A NEW STRATEGY BASED ON SPATIOGRAM SIMILARITY ASSOCIATION FOR MULTI-PEDESTRIAN TRACKING

Nabila MANSOURI[1 5], Yousra BEN JEMAA[2], Cina MOTAMED [3], Antonio PINTI [4] and Eric WATELAIN[1 6]
[1] University of Lille North of France, UVHC, LAMIH laboratory
e-mail: nabila.elmansouri@gmail.com, eric.watelain@univ-valenciennes.fr
[2] University of Sfax-Tunisie, U2S laboratory
e-mail: Yousra.BenJemaa@enis.rnu.tn
[3] University of Lille North of France, ULCO, LISIC laboratory
e-mail: motamed@lisic.univ-littoral.fr
[4] University of Orleans -France, I3MTO laboratory
e-mail: Antonio.Pinti@univ-valenciennes.fr
[5] University of Sfax-Tunisie, ReDCAD laboratory
[6] University of south Toulon-Var, HandiBio laboratory

***Abstract*— Multiple pedestrian tracking is an active and challenging research topic that many different approaches have addressed it. Since Human stick changes over time and person usually moving in random way, the identity association remains a hard task. In this paper we propose a new method for coupling detections over all the frames of the video sequence in order to make a performant tracker. The Pedestrian detection is ensured using the Dalal and Triggs's human detector. In order to overcome the problem of missing detections caused by occlusion, we propose to use an interpolation process based on average speed. Then, all the previously detections are organized over a tree structure, where each frame represents a tree level. All detections in level 'i' are linked to the next level by an arc characterized by a cost representing the spatiogram similarity between these 2 detections. After trajectories refinement is done based on Euclidean distance to palliate the false detection association.**

An experimental study conducted in 2 datasets (CAVIAR and our database) proves the good performance of our proposed method in term of tracker precision and tracker accuracy.

***Keywords*— Multi-pedestrian tracking, Spatiogram, Coupling detection, Tree structure, Interpolation, Human detector.**

I. INTRODUCTION

Pedestrian tracking is an active research area that can be applied in different fields such as in video surveillance, people behavior analysis, road crossing security. Tracking pedestrian remains a challenging problem in computer vision works given the highly articulated human body. It consists of 2 main stages: 1) pedestrian detection and 2) association of the detected pedestrian windows [1].

1) Pedestrian detection consists on differentiating people from others objects in the image. In this context several works have been proposed in the literature. They can be divided into 3 categories: i) Appearance based approaches, ii) Motion based approaches, iii) Shape based approaches. Works done in [2],[3] detect pedestrian by finding the occurrence of skin color [2] or texture [3]. Although they are not complex and intuitive, these approaches are very sensitive to occlusion, shadow and lighting variations. Whereas, in [4], [5] authors use a global or local periodicity of pedestrian motion to discriminate people. Since pedestrian motion is very sensitive to point of view variations and needs a long video sequence in order to detect the periodicity of the displacement, people detection by motion gives unsatisfactory results. The shape approaches are based on a morphological model such as, cylindrical, elliptical, stick models. A probability of similarity is calculated between the predefined model (learning model) and the query one [6]. The major limit of this method is the point of view variation between the learning model and the query one that gives a false response. Recently Dalal and Triggs [7] propose a more satisfactory technique to people detection by using the Histogram of Oriented Gradient (HOG) descriptor to represent the target object and a linear Support Vector Machine (SVM) as a classifier. Results given by this approach are insensitive towards color and lighting. But they are sensitive to occlusion, so a pedestrian can be dismissed in some frames.
2) Pedestrian tracking consists on estimating the trajectory of the target previously detected. The motion model like a kalman filter [8] and particle filter [9] estimates the current state based on the past state recursively. In fact particles filter overcomes the linear motion limitation of kalman filter. But in multi-objects tracking the number of required particles grows considerably and become hardly to do with in practical situations. Also motion models can't handle the trajectory overlap and the identity switch because they haven't any information to discriminate between peoples. For this reason some works propose to merge between the motion and appearance models. The more popular appearance model used in

the literature is the color histogram. In fact the approach proposed in [10] use kalman filter to predict pedestrian position and 4D color histogram, presented in 'Lab' color space, to ensure identity coupling. Since histogram represents only color distribution, two pedestrians wearing a similar color clothes have an important likelihood to be confused. In [11], authors propose to combine color histograms defined in 'HSV' space and a particle filter. Given that 'HSV' color space is significantly power compared to others ones the tracker performances are ameliorates. The downside of this method is the important computational cost of the particle filter. The solution adopted by recent works [10], [12], [13] is to associate the previews detections. In fact in [12] authors provide a continuous trajectory model based on cubic B-splines. For this purpose, they need to know the start and the end positions. Also to associate inter-frames detections, this method uses only the distance between consecutive positions that remains a considerable shortcoming especially with the trajectories' overlap and (the direction changes). However in [13], the authors base their method on the foundations of Bayesian estimation theory. They have only the visibility on the single current object and not on other objects.

In this paper, we search to overcome some challenging problems of multi-pedestrian tracking such as: identity switch, and re-identification after occlusion and pedestrians' trajectories refinement. Our proposed solution is based on coupling detections and data association. Indeed, our main contribution is to re-address the coupling detection process in order to make a multi-pedestrian tracker. So we introduce the tree structure representation of frame by frame detections and exploit the spatiogram descriptor, powerful to ensure a good identity association.

This paper is organized as follows. Section 2 detailed our proposed method based on tracking pedestrian by coupling the detections that maximize the spatiogram similarity, passing by short introduction of Dalel and Trigges' person detection method. Section 3 explains the test protocol and analyzes the performance of our tree research method by presenting some results. Finally, conclusions and future works are discussed in Section 4.

II. COUPLING DETECTIONS BY MAXIMIZING THE SPATIOGRAM SIMILARITY

A. People detection

Our pedestrian tracking process uses Dalal and Triggs's [7] method to detect humans in each video frame. This approach consists of several steps. Firstly, color normalization is done in order to reduce the influence of shadows and illumination changes. The intensity gradient at each pixel is then computed for each color channel. The next step is to acquire a histogram of the gradient orientations at each spatial region or (cell). A range of adjacent cells are grouped into larger blocks so as to normalize edge contrast and illumination. The last stage is to concatenate all of the normalized block vectors over each detection window to form a final window descriptor. Then a linear SVM classifier is applied to successfully classify detection windows into 'non-pedestrian' and 'pedestrian'.

Dalal and Trigss's approach can't detect all pedestrian in each frame because of partial or total occlusion. Indeed we present in the next our contribution in people detection and our coupling process.

B. Coupling detections

1) Spatiogram similarity calculation: Since spatiogram descriptor has good performances according to a comparative study between color, texture and shape descriptors for multi-pedestrians identification done in [14], we assume that the use of the spatiogram as an appearance model to check the similarity between targets can improve tracking pedestrian performances. The spatiogram represents an extension of color histogram incorporating spatial information. In this work, we use the Cornaire [15] spatiogram known for its great representational power. First, a quantization phase is performed on the original image to reduce color variation. Second, a color histogram (denoted η_b) was computed, that corresponds to assigning a number of pixels to each color intensity. Then the algorithm calculates for each intensity existing in the quantized image its spatial distribution represented by the mean μ and the covariance σ of the spatial position of all pixels that fall into each histogram bin as explained in equation (1)

$$\eta_b = \sum_{i=1}^{N} \delta_{ib} \tag{1}$$

$$\mu b = \frac{1}{\sum_{j=1}^{N} \delta_{jb}} \sum_{i=1}^{N} X_i^T \delta_{ib}$$

$$\delta_{ib} = \frac{1}{\sum_{j=1}^{N} \delta_{jb}} \sum_{i=1}^{N} (X_i - \mu_b)(X_i - \mu_b)^T \delta_{ib}$$

Where $\delta_{ib} = 1$ if the pixel i falls into bin b and 0 otherwise; C is a normalizing constant ensuring that the sum of all bins is equal to 1 and $X_i = (x_i, y_i)^T$ is the spatial position of each pixel.
To compare two persons spatiograms, $S(\eta, \mu, \sigma)$ and $S^{'}(\eta^{'}, \mu^{'}, \sigma^{'})$ each containing B bins, the following similarity measure presented in equation (2) is used.

$$similarty = \sum_{b=1}^{B} \psi_b \sqrt{\eta_b \eta_b^{'}} \tag{2}$$

where

$$\psi_b = \eta exp(\frac{-1}{2}(\mu_b - \mu_b^{'})\sigma_b^{'-1}(\mu_b - \mu_b^{'}))$$

$$\sigma_b^{'-1} = \sigma_b^{-1} + (\sigma^{'})_b^{-1}$$

with σ_b represents the spatial similarity measure and η is the Gaussian normalization term.

2) Proposed process for coupling detection: People detection is done frame by frame along the video sequence. Each detection in each frame represents a level in the tree. Every detection in frame $'i'$ is initially linked to all the detections in frame $'i+1'$ by an oriented arc. The arc cost represents the spatiograms similarity between these 2 linked pedestrians (detections). Spatiograms similarity is calculated as shown previously according to equation (2). The pedestrian trajectory is the longest path that connects similar detections. Association of the correct detection to the correct trajectory must ensure the constraint that, in every level the same detection can't be associated to more than one trajectory in the same time.

Given that a frame by frame HOG detection can lost some pedestrians in some frame because of occlusion, our main contribution in people detection in video sequence is to ensure a continuous detection and overcome occlusion problem.
In fact, we seek the lost pedestrian position by interpolation process based on average speed. Based on study made in [16], walking speed ranging is generally between 3.0 to 4.95 feet-per-second. So a pedestrian make approximately an average displacement of $0.05m$ every video frame. Consequently, we model interpolation process as a translation, with vector $\overrightarrow{u}$, of the last detected pedestrian windows as shown in figure 1. In fact Bounding box including pedestrian is characterized by a quadruple (x, y, w, h) where $'(x, y)'$ is the begin point and $'w'$ and $'h'$ are respectively the width and height of the bounding box. The translation consists of translating the point $'P(x, y)'$ to $'P_i(x_i, y_i)'$ in the same direction and using a vector $\overrightarrow{u}$ equal to the average distance calculated referred to the average speed by keeping the same dimensions $(w$ and $h)$.

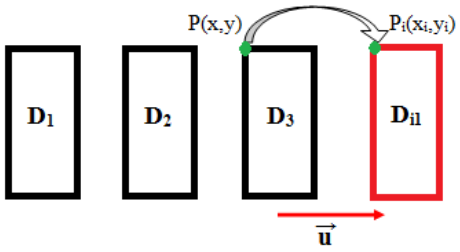

Fig. 1. Finding lost detection by interpolation.

Since the false detection will have neither previous nor next similar one.
All the coupling detection process is shown in figure 2.

III. PERFORMANCE ANALYSIS

In this section, we present the used datasets and the test protocol. The performance of the proposed system is then presented and analyzed.

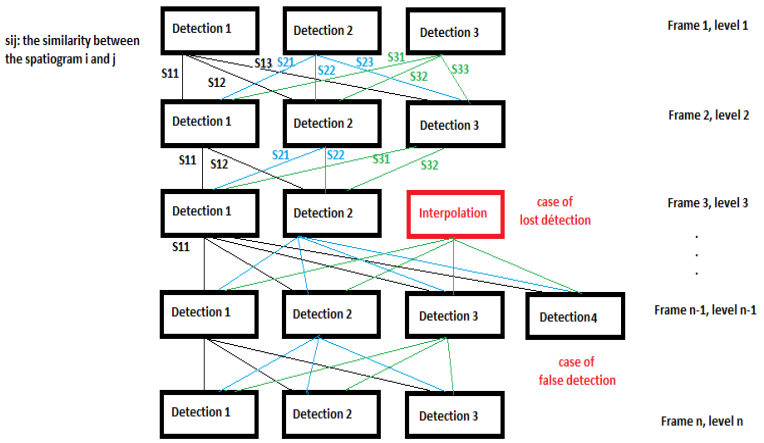

Fig. 2. The proposed pedestrian tracking by coupling detections process.

A. Used datasets

To evaluate our system performances, we conduct an experimental study on two datasets. The 'CAVIAR' video database [17] having 25 frames/second and 384×288 pixels per frame. It integrates 3 peoples walking together along the corridor and sometimes crossing with others with respect to the camera field of view.

The second dataset is ours. We have filmed 2 scenarios. The first contains 2 pedestrians crossing a street on a crosswalk in the same direction and the second contains 3 pedestrians with 2 in the same direction and the third is in opposite direction. These video sequences have 25 frames/second and 720×576 pixels per frame. Our database contains voluntarily difficult occlusion situations where people were occulted by vehicles or by other people. Also, it is captured in outdoor context with important lighting changes.

B. Test protocol and results

The test protocol aims to quantify the right coupling detection during the pedestrian displacement. For that, we will introduce some detection coupling in different time of the video sequence. In fact, each pedestrian will be denoted by a unique number $(1, 2, 3, ...)$ during the sequence. The perfect result is found when the pedestrian's associated number remains the same during all the tracking phase. Figure 3 (a) shows the tracking results on CAVIAR database whereas figure 3 (b,c) illustrates the results on our database.

Since the spatiogram performance in person's identification isn't 100% so some false association will happened. In fact in figure 4, blue (and red)points represent respectively the trajectories of the first pedestrian (denoted $p1$) and the second pedestrian (denoted $p2$). We remark some blue points in the pedestrian 2 's trajectory, this is caused by some confused identity association. In order to overcome this shortcoming, we conduct a verification process based on natural constraint of human displacement. In fact, a pedestrian walking in the normal condition, can't suddenly change his spatial position.

So the detection point location will be associated to the nearest trajectory if the Euclidean distance between each successive detections associated with the same trajectory bypass the threshold distance as shown in figure 4. The threshold distance is defined as the mean distance between all of the trajectory points. This trajectory filtering gives satisfactory results and corrects the association mistakes.
A quantitative evaluation presented above illustrates also the good performance of our trajectories refinement especially in the 'MOTP' value before and after fitting.

C. Quantitative experimental study

An ideal multiple object trackers, should at all points in time, find the correct number of objects present in the video and estimate the position of each object precisely. It should also keep consistent track of each object over time: Each object should be assigned a unique track ID which stays constant throughout the sequence [18]. So for a quantitative evaluation we use a widely used metrics proposed by [18] such as:

1) The Multiple Object Tracking Precision (MOTP): indicates the coupling precision. MOTP is calculated as explained in equation (3)

$$MOTP = \frac{\sum_{i,t} d_{it}}{\sum_t c_t} \tag{3}$$

Where $'d'_{it}$ is the distance between ground-truth trajectories and automatically generated trajectories' positions for each frame $'t'$ and $'ct'$ is the total number of matches made in this same frame $'t'$.
2) Multi-Object Tracking Accuracy (MOTA): is the average error. It includes the false positives (f_{pt}), missed targets (m_t) and identity mismatches (mme_t). MOTA is calculated using equation (4).

$$MOTA = 1 - \frac{\sum_{i,t}(f_{pt}, m_t, mme_t)}{\sum_t g_t} \tag{4}$$

We add also other metric to evaluate the performance of detection process with interpolation technique. The Multiple Object Detection Accuracy (MODA) that checks for missed targets and false positives in each frame t as explained in equation (5).

$$MODA = 1 - \frac{\sum_{i,t}(f_{pt}, m_t)}{\sum_t g_t} \tag{5}$$

All results are presented in table 1.

Table 1. Quantitative results of the proposed tracking system in term of MOTP, MODA and MOTA.

	MOTP		MODA	MOTA
	Before	After		
CAVIAR			73.74%	55.33%
Our dataset (scenario 1)	29.89 cm	10cm	91.67%	86.41%
Our dataset (scenario 2)	31.4 cm	14.34 cm	92.06%	85.56%

According to table 1, we can conclude:

1) Our Interpolation processes based on an average velocity parameter, bypasses the lost detection problem. This was proven by the important MODA's values. In fact we got a 73.74%, 91.67% and 92.06% detection precision respectively for CAVIAR and our dataset's scenarios.
2) All detections are organized in a tree structure and each frame represents a level in this tree. The detection at level $'i1'$ is well associated with the detection in level $'i+1'$ by maximizing the similarity value. In association process each detection at level $'i'$ should be associated with one detection at level $'i+1'$. So a new coming person can be simply detected and integrated in the tracking system given that isn't similar to any other previous detection but is similar to next one. However, in the case of false detection, it will be removed from the next level because it will not have a similar one. Association performance is illustrated by the MOTA's important values. In fact in a sequence of 114 frames (senario1 of our dataset) we have a 86.41% as MOTA result, so the number of identity mismatches is only 6 per 114. This proves the robustness of our coupling process.
3) Another contribution is proposed in this paper concerning the trajectory refinement. In fact, we ameliorate the trajectory estimation and correct some false associated detections to some trajectories. Our solution is based on the calculus of the Euclidean distance between pedestrian's position in level 'i' and her position in level 'i+1'. If this distance, exceeds the threshold distance the detection will be associated to the nearest trajectory. A trajectories refinement step ameliorates considerably our tracker precision. In fact the MOTP values are $29.89cm$, $10cm$ respectively before and after trajectories refinement. A decrease of $20cm$ of the average distance separating the ground truth and estimated trajectories proves the important and the good impact factor of this step.

IV. CONCLUSION AND FUTUR WORKS

In this paper, we have introduced a new approach for multi-pedestrians tracking based on coupling detections. We maintain the good person identity during video sequences by increasing the spatiogram similarity between associated detections.
Thanks to our approach, the problem of lost detections caused by occlusion was solved. To ensure this objective, we use an interpolation process based on pedestrian average speed. It is proven that interpolation can proportionally overcomes occlusion impact and gives a good MODA values. Also the detection organization inspired from the tree theory represents a good strategy to perform the global association of pedestrian's identity.
An experimental study conducted in 2 datasets (CAVIAR and our database) illustrates the good performances obtained by the proposed method which can reduce the identity confusion

problem in multi-pedestrian tracking and gives enhanced fitted trajectories.

Future work will focus on developing a robust behavioral analysis module in order to integrate it in the proposed system which use only appearance descriptor (spatiogram). The goal is to improve the global people tracking performances especially over the pedestrian identities estimation in the case of visually similar pedestrians.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah. Object Tracking: A Survey *ACM Computing Surveys*, 38(4), 2006.

[2] C.Dai, Y.Zheng and X. Li. Layered representation for pedestrian detection and tracking in infrared imagery. In *Second Int. Conference on Computer Vision and Pattern Recognition (CVPR'02)*, proceedings, San Diego-CA-USA, June 2005.

[3] S. Munder, C. Schnorr and D. Gavrila. Pedestrian detection and tracking using a mixture of view-based shape-texture models. *IEEE Trans. on Intelligent Transportation Systems*, 9 (2): 333-3439, 2008.

[4] H. Sidenbladh. Detecting human motion with support vector machines. In *seventeenth Int. Conference on Pattern Recognition (ICPR'17)*, proceedings, Cambridge-England, August 2004.

[5] A. Shashua, Y. Gdalyahu and G. Hayun. Pedestrian detection for driving assistance systems : Single-frame classification and system level performance. In *fourth IEEE. Intelligent Vehicles Symposium (IV'04)*, proceedings, Parma-Italy, June 2004.

[6] H.L. Eng, J. Wang,A.H. Kam and W.Y. Yau. A bayesian framework for robust human detection and occlusion handling using human shape model. In *seventeenth Int. Conference on Pattern Recognition (ICPR'17)*, proceedings, Cambridge, UK, August 2004.

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *second Int. Conference on Computer Vision and Pattern Recognition(CVPR'02)*, proceedings, San Diego-CA-USA, June 2005.

[8] M. Bertozzi, A. Broggi, A. Fascioli, A. Tibaldi, R. Chapuis, F. Chausse. Pedestrian Localization and Tracking System with Kalman Filtering. In *fourth IEEE. Inteligenet Vehicles Symposum (IV'04))*, proceeding, Parma-Italy, June 2004.

[9] T. Tung and T. Matsuyama. Human Motion Tracking using a Color-Based Particle Filter Driven by Optical Flow. In *first Int. Workshop on Machine Learning for Vision-based Motion Analysis (MLvMA'01)*, proceeding, Marseille-France, October 2008.

[10] Z. Jiang, D.Q.Huynh, W. Moran, S. Challa, N. Spadaccini. Multiple Pedestrian Tracking using Colour and Motion Models. In *first Int. Conference on Digital Image Computing: Techniques and Applications (DICTA'01)*, proceedings, Sydney-Australia, December 2010.

[11] P. PÃ©rez, C. Hue, J. Vermaak, and M. Gangnet. Colorbased probabilistic tracking. In *seventh European. Conference on Computer Vision (ECCV'07)*, proceedings, Copenhagen-Denmark , May 2002.

[12] A. Andriyenko, K.Schindler, S. Roth. Discrete-Continous Optimization for multi-Target Tracking. In *ninth Int. Conference on Computer Vision and Pattern Recognition (CVPR'09)*, proceedings, RI-USA, June (2012).

[13] J. Kang, I. Cohen, and G. Medioni. Object Reacquisition using Invariant Appearance Model. In *Int. Conference on Pattern Recognition(ICPR'17)*, proceedings, Cambridge, UK, August 2004.

[14] A. derbel, Y. Ben Jemaa, R. Canals, B. Emile, S. Treillet, A. Ben hamadou. Comparative study between Color Texture and Shape descriptors for multi-camera pedestrians identification. In *twenty-third. Symposium on Signal and Image Processing (GRETSI'23)*, proceedings, Bordeaux-France, September 2011.

[15] C. O Conaire, N. E. O'Connor and A. F. Smeaton. An improved spatiogram similarity measure for robust object localization. In *thirty-first Int. Conference on Acoustics, Speech, and Signal Processing (ICASP'31)*, proceedings, Honolulu-USA, April 2007.

[16] K. Ismail, T. Sayed, N. Saunier. Automated Collection of Pedestrian Data Using Computer Vision Techniques. In *Video Surveillance and Transportation Imaging Applications SPIE*, (8664) San Francisco, California-USA, February-2014.

[17] *http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.*

[18] K. Bernardin, A. Elbs, R. Stiefelhagen. Multiple Object Tracking Performance Metric and Evaluation in a Smart Room Environment. In *sixth Int. Workshop on Visual Surveillance (VS'06)*, proceeding, Graz-Austria, May 2006

Fig. 3. Tracking results: (a) CAVIAR database, (b) Our dataset: scenario 1, (c) Our dataset: scenario 2.

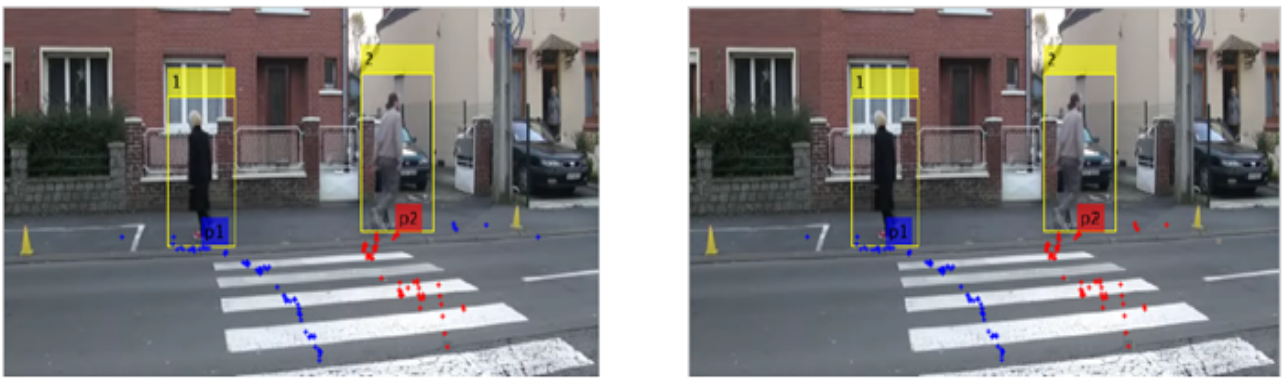

Fig. 4. Trajectories refinement.