Bulletin de la Société Royale des Sciences de Liège, Vol. 85, 2016, *p.* 1211 - 1218

Tracking Pedestrians based on Patch Matching

Seyed Masoud JAMSHIDI[*], Sassan AZADI[†]
Department of Electrical and Computer Engineering, Semnan University

Abstract

One of the challenging issues in machine vision over the last decade involves how to track people in consecutive video frames. Nowadays, scholars are greatly interested in tracking people in videos due to a need to track moving objects. In this paper we track pedestrians through the Dollár's detection [1] technique combined with patch matching from detected objects. At this stage, the patches are matched separately through Bhattacharyya's method in order to increase the tracking precision. We use color histogram with background subtraction. Desirable results are obtained in estimating displacement and patch search space through the combination of displacement vectors with the detection results from weighted vector median filter (WVMF). Our new technique indicates that we were able to track pedestrians more accuracy.

Keywords: pedestrian tracking, patch matching, WVMF

Introduction:

Object detection and tacking are unique characteristics enabling humans to distinguish different pedestrians on highly crowded sidewalks in a sequence of images. The implementation of tracking individuals in consecutive video frames has been a controversial topic in machine vision over the last decade. This branch of machine vision has various applications in human-computer interaction, automatic monitoring systems, and robotics.

In general, tracking objects or people in consecutive video images come with certain obstacles such as:

- Noise in images
- Overlapping of people or objects either partially or totally
- Complex movements in the image
- Lighting variations in the image

In this paper, pedestrians are tracked based on patch matching elaborated in [2], to which new features have been added.

- Tracking multiple individuals in images in an automatic procedure where they are initially detected, converted into patches and separately matched in the images.
- Prediction of the target's (pedestrian) position can tremendously help identify him/her after total or partial overlapping, taking into account the target's displacement history.

In the next section, we present review the literature regarding the pedestrians tracking.

[†] Email address: sazadi@semnan.ac.ir

Literature review

Object tracking is a charming research field in machine vision, where numerous solutions and challenging barriers are involved. This section explores a number of previous studies on pedestrian tracking.

In every tracking method, an object detection mechanism is essential either in every frame or when the object first appears in the video. The most commonly used method for object detection is to use information in a single frame, where in the video is divided into multiple sequences of frames. The first step in many computer vision applications is identifying region of interest. An image from video sequence is divided into two complimentary sets of pixels. The first set comprises the pixels which correspond to foreground objects whereas the second set contains the background pixels. This output is represented as a binary image or as a mask. It is difficult to specify an absolute standard with respect to what should be identified as foreground and what should be marked as background because this definition is somewhat application specific.

A conventional technique for detecting subjects through a fixed camera involves removing the background where a mathematical model of the image fixed background is obtained and each new frame is compared with the model until the background (objects) are detected [2]. For instance, the W_4 [3] algorithm uses the segmentation of image background as well as the combination of shape and texture data for real-time processing. Generally, foreground objects are moving objects such as people, cars and boats and consequently everything else considered to be background.

Basic steps for tacking an object are: Object Detection, Object Classification, and Object Tracking

Object detection identifies objects of interest in the video sequence and clusters pixels of these objects. Object detection can be done by different techniques such as optical flow, Background subtraction and Frame differencing. Object can be classified as vehicles, birds, floating clouds and other moving objects. Different approaches to classify the objects are Shape-based classification, Color based classification, Motion-based classification, and texture based classification. Tracking, defined as the problem of approximating the path of an object in the image plane as it moves around a scene. The methods of tracking the objects are point tracking, kernel tracking and silhouette.

The point trackers involve detection in every frame whereas kernel based tracking or contours-based tracking require detection only when the object first appears in the scene. In the structure of an image, during tracking moving objects are represented by their feature points. A set of mathematical equations to provide an efficient computational (recursive) means for estimating the process state in several ways is required. It also should support the estimations of past, present, and even future states, and can do the same even after the precise nature of the modeled system is not known. The Kalman filter estimates a process by making use of feedback control.

The filter also estimates the process state at particular time and then obtains feedback in the form of noisy measurements.

The particle filtering generates all the models for one variable before moving to the next variable. Algorithm has an advantage when variables are generated dynamically and may have numerous variables. The other algorithm is Multiple Hypothesis Tracking.

Several frames have been observed for better tracking outcomes. MHT, an iterative algorithm and iterates with a set of existing track hypotheses. Each hypothesis is a group of mutually separate tracks. For each hypothesis, a prediction of object's position in the succeeding frame is made. These predictions are then compared by calculating a distance measure.

Kernel tracking [9] is usually performed by computing the moving object, which is represented by an embryonic object region. These algorithms diverge in terms of the presence representations used, objects tracked, and the method used for estimation of the object motion.

Template matching [9] [10] is a brute force method of examining the Region of Interest in the video. In template matching, a reference image is verified with the frame that is separated from the video.

The color information can be employed to model pedestrians (objects) so as to enhance the tracking efficiency. For instance, some algorithms adopt color histograms as a feature for matching video frames [4]. Mean-shift tracking tries to find the area of a video frame that is locally most similar to a previously initialized model. The image region to be tracked is represented by a histogram. A vintage procedure is used to move the tracker to the location that maximizes a similarity score between the model and the current image region.

Support Vector Machine (SVM) [11] is a broad classification method which gives a set of positive and negative training values. For SVM, the positive samples contain tracked image object, and the negative samples consist of all remaining things that are not tracked. It can handle single image, partial occlusion of object but necessity of a physical in initialization and necessity of training.

Contour tracking methods [9], iteratively progress a primary contour in the previous frame to its new position in the current frame. This contour progress requires that certain amount of the object in the current frame overlay with the object region in the previous frame. Contour Tracking can be performed using two different approaches. The first approach uses state space models to model the contour shape and motion. The second approach directly evolves the contour by minimizing the contour energy using direct minimization techniques such as gradient descent. The most significant advantage of silhouettes tracking is their flexibility to handle a large variety of object shapes.

In addition to choose the desirable features, it is critical to appropriately position the objects when they overlap. The histogram of oriented gradients (HOG) and Kanade-Lucas-Tomasi (KLT) techniques detect people and track their motions between detections so as to obtain the final trajectories.

There are other techniques functioning as tracking by detection. In tracking-learning-detection (TLD), there is a combination of tracking, learning and detection, where the tracker focuses on the object in consecutive frames, the detector focuses on all observed appearances, and learning estimates the errors found in the detector, which is updated to avoid future errors.

Patch matching algorithm

The pedestrians first detected automatically in this algorithm. After the position of each object is specified in the image (red bounding box), each subject is converted into patches (six patches for each pedestrian), which are separately matched in the frames. Then, ideal accuracy can be achieved and the overlapping can be handled by adding the predicted displacement vectors for detecting the pedestrians through weighted vector median filter (WVMF).

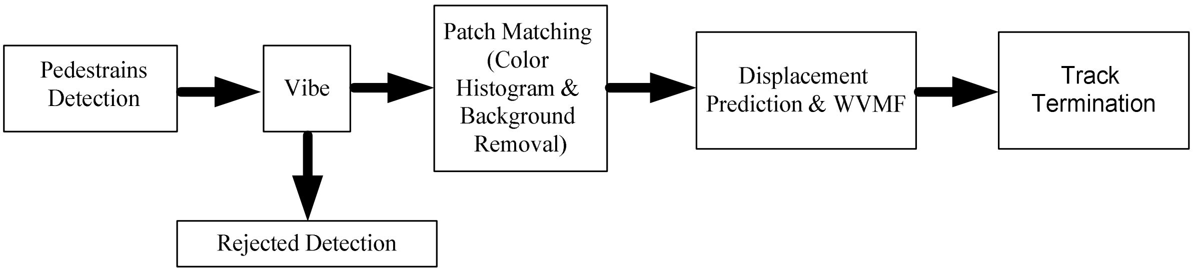
Pedestrains Detection
Vibe
Patch Matching (Color Histogram & Background Removal)
Displacement Prediction & WVMF
Track Termination
Rejected Detection

Fig. 1: Our proposed algorithm flowchart

Pedestrian detection

Numerous algorithms have been proposed for pedestrian detection. We considered Dollár's algorithm for pedestrian detection in this paper, since it provides an ideal balance between accuracy and speed.

When this algorithm is applied to a frame, bounding boxes display the position of each object as output. The algorithm's accuracy in detection should be determined by assessing the bounding boxes. In this procedure, we applied ViBE algorithm is implemented on the output of the previous stage. The detection accuracy of the Dollár's algorithm can be assessed by removing the background images and obtaining foreground blobs in each frame.

The procedure of our work are as follows:

If the value of background pixels was lower than a threshold in the bounding box detected in the previous stage (in our experiments we used 20%), the detection was rejected. Otherwise, the box containing the detected pedestrian was considered true, and then was sent to the next stage, i.e. creating patches.

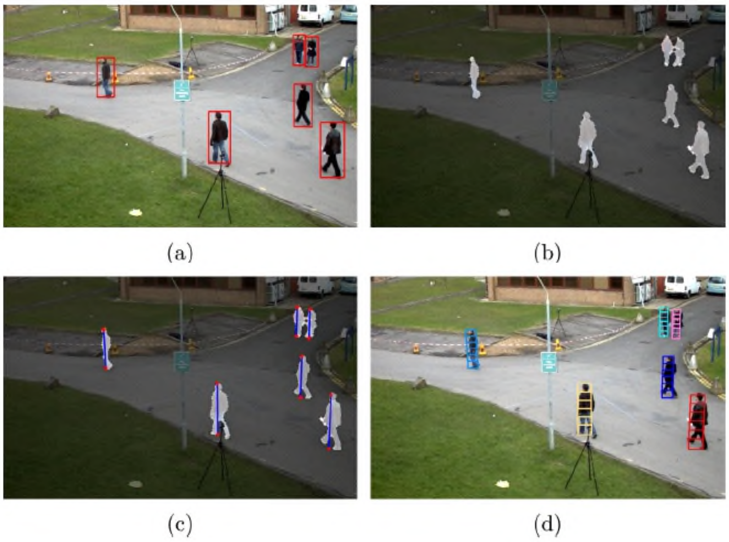

Fig.2: a) Detection with Dollar algorithm, b & c) applying the Vibe algorithm and d) creating and matching patches

Matching the patches

Each pedestrian is characterized by a set of patches. In order to assess the similarity between two patches, there are several techniques involving a variety of features. This paper employed the color histogram. In order to better handle the lighting variations in the image, two channels of *ha* and h_b were considered for each image in the color space of *CIELAB*. In fact, $h^b{}_{\mathrm{m}}$ and $h^a{}_m$ histograms were assigned to the model, while $h^a{}_i$ and $h^b{}_i$ histograms were assigned to the candidate region. The Bhattacharyya distance was calculated according to the following equation [7]:

$$b_i = \frac{1}{2}\left\{\sqrt{1 - BC(h_m^a, h_i^a)} + \sqrt{1 - BC(h_m^b, h_i^b)}\right\} \tag{1}$$

$$BC(h_1, h_2) = \sum_{j=1}^{N_b} \sqrt{h_1(j) h_2(j)} \tag{2}$$

A model patch was matched against a candidate patch so that the equation above will attain minimum Bhattacharyya distance. A simple way to identify the candidate patches is to consider a fixed region around the previous position of the patch, where the patch can be searched and matched. A variable search space, however, can be considered based on the camera parameters as well as the maximum displacement of each object.

Assuming the maximum displacement of each pedestrian is represented by s_{max} and the number of frames per second is f_r, then the maximum displacement of each pedestrian (s_{max}) can be achieved through the following equation:

$$r = \frac{S_{max}}{F_r} \tag{3}$$

Thus, the search space can be considered a *2r* by *2r* area, where:

$$\hat{r} = (1 + \alpha_r)r \qquad (4)$$

$s_{max,}$ r, $r^{\wedge}$, α_r are maximum displacement of each pedestrian, radius of search area and Smoothing factor.

In this paper, the maximum displacement of each pedestrian was assumed to be 1.5 meters per second and α was 0.5.

Displacement prediction and weighted vector median filter (WVMF)

In the previous stage, the *2r* by *2r* search space was achieved. Instead of inserting the centers of the search window on the previous position of the pedestrian, the displacement prediction vector was calculated based on the history of displacement. This displacement vector was added to the previous position of the pedestrian so as to find the new position for the search space.

With the displacement vectors *D(t)*, the displacement vector of the pedestrian can be predicted at *t+1*. This paper involved the double exponential smoothing technique [8], whose efficiency is equal to the Kalman Filter.

Now assume a pedestrian at position *x(t)* detected in the previous stages in frame *t*. For this pedestrian, there are as many patches (N_p) as the displacement vectors (D_i), where each vector belongs to a specific patch. There is a prediction displacement vector calculated in the previous stage $D_{np}+1 = D_p\ (t+1)$.

It should be noted that all displacement vectors should be similar. The WVMF is a tool to calculate the weighted means of vectors and eliminate the effect of values deviated from the mean. Hence, the WVMF was used to combine displacement vectors concerning the patches as well as the displacement vectors predicted a follows:

$$s_j = s(D_j) = \sum_{i=1}^{N} \| D_j - D_i \| \qquad j = 1, \dots, N \qquad (5)$$

$$D_f = \frac{\sum_{i=1}^{N} w_i D_i}{\sum_{i=1}^{N} w_i} \qquad (6)$$

Finally, the pedestrian's position is obtained at *t+1* as follows:

$$x(t+1) = x(t) + D_f \qquad (7)$$

Background removal

In this paper, to obtain a better result in overlapping with obstacles, the background removal technique was combined with the Bhattacharyya's matching technique which is applied through color histogram. In fact, there is another measure in addition to the color histogram during patch matching, which is based on the values of the background pixels found in the image. Among the available candidates, a patch is selected to entail the color histogram matching as well as the maximum value of background pixel.

Track termination

In multi-object tracking, one crucial discussion revolves around how to terminate the tracking process. In certain cases, there are criteria associated with track termination. For instance, the tracking process terminates when the pedestrian walks out of the camera's screen.

In this paper, we considered a criterion for the quality of tracking. If this criterion is assessed to be weak for a long period (T_n frames) for a particular target, then the tracking process is said to be terminated. This criterion works equally well when the pedestrian walks out of the camera's screen.

Experiments

This paper applied Pet2009, and Sequences2l1 database to test the algorithms. It obtained all ground truths and performed the experiments in a certain number of frames. The study by Gustavo Fuhr [2] offered good results, and compared to which the method adopted in this paper yielded better results than Fragtrack and Tld methods.

According to the obtained results, our technique outperformed that of Gustavo Furth.

The MOT was used as a metric to assess the results obtained by the experiments. The MOT can determine the detection performance of multiple targets. It is characterized by two parameters including MOTA and MOTP defined as follows:

$$MOTA = 1 - \frac{\Sigma_f (m_t + fp_t + mme_t)}{\Sigma_t g_t} \tag{8}$$

M_t, f_{pt} and mme_t and G_t, are the number of misses, false positives, mismatches and the number of objects in the frame at t, respectively.

This paper was implemented on a MATLAB code, the results of which indicates that the Furth` method based on MOTA accuracy is 45% while our proposed method have an accuracy of 54% which is much better.

Conclusion

In this paper, a new method was proposed for pedestrian detection. Attempts were made to integrate the color histogram and background removal for patch matching. Our technique not only demonstrated good tracking performance, but also handled the overlapping better than previous techniques. Our new technique indicates that we were able to track pedestrians more accuracy (9%).

References

[1] P. Dollár, S. Belongie and P. Perona, *The fastest pedestrian detector in the west,* British Machine Vision Conference, pp. 68.1–68.11, 2010.

[2] G. Führ and C.R. Jung , *Combining patch matching and detection for robust pedestrian tracking in monocular calibrated cameras,*Pattern Recognition Letters 39 (2014) 11–20, 2014.

[3] A. Yilmaz,O. Javed and M. shah.*Object Tracking: A survey,Acm computer* surveys 38 (4),1-45,2006.

[4] D. Ramanan., D. Forsyth and A. Zisserman. *Tracking people by learning their appearance*. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1), 65–81, 2007.

[5] B. Benfold and I. Reid, *Stable multi-target tracking in real-time surveillance video*, IEEE Conference on Computer Vision and, Pattern Recognition, pp. 3457–3464, 2011.

[6] G. Führ and C.R. Jung, Robust *patch-based pedestrian tracking using monocular calibrated cameras*, Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 166–173. (SIBGRAPI), 2012, pp. 166–173,2012.

[7] A. Adam, E. Rivlin and I. Shimshoni , *Robust fragments-based tracking using the integral histogram*, IEEE Conference on Computer Vision and Pattern Recognition,vol. 1, pp. 798–805, 2006.

[8] W. Choi and S. Savarese, Multiple *target tracking in world coordinate with single,minimally calibrated camera*, Proceedings of the 11th European conference on Computer vision: Part IV, pp. 553–567, 2010.

[9] J. Joshan Athanesious and p. Suresh, *Implementation and Comparison of Kernel and Silhouette Based Object Tracking*, International Journal of Advanced Research in Computer Engineering & Technology, March 2013.

[10] S. Saravanakumar,A. Vadivel and S. Ahmed, C.G., *Multiple human object tracking using background subtraction and shadow removal techniques*, Signal and Image Processing (ICSIP), 2010 International Conference on , vol, no, Dec. 2010.

[11] R. Mishra, Mahesh K. C houhan and Dr. Dhiiraj Nitna, *Multiple Object Tracking by Kernel Base d Centroid Method for Improve Localization*, International Journal of Advanced Research in Computer Science and Engineering , July-2012.