Pedestrian Detection and Tracking Using Deformable Part Models and Kalman Filtering

Shubham Mittal, Twisha Prasad, Suraj Saurabh
Dept. of Electronics & Electrical Engineering
IIT Guwahati
Guwahati, India
{m.shubham, twisha, s.suraj}@iitg.ernet.in

Xue Fan, Hyunchul Shin
Dept. of Electronics and Communication Engineering
Hanyang University
Sangnok-gu, Korea
fanxue@digital.hanyang.ac.kr, shin@hanyang.ac.kr

Abstract—Both detection and tracking people are challenging problems, especially in complex real world scenes that commonly involve multi-person, complicated occlusions, and cluttered backgrounds. In this paper, we propose a novel approach for multi-person tracking-by-detection using deformable part models in Kalman filtering framework. The Kalman filter is used to keep track of each person and a unique label is assigned to each tracked individual. Based on this approach, people can enter and leave the scene at random. We test and demonstrate our results on the Caltech Pedestrian benchmark, which is the largest available dataset and consists of pedestrians varying widely in appearance, pose and scale. Complex situations such as people merging together are handled gracefully and individual persons can be tracked correctly after a group of people split. Experiments demonstrate the real-time performance and robustness of our system working in complex scenes. Our tracking model gives a tracking accuracy of 72.8% and a tracking precision of 82.3%.

Keywords-Multi-Person Tracking; Part-Based Models; Data Association; Kalman Filter; Pedestrian Detection; Tracking-by-Detection

I. INTRODUCTION

The goal of this paper is to enable reliable multi-person tracking from a moving platform in real-world scenarios. In this paper, we present an integrated tracking-by-detection framework which is able to detect and track multiple persons even in challenging scenarios such as instances where the objects are partially occluded for extended periods of time. We integrate deformable part based models with Kalman filtering. We use the detector of [1] and modify the source code to train it on Caltech Pedestrian Dataset [2].

The main challenge when using the pedestrian detector for tracking is that the detector output is unreliable. The detection output sometimes consists of some false positives and missed detections. Thus, the resulting association problem between detections and targets is difficult. Several recent algorithms address this problem by optimizing detection assignments over a large temporal window in an offline step [3], [4]. With the improvement of detection algorithms [1], both in accuracy and computational feasibility, tracking-by-detection is one of the most popular concepts for tracking [5]. Due to the high amount of false positives and missing detections in the output of the detector, it is necessary to incorporate temporal context. Recent tracking approaches [6] try to associate the detections and track objects from uncalibrated single camera.

Recently, pedestrian detection methods have achieved great improvements [7], [8] which have allowed the scientific community to focus on associating tracklets which are initially linked by detection responses in consecutive frames. Next, short tracklets are associated to form longer ones by maximizing the probabilities globally using the similarity and discrimination scores both from appearance and motion models. By using the information of the previous, current, and future frames, these methods rely on the spatio-temporal restriction to improve the results with different types of features from simple object signatures such as height, width to more complicated ones like local gradient intensity features. The most important factor to improve the tracking results is building the affinity scores between two tracklets to associate tracklets to the right target.

Our paper makes the following contributions: (1) We test our model on the Caltech Pedestrian Dataset (the biggest pedestrian dataset available) and demonstrate impressive results even in heavy occluded scenes; (2) We show that our method achieves high tracking accuracy by successfully associating data within consecutive frames to overcome the problem of unreliable detection.

The rest of the paper is organized as follows. In Section II, a brief review of the related work is presented. Deformable part models are shown in section III. Kalman Filtering is presented in Section IV. Data Association is presented in Section V. Results are shown in Section VI. Finally, the conclusions and possible future works are summarized in Section VII.

II. RELATED WORKS

A. Tracking-by-Detection

Many tracking methods rely on background subtraction from one or several static cameras [9], [10]. But, recent improvements in object detection have stimulated the interest in combining tracking and detection. The combination of object detectors and Kalman filtering results in algorithms that are more suitable for time-critical, online applications. On the other hand, state-of-the-art object detectors all build up some form of confidence density as one stage of their pipeline, which can be used instead as a graded observation model to handle difficult situations more robustly. Previous algorithms that
exploit this intermediate output have been developed primarily for single-target tracking (mostly of faces) and have not been evaluated thoroughly for multiple interacting targets. Relying on the detector confidence in every situation can cause tracking errors, particularly during occlusions between interacting targets and in complex cluttered scenes. This work presents a method to use this unreliable information source for robust multi-person tracking.

B. Data Association

Using independent trackers require solving a data association problem to assign detections to targets. A large number of strategies are available to solve the data association problem. Classical approaches include the Joint Probabilistic Data Association Filter (JPDAF) [11] and Multi Hypotheses Tracking (MHT) [12]. MHT attempts to keep track of all the possible association hypotheses over time. This is an NP-hard problem, since the number of association hypotheses grows exponentially over time. Thus methods are required to reduce the computational complexity. JPDAFs instead try to make the best possible assignment in each time step by jointly considering all possible associations between targets and detections, to the cost of an exponentially increasing complexity. Alternatively, the Hungarian algorithm [13] can be used to find the best assignment of possible detection-tracker pairs in a runtime that is cubic in the number of targets. Another alternative is to use a greedy approach, as pointed out by [14].

III. DEFORMABLE PART MODELS

In this paper we use a pedestrian detection system based on mixtures of multi-scale deformable part models [1]. These models are trained using a discriminative procedure that only requires bounding boxes for the objects in a set of images. The system relies heavily on new methods for discriminative training of classifiers that make use of latent information. It also relies heavily on efficient methods for matching deformable models to images.

The core ideas of deformable part-based models boil down to three factors: (1) A deformable part representation for pedestrian; (2) An efficient matching process; (3) Latent SVM, a discriminative training method. A star-structured part-based model is defined which is composed of a root filter, n (usually six) part filters, and associated deformation parameters. An efficient matching process based on dynamic programming and generalized distance transforms is proposed. To detect objects in an image we compute an overall score for each root location according to the best possible placement of the parts. Finally, a latent SVM training process is formulated to train a mixture of star models from bounding box ground truth.

Rather than trying to capture a global pattern of an object with one template, part-based models focus on parts of an object, thereby, providing more flexible and robust representations. The resulting system is both efficient and accurate, achieving state-of-the-art results on the PASCAL VOC benchmarks and the INRIA Person dataset.

IV. KALMAN FILTERING

The Kalman filter has been used extensively for tracking in interactive computer graphics. It is essentially a set of mathematical equations that implement a predictor-corrector type estimator that is optimal in the sense that it minimizes the estimated error covariance recursively.

In Kalman filer, we consider a behavior of the object, where subscript k indicates the discrete time. The objective is to compute posteriori state estimate $\hat{x}_k$ from the measurement $z_k$ and a priori state estimate $\tilde{x}_k^-$. In the following we give a mathematical description of Kalman filter.

A. The process to be estimated

The Kalman filter addresses the general problem of trying to estimate the state $x \in R^n$ of a discrete-time controlled process that is governed by the linear stochastic difference equation

$$x_k = Ax_{k-1} + w_{k-1},$$

with a measurement $z \in R^m$ that is

$$z_k = Hx_k + v_k,$$

where $A$ represents the state transition model and $H$ represents the observation model. The random variables $w_k$ and $v_k$ represent the process and measurement noise respectively. They are zero-mean mutually dependent white Gaussian noise vectors with covariance matrices $Q$ and $R$.

B. Time update

We define $\tilde{x}_k^- \in R^n$ to be our a priori state estimate at step k given knowledge of the process prior to step k, and $P_k^-$ to be priori estimate error covariance. $\tilde{x}_k^- \text{ and } P_k^-$ are obtained for the next time step in following equations:

$$\tilde{x}_k^- = Ax_{k-1} + w_{k-1},$$

$$P_k^- = AP_{k-1}A^T + Q.$$  

These equations show how to project the state and covariance estimates forward from time step k-1 to step k.
C. Measurement update
We define $\hat{x}_k \in \mathbb{R}^n$ to be our a posteriori state estimate at step $k$ given measurement $z_k$ and a priori estimate $\hat{x}_k^-$, and $P_k$ to be posteriori estimate error covariance. The objective is to find an equation that compute a posteriori estimating $\hat{x}_k$ which is a linear combination of the a priori estimate and the new measurement $z_k$. These equations are given below:

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1},$$  \hspace{1cm} (5)$$

$$\hat{x}_k = \hat{x}_k^- + K_k (z_k - H\hat{x}_k^-),$$ \hspace{1cm} (6)$$

$$P_k = (1 - K_k H) P_k^-,$$ \hspace{1cm} (7)$$

where $K_k$ is the Kalman gain. After that a posterior state estimate $\hat{x}_k$ and a posterior error estimate $P_k$ is computed. The time and measurement equations are calculated recursively with previous a posteriori estimates to predict new a priori estimate. This recursive behavior of estimating the states is one of the highlights of the Kalman filter.

V. DATA ASSOCIATION
Data association is a problem of great importance for multiple target tracking applications. In order to decide which measurement should associate with which target, we use data association to get the correspondence to assign at most one detection to one target. Then data association techniques are used to combine detections after $k$ consecutive frames to form tracklets. Each tracklet is then associated with a Kalman filter. To eliminate duplicate tracking results, we use several similarity measurements based on appearance features as well as geometry reasoning.

A. Short tracklet generation
The data-assocation between people detections in different frames is highly challenging and ambiguous. To address this, we first extract people-tracklets (people detections consistent over frames) from a small number of consecutive frames. As any single person might be detectable only for a small number of frames, the extraction of people-tracklets has to be highly robust.

B. Algorithm for JPDA
The Joint Probability Data Association (JPDA) algorithm is considered as the most widely and successful strategy for multi-target tracking under data association uncertainty. The JPDA is an extension of the Probability Data Association (PDA) algorithm to track multiple targets. Whereas the PDA algorithm computes the probability that measurement $j$ originated from target $t$, separately at each target, under the assumption that all measurements not associated with target $t$ are false, the JPDA algorithm computes jointly across the set of $T$ targets and clutters. To compute the association probability, we should note that:

1) Each target state has to be assigned to a measurement. Indeed, at each time step the sensor provides a set of measurements. The source of these measurements can be the targets or the disturbances also known as clutters. Therefore, a special procedure is needed to assign each measurement to its associated target.

2) Similarly, since the JPDA algorithm updates the estimated states sequentially, a recursive solution should be applied to update the states at each sample time.

In JPDA, infeasible hypotheses are pruned away using a gating procedure at each time step. A filtering estimate is then computed for each of the remaining hypotheses, and combined in proportion to the corresponding hypothesis probabilities. We use $\omega_{jt}$ denoting whether the $j^{th}$ valid measurement found in the validation gate of target $t$ at time $k$:

$$\omega_{jt} = \{ 1 \text{ if the } j^{th} \text{ measurement found in the validation gate of } t \},$$ \hspace{1cm} (8)$$

$\theta_{jt}$ is defined as the association event, denoting that valid measurement $Z_j(k)$ is originated from target $t$. The function of the joint association event $\theta$ as follows:

$$\theta = \bigcap_{j=1}^{m_k} \theta_{jt},$$ \hspace{1cm} (9)$$

where $m_k$ is defined as the number of valid measurements. We use $\hat{\omega}_j(\theta)$ denote the feasible joint event if each measurement is originated from one source and each target has only one valid measurement. Then we can get the association probability $\beta_{jt}$ as follows:

$$\beta_{jt} = \sum \hat{\omega}_j(\theta) \hat{\omega}_j(\theta).$$ \hspace{1cm} (10)$$

VI. EXPERIMENTAL RESULTS
We apply our joint probability based data association framework to the multiple person tracking problems using Kalman filtering. We test and demonstrate our results on the Caltech Pedestrian benchmark, which are two orders of magnitude larger than any existing dataset.

The metrics used for the evaluation involves the followings:

1) The Multiple Object Tracking Precision (MOTP).

$$\text{MOTP} = 1 - \frac{\sum_k d_{t,k}}{\sum_k c_k},$$ \hspace{1cm} (11)$$

where $d_{t,k}$ is the position error for target hypothesis pairs, at time $k$, $\sum_k c_k$ is the total number of matches made. It shows the ability of the tracker to estimate precise object positions, independent of its skill at recognizing object configurations, keeping consistent trajectories, etc.

2) The Multiple Object Tracking Accuracy (MOTA).

$$\text{MOTA} = 1 - \frac{\sum_k m_{k,t} f_{p_k, mme_k}}{\sum_k g_k},$$ \hspace{1cm} (12)$$

where $m_{k,t}$, $f_{p_k}$ and $mme_k$ are the number of misses, false positives and mismatches respectively, at time $k$. $\sum_k g_k$ is the total number of pedestrians over all frames. The MOTA can be seen as composed of 3 error ratios:

a) the ratio of false positives

$$fp = \frac{\sum_k f_{p_k}}{\sum_k g_k}$$ \hspace{1cm} (13)$$

-3-
b) the ratio of misses in the sequence, computed over the total number of objects present in all frames

\[ m = \frac{\sum \text{mme}_k}{\sum \theta_k} \]  

(14)

c) the ratio of mismatches

\[ \text{mme} = \frac{\sum \text{mme}_k}{\sum \theta_k} \]  

(15)

Figure 2 displays our tracking results obtained on a test video of Caltech Dataset. The true detections are shown in green bounding boxes. As can be seen, our model successfully tracks multiple persons in challenging scenarios such as heavy occlusions. The accuracy of our tracking model (MOTA) is 72.8 % and the precision (MOTP) obtained is 82.3 %.

VII. CONCLUSIONS

We presented a framework to track multiple persons on Caltech dataset which comprises of difficult scenes. Our method successfully integrated the deformable part-based model with Kalman filtering based on Joint Probability Data Association (JPDA) for tracking. Good results are obtained on Caltech Dataset. For future works, we plan to integrate more complex motion models and also introduce new models such as to detect people from arbitrary view-points. We also plan to focus on estimating the 3 dimensional trajectory of each moving object using multi-camera data fusion, analyzing interactions between different people and thereby detecting suspicious behaviors.

ACKNOWLEDGMENT

This work was supported by the Ministry of Knowledge Economy (MKE) through IDEC Platform center (IPC) at Hanyang University.

REFERENCES