ONLINE SUBJECTIVE FEATURE SELECTION FOR OCCLUSION MANAGEMENT IN TRACKING APPLICATIONS

*Loris Bazzani, Marco Cristani, Manuele Bicego, and Vittorio Murino**

Dipartimento di Informatica, University of Verona - Strada Le Grazie 15, 37134, Verona, Italy

ABSTRACT

Most of the state-of-the-art tracking algorithms are prone to error when dealing with occlusions, especially when the involved moving objects are hardly discernible in appearance. In this paper, we propose a multi-object particle filtering tracking framework particularly suited to manage the occlusion problem. The presented solution consists in the introduction of a *online subjective feature selection* mechanism, which highlights and employs the most discriminant features characterizing a single object with respect to the neighbouring objects. The policy adopted fits formally in the observation step of the particle filtering process, it is effective and not computationally costly. Trials carried out on illustrative synthetic data and on recent challenging benchmark sequences report compelling performances and encourage further development of the technique.

***Index Terms*—** Tracking, feature extraction.

1. INTRODUCTION

Tracking multiple persons in video sequences is a classical computer vision open problem, far from being conclusively solved. Among the realm of the tracking strategies, an important role is played by the particle filtering approaches [1], which essentially perform a three-step on-line procedure at each instant. First, several events which describe the state of the system, *i.e.*, the displacement of the objects, are hypothesized by sampling a candidate probability distribution (pdf) over a state space (sampling step); the candidate pdf is thus approximated by a set of weighted samples, where each sample is an event whose weight mirrors the associated probability. Second, dynamics is applied to the objects (dynamical step), and, third, the hypotheses that agree at best with an observation process are awarded (observational step), so avoiding a brute-force search in the prohibitively large state space of the possible events. After that, the candidate pdf is refined for the next step. Particle filtering was born originally for single-object tracking [2], and later extended in a multi-object tracking scenarios [3]. Multi-object particle filters follow different strategies to achieve strong tracking performances avoiding huge computational burden, due primarily to the required large number of particles to consider, which is (in general) exponential in the number of objects to track [3]. Recently, an interesting yet general solution is proposed in the Hybrid Joint-Separable (HJS) filter [4], that maintains a linear relationship between number of objects and particles.

In this paper, we will employ the HJS filter framework as multi-object particle filtering platform for tracking and we extend it by improving the robustness to occlusions. Despite the proved versatility of particle filter-based approaches for tracking, occlusion management still remains a hard issue to solve, especially when the moving objects are hardly discernible in appearance. For example, in [5] an appearance-based model updating method evolves the model in a selective way, based on the type of occlusion (classified at sub-region level).

We propose an extension of the HJS filter which satisfactorily faces the occlusion problem, increasing substantially the tracking accuracy. We introduce a mechanism of *online subjective feature selection*, that selects and employs the most discriminant features characterizing a single target with respect to the neighbouring objects. Our approach takes inspiration from [6], consisting in a feature selection policy for discriminating effectively foreground (the moving objects) and background (the static scene) in a video surveillance context. The improvement here consists in devising a feature selection technique which operates among different foreground objects, whose number and appearance may change dramatically over time. Further, attention is devoted to maintain the computational effort limited while still achieving performances higher than those of the original framework. Therefore, the feature selection process is activated only in case of proximity among objects. When an occlusion occurs, the mechanism is frozen and only the previously selected features are used into the observational step. A thoroughly experimental evaluation on synthetic data and on actual challenging datasets has been carried out, which shows very promising results.

The rest of the paper is organized as follows. Section 2 describes HJS filter for tracking. The proposed feature selection method is analyzed in Section 3. Section 4 provides quantitative results on synthetic videos and qualitative results on real-world datasets. Finally, the conclusions are summarized

*This research was partially funded by the EU-Project FP7 SAMURAI (Grant No. 217899). V.Murino, M.Bicego and M.Cristani are also with IIT, Istituto Italiano di Tecnologia, Via Morego, 30, 16163 Genova, Italy.

in Section 5.

2. HJS FILTER

Particle filters offer a probabilistic framework for recursive dynamic state estimation [1]. The goal is to determine the posterior distribution $p(x_t|z_{1:t})$, where x_t is the current state, z_t is the current measurement, and $x_{1:t}$ and $z_{1:t}$ are the states and the measurements up to time t, respectively. We refer as x_t the state of a single object, and $\mathbf{x}_t = \{x_t^1, x_t^2, \ldots, x_t^K\}$ the joint state (ensemble of objects). Finally, the posterior distribution $p(x_t|z_{1:t})$ is approximated by a set of N weighted particles, *i.e.* $\{(x_t^n, w_t^n)\}_{n=1}^N$.

The HJS approach represents a theoretical grounded compromise between dealing with a strict joint process (as [3] does) and instantiating a single, independent tracking filter for each distinct object. Roughly speaking, HJS alternates a separate modeling during the sampling step and a joint formulation using a hybrid particle set in the dynamical and observational steps. The rule that permits the crossing over joint-separable treatments is based on the following approximation (see [4] for rigorous math details):

$$p(\mathbf{x}_t|z_{1:\tau}) \approx \prod_k p(x_t^k|z_{1:\tau}) \tag{1}$$

that is, the joint posterior could be approximated via product of its marginal components (k indexes the objects). This assumption enables us to sample the particles in the single state space (requiring thus a linear proportionality between the number of objects and the number of samples), and to update the weights in the joint state space. The updating exploits a joint dynamical model which builds the distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ (explaining how the system does evolve) and a joint observational model that provides estimates for the distribution $p(z_t|\mathbf{x}_t)$ (explaining how the observations are related to the state of the system). Both models take into account the interactions among objects; in particular $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ accounts for physical interactions between the targets, thus avoiding track coalescence of spatially close targets. The joint observational model $p(z_t|\mathbf{x}_t)$ quantifies the likelihood of the single measure z_t given the state $\mathbf{x}_t$, considering inter-objects occlusions.

The joint observational model builds upon the representation of the targets, that here are constrained to be human beings. The employed person representation is based on [3], which divides the human body in three parts: head, torso and legs. For the sake of clarity, we assume the body as a whole volumetric entity, described by its position in the 3D plane, with a given volume and appearance captured by HSV intensity values. The joint observational model works by evaluating a separate appearance score for each object, encoded by a distance between the histograms of the model and the hypothesis (a particle), involving also a joint reasoning captured by an *occlusion map*. The occlusion map is a 2D projection of the 3D scene which focuses on the particular object under analysis, giving insight on what are the expected visible portions of that object. This is obtained by exploiting the hybrid particles set $\{x_p\}_{p=1}^{NK}$ in an incremental visit procedure on the image plane: the hypothesis nearest to the camera is evaluated first, its presence determines an occluding cone in the scene, where the confidence of the occlusion depends on the observational likelihood achieved. Particles farther in the scene which fall in the cone of occlusion of other particles are less considered in their observational likelihood computation. The process of map building is iterated as far as the farthest particle in the scene. In formulae, the observation model is defined as

$$p(z_t|x_p) \propto \exp\left(-\frac{\mathrm{fc}_p + \mathrm{bc}_p}{2\,\sigma^2}\right) \tag{2}$$

where fc_p is the foreground term, *i.e.*, the likelihood that an object matches the model considering the visible parts, and bc_p, the background term, accounts for the occluded parts of an object. However, it is worth noting that the most similar are the appearances of the objects, the most difficult results the building of an informative observation model.

3. THE PROPOSED APPROACH

In this paper, we extend substantially the joint observational model of the HJS framework, stressing the fact that the contribute is easily generalizable to whatever multi-object particle filtering environment, in which the observations are evaluated in a joint space (*i.e.*, taking into account dependencies among all the tracked objects). The extension translates in a new term in the observational model, the *foreground feature discrimination term* ff_p:

$$p(z_t|x_p) \propto \exp\left(-\frac{\mathrm{ff}_p}{2\,\sigma_f^2} - \frac{\mathrm{fc}_p + \mathrm{bc}_p}{2\,\sigma^2}\right). \tag{3}$$

The foreground feature discrimination term is introduced in occlusion cases in order to help appearance disambiguation among similar targets which stand spatially close, finding the most discriminative parts of an object with respect to the other surrounding objects. In the following, we explain how this term is evaluated in a two-objects scenario, generalizing then to a higher number of objects.

3.1. Two-objects case

The first step is to choose a set of candidate features. Here, we use a small set of M features based on RGB color histogram, which it has been shown in [6] to be experimentally appropriate for tracking applications. The feature set contains the linear combination of R, G, B pixel values: $\mathcal{F} = \{w_1R + w_2G + w_3B \mid w_* \in [-2, -1, 0, 1, 2]\}$, pruning

out redundant combinations. Such class of features is computational fast to manage, and shows adequate expressiveness. Then, M histograms of features (b bins) have been built considering each of the two objects' appearances.

Second, the histograms of features are combined together to distill a combined feature, tuned to discriminate between the two objects in the current frame. In particular, the log-likelihood ratio has been computed as $L = \log \frac{p}{q}$, where p and q are the histograms of a single feature for the first and second object, respectively. Log-likelihood expresses naturally discriminativeness; actually, thresholding L at zero is equivalent to use a maximum likelihood rule to classify the two objects. This feature permits thus to rewrite the possible multimodal distributions p and q into a unimodal distribution. Finally, we introduce an evaluation criterion which measures the separability that feature L induces between the two classes. Likewise [6], we employ the two-class variance ratio, which, given two class distributions q and p (their histograms), is defined as:

$$\mathrm{VR}(L; p, q) = \frac{\mathrm{var}(L; (p+q)/2)}{\mathrm{var}(L; p) + \mathrm{var}(L; q)}. \quad (4)$$

The denominator enforces that the within-class variances should be small for both objects' classes, while the numerator rewards cases where values associated with two different objects are widely separated. At the end of the process we have, for each moving object, a new feature set $\mathcal{F}_s = \{f_1, \ldots, f_s\} \subseteq \mathcal{F}$, built by selecting the the top N_f most discriminative individual features (ordered by decreasing VR). For each frame of the video we compute the set $\mathcal{F}_s$ only if the two objects are very close.

In [6] the feature selection method is embedded in a mean-shift tracking system. Vice versa, our method uses the selected features in order to build a map for each object k involved in an occlusion, called *discrimination map* F_k, that favors the pixels corresponding to the discriminative parts of an object. Such map is obtained fusing the rank of the discriminative features as a weighted sum, *i.e.*,

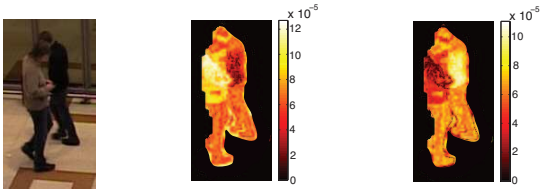
$$F_k = \sum_{s=1}^{N_f} \mathrm{VR}_s \; g(L_s, I) \quad (5)$$

where s indexes the log-likelihood ratio features $\{L\}$, and g is the function that maps the 2D rendering I of a person to the discrimination map, assigning the values of L_s to the image I. An example of discrimination maps are shown in Figure 1.

During an iteration of the filtering step, feature selection is stopped when an occlusion occurs. A discrimination map is built for each sample hypothesis, using the feature set $\mathcal{F}_s$ selected at previous time, and employed to assign a reasonable weight in the observational step. Given a particle x_p, the foreground feature discrimination term ff_p is computed by the discrimination map F_k:

$$\mathrm{ff}_p = 1 - p(z_t|\mathcal{F}_s, x_p) = 1 - \sum_u F_k(u) \; \delta(B_t(u)) \quad (6)$$

where B_t is a binary map given from FG/BG subtraction [7], δ is the Kronecker delta and u indexes the pixels. This term will be higher for the particles far from the discriminant parts of the object and vice versa.



(a) Image (b) Object 1 (c) Object 2

Fig. 1. Discrimination maps for two objects during an occlusion; brighter pixels are those whose values ensure higher discrimination.

3.2. Multi-objects case

Our goal is to select the features that discriminate between a particular object h and the surrounding, multiple, objects. We decompose such task as that of finding a set of ranked features for a single pairwise discrimination, repeating the process for all the couple of objects that include h. In particular, assuming that the discriminative features for an object with respect to the surrounding single objects are given, we can group the set $\{\mathcal{F}_s^{h,1}, \ldots, \mathcal{F}_s^{h,K}\}$, where the $\mathcal{F}_s^{h,k}$ is the set of discriminative features related to object h with respect to the object k. The features for the multiple discrimination are retrieved by exploiting an intersection operation:

$$\mathcal{F}_\cap^h = \bigcap_{k=1}^{K} \mathcal{F}_s^{h,k}. \quad (7)$$

The new feature set $\mathcal{F}_\cap^h$ contains the common features of every single set $\mathcal{F}_s^{h,k}$, that is the set of the best discriminative features for the object h as compared to all the other surrounding entities.

4. EXPERIMENTAL RESULTS

The testing session has been focused on the PETS dataset (edition 2006[1], and 2007[2]), in order to certify the suitability of the proposed technique with public and challenging data. As comparative tracking framework, we consider the original version of the HJS filter proposed in [4]. In a wider sense, such comparison is highly valuable, being HJS a tracker with

[1] http://www.cvg.rdg.ac.uk/PETS2006/index.html
[2] http://www.pets2007.net

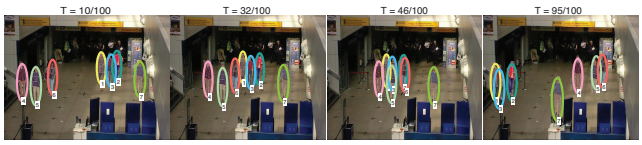


Fig. 2. Synthetic video example tracked by our method. Occlusions occur among different objects are correctly handled.

strong performances [4]. Since PETS videos do not contain ground truth information (GT) on the real position of the objects, as preliminary test and quantitative analysis we create a set of *synthetic* videos. In particular, 8 synthetic sequences (of 100 frames each) have been built by superimposing different static pedestrian images on a static background, mimicking the 3D scenario by applying scaling on the silhouette. The synthetic pedestrians move in the scene, with a dynamics similar to the that of the PETS videos. The comparison of our approach with HJS on this synthetic dataset (see Fig. 2) has been performed using a different number of moving objects in the range $[2, 7]$. Two error measures are considered: 1) the estimation of the Mean Error (ME), defined as the distance between the GT position of an object and the location estimated by the tracker on the floor (in meters); 2) the estimation on the image plane in terms of False Positives (FP), Multiple Objects (MO), False Negatives (FN), Multiple Trackers (MT), and Tracking Success Rate (TSR) (see [8] for further details). The results, averaged over all the experiments and for all the moving objects, are summarized in Table 1 (the lower the better, except TSR): it can be noted that the proposed approach outperforms HJS, especially in terms of ME and TSR.

	ME	FP	MO	FN	MT	TSR
HJS	0.64	0.12	0	0.12	0	0.87
Our Method	**0.53**	**0.09**	0	**0.09**	0	**0.91**

Table 1. Results comparison on synthetic videos.

Concerning the real dataset, qualitative evaluations have been carried out exploiting different videos of varying length. The achieved results mirror those gathered on the synthetic trials. Actually, our feature selection strategy provides tracking performances that qualitatively and in average are comparable to those obtained by HJS, outperforming the latter in the case of occlusions. Two examples are shown in Fig. 3 and 4.

5. CONCLUSIONS

In this paper, we proposed an online feature selection strategy embedded in a multi-object tracking framework. The strategy is repeatedly applied in order to distill a pool of features discriminating one object with respect to the surrounding ones, that permits to deal with occlusions among multiple persons. The effectiveness of the method is proved by testing it on a set of trials consisting in synthetic and real standard datasets,

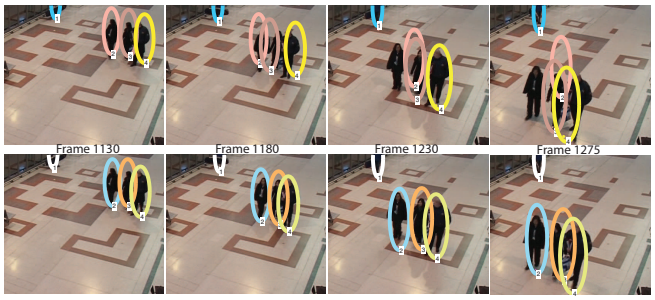


Fig. 3. A comparison of HJS (first row) and our method (second row): frames 1130 to 1300, seq. S5-T1-G, PETS '06.

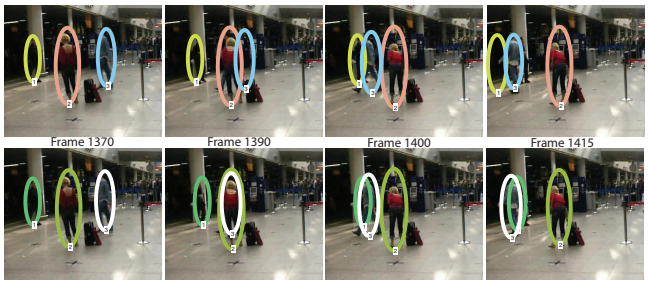


Fig. 4. A comparison of HJS (first row) and our method (second row): frames 1370 to 1470, seq. S07, PETS '07.

promoting its use as a standard step in the observational phase of any particle filtering tracker.

6. REFERENCES

[1] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing*, 2002.

[2] M. Isard and A. Blake, "Condensation: Conditional density propagation for visual tracking," *Int. J. of Computer Vision*, 1998.

[3] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *IEEE Int. Conf. on Computer Vision*, 2001.

[4] O. Lanz, "Approximate bayesian multibody tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2006.

[5] R. Vezzani and R. Cucchiara, "Ad-hoc: Appearance driven human tracking with occlusion handling," in *Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences*, 2008.

[6] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2005.

[7] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1999.

[8] K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba, "Evaluating multi-object tracking," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2005.