature pooling for image classification and action recognition" provides a new feature pooling algorithm in mid-level feature representation. This work generalizes previous pooling methods toward a weighted ℓ_p-norm spatial pooling function tailored for class-specific feature spatial distribution. Optimizing such a pooling function toward discriminative class separability that is subject to a spatial smoothness constraint yields a so-called geometric ℓ_p-norm pooling (GLP) method. Moreover, to handle the variation of object scale/position, an effective self-alignment step during both learning and testing to adaptively adjust the pooling weights for individual images, which leads to a robust version of GLP (RGLP).

The article "Action recognition by joint learning" proposes a novel action recognition method that simultaneously learns middle-level representation and classifier by jointly training a multi-nominal logistic regression (MLR) model and a discriminative dictionary. In this work, mid-level representations obtained by sparse coding by treating as latent variables of MLR, captures the structure of low-level features and thus is more discriminate. Dictionary learning and the MLR model training are integrated into one objective function for considering the information of categories, which achieves a discriminative dictionary modulated by MLR for improved performance.

The article "Dynamic texture recognition with video set based collaborative representation" proposes a novel video set based collaborative representation for dynamic texture classification. In this work, a regularized collaborative representation model is proposed to code the LBP histograms of the query video sets over the LBP histograms of the training video sets.

The distance between the query video set and the training video sets can be calculated based on the coding coefficients for classification.

2.3. Deep learning

While handcrafted features are still widely used for human action representation, deep learning based feature representation has also started to be explored for human action recognition.

The article "3D-based Deep Convolutional Neural Network for action recognition with depth sequences" introduces a new deep learning based method for action recognition in depth sequences. In this work, a 3D-based Deep Convolutional Neural Network (3D2CNN) to directly learn spatio-temporal features from raw depth sequences. A joint based feature vector named JointVector is computed for each sequence by taking into account the simple position and angle information between skeleton joints. The support vector machine (SVM) classification results from 3D2CNN learned features and JointVector are fused to achieve final action recognition.

The article "Deep and fast: Deep learning hashing with semi-supervised graph construction" proposes a semi-supervised deep learning hashing (DLH) method for fast multimedia retrieval. Visual and label information is combined to learn an optimal similarity graph that can more precisely encode the relationship among training data, bases on which hash codes are generated. A deep convolutional network is applied to simultaneously learn a good multimedia representation and a set of hash functions.

2.4. Cross-domain transfer learning

Cross-domain transfer learning as one of the most important machine learning techniques has also been investigated for human action recognition, which is especially suitable in multi-view and cross-dataset scenarios.

The article "Cross-view action recognition by cross-domain learning" proposes a novel cross-view human action recognition method by discovering and sharing common knowledge among different video sets captured in multiple viewpoints. In this work, a specific view is treated as target domain and the rest as source domains, and the cross-view action recognition is cast into a cross-domain learning framework. Based on the BoW model based representation, two transformation matrices are learned to transform original action feature from different views into one common feature space. The method has been evaluated on two datasets: IXMAS and TJU, showing competitive performance.

The article "Cross-domain action recognition via collective matrix factorization with graph Laplacian regularization" presents a cross-domain action recognition framework by utilizing some labeled data from other data sets as the auxiliary source domain. To map data from different domains into the same abstract space and boost the action recognition performance, a method named collective matrix factorization with graph Laplacian regularization (CMFGLR) is proposed. The approach is built upon techniques of collective matrix factorization and an optimal linear classifier. Moreover, label consistency across different domains and the local geometric consistency in each domain are explored as a graph Laplacian regularization term to enhance the discrimination of learned features.

The article "Dual many-to-one-encoder-based transfer learning for cross-dataset human action recognition" presents a new transfer learning method for cross-dataset human action recognition, which learns generalized feature representations across datasets. A novel dual many-to-one encoder architecture is proposed to extract generalized features by mapping raw features from source and target datasets to the same feature space. Experiments on pairs of benchmark human action datasets demonstrate state-of-the-art performance.

2.5. Metric learning

Metric learning has also been a active topic in computer vision and machine learning. The article "Statistical adaptive metric learning in visual action feature set recognition" proposes a statistical adaptive metric learning (SAML) method by exploring various selections and combinations of multiple statistics in a unified metric learning framework. In this work, multiple statistics, include means, covariance matrices and Gaussian distributions, are explicitly mapped or generated in the Riemannian manifolds. Experimental evaluations are conducted on human action recognitions in both static and dynamic scenarios. Promising results demonstrate that the proposed method performs effectively for human action recognitions in the wild.

2.6. Other methods

The article "Using the conflict in Dempster-Shafer evidence theory as a rejection criterion in classifier output combination for 3D human action recognition" proposes a comprehensive solution to 3D human action recognition including feature extraction, classification, and multiple classifier combination. Two feature extraction methods, four different types of well-known

classifiers, and four multiple classifier combination strategies including a specially designed belief based method are presented. To further enhance the performance, a new rejection criterion is proposed based on the conflict from the information sources: the classifier outputs. The method has been evaluated on the MSRAction 3D dataset, which shows superior results than other combination approaches.

Acknowledgement

We would like to thank all the authors for their contributions to this special issue, and reviewers for their timely and insightful reviews. We thank Professors J.-M. Frahm and M. Pantic, Editor-in-Chief of the Image and Vision Computing Journal for giving us the opportunity to guest edit this special issue, and the Elsevier staff, Yanhong Zhai for her great support to this special issue.

Dr. Xiantong Zhen
Department of Medical Biophysics,
University of Western Ontario, London, ON, Canada.

Prof. Ling Shao
Department of Computer and Information Sciences,
Northumbria University, Newcastle upon Tyne, United Kingdom.

Prof. Stephen J. Maybank
School of Computer Science and Information Systems,
Birkbeck College, University of London, United Kingdom.

Prof. Rama Chellappa
Department of Electrical and Computer Engineering,
Center for Automation Research,
University of Maryland,
College Park, MD, USA.