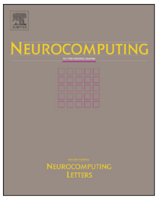
Neurocomputing 167 (2015) 359–370

Contents lists available at ScienceDirect

Neurocomputing

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



Context and locality constrained linear coding for human action recognition

Yi Tian [a,b,*], Qiuqi Ruan [a,b], Gaoyun An [a,b], Wanru Xu [a]

[a] *Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China*
[b] *Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China*

ARTICLE INFO

Article history:
Received 28 September 2014
Received in revised form
27 March 2015
Accepted 19 April 2015
Available online 2 May 2015

Keywords:
Human action recognition
Sparse representation
Context and locality constrained linear coding (CLLC)
Group-wise sparse representation based classification (GSRC)

ABSTRACT

Bag of Words (BOW) method with spatio-temporal local features has achieved great performance in human action recognition. However, most of the existing BOW approaches based on vector quantization (VQ) neglect the contextual information of each descriptor, and suffer serious quantization error. There are two main reasons for these: in the first, each local feature is only assigned to one label and second, the information about the spatial layout of the features is disregarded. In this paper, we present a novel and effective coding method called context and locality constrained linear coding (CLLC) to overcome these limitations, in which the relationships among local features and their structural information are preserved. After that, a group-wise sparse representation based classification (GSRC) method is implemented to assign the query sample into one category which yields the smallest reconstruction error. Our method is verified on the challenging databases and achieves comparable performance with state-of-the-art methods.

1. Introduction

Human action recognition is one of the most popular problems in the computer vision field, which has received lots of attentions. It has rich potential applications in varied areas such as: human–computer interaction, image retrieval and video indexing. Over the past few years, many methods have been suggested to solve the problem and shown impressive results. However, because of the occlusion, illumination and camera movements, it remains a puzzling problem. These methods can be roughly classified into three categories: template based methods, dynamic model based methods and local feature based methods.

Template based methods are the traditional methods in action recognition which represent each video as a space-time template. Bobick et al. [1] used motion history image (MHI) to describe different actions which accumulates the foregrounds along the temporal axis with a decay factor. Based on the MHI, Shao et al. [2] applied a shape-based feature descriptor to represent the actions. However, to the template based methods, foreground estimation is still a challenge which is sensitive to noise, viewpoint, occlusions, and illumination. Meanwhile, the dynamic model based methods which represent an activity as a particular sequence of observations always come with complex models and high computation cost. There are various methods in this group [3–5], including Hidden Markov Models (HMMs), Dynamic Bayesian Networks (DBNs), and Conditional Random Fields (CRFs).

Compared to the methods mentioned above, local feature based methods [6,7] which have no need of any preprocessing operation are more stable and robust. In particular, the BOW (bag-of-words) model [8] together with the local spatio-temporal features has shown optimal performance on human action recognition. These existing BOW-based methods quantize each local feature to the corresponding visual word, and then each video sequence is summarized by a histogram of the local features' occurrences. However, the traditional BOW-based methods have two key limitations which bring ambiguities to the action representation. One of the most significant limitations is that, the large quantization errors are easy to be generated during assigning each local feature with one label. As each local feature is only represented by a single word, masses of information lose during calculating histogram for each video. Another limitation is that, it fails to capture adequate spatial and temporal relationships between each feature pair which are definitely helpful to enhance the discriminability of the features.

To address these two problems, various methods have been proposed. Savarese et al. [9] used ST-correlagrams to capture the relationships between the center feature and its neighbors in a local

[*] Corresponding author.
E-mail addresses: 11112062@bjtu.edu.cn (Y. Tian), qqruan@bjtu.edu.cn (Q. Ruan), gyan@bjtu.edu.cn (G. An).

http://dx.doi.org/10.1016/j.neucom.2015.04.059
0925-2312/© 2015 Elsevier B.V. All rights reserved.

3-D region. Kovashka et al. [10] learned a hierarchy of discriminative space-time neighborhood features for human action recognition. Yan et al. [11] used effective human body regions to find relations to compensate for the traditional BOW method. Ballan et al. [12] employed radius-based clustering method with soft assignment to create a rich codebook which accounted for the high variability of human actions. Guha et al. [13] firstly introduced the idea of sparse representation into human action recognition in videos which represented each local descriptor by a linear combination of a small number of visual words in an overcomplete codebook. It offered a more compact and richer representation for each video sequence. After that, sparse representation has been emerged as an extremely successful tool for human action recognition, which achieved a better reconstruction for each local feature. Wang et al. [14] utilized locality constraints to project each descriptor into its local-coordinate system and captured the correlations between similar local features by sharing words. Zhang et al. [15] proposed the context-constrained linear coding which used several nearest words to encode each local descriptor on the basis of contextual distance.

In this paper, we propose a novel coding method named context and locality constrained linear coding (CLLC) for human action recognition, which considers both the local and contextual information of each local feature. The locality-constraint projects each descriptor into local-coordinate system. Meanwhile, different from the existing approaches that treat each local feature as an independent individual, our encoding method incorporates the spatio-temporal contextual constraint into the object function to improve the features' discriminability. After that, we use the group-wise sparse representation based classification (GSRC) method which was proposed by Wei et al. [19] to classify each action sequence. GSRC assigns each query video sequence to the class which generates the smallest reconstruction error in terms of its corresponding sparse encoding.

The rest of the paper is structured as follows. Section 2 reviews related works of the BOW-based methods and sparse representation in action recognition. In Section 3, our context and locality constrained linear coding method and group-wise sparse representation based classification method are introduced in detail. Experiments and discussions are shown in Section 4. Finally, Section 5 concludes the paper and looks into the future.

2. Related work

The bag-of-words (BOW) method [8] is one of the most classical local feature based methods, and has shown superior performance in image classification and object recognition in the recent years. The popular BOW-based approaches together with the local spatio-temporal features also achieve promising results in human action recognition. However, the traditional methods which encode the descriptors in vector quantization (VQ) only label each local feature with its nearest visual word. The restrictive constraints lead to heavy quantization error which directly degrades the discriminability of the representation. To ameliorate the deficiencies, standard sparse coding is applied whose constraint is relaxed by a sparsity regularization term. Given a set of D-dimensional local descriptors $X = [x_1, x_2, ..., x_N] \in R^{D \times N}$ and an over-complete codebook$B = [b_1, b_2, ..., b_M] \in R^{D \times M}$, we can calculate the sparse encodings S of the local features as follows:

$$\min_S \|X - BS\|^2 + \lambda \|S\|_1 \tag{1}$$

where the parameter λ controls the sparsity of S. Superior to vector quantization, sparse coding allows a few more visual words to participate in the approximation process, which is more objective and discriminative. Inspired by the sparsity regularization term, a massive of reconstruction based encoding methods using different regularization terms have been emerged in recent literatures. Gao et al. [17] incorporated Laplacian matrix constructed by the histogram intersection based k-Nearest Neighbor (KNN) method into the function (1) to preserve the consistence in sparse representation of the similar local features. With the using of the Laplacian regularization term, the quantization error of the local features could be substantially reduced. Moreover, the Laplacian matrix maximally preserved the similarity between each similar local feature so as to generate more discriminative sparse encodings. However, it only considered the relationship between each feature pair, yet it ignored the distribution of its neighbors. Zheng et al. [18] also incorporated the Laplacian regularization term into the object function. He took the local manifold structure into consideration and proposed a graph based algorithm, called graph regularized sparse coding. If two local features were close to each other in the intrinsic geometry of the feature distribution, the representations of them in the new projection subspace were also close to each other. Based on the manifold assumption, a weighted graph G was constructed for each local feature in order to capture the relationships between the feature and its neighbors. Nevertheless, it failed to take locality into consideration and resulted in astronomical computations due to the $l1$-regularization term. To solve the problem, Wang et al. [14] proposed the locality-constrained linear coding (LLC) for the image classification, which replaced the $l1$-regularization term with the locality-constraint. The locality-constraint not only ensured similar features sharing similar bases, but also implied the sparsity of the encodings. Furthermore, its approximated solution owned lower computation complexity than the $l1$-regularization methods. Similar to the previous methods, LLC did not take the contextual information of each local descriptor into account which was helpful to enhance the discriminability. To solve the common problem, Zhang et al. [15] proposed the context-constrained linear coding (CLC) on the basis of LLC. During searching the k nearest visual words to construct the input feature, the traditional Euclidean distance was replaced by the contextual distance, which considered the influence of the neighbor local descriptors of the input feature. However, in CLC, the VQ method was applied to label each local feature with one visual word before the encoding procedure, which brought distortion of information. Moreover, Zhao et al. [25] presented an optimized version of 3D shape context to encode the layout information of the local features. Three atomic histograms were formed to describe the log distance between any of the local features and the reference feature (m bins), azimuth (n bins) and inclination angles (s bins), respectively. Consequently, the optimized 3D shape context descriptor ($m*n*s$ bins) which was formed by the atomic histograms led to high computation cost and its performance was sensitive to the confusion method of the three atomic histograms. In recent years, the super vector based encoding methods [26,27] yielded effective representation for the retrieval and classification tasks. These popular methods emerged to encode descriptors with respect to the residuals between visual words and features. Yang et al. [27] aggregated the low-level features into super descriptor vector (SDV) and proposed the novel super location vector (SLV) to incorporate the locations of local features. In the end, SDV and SLV were combined into the super sparse coding vector (SSCV), which jointly modeled the motion, appearance, and location cues. Because of the aggregating of the high order statistics, the super vector based encoding methods yielded much higher dimensional representations than the reconstruction based encoding methods with the same codebook.

In this paper, we succeed the characteristics of the two above mentioned methods (LLC and CLC) while overcome their deficiencies and eventually propose a novel and efficient encoding method named context and locality-constrained linear coding (CLLC) for human