Pattern Recognition 47 (2014) 3841–3851

Contents lists available at ScienceDirect

Pattern Recognition

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

Video object matching across multiple non-overlapping camera views based on multi-feature fusion and incremental learning

Huiyan Wang [a], Xun Wang [a,*], Jia Zheng [a], John Robert Deller [b], Haoyu Peng [a], Leqing Zhu [a], Weigang Chen [a], Xiaolan Li [a], Riji Liu [a], Hujun Bao [c]

[a] School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
[b] Electrical and Computer Engineering Department, Michigan State University, East Lansing 48824, USA
[c] The State Key Lab of CAD and CG, College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form
19 May 2014
Accepted 19 June 2014
Available online 28 June 2014

Keywords:
Video object matching
Non-overlapping multi-camera views
CMFH
Incremental learning
Video surveillance system

ABSTRACT

Matching objects across multiple cameras with non-overlapping views is a necessary but difficult task in the wide area video surveillance. Owing to the lack of spatio-temporal information, only the visual information can be used in some scenarios, especially when the cameras are widely separated. This paper proposes a novel framework based on multi-feature fusion and incremental learning to match the objects across disjoint views in the absence of space–time cues. We first develop a competitive major feature histogram fusion representation (CMFH[1]) to formulate the appearance model for characterizing the potentially matching objects. The appearances of the objects can change over time and hence the models should be continuously updated. We then adopt an improved incremental general multicategory support vector machine algorithm (IGMSVM[2]) to update the appearance models online and match the objects based on a classification method. Only a small amount of samples are needed for building an accurate classification model in our method. Several tests are performed on CAVIAR, ISCAPS and VIPeR databases where the objects change significantly due to variations in the viewpoint, illumination and poses. Experimental results demonstrate the advantages of the proposed methodology in terms of computational efficiency, computation storage, and matching accuracy over that of other state-of-the-art classification-based matching approaches. The system developed in this research can be used in real-time video surveillance applications.

1. Introduction

Visual tracking is a challenging problem in applications such as intelligent video surveillance, human motion analysis, human–computer interaction and augmented reality. The main task of a tracking system is to assign consistent labels to tracked objects in a stream of video frames. In video surveillance, it is not possible to monitor a wide area by using only a single camera because the field of view (FOV) of a camera is finite and limited by the scene structure. Therefore, a surveillance system for a wide area has to be deployed in a network of cameras and has to track objects across multiple cameras.

Visual cameras are widely installed in many large public places, such as airports, subway stations, and squares. The use of multiple cameras makes it possible to track objects over long distances. In practice, considering the cost and computational complexity, the cameras are often non-overlapping (disjoint). Thus, object matching across disjoint camera views is now becoming one of the most important tasks for any surveillance system. The ultimate goal of object matching is to associate the tracks of objects observed in different camera views, and re-identify the targets at different locations and time instants. This task poses considerable challenge because the same object observed in different camera views may be widely separated in space and time. Moreover, the object may undergo significant changes in illumination, viewpoints, and poses.

In a large public area monitored by multiple non-overlapping cameras, the topology of the cameras is not available and the cameras are not calibrated. Camera calibration is a very time-consuming task,

* Corresponding author. Tel.: +86 57128008302, fax: +86 57128008303.
E-mail addresses: cederic@zjgsu.edu.cn (H. Wang), wx@zjgsu.edu.cn (X. Wang), zj10102@sina.com (J. Zheng), deller@egr.msu.edu (J.R. Deller), phy@zjgsu.edu.cn (H. Peng), zhuleqing@zjgsu.edu.cn (L. Zhu), gary302a2003@yahoo.com.cn (W. Chen), lixiaolana@hotmail.com (X. Li), liuriji@qq.com (R. Liu), bao@cad.zju.edu.cn (H. Bao).

[1] CMFH is the abbreviation of Competitive Major Feature Histogram fusion representation.

[2] IGMSVM is the abbreviation of Incremental General Multicategory Support Vector Machine learning algorithm.

http://dx.doi.org/10.1016/j.patcog.2014.06.019
0031-3203/

without which the spatio-temporal relationships among cameras cannot be identified. The observations in large public spaces are widely separated and continuous tracking information for objects is unavailable. Therefore, the traditional object matching methods based on trajectory tracking are not applicable. In this situation, object matching becomes difficult because a model has to be built only from visual appearance features without spatio-temporal reasoning. Object matching based on appearance features is a classical method to solve the problem of consistent labelling across disjoint camera views.

The visual appearance features are mainly extracted from the clothing and shapes of objects, and are usually characterized by their colour, shape, and texture. In most situations, a single type of feature is not sufficient to represent the subtle differences between all object pairs. In existing appearance-based methods, colour histograms [1,2], histograms of oriented gradients (HOG) [3], scale-invariant feature transforms (SIFT) [4] or other representative features are used. However, traditional colour histograms are sensitive to changes in illumination and photometric settings of cameras and often fail to distinguish a large number of objects in long distance tracking. The SIFT features are robust to illumination changes and affine distortion, and have a strong ability in location of feature points. However, a SIFT descriptor is a 128-dimensional vector and the number of descriptors in each frame is often more than thousands. It is difficult to directly use the high-dimensional features for object matching. Recently, Teixeira and Corte-Real [5] proposed a novel method to reidentify objects by using a vocabulary tree to quantize SIFT features and represent the object as a multi-dimensional vocabulary vector. This method was proved to be effective in discriminating visual objects. However,the feature dimension in this method is very high, which may impair the efficiency and accuracy of object recognition. For example, suppose the size of the feature matrix of one object is 100×128 (100 is the number of features and 128 is the dimension of a SIFT descriptor), the features are quantized by using a vocabulary tree with K cluster centres at each of L levels. To improve the accuracy, K and L need to be set to larger values. When K is set to 10 and L to 4, the dimension of features goes to $1 \times 11\,110$.

To solve the above problem and improve the matching accuracy (which is 76.2% in [5]), we propose a new framework based on multi-feature fusion and a generic support-vector-machine based incremental learning algorithm for object matching across disjoint views when spatio-temporal information is not available for long distance tracking.

The major contribution of this paper is two-fold. First, a novel appearance model is developed by employing the competitive major feature histogram fusion representation (CMFH) algorithm. CMFH combines the local texture descriptors with colour features to form the original feature space. Following this a competition mechanism is introduced and a kernel-based algorithm is used to fuse the features, This can reduce the dimension and obtain the important features with a high discriminative power. Second, we provide an efficient incremental algorithm for updating the appearance model online, and perform matching based on a classification method. We called our algorithm Incremental General Multicategory Support Vector Machine learning algorithm (IGMSVM). The proposed framework can meet real-time matching and low computation storage requirements, as well as achieve a higher object recognition performance than existing state-of-the-art classification-based matching algorithms.

The paper is organized as follows. We begin by introducing the relevant work in the next section. Section 3 describes the proposed construction method of appearance model in detail. Section 4 presents the model update scheme and the procedure of our proposed framework for object matching. The experiment results and discussion are given in Section 5. Section 6 provides a conclusion.

2. Related work

Multi-camera object matching in disjoint views is a very challenging task. The same objects observed with different cameras undergo large visual appearance changes caused by differences in illumination, different camera viewpoints, poses, occlusions, and other environmental conditions. Moreover, the objects captured are often small in size and some important visual details may be indistinguishable. In order to address the difficulties, a great deal of research in appearance feature extraction has been conducted in recent years. A study done by Javed et al. [6] constructed a brightness transfer function (BTF) from a given camera to another, and effectively eliminated object appearance changes caused by differences in illumination and optical characteristics of cameras. A machine learning (ML) framework for tracking by integrating space-time cues and appearance schemes is proposed in [7]. In this study, kernel density estimation is used to train and learn the inter-camera space–time relationships, and BTF is employed to address appearance changes. The approach taken in [8] involved a tracking method for large-scale scenes based on key frames, which can effectively reduce ambiguity and redundancy in object matching. Montcalm and Boufama [9] used multiple features that were dynamically weighted for matching moving objects across multiple-cameras. Chen et al. [10] used a categorical feature indicating the entry/exit of cameras to handle different patterns of appearance changes. They also used an AdaBoost classifier to solve the problem of multi-object recognition. Zheng et al. [11] presented a probabilistic relative distance comparison model to learn the optimal distance that can maximize matching accuracy. Arth et al. [12] generated unique object signatures based on principal component analysis (PCA)-SIFT features and vocabulary trees and then used the signatures to perform object reacquisition in large-scale traffic scenarios. Liu et al. [13] proposed an adaptive feature-fusion algorithm to fuse four types of features (colour histogram, UV chromaticity, major colour spectrum, and SIFT) for object matching in non-overlapping scenes.

In general, object matching can be conducted based on classifiers or distance metrics, such as Euclidean distance and Mahalanobis distance. To adapt to appearance changes, the object feature model needs to be learned and updated adaptively. Moreover, in order to detect new objects and adapt to the morphological diversity of objects, incremental learning classifiers are good choices. Incremental learning is a machine learning paradigm in which a learning process is carried out to update the model when a new sample is added. Unlike traditional machine learning methods, incremental learning does not assume the availability of a sufficient training set before the learning process. On the contrary, the training examples are augmented over time. In recent years, a number of incremental-learning-based models were developed. In [14], an incremental Bayesian approach was proposed, which makes use of prior information and assembles from (unrelated) object categories that were previously learnt. In addition, a generic probabilistic model was developed and the parameters of the model were learnt incrementally. Polikar et al. [15] presented an incremental learning method called Learn++, and an improved algorithm called Learn++.MF was proposed in [16]. The Learn++.MF algorithm is based on an ensemble of classifiers and utilizes random subspace selection to deal with the missing feature problem in supervised classification. Gong-De et al. [17] presented an incremental k nearest neighbour (KNN) model, in which a few clusters were constructed for new incoming data to optimize the previous KNN model. A number of incremental-learning-based methods have been applied to visual tracking. Ross et al. [18] proposed a tracking method that can incrementally learn a low-dimensional subspace representation and adapt to