Computer Vision and Image Understanding 144 (2016) 36–45

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Recognising occluded multi-view actions using local nearest neighbour embedding

Yang Long[a], Fan Zhu[b], Ling Shao[c,*]

[a] Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 3JD, UK
[b] Department of Electrical and Computer Engineering, New York University, Abu Dhabi, UAE
[c] Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, UK

ARTICLE INFO

Article history:
Received 24 December 2014
Accepted 13 June 2015

Keywords:
Occlusion
Action recognition
Multi-view fusion
Naive bayes
Local nearest neighbour
Embedding

ABSTRACT

The recent advancement of multi-sensor technologies and algorithms has boosted significant progress to human action recognition systems, especially for dealing with realistic scenarios. However, partial occlusion, as a major obstacle in real-world applications, has not received sufficient attention in the action recognition community. In this paper, we extensively investigate how occlusion can be addressed by multi-view fusion. Specifically, we propose a robust representation called local nearest neighbour embedding (LNNE). We then extend the LNNE method to 3 multi-view fusion scenarios. Additionally, we provide detailed analysis of the proposed voting strategy from the boosting point of view. We evaluate our approach on both synthetic and realistic occluded databases, and the LNNE method outperforms the state-of-the-art approaches in all tested scenarios.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Human action recognition has received increasing attentions during the past decades. It has a wide range of applications such as medical surveillance [1], smart home [2] and human-machine interaction [3]. However, how to recognise multiple, complex human actions or activities remains a challenging problem [4]. So far, the majority of action recognition systems are only restricted to a finite number of well-defined action categories, and the performance is evaluated on actions cropped by detected bounding boxes [5,8]. For realistic applications, current methods are still very sensitive to trivial environmental variations, e.g., gender, body size, viewpoint and illumination variations, and occlusions [6,7]. Among these problems, view-variation and occlusion are two main inevitable hurdles of action recognition. As a pessimistic conclusion claimed in [9], the monocular computer vision systems are not competent enough for surveillance applications. Fortunately, the progressive visual technologies have made it possible to solve the action recognition problem using multi-view or range sensors [10,11]. Hence, extensive studies are conducted on view-invariance and transferable representations [6,13,14,16–20]. Other works also consider multi-descriptor fusion approaches [22,23]. Nonetheless, only few techniques such as [15] tackle the occlusion problem. Therefore, dealing with the occlusion problem remains an imminent research area to bridge the gap between existing action recognition algorithms and realistic applications [10].

Intuitively, the occlusion problem can be solved by a multi-view system, as shown in Fig. 1. If actions captured from a viewpoint are occluded, the information loss can be compensated by data from other views which are not occluded, thus, the occlusion problem is transformed to a view-disparity problem. However, such a strategy leads to two main difficulties. The first one is how to suppress the intra-class distance caused by viewpoint variations. For this concern, it is widely acknowledged that local descriptors are less susceptible to intra-class variations [24–28], which are generally fused with holistic representations [29–31]. The second difficulty is that, in real-world applications, occlusions appear unpredictably in both training and testing data, and, as a result, break the consistency of the holistic models in the two datasets.

In order to overcome these problems, this paper is devoted to investigating multi-view methods that can incorporate local descriptors and are robust to occlusions in both training and testing action datasets. Specifically, we adopt the dense trajectories (DT) [24], which are further transformed to a robust higher-level representation and then used for multi-view fusion. We conclude our main contributions in the following 3 aspects: (1) we propose a robust learning-free algorithm: local nearest neighbour embedding (LNNE); (2) we introduce 3 multi-view fusion scenarios to test the LNNE method; (3)we conduct extensive experiments on two multi-view action data sets with

[*] Corresponding author.
E-mail address: ling.shao@ieee.org (L. Shao).

http://dx.doi.org/10.1016/j.cviu.2015.06.003
1077-3142/© 2015 Elsevier Inc. All rights reserved.

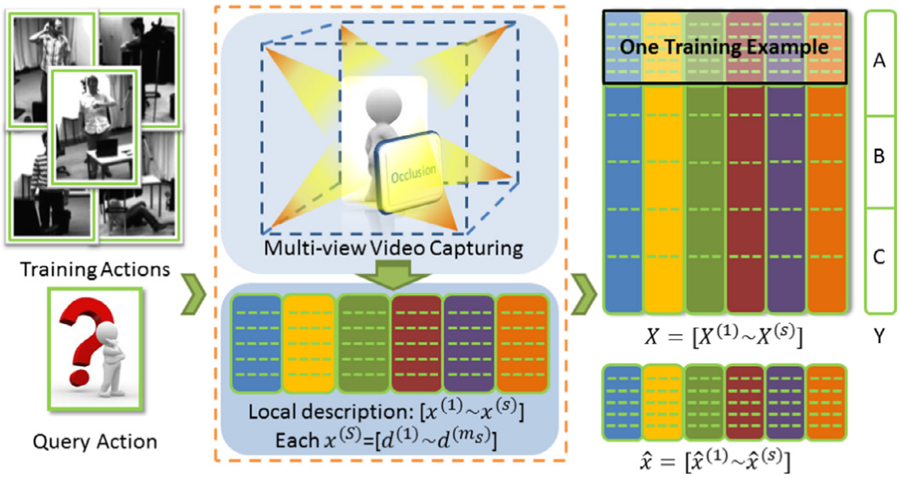
Training Actions
Query Action
Multi-view Video Capturing
Occlusion
Local description: $[x^{(1)}\sim x^{(s)}]$
Each $x^{(S)}=[d^{(1)}\sim d^{(m_s)}]$
One Training Example
A
B
C
$X = [X^{(1)}\sim X^{(s)}]$
Y
$\hat{x} = [\hat{x}^{(1)}\sim \hat{x}^{(s)}]$

Fig. 1. An illustration of supervised multi-view action recognition with occlusion. In general, visual features are extracted in advance. Occlusion can exist in both training and query actions, and the descriptions from occluded views are usually merged into clear views without any clue to distinguish.

occlusions, where the LNNE method achieves significant performance improvements on all scenarios.

The following sections are arranged as follows: We introduce related works in Section 2. In Section 3, we explicitly describe the LNNE method. We then illustrate the structures of the 3 fusion pipelines in the Section 4. Detailed experimental results are presented with analysis and discussions in Sections 5 and 6. Finally, we conclude our work in Section 7.

2. Background

We review previous works from two main aspects. In the first aspect, we review the basis of feature embedding techniques, and we aim at providing an intuitive and generalised view of embedding. Also, we discuss their relations to our LNNE method. In the second aspect, we compare the proposed fusion scenarios with existing multi-view action recognition scenarios.

2.1. Feature embedding

The primary motivation of feature embedding is inspired by the multi-sensor robotic control system [32]. The target of such a system is to dynamically determine the current state from coarse finite belief states (all of which form a belief space (BS)) under partially observable evidence. Each partial evidence is assumed achieving a proportional contribution to the overall probability distribution. Also, the control system assumes that accumulating the past likelihoods of the observation can result in one state with maximum likelihood in the belief space.

Accordingly, in a multi-view action recognition system, each individual sensor denotes a single camera viewpoint. The dimension of the BS is equivalent to the number of known query action categories. Each action category corresponds to one of the finite states of the belief space. Moreover, each local descriptor can be treated as a partial observation of the whole action sequence. Therefore, similar to the control system, we aim at finding a projection between the local feature space and the belief space in order to maximise the posterior probability. By accumulating the posterior probabilities of all local descriptors, the final static state can be determined as the overall recognition output. As a typical example, Weinland et al. [20,33] propose a class-to-query distance projection technique, where each action category $C \in [1, \ldots, c]$ is represented by a set of exemplars: $\dot{X}^C = [x^{(1)}, \ldots, x^{(n)}]$, where $\dot{X}$ denotes the training set of the category. Each query action is represented by local descriptors: $\hat{x} = [\hat{d}^{(1)}, \ldots, \hat{d}^{(t)}]$. The distance between each exemplar $x^{(i)}$ and the nearest local descriptor $\hat{d}^{(j)}$ of the query action is estimated. Then all estimated distances are concatenated into a long vector:

$$\begin{aligned} D(\hat{x})^C &= (dist_1(\hat{x}), \ldots, dist_n(\hat{x})) \in R^n \quad where\ dist_i(\hat{x}) \\ &= \min_j dist(x^{(i)}, \hat{d}^{(j)}) \end{aligned}$$

Note that D is an embedding procedure which maps the descriptors of the query action into an n-dimensional new representation for each category, each dimension of which accounts for the posterior probability of each exemplar in class C to the closest partial observation in the query $\hat{x}$. Thus, the final static state can be determined by the typical maximum likelihood criterion:

$$g(D) = \arg\max_C p(D|C)p(C)$$

On the other hand, such exemplar-based approaches also receive massive criticisms. Boiman et al. [41] emphasise that the quantisation error may decay the performance, which is inevitable in the exemplar-based methods because of clustering. They claim the importance of using the original local features. However, as a classical local feature-based approach, NBNN is also criticised by Timoftem et al. [44] for another two issues. The first issue is that the independence assumption of the descriptors may not hold. Secondly, the decision rule of NBNN is close to a hard-assignment of each category. A number of related methods attempt to extend NBNN [44]. These recent approaches try to improve the nearest neighbour formulas with more advanced solutions. kNN [43] expands the number of nearest neighbours and therefore makes the decision more fuzzy. LLE [35] aims at richer representations on l_1 constrained least squares so that the feature structure can be more discriminative. INN combines l_1-regularised least squares with LLE and adopts the simplicity of kNN. Besides, a number of related subspace learning methods are proposed, such as [34,36–40].

Nonetheless, most of the above approaches do not find an embedding from the original feature space directly to a belief space (which has an identical number of dimensions as the number of categories for query). In comparison, our method will directly achieve the predicted labels after embedding whereas other methods only focus on learning the data structure. Even though existing feature embedding techniques are equipped with strong theoretical supports, they also suffer from the above stated problems. In order to overcome these problems, we follow four principles to design our method: (1) we utilise soft-assignments so that some ambiguous local features can contribute to multiple categories; (2) we directly map from the feature space to the belief space in order to substantially utilise the supervision; (3) our method is training-free, and requires only one

Download English Version:

https://daneshyari.com/en/article/6937649

Download Persian Version:

https://daneshyari.com/article/6937649

Daneshyari.com