Human Detection and Tracking under Complex Activities

B. Cancela, M. Ortega and M. G. Penedo

VARPA Group, University of A Coruña, Campus de Elviña s/n, A Coruña, Spain

Keywords: Background Subtraction, Cascade Classifier, Histogram of Oriented Gradients, Particle Filter, Collision Detection.

Abstract: Multiple-target tracking is a challenging question when dealing with complex activities. Situations like partial occlusions in grouping events or sudden target orientation changes introduce complexity in the detection which is difficult to solve. In particular, when dealing with human beings, often the head is the only visible part. Techniques based in upper body achieve good results in general, but fail to provide a good tracking accuracy in the kind of situations mentioned before. We present a new methodology for provide a full tracking system under complex activities. A combination of three different techniques is used to overcome the problems mentioned before. Experimental results in sport sequences show both the speed and performance of this technique.

1 INTRODUCTION

In complex scenes, human detection and tracking is a challenging question far to be solved. Since the use of full-body human detections (Desai et al., 2009) is questionable due to possible occlusions, novel systems try to focus into the head region.

For instance, Rodriguez et al. (Rodriguez et al., 2011) introduce a density function to detect every head into a crowded scene. However, tracking system based in local points does not take into account complex activities, like quick human spins. On the other hand, Li et al. (Li et al., 2009) use the head-shoulder omega shape feature to perform a human detection technique. This system cannot track people when exist human interaction events, like people grouping, or when sudden orientation changes occur. It also obtain poor results in presence of noise within the image. Many other different techniques can be seen in (Zhan et al., 2008).

In this paper we present a new methodology for human detection and tracking people under complex activities. Based on the work of Li et al., we introduce additional information about the motion within the scene to reduce the region to process. In our work, a combination of a Viola-Jones type classifier with a HOG feature based SVM is used to detect every person in the scene, while a particle filter system, supported by the human detection system and a adaptive filter technique to predict the target position, performs the tracking methodology, achieving good results in both areas.

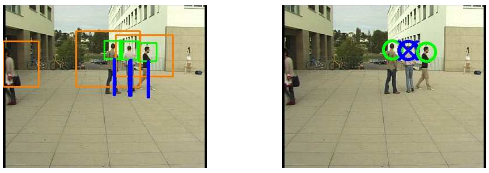
This paper is organized as follows. Section 2 explains the detection system; section 3 describes the tracking system; finally, section 4 shows some experimental results and section 5 offers some conclusions.

2 DETECTION SYSTEM

To detect every human in the scene, we perform a method based in the omega-shape feature (Li et al., 2009). We combine a Viola-Jones type classifier with a Histogram of Gradients (HOG) feature based SVM. The main reason to introduce two different classifiers in related with the processing speed. The SVM is a better classifier, but the time needed to compute the HOG feature and to classify it is high, making it unable to be used in real-time scenarios. On the other hand, the Viola-Jones type classifier speed is high, but lacks in the classification accuracy, since it introduces several false positive detections.

First, the Viola-Jones type classifier is used to detect every possible human being in the scene. This results in a pool of different patches, which are going to be confirmed as human beings using the HOG based feature SVM. However, using the Viola-Jones along the whole scene also involves a large amount of time. Our detection system is divided into three different steps: first, a background subtraction technique is used to detect every moving pixel in the scene;

Cancela B., Ortega M. and Penedo M..
Human Detection and Tracking under Complex Activities.
DOI: 10.5220/0004201503700374
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 370-374
ISBN: 978-989-8565-48-8



(a) (b)

Figure 1: Detection system example. Once the background subtraction technique is applied, we select a poo of bounding boxes containing the blobs in the scene (a). Viola-Jones type classifier, in combination with a torso estimation, are used to select the possible detections (a), which are finally evaluated using the HOG feature based SVM (b).

then, both the Viola-Jones type classifier and the HOG based feature SVM are used to detect every moving person. We decided not to use optic flow to detect every moving pixel in the scene because of the processing speed. Since information about the module or the orientation of the movement are not going to be used in this methodology, a background subtraction technique with online updating system is used. We use the Mixture of Gaussians (MoG) algorithm (Stauffer and Grimson, 1999) with a short window history, in order to quickly consider as background stopped targets. This is also done to make the system more robust to sudden illumination changes. After that, opening and closing morphological operators are user to reduce the image noise.

Before using the Viola-Jones type classifier, we have to reduce the processing regions. Having the background information technique, we perform a blob detection and, for each blob, a bounding box is created containing all its foreground pixels.

After we have the bounding boxes, we use the Viola-Jones type classifier under these regions. Details of how to train the classifier can be seen in (Viola and Jones, 2001). Finally, a pool of different patches are obtained (green rectangles in Fig. 1-(a)). Before we apply the HOG feature based SVM, we introduce a restriction based in the human movement knowledge. A head movement detected by the background subtraction technique is always followed by a torso movement. So, using a small rectangle at the bottom of the patch we can check the foreground pixels to see if a movement occurs down the head. If no movement is found, the patch is discarded. The height of this rectangle is related with the patch size we are checking. Blue rectangles in Fig. 1-(a) shows an example of the rectangles we use to check the torso.

The HOG Feature Based SVM is applied to every remaining patch to confirm it is related to a person. We perform a classic HOG technique. HOG feature extraction details can be seen in (Dalal and Triggs, 2005). The computational cost to obtain this feature is high, but we can improve the performance using the integral histogram technique (Porikli, 2005). In Fig. 1-(b) we can see both accepted and rejected patches.

3 TRACKING SYSTEM

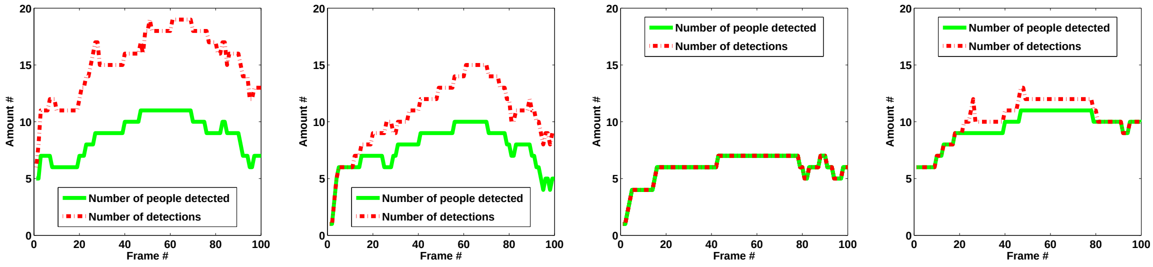
Although tracking system is mainly focused on using a particle-filter system, that method decreases its quality during long-time scenes, since an error in the estimation is carried along the frames without any possibility to correct them. Our tracking system is defined as follows: first, we use the detection system explained before to also track the target. If this method fails, we perform a particle-filter system to locate the best target state into the new frame. Finally, we introduce the new elements detected by the detection system.

After using the detection system, we have a pool of different accepted patches. We also have, at time t, a set containing all the targets tracked by the system in previous frames. Thus, we create, for every new patch, a set containing all the targets that could fit with the new detection, using the euclidean distance. If we have only one target candidate, we assign the new patch with the target contained into the set. If there is more than one candidate. a collision occurs. Then, we perform a comparison between their respective HOG features using the Bhattacharyya coefficient, which is a good method for tracking nonrigid objects (Comaniciu et al., 2000). The target which obtains a better coefficient is assigned to the new patch position.

We compute a predicted position for the targets we have stored, using a bunch of linear filters (Adalines) to estimate the target speed, because of its simplicity and its performance under noisy images (Cancela et al., 2011).

We propose to use a particle-based system, choosing the extracted local HOG features, to model its appearance. Using the linear filters to predict the velocity of each target, we can assume the new target position is described as $z^t_j = \tilde{z}^t_j + \omega$, where $\tilde{z}^t_j$ is the predicted position of the target z_j at time t and $\omega \sim N(0,\Sigma)$ is a Gaussian Noise. In our particle filter system, we add Gaussian noise to the predicted position to generate a bunch of different particles. The local HOG feature is extracted for each particle and, using the Bhattacharyya coefficient we choose the position which obtains the best value.

Once the position is located, we have to update the model appearance. Because of a bad chosen particle, the model should maintain information about previous appearances. So, having the target model $\hat{\mathbf{O}}^t$ and



(a) (b) (c) (d)

Figure 2: Detection system stats. With no background subtraction technique (Li et al., 2009), false positive rate is high (a). Including background subtraction technique we reduce the number of total and false detections (b). Including the torso estimation, the system locate every moving person in the scene avoiding wrong detections (c), while stopped persons cannot be reached. Initializing the system having disabled the restrictions under a few frames obtains the better results (d).

the HOG feature of the winning particle, $\mathbf{O}^{t+1}$, the model is updated using a learning parameter, α (set to 0.05 in our case).

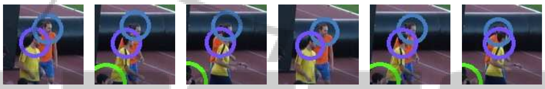
We always accept the best particle. Thus, we introduce a threshold τ_B to detect lost targets. If, along consecutive frames, the Bhattacharyya coefficient of the best particle is below to that frame, we declare the target as lost.

Once we have successfully tracked all the targets we have detected in previously frames, we introduce new targets appearing within the scene. Basically, we define as new target every patch that has no overlapping region with any of the targets tracked with the other two techniques mentioned before.

4 EXPERIMENTAL RESULTS

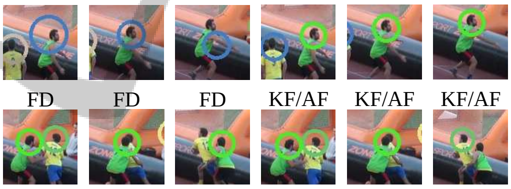
To perform our experiments we have recorded a sport sequence, consisting in a 3 × 3 basket match. More than 15000 frames were recorded in this video, having a 640 × 368 resolution running at 25 frames per second. We divide the test results in two big groups: one involving the detection performance and another focused in the tracking system.

Since our method for adding new targets is embedded in the tracking system, we use the whole system to test the detection, without taking into account any information about the identification system. To train both Viola-Jones type classifier and the RBF kernel SVM we use the generic dataset introduced in (Li et al., 2008). To validate the advantage of our method on the detection task, we compared it with two different configurations, in terms of detection quality and processing speed. First, Li et al. (Li et al., 2009) implementation is used. Also, we test out method with and without torso estimation. We evaluate the detection system during a sequence. Fig. 2 shows the results obtained.



(a) (b) (c) (d) (e) (f)

Figure 3: Tracking system example. Without taking into account information about the human detection system, the accuracy of the tracking decreases when partial occlusion events occur (a, b, c). More accurate solution can be achieved introducing that information (d, e, f).



FD FD FD KF/AF KF/AF KF/AF
KF KF KF AF AF AF

Figure 4: Examples of tracking using different predicted position techniques. Using a simple finite difference scheme (FD) lacks in presence of noise, while both kalman filter (KF) and adaptive filters (AF) can deal with it. When occlusion occurs, AFs performs better when recovering a moving object, while KFs obtain better results tracking the stopped one.

The idea of using both background subtraction information and the torso estimation reduces the sensitivity of the system. However, in video sequences, it is very important to avoid, as much as possible, the false positive rate. Although our system can reject good patches, the specificity is close to 100%. The impact due to the sensitivity decreasing is limited, since video sequences provide good chances to detect that patches into the successive frames. Table 1 shows the speed processing results using a Pentium Quad Core, running at 2.40GHz with 4 RAM GB in a Linux Operative System. In these images, we can increase the number of frames processed per second from 1 up to 4, while we improve the accuracy in the detection procedure.

According to the prediction system, three differ-

Table 1: Time consuming on a 640 × 368 image by the Viola-Jones type classifier in combination with the HOG feature based SVM, and by the tracking system, with or without taking into account the detection system information.

	Detector Type	Time per frame
Detection	No background subtraction (dense scan)	962ms
	Background subtraction	255ms
	Background subtraction + torso estimation	245ms
Tracking	Particle filter	32ms
	Particle filter + detection information	16ms

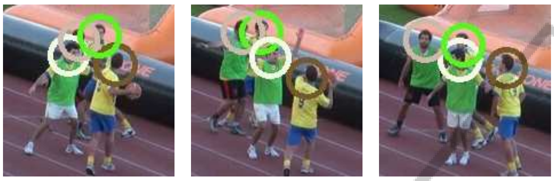


Figure 5: Partial occlusion example. The system is able to detect every person in the collision, even when sudden orientation changes occur.

ent approximations were tested: a simple finite difference scheme, a Kalman filter (Kalman, 1960) and the adaptive filters. As we can see in Fig. 4, finite difference scheme is too sensible to noise. Similar results were obtained using both Kalman and adaptive filters. Since the behavior is similar, we decide to use the adaptive filter, as it is less memory and computation demanding.

The system can successfully track multiple targets under sudden movement changes. Also can detect every person within a scene under partial occlusion events. In Fig. 5 we can see four different players under partial occlusion circumstances. Every frame during that collision, the system is able to successfully detect all the targets involving.

5 CONCLUSIONS

In this paper we present a new methodology for tracking people under uncontrolled and complex scenarios. Three different techniques were tested. We can conclude the system performs a better detection technique.

A sport sequence is used to test our algorithm, including several challenging situations, like occlusions, grouping events and sudden speed and orientation change movements. While previous approaches failed to deal with these problems, the results show our system is able to maintain a good tracking quality during the sequences, while we increase their algorithm speed. In future researches, we plan to introduce a full tracking system with total occlusion recovery.

ACKNOWLEDGEMENTS

This paper has been partly funded by the Consellería de Industria, Xunta de Galicia through grant contracts 10/CSA918054PR and 10TIC009CT.

REFERENCES

Cancela, B., Ortega, M., Penedo, M. G., and Fernández, A. (2011). Solving multiple-target tracking using adaptive filters. In *Lecture Notes in Computer Science (ICIAR 2011)*, volume 6753, pages 416 – 425.

Comaniciu, D., Ramesh, V., and Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142 –149 vol.2.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 – 893.

Desai, C., Ramanan, D., and Fowlkes, C. (2009). Discriminative models for multi-class object layout. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 229 –236.

Kalman, R. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Li, M., Zhang, Z., Huang, K., and Tan, T. (2008). Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4.

Li, M., Zhang, Z., Huang, K., and Tan, T. (2009). Rapid and robust human detection and tracking based on omega-shape features. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 2545 – 2548.

Porikli, F. (2005). Integral histogram: a fast way to extract histograms in cartesian spaces. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 829 – 836.

Rodriguez, M., Sivic, J., Laptev, I., and Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511 – I–518.

Zhan, B., Monekosso, D., Remagnino, P., Velastin, S., and Xu, L.-Q. (2008). Crowd analysis: a survey. *Machine Vision and Applications*, 19:345–357.