Pattern Recognition 47 (2014) 1791–1799

Contents lists available at ScienceDirect

Pattern Recognition

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

Sparse representation for robust abnormality detection in crowded scenes

Xiaobin Zhu [a,b], Jing Liu [b,*], Jinqiao Wang [b], Changsheng Li [c], Hanqing Lu [b]

[a] School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China
[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[c] IBM Research-China, Beijing, Beijing 100193, China

ARTICLE INFO

Article history:
Received 5 June 2013
Received in revised form
9 November 2013
Accepted 20 November 2013
Available online 1 December 2013

Keywords:
Nonnegative matrix factorization
Crowded scene
Abnormality detection
Sparse coding
Earth mover's distance
Wavelet EMD

ABSTRACT

In crowded scenes, the extracted low-level features, such as optical flow or spatio-temporal interest point, are inevitably noisy and uncertainty. In this paper, we propose a fully unsupervised non-negative sparse coding based approach for abnormality event detection in crowded scenes, which is specifically tailored to cope with feature noisy and uncertainty. The abnormality of query sample is decided by the sparse reconstruction cost from an atomically learned event dictionary, which forms a sparse coding bases. In our algorithm, we formulate the task of dictionary learning as a non-negative matrix factorization (NMF) problem with a sparsity constraint. We take the robust Earth Mover's Distance (EMD), instead of traditional Euclidean distance, as distance metric reconstruction cost function. To reduce the computation complexity of EMD, an approximate EMD, namely wavelet EMD, is introduced and well combined into our approach, without losing performance. In addition, the combination of wavelet EMD with our approach guarantees the convexity of optimization in dictionary learning. To handle both local abnormality detection (LAD) and global abnormality detection, we adopt two different types of spatio-temporal basis. Experiments conducted on four public available datasets demonstrate the promising performance of our work against the state-of-the-art methods.

1. Introduction

Abnormality detection plays an important role in video surveillance context. Recently, there has been an increasing interest in modeling activities and detecting abnormal events in crowded scenes. One of the most challenging tasks in computer vision is event analysis (behavior analysis) in crowded scenes, because it requires monitoring an excessive number of individuals and their activities, retaining structural information regarding the entire scene. A central task of event analysis involves detecting or even predicting abnormal events, i.e., finding new patterns in data that do not conform to the expected case. Fig. 1 shows four examples of abnormal events: (a) and (b) are abnormalities in global scale; (c) and (d) are abnormalities in local scale.

Abnormality detection in crowded scenes is thoroughly studied by the computer vision communities, where some well established models have been developed. As a one-class learning problem, most of algorithms [1–6] intended to detect query sample with lower probability as abnormality by fitting a probability model over the training samples. However, as a high-dimensional feature is essential to better represent the event and the required number of training sample increases exponentially with the feature dimension, it is unrealistic to collect enough data for density estimation in practice. Recently, Cong et al. [7] proposed to detect abnormal events via sparse reconstruction over the normal bases (dictionary). Sparse representation is suitable to model high-dimensional samples [7,8]. Normal event is likely to generate sparse reconstruction coefficients with a small reconstruction cost, while abnormal event is dissimilar to any of the normal basis, thus generates a dense representation with a large reconstruction cost. In crowded scenes, low-level visual information (e.g. optical flow, spatio-temporal interest point) extracted is inevitably uncertain and noisy. Recent works [7,1,5,6,9,10] had shown the power of machine learning or statistic approaches can be used to implicitly handle noisy and uncertainty. However, none of previous works had explicitly tackled with feature's noisy and uncertainty.

To address the above issue, we propose a new approach for abnormality detection in crowded scenes based on sparse coding framework, which is specifically tailored to cope with features' uncertainty and noisy. Different from previous works [7,9], we model the task of dictionary learning as a non-negative matrix factorization problem. There is a subtle difference between dictionary learning and matrix factorization. Matrix factorization can be thought of as a special case of dictionary learning, where the size of the dictionary is constrained to be less than or equal to the observed data dimension [11]. Non-negative matrix factorization (NMF) provides an elegant framework to achieve sparsity on the

* Corresponding author. Tel.: +86 10 82544507.
E-mail addresses: xbzhu@nlpr.ia.ac.cn (X. Zhu), jliu@nlpr.ia.ac.cn (J. Liu), jqwang@nlpr.ia.ac.cn (J. Wang), csli@nlpr.ia.ac.cn (C. Li), luhq@nlpr.ia.ac.cn (H. Lu).

0031-3203/$ - see front matter
http://dx.doi.org/10.1016/j.patcog.2013.11.018

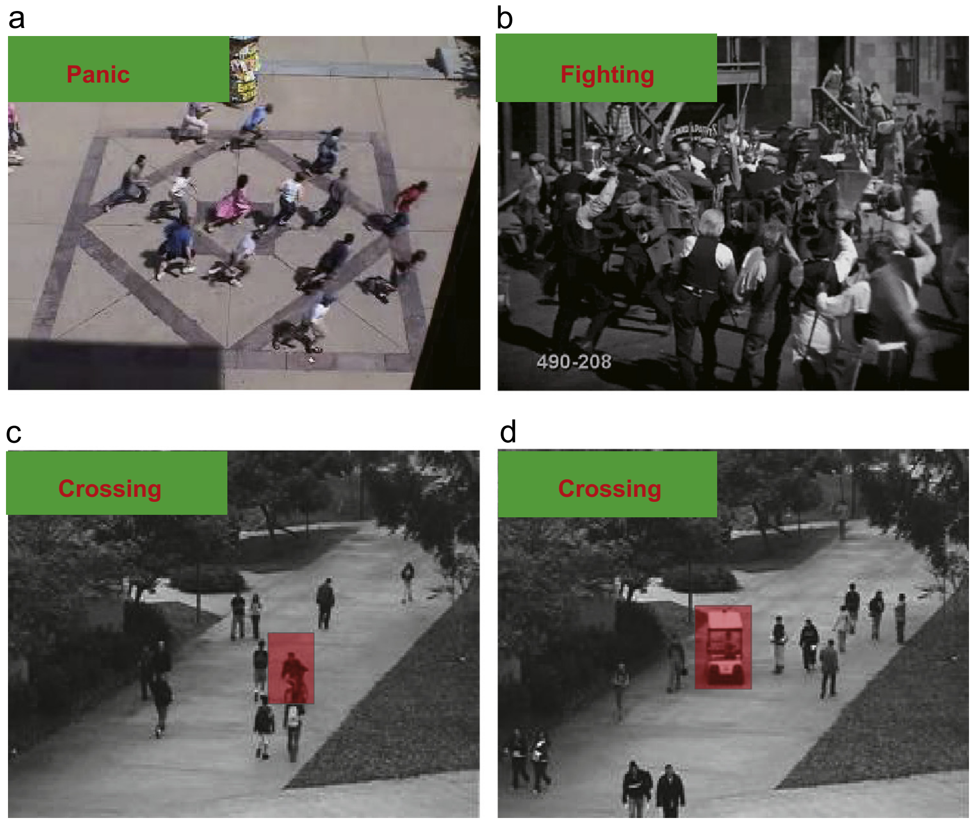

Fig. 1. Abnormal event examples: (a) people are running away in panic; (b) people are fighting; (c) bicycle is crossing the sidewalk; (d) car is crossing the sidewalk. The red masks indicate where the abnormality is. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

basis or coefficient matrices by using corresponding regularizer. In our algorithm, we adopt the Earth Mover's Distance (EMD) as objective cost function, which is a well-known robust metric in the case of noisy histogram comparison. To further reduce the influence of noisy data, we constrain the feature sample as a sparse linear combination of the elements in training dictionary by a weighted $l_{2,1}$ minimization on weight matrix in dictionary learning. The problem with the EMD is its expensive computation, which prohibits its applications in many vision problems. To tackle this problem, we introduce approximate EMD [12] into our approach to greatly degrade the computation complexity without lose of performance. In addition, the combination of approximate EMD and our approach guarantees the convexity of the optimization problem. To the best of our knowledge, no previous works have addressed abnormality detection for crowed scenes under a non-negative matrix factorization framework with EMD metric. Our method has been tested on four video datasets, all of which are publicly available. The experimental results show that our approach successfully detects abnormality both in global and local crowded scenes, and it is very competitive with respect to state-of-the-art algorithms [7,1].

The rest of the paper is organized as follows. Section 2 overviews the related works. In Section 3, we introduce the details of Earth Mover's Distance. In Section 4, we elaborate our method. The experimental evaluation is given in Section 5, and we draw conclusion in Section 6.

2. Related work

Abnormality detection is an active area of research on its own [13]. Various approaches have been proposed for both crowded and uncrowded scenes. Abnormality in crowded scenes is very challenging, due to the diversity of events and the noise in the scenes. In the following, we will focus on the case of crowded scenes. In crowed scenes, object occlusion can severely affect the accuracy of segmentation or tracking, which can heavily degrade the performance of detection. Additionally, the computational cost will also be tremendous when various objects exist. Consequently, the most popular works of abnormality detection extract motion or spatio-temporal interest points from local 2*D* patches or local 3*D* bricks to avoid tracking [14,7,1,4,15–18].

Depending on applications, the works relevant to the crowded scenes can be broadly divided into two categories: the local abnormality detection (LAD) and the global abnormality detection (GAD) [7]. For LAD, the local behavior is different from its spatio-temporal neighborhoods. Kratz and Nishino [14] proposed a spatial-temporal model, in which a coupled distribution-based Hidden Markov Model (HMM) is used to detect abnormalities in densely crowded scenes. However, this method may work only for one kind of normal behavior type in the scene. Mahadevan et al. [19] modeled the normal crowded behavior using mixtures of dynamic textures (MDT). Abnormality detection is formulated as an outlier detection problem, where temporal anomalies equated to events of low-probability, and spatial anomalies equated to events of discriminant saliency. In [20], a probabilistic framework based on Neyman–Pearson decision rule was developed to detect local anomalies that have infrequent patterns with respect to their neighbors. Kim et al. [6] modeled local optical flow with a mixture of probabilistic PCA models and enforce the consistency by Markov Random Field. Antic [5] parsed video frames by establishing a set of hypotheses that jointly explain all the foreground while, at the same time, trying to find normal training samples that explain the hypotheses. Zhao et al. [9] proposed a fully unsupervised dynamic sparse coding approach for detecting unusual events in videos based on online sparse re-constructibility of test samples from an