Pattern Recognition 46 (2013) 2711–2723

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

M^{4}L: Maximum margin Multi-instance Multi-cluster Learning for scene modeling

Tianzhu Zhang [a,b,*], Si Liu [a,b], Changsheng Xu [a,b], Hanqing Lu [a,b]

[a] *National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China*
[b] *China-Singapore Institute of Digital Media, Singapore 119615, Singapore*

ARTICLE INFO

Article history:
Received 14 October 2012
Received in revised form
10 January 2013
Accepted 15 February 2013
Available online 7 March 2013

Keywords:
Scene understanding
Maximum margin clustering
Multiple instance learning (MIL)
Gaussian Mixture Model (GMM)
Constrained Concave–Convex Procedure (CCCP)

ABSTRACT

Automatically learning and grouping key motion patterns in a traffic scene captured by a static camera is a fundamental and challenging task for intelligent video surveillance. To learn motion patterns, trajectory obtained by object tracking is parameterized, and scene image is spatially and evenly divided into multiple regular cell blocks which potentially contain several primary motion patterns. Then, for each block, Gaussian Mixture Model (GMM) is adopted to learn its motion patterns based on the parameters of trajectories. Grouping motion pattern can be done by clustering blocks indirectly, and each cluster of blocks corresponds to a certain motion pattern. For one particular block, each of its motion pattern (Gaussian component) can be viewed as an instance, and all motion patterns (Gaussian components) constitute a bag which can correspond to multiple semantic clusters. Therefore, blocks can be grouped as a Multi-instance Multi-cluster Learning (MIMCL) problem, and a novel Maximum Margin Multi-instance Multi-cluster Learning (M^{4}L) algorithm is proposed. To avoid processing a difficult optimization problem, M^{4}L is further relaxed and solved by making use of a combination of the Cutting Plane method and Constrained Concave–Convex Procedure (CCCP). Extensive experiments are conducted on multiple real world video sequences containing various patterns and the results validate the effectiveness of our proposed approach.

1. Introduction

Video scene understanding by unsupervised inference of motion patterns in static camera scenes is a very important task in visual surveillance. With the increasing of the surveillance systems, large amounts of video data are created every day, and it is difficult and time-consuming to label and organize these videos manually. Therefore, various methods have been made to understand the videos automatically [1–3]. By clustering the motion patterns, we can obtain information along which path or direction the vehicles or pedestrians should move or walk. Based on such information, it is very convenient to detect abnormal activities and meet the great needs for traffic management systems. However, the motion of pedestrians and vehicles is complex and their motion patterns are different from each other. Thus, automatically clustering the motion patterns for video understanding is a challenging problem in computer vision and pattern recognition.

The standard approach for analysis of video sequences involves four primary parts: (1) moving object detection; (2) object classification; (3) motion pattern learning and clustering and (4) activity analysis. There is a lot of progress made in each of the modules. For moving object detection, the Gaussian Mixture Model (GMM) [1] is frequently adopted to model background. For classification of objects into different categories (e.g. a vehicle, a person), scene context features (such as position, area in pixels, and velocity) [4] are used to cluster trajectories into different types (vehicles vs. pedestrians), and show effective performance by experimental results. However, due to low resolution, shadow, and different viewing angles, object classification only using these features is not enough in video surveillance. Therefore, a co-training based method [5] is proposed to train classifiers with multiple different kinds of features. For motion pattern learning and clustering, the most commonly used features are low level motion and appearance features [6–8]. Examples of such features include sparse or dense optical flows [9], spatiotemporal gradients [10], and object trajectories obtained after detection and tracking [11,5,3]. Based on different descriptions, various learning and clustering methods are adopted [12–14]. For activity analysis, existing approaches [1,4,5] use clustered motion patterns to recognize the abnormal activities of objects in video sequences.

[*] Corresponding author at: National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China. Tel.: +86-10-82631857.
E-mail address: tzzhang10@gmail.com (T. Zhang).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.02.018

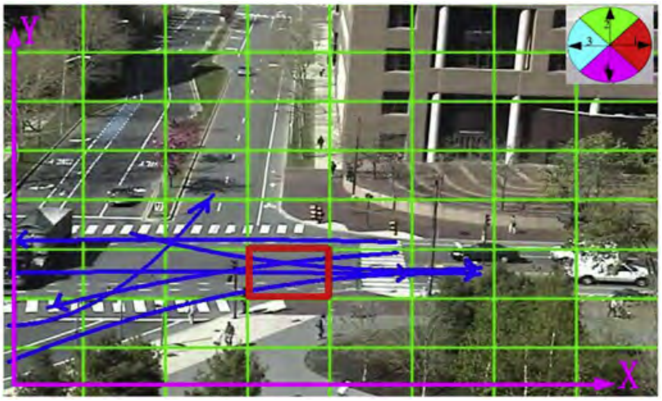

Fig. 1. Multiple primary motion patterns may exist in a certain block as shown in the red rectangle. Here, each trajectory represents a motion pattern. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this Article.)

In this work, we mainly focus on the third part about motion pattern learning and clustering, i.e., how to learn the motion pattern and design a suitable clustering algorithm. As shown in Fig. 1, object trajectory can be obtained by tracking algorithm [1], and each trajectory is a potential motion pattern in the scene image. Distribution of trajectories in a certain region can be viewed as a Gaussian distribution from the statistic point of view. The scene image can be divided into multiple cell blocks, and each block may have multiple distributions (multiple clusters of trajectories). Therefore, the Gaussian Mixture Model (GMM) is adopted to learn spatial distributions of trajectories in each block, and each Gaussian component is one of the underlying motion pattern. Then, how to cluster the motion patterns of all blocks in the scene image? Since each cluster of blocks is corresponding to a certain motion pattern, grouping motion patterns can be done by clustering blocks indirectly. For each block, each motion pattern (Gaussian component) can be viewed as an instance, and all motion patterns (Gaussian components or instances) constitute a bag which may contain multiple semantic clusters simultaneously. In this way, each block (bag) is associated with not only multiple instances but also multiple clusters. Therefore, we formulate the blocks grouping task as a Multi-instance Multi-cluster Learning (MIMCL) problem. There is little work directly dealing with this problem. Most of the existing clustering methods [15–18] are designed to solve traditional Single-instance Single-cluster Learning (SISCL) problems, and some methods [19,20] only consider a bag including one semantic cluster and deal with Multi-instance Single-cluster Learning (MISCL) problems. In many real-word cases, such as block clustering, a bag (block) may belong to more than one clusters. Therefore, it is unreasonable to adopt SISCL or MISCL formulation and assign a bag to only one cluster.

Considering the MIMCL formulation and the maximum margin clustering criterion [18,20,21], we propose a novel algorithm named M^4L, i.e. Maximum Margin Multi-instance Multi-cluster learning, to cluster motion patterns for video scene modeling. Briefly, M^4L assumes a linear model for each cluster, where the output of a bag on one cluster is set to be the maximum prediction scores of all the instances. Subsequently, the outputs on all possible clusters for all instances of a bag are adopted to define the margin. That is, for a certain cluster, the corresponding margin of a bag is defined by using the output of the most discriminative instance, and the margin of the bag with respect to the clustering system is set to be the minimum margin of the bag over all possible clusters. Obviously, each instance is adopted to determine the output on each possible cluster and the correlations between different clusters are also considered in the combination phase. Therefore, the connections between the instances and the clusters are explicitly exploited by M^4L. Compared with the existing approaches, the contributions of our work can be summarized as follows.

1. We formulate motion pattern clustering in video surveillance as a Multi-instance Multi-cluster Learning (MIMCL) problem, which provides a new perspective and is very suitable to group motion patterns.
2. To group motion patterns efficiently, we propose a novel clustering algorithm denoted as M^4L: Maximum Margin Multi-instance Multi-cluster Learning, which adopts the theory of support vector machine and aims at finding the maximum margin hyperplane to separate the data from different classes.
3. To solve the nonconvex M^4L algorithm, we make use of combination of Constrained Concave–Convex Procedure (CCCP) and the Cutting Plane method for efficient optimization solution.

The rest of the paper is organized as follows. In Section 2, we introduce some related work to this paper. The proposed approach is described in details in Section 3 including motion pattern learning and motion pattern clustering. Experimental results are reported and analyzed in Section 4. Finally, we conclude the paper with future work in Section 5.

2. Related work

The problem of scene modeling in visual surveillance is not new [4,1,2,5,22,14,23–25]. In general, the task is to lay out the structure of traffic scenes (e.g., roads, sidewalks, intersections), or learn motion patterns (e.g., pedestrian crossings, vehicles turning). The proposed work is an attempt to learn and cluster motion patterns from static camera videos without any user intervention. The most related work to our method is scene understanding in visual surveillance, and clustering relevant methods. We review the state-of-the-arts of these two topics, respectively.

2.1. Video scene understanding

In video surveillance, many methods attempt to learn motion patterns for video scene understanding. Stauffer and Grimson [1] use a real-time tracking algorithm in order to learn patterns of motion (or activity) from the obtained tracks. Due to the use of co-occurrence matrix from a finite vocabulary, these approaches are independent from the trajectory length. However, the vocabulary size is limited for effective clustering and time ordering is sometimes neglected. Hu et al. [2] generate trajectories using fuzzy k-means algorithms for detecting foreground pixels. Trajectories are then clustered hierarchically and each motion pattern is represented with a chain of Gaussian distributions. However, the number of clusters must be given manually and the data must be of equal length, which weakens the dynamic aspect. Wang et al. [4] propose a trajectory similarity measure to cluster the trajectories and then learn the scene model from trajectory clusters. Basharat et al. [26] learn patterns of motion as well as patterns of object motion and size. These approaches [4,26] are adapted to real-time applications and time-varying scenes because the number of clusters is not specified and they are updated over time. However, it is difficult to select a criterion for new cluster initialization that prevents the inclusion of outliers