THIS PAPER APPEARED IN IEEE INT. CONF. ON ADVANCED VIDEO AND SIGNAL-BASED SURVEILLANCE (AVSS), KLAGENFURT, 2011

Combined Estimation of Location and Body Pose in Surveillance Video

Cheng Chen
cchen@idiap.ch

Alexandre Heili
aheili@idiap.ch

Jean-Marc Odobez
odobez@idiap.ch

Idiap Research Institute – CH-1920, Martigny, Switzerland[*]

Abstract

In surveillance videos, cues such as head or body pose provide important information for analyzing people's behavior and interactions. In this paper we propose an approach that jointly estimates body location and body pose in monocular surveillance video. Our approach is based on tracks derived by multi-object tracking. First, body pose classification is conducted using sparse representation technique on each frame of the tracks, generating (noisy) observation on body poses. Then, both location and body pose in 3D space are estimated jointly in a particle filtering framework by utilizing a soft coupling of body pose with the movement. The experiments show that the proposed system successfully tracks body position and pose simultaneously in many scenarios. The output of the system can be used to perform further analysis on behaviors and interactions.

1. Introduction

Tracking people is a very important task in surveillance environments, and is beneficial to many applications such as behavior recognition, group and interaction detection, and facility usage analysis. However, people tracking is also a challenging task. The difficulty comes from the low quality and resolution of the surveillance video, the occlusion, the cluttered background, and so on. Much work has been done in tracking the location of people over time. Recent work investigates robustness and online learning issues [1][2], and tracking by detection for multi-object tracking [3][4].

Unlike most techniques which focus on tracking the location, our aims is to characterize more precisely people's behavior by estimating more precise cues like head orientation, visual focus and body orientation. More specifically, in this paper we focus on the joint estimation of location and body pose (orientation). Including body pose cue can help the surveillance systems in many aspects. For example, from the body orientation we know where the person is probably looking at. This is especially useful in surveillance videos where the low resolution only allows coarse head pose or gaze tracking. It is also important for group and interaction detection. For example, when people are interacting, they typically face towards each other (especially when they are static). Such group analysis cannot be performed well with the location information alone, and body pose introduces valuable complement cues (See Fig. 1).

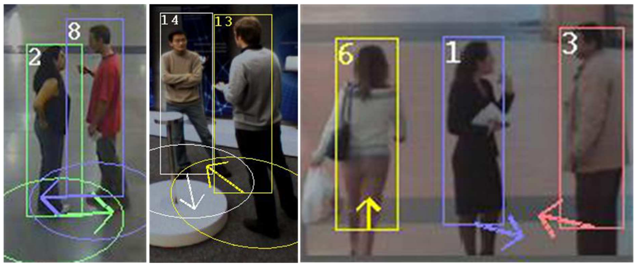


Figure 1. Body pose provides important information to detect interactions, while location alone is not sufficient (right figure).

The workflow of our approach is as follows. We use as input the tracks generated by a multi-object tracker, where each track contains a noisy bounding box sequence in the image for one person identity. For each track, we first perform body pose classification on each frame separately using multi-level HoG (Histogram of Oriented Gradients) feature and sparse representation technique [5]. Then, we perform a joint estimation of the true states (location, velocity and body pose) in a particle filtering framework using the noisy location and pose observation. We also propose to use a soft coupling between the movement and body pose conditioned on the speed (i.e. when the person is moving fast, the body orientation is more aligned to the movement direction, and vice versa).

Some other work has addressed the issue of body pose in surveillance videos. For example, [6] estimated body pose discretized in eight directions. However, the dependency

[*]This work was supported by the Integrated Project VANAHEIM (248907) supported by the European Union under the 7th framework program.

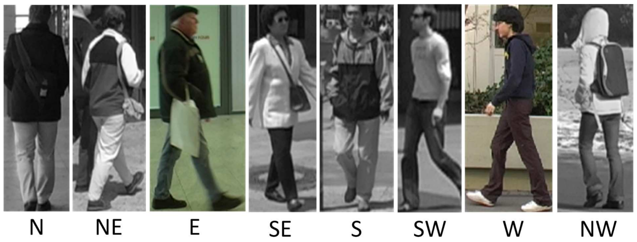


Figure 2. Eight body pose classes.

between pose and velocity is not exploited in their temporal filtering stage. The coupling between movement direction and pose has been exploited in previous work [7] and [8], but problems remain when people are static or have only slow movement. In [8], the coupling is constant regardless of the speed, and thus provides bad information at low magnitude since speed orientation is highly noisy in such cases. In contrast, [7] exploited a loose coupling at low speed, but due to the absence of a discriminative model for pose estimation, the system is almost blind to body pose information when people are (almost) static.

In summary, the contribution of the paper are as follows:

- A framework for joint location and body pose estimation.
- A sparse representation for body pose estimation.
- A soft coupling between body orientation and velocity which works when people are moving or static.

In the following, we introduce our static body pose estimation method in Section 2. Then we show the combined estimation in Section 3. Experimental results are presented in Section 4 and we conclude the paper in Section 5.

2. Static Body Pose Classification

We use multi-level HoG feature and sparse representation for pose classification.

Body pose representation

Given the low resolution images, we quantize the body orientation in the image into eight directions (See Fig. 2): N, NE, E, SE, S, SW, W, NW [1]. Note that at this stage body pose classification is performed in the 2D images, and no 3D information (e.g. the camera's tilting angle) is inferred. We make a reasonable assumption that the camera tilt is not too large ($< 30^o$) and that the pose classification can be conducted without explicitly considering the tilt.

For each human bounding box in the image, we calculate a multi-level HoG feature vector [9]. The bounding box is divided into non-overlapping blocks at three different levels: 1×3, 2×6 and 4×12. Each block in turn consists of 4 cells. We quantize the gradient orientation into 9 unsigned directions, and each pixel votes to the corresponding direction using the gradient magnitude as weight. In this way, for each human region we end up with a 2268 dimensional feature vector.

Pose classification by sparse representation

To perform pose classification on images, we use sparse representation technique, which has been shown to be effective in image analysis and face recognition [5]. Let $\{(\mathbf{f}_i, l_i)\}\,(1 < i < N)$ denote the training data, where each $\mathbf{f}_i$ is a multi-level HoG feature vector, and l_i is the corresponding pose label. For a novel feature vector $\mathbf{f}'$, the pose label l' can be inferred from its relation to the training data. Specifically, $\mathbf{f}'$ can be approximated as a linear combination of the training features:

$$\mathbf{f}' \approx a_1\mathbf{f}_1 + ... + a_N\mathbf{f}_N = \mathbf{F}\mathbf{a}, \tag{1}$$

where $\mathbf{F} = [\mathbf{f}_1, ..., \mathbf{f}_N]$, and $\mathbf{a} = [a_1, ..., a_N]^T$ is the reconstruction weights vector subject to non-negative constraint $a_i \geq 0$. It is reasonable to assume that $\mathbf{f}'$ will be well approximated by using only the part of training data with the same inherent pose label, which means the reconstruction vector $\mathbf{a}$ is sparse. To seek for the sparse solution, L_1 term is used for regularization and our goal is to find:

$$\mathbf{a}^* = \arg\min \|\mathbf{f}' - \mathbf{F}\mathbf{a}\|_2^2 + \gamma\|\mathbf{a}\|_1, \tag{2}$$

where $\|.\|_2$ and $\|.\|_1$ are the L_2 norm and L_1 norm, respectively, and γ is the parameter controlling the importance of the L_1 regularizor.

The non-zero elements of $\mathbf{a}^*$ will be concentrated on the training data point with the same label as $\mathbf{f}'$, and we can perform pose classification accordingly. We define the probability of the pose label being k as the partial L_1 norm of $\mathbf{a}$ associated with class k, i.e.:

$$\rho_k\left(\mathbf{f}'\right) = \sum_{l_i=k} a_i^* \Big/ \|\mathbf{a}^*\|_1 \tag{3}$$

3. Combined Estimation of Position and Pose

Up to now, we have the position information encoded in the tracks, and the body pose estimation generated as in the previous section. As we will see in the experiments, both types of information are quite noisy. The body position jumps because of the uncertainty introduced by the human detector, and the pose estimation is not very accurate due to poorly localized bounding boxes or occlusion. To improve estimation accuracy, we consider temporal consistency and the consistency between pose and location (movement) information.

[1] The naming of these directions is just for convenience and has nothing to do with the real directions such as north/south.

Our estimation problem is formulated in a Bayesian framework, where the objective is to recursively estimate the filtering distribution $p(\mathbf{s}_t|\mathbf{z}_{1:t})$ where $\mathbf{s}_t$ is the state at time t and $\mathbf{z}_{1:t}$ denotes the set of measurements from time 1 to time t. Under standard assumptions, the recursion is given by:

$$p(\mathbf{s}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{s}_t) \int p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{t-1}|\mathbf{z}_{1:t-1})d\mathbf{s}_{t-1}. \quad (4)$$

In non-linear non-Gaussian cases, it can be solved using sampling approaches, also known as particle filters (PF). The idea behind PF consists of representing the filtering distribution using a set of weighted samples (particles) $\{\mathbf{s}_t^n, w_t^n, n = 1, ..., N\}$ and updating this representation when new data arrives. Given the particle set of the previous time step, configurations of the current step are drawn from a proposal distribution $\mathbf{s}_t \sim q(\mathbf{s}|\mathbf{s}_{t-1}^n, \mathbf{z}_t)$. The weights are then computed as $w_t \propto w_{t-1}^n \frac{p(\mathbf{z}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{s}_{t-1}^n)}{q(\mathbf{s}_t|\mathbf{s}_{t-1}^n, \mathbf{z}_t)}$.

In this work, we use the Boostrap filter, in which the dynamics is used as proposal. Then, three terms which are defined below are important to define our filter: the state model defining our abstract representation of our object, the dynamical model $p(\mathbf{s}_t|\mathbf{s}_{t-1})$ governing the temporal evolution of the state, and the likelihood $p(\mathbf{z}_t|\mathbf{s}_t)$ measuring the adequacy of the observations given our state configuration.

State space: The state vector is defined as $\mathbf{s}_t = [\mathbf{x}_t, \dot{\mathbf{x}}_t, \theta_t]^T$, where $\mathbf{x}_t = [x_t, y_t]$ is the body position in the 3D world coordinate frame, $\dot{\mathbf{x}}_t = [\dot{x}_t, \dot{y}_t]$ is the velocity, and $\theta_t (0 \leq \theta_t < 2\pi)$ is the body orientation angle on the ground plane.

Dynamical model: We use a first-order dynamical model which decomposes as follows, given adequate conditional independence assumptions:

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = p(\mathbf{x}_t, \dot{\mathbf{x}}_t|\mathbf{x}_{t-1}, \dot{\mathbf{x}}_{t-1})\, p(\theta_t|\theta_{t-1}, \dot{\mathbf{x}}_t). \quad (5)$$

The first term of Eq. (5) describes the position and velocity evolution, and for this we use a linear dynamical model:

$$p(\mathbf{x}_t, \dot{\mathbf{x}}_t|\mathbf{x}_{t-1}, \dot{\mathbf{x}}_{t-1}) = \mathcal{N}(\tilde{\mathbf{x}}_t; \mathbf{H}\tilde{\mathbf{x}}_{t-1}, \mathbf{Q}_\mathbf{t}), \quad (6)$$

where $\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$ is the Gaussian probability distribution function (pdf) with mean μ and variance $\mathbf{\Sigma}$, $\tilde{\mathbf{x}}_t = [\mathbf{x}_\mathbf{t}, \dot{x}_t]^T$ is the composite of position and velocity, $\mathbf{H}$ is the 4×4 transition matrix corresponding to $\mathbf{x}_t = \mathbf{x}_{t-1} + \dot{\mathbf{x}}_{t-1}\delta_t$ (with δ_t the time interval between successive frames), and $\mathbf{Q}_\mathbf{t}$ is the system variance.

The second term of Eq. (5) describes the evolution of body pose over time. It is in turn decomposed as:

$$p(\theta_t|\theta_{t-1}, \dot{\mathbf{x}}_t) = \mathcal{V}(\theta_t; \theta_{t-1}, \kappa_0)\, \mathcal{V}(\theta_t; \text{ang}(\dot{\mathbf{x}}_t), \kappa_{\dot{\mathbf{x}}_t}), \quad (7)$$

where $\text{ang}()$ is the angle of the velocity vector (in ground plane), and $\mathcal{V}(\theta; \mu, \kappa)$ is the pdf function of the von Mises distribution parameterized by mean orientation μ and concentration parameter κ:

$$\mathcal{V}(\theta; \mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi I_0(\kappa)}, \quad (8)$$

where I_0 is the 0^{th} order modified Bessel function. Eq. (7) puts two constraints on the dynamics of body pose. The first term says that the new body pose at time t should be distributed around the pose at previous time $t-1$. The second term imposes that the body orientation should be somewhat aligned with the moving direction of the body. The concentration of the second term, $\kappa_{\dot{\mathbf{x}}_t}$, is dependent on the magnitude of velocity and is defined as:

$$\kappa_{\dot{\mathbf{x}}_t} = \begin{cases} 0, & \text{if } \|\dot{\mathbf{x}}_t\| < \tau \\ \kappa_1 \|\dot{\mathbf{x}}_t\|, & \text{otherwise} \end{cases}, \quad (9)$$

This means that if the speed is below some threshold τ, then the person is treated as static and the prior on body pose from velocity is completely flat. When the speed is above τ, however, a larger speed introduces a tighter coupling of the body pose around the moving direction. MCMC (Markov Chain Monte Carlo) is employed to sample from the dynamical model in Eq. (7).

Observation model: We have two types of observations $\mathbf{z}_t = [\mathbf{z}_t^{\text{loc}}, \mathbf{z}_t^{\text{pose}}]$, where $\mathbf{z}_t^{\text{loc}} = [u_t, v_t]$ is the location of the bottom-center of the human bounding box in the image plane, and $\mathbf{z}_t^{\text{pose}}$ is the multi-level HoG feature described in Section 2. Under observation independence assumptions, we have: $p(\mathbf{z}_t|\mathbf{s}_t) = p(\mathbf{z}_t^{\text{loc}}|\mathbf{s}_t)\, p(\mathbf{z}_t^{\text{pose}}|\mathbf{s}_t)$. The location observation likelihood is calculated as:

$$p(\mathbf{z}_t^{\text{loc}}|\mathbf{s}_t) = \mathcal{N}([u_t, v_t]; \mathbf{C}(\mathbf{x}_\mathbf{t}), \mathbf{\Sigma}_{\text{loc}}) \quad (10)$$

where $\mathbf{C}$ is the homography from ground plane to image plane, and $\mathbf{\Sigma}_{\text{loc}}$ is the uncertainty of the detected location (in pixels) in the image plane. As for the pose observation, we first transform the eight angles α_k corresponding to the eight orientation labels (See Fig. 2) into their equivalent orientation $\alpha_k^{gp}(\mathbf{x})$ in the ground plane frame at location $\mathbf{x}$[2]. Then, we define the pose observation likelihood as the probabiliy of the closest orientation class, i.e.:

$$p(\mathbf{z}_t^{\text{pose}}|\mathbf{s}_t) = \rho_{k^{\text{clo}}(\mathbf{s}_t)}(\mathbf{z}_t^{\text{pose}}) \quad (11)$$

where ρ has been defined in Section 2, and $k^{\text{clo}}(\mathbf{s}_t)$ returns the orientation label k whose orientation $\alpha_k^{gp}(\mathbf{x}_t)$ is the closest to the state orientation θ_t.

[2] In other words, two persons facing the camera from different positions will be oriented differently in the ground plane reference frame.

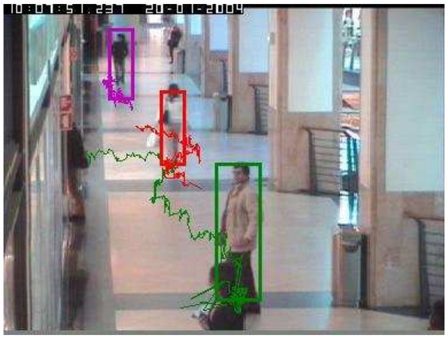


Figure 3. An example result on tracks.

4. Experimental Results

4.1. Deriving tracks

We implemented a detection-based tracker to generate the tracks from videos. Specifically, the output of a human detector [10] is used at every frame as input to our tracker[3]. The task of tracking in this case is to assign temporally consistent identity labels to the detections. We consider all pairwise associations within a short time window and attribute a cost to each pair. The pairwise cost is based on a similarity measure involving proximity both in ground plane position and in terms of HSV color histogram. The label field is then obtained by optimization.

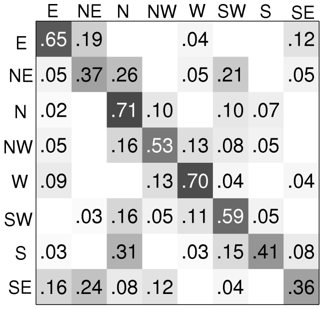
Fig. 3 shows the generated tracks on one clip. The bounding boxes and trajectories of three tracks are shown. Note that the position information is quite noisy.

4.2. Evaluating body pose classification

We use the TUD Multiview Pedestrians dataset [6] to evaluate our novel sparsity-based body pose classification algorithm. For sparse representation, we use the toolbox as in [11] which utilizes truncated Newton interior-point method.

We compare our performance with that reported in [6]. We also tried multi-class SVM on multi-level HoG features (denoted by Multi-SVM). Table 1 shows the details. As in [6], we report two measurements: ”accuracy 1” where we only consider exact hit, and ”accuracy 2” where adjacent hit is also considered as correct. Our method generates more concentrated pose predictions than others.

Fig. 4 shows the classification confusion matrix of our method, where each row is the distribution of predicted labels over a ground-truth group. There is a concentration on the diagonal of the confusion matrix. Note that quite a few errors are introduced by assigning an adjacent rather than the exact direction (which is not a complete error). Also, there are some confusions between symmetric poses (i.e. N and S, NE and SW).



	E	NE	N	NW	W	SW	S	SE
E	.65	.19			.04			.12
NE	.05	.37	.26		.05	.21		.05
N	.02		.71	.10		.10	.07	
NW	.05		.16	.53	.13	.08	.05	
W	.09			.13	.70	.04		.04
SW		.03	.16	.05	.11	.59	.05	
S	.03		.31		.03	.15	.41	.08
SE	.16	.24	.08	.12		.04		.36

Figure 4. Confusion matrix for pose classification.

Table 1. Comparison on pose classification

	Accuracy 1	Accuracy 2
[6] (SVM)	0.42	0.70
[6] (SVM-adj)	0.35	0.76
Multi-SVM	0.48	0.75
Ours	0.55	0.76

4.3. Combined tracking results

We conduct experiments on three datasets: indoor video clips captured by ourselves where people may stay still, move or interact (Fig. 5a), surveillance videos acquired in a metro station (Fig. 5b), and videos of the corridor inside a shopping mall from the Caviar dataset (Fig. 5c). In all experiments we use 500 particles in the recursive sampling.

Fig. 5c also shows the trajectory after the joint estimation. Comparing it to Fig. 3, we see that our method effectively filters out the location noise and gets smoothed tracks.

Figs. 6-8 show our body pose estimation results[4]. To save space, the images are cropped and only the region around the active person is shown. We show the bounding boxes as dash rectangles. To provide more 3D sense, we also display a circle of radius 50cm on the ground plane centered at the bottom of the person in the 3D space. The body pose (in 3D space) is shown using radial lines within the circle, where the thick line with arrow is the final estimation, and the thin line without arrow is the pose observation (i.e. the mode of the eight pose classes as in Section 2).

Fig. 6 shows the results on an indoor clip which contains 350 frames. The two persons talk to each other, and finally they leave the room. The first row of Fig. 6 is our results, and the second row shows the results if we directly use movement direction as the body orientation. At $t = 50$ and $t = 150$, the two persons are talking without significant movement, and from the second row we can clearly see that in this case the movement direction is not a good sign for body pose, while our method is able to keep track on the

[3] Code avalable at http://www.idiap.ch/ odobez/human-detection

[4] Video online at: http://www.idiap.ch/paper/2074/

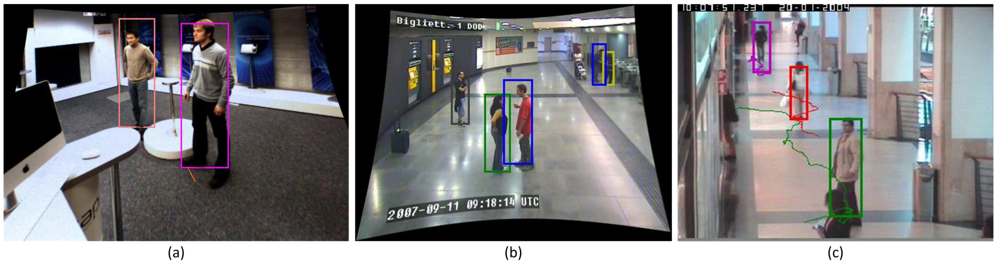


Figure 5. Datasets used in the experiments. (a): indoor scenario; (b): Torino metro station; (c): Caviar dataset

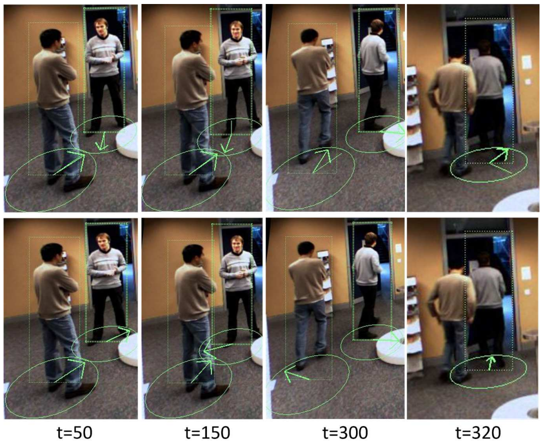


Figure 6. Results on one indoor clip. First row: our results. Second row: using movement direction directly as body orientation.

body pose. This situation continues until around $t = 280$ when the persons start to move and movement direction becomes more reliable for body pose afterwards.

Fig. 7 shows the results on a metro station clip, where two persons walk, meet, discuss and then separate. We show the results on the two persons separately, and compare our method with ”No coupling” (where no coupling between movement and body pose is used, i.e. the second term of Eq. (7) is removed) and ”Only movement” (where the movement direction is directly used as the body pose). At $t = 0, 220, 330$ no detection exists for the man either because he is outside the camera range or the detection is lost due to occlusion. During $t = 0-100$ and $t = 220-300$, the people are moving with significant speed, and the velocity direction provides a good prior on the body pose. During $t = 100 - 220$ they are talking without notable movement and the influence of velocity on the body pose should diminish. We can see that out method generates good results throughout the clip compared to both the ”No coupling” and ”Only movement” approaches.

Fig. 8a shows our result on a clip from the Caviar dataset, where a person walks, wait for someone, and then walks again. At $t = 250$ and $t = 310$, the person is partially occluded but the estimated body pose is correct. During $t = 370 - 1030$, the person remains static in position and keeps on turning his body in different directions, and our method estimates the correct pose. During $t = 10901150$, the poses estimated by our algorithm are wrong. After examination, the problem appeared to be the pose likelihood which (erroneously) favored too strongly the backward pose compared to the front pose, therefore overruling the information provided by the coupling with the motion direction. Solving such a problem can come from the improvement of the body pose classifier, or from the use of other cues (e.g. face orientation could help in this case).

Fig. 8b shows another example from the Caviar dataset. In this video the man remains static and discuss with other person during $t = 0 - 420$, and then he walks away. Our method generates correct poses at most times. Failures at $t = 240, 300, 360$ can be atributed to poor bounding box location or severe occlusion.

5. Conclusion

We have presented a method for jointly estimating body position and pose in surveillance videos. First, we perform body pose classification using sparse representation. Then, we jointly estimate position and body pose by particle filtering, using a soft coupling between velocity and body pose. In the future, we would like to investigate in jointly extracting more behavior cues, such as head pose. We will also study on multi-camera scenarios.

References

[1] N. Alt, S. Hinterstoisser, and N. Navab. Rapid Selection of Reliable Templates for Visual Tracking. in CVPR, 2010. 1

[2] C. Aeschliman, J. Park, and A. Kak. A Probabilistic Framework for Joint Segmentation and Tracking. in CVPR, 2010. 1

[3] B. Song, T. Jeng, E. Staudt, and A. Roy-Chowdhury. A Stochastic Graph Evolution Framework for Robust Multi-Target Tracking. in ECCV, 2010. 1

[4] C. Kuo, C. Huang, and R. Nevatia. Multi-Target Tracking by On-Line Learned Discriminative Appearance Models. in CVPR, 2010. 1

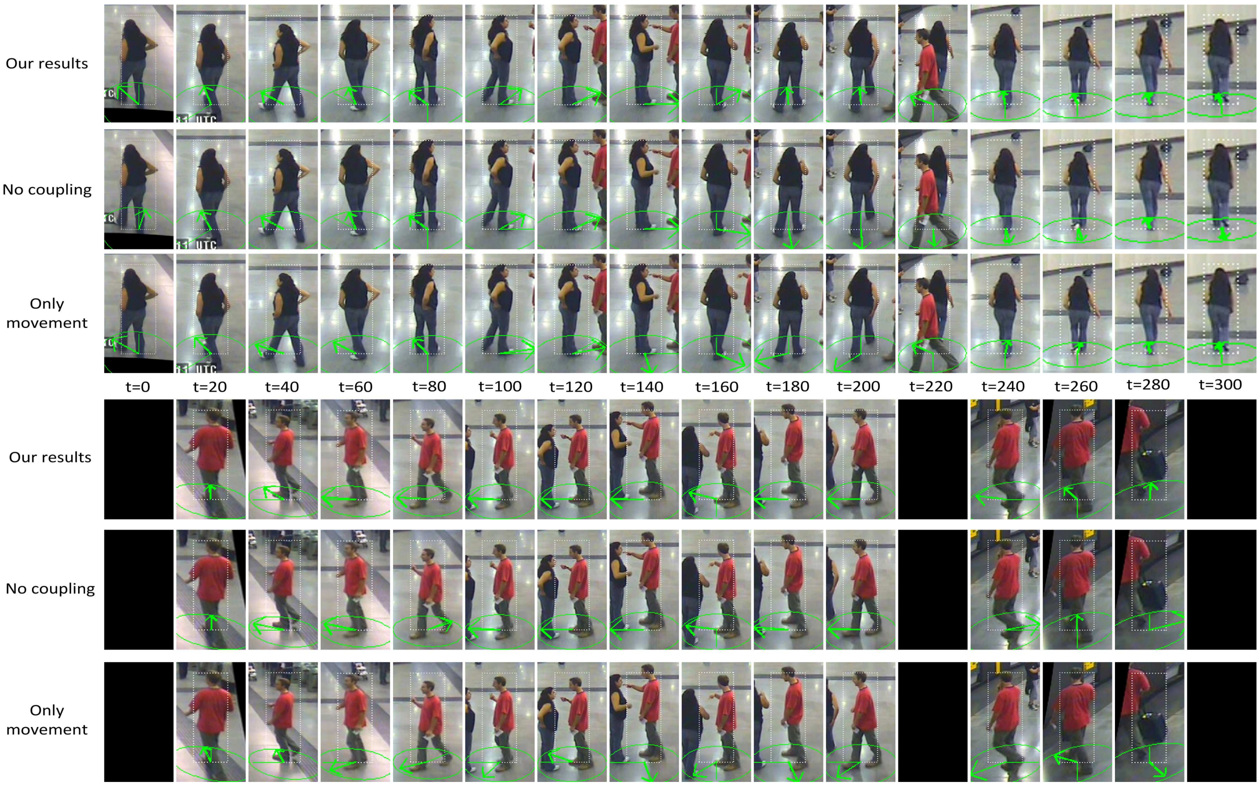


Figure 7. Results on a metro station clip.

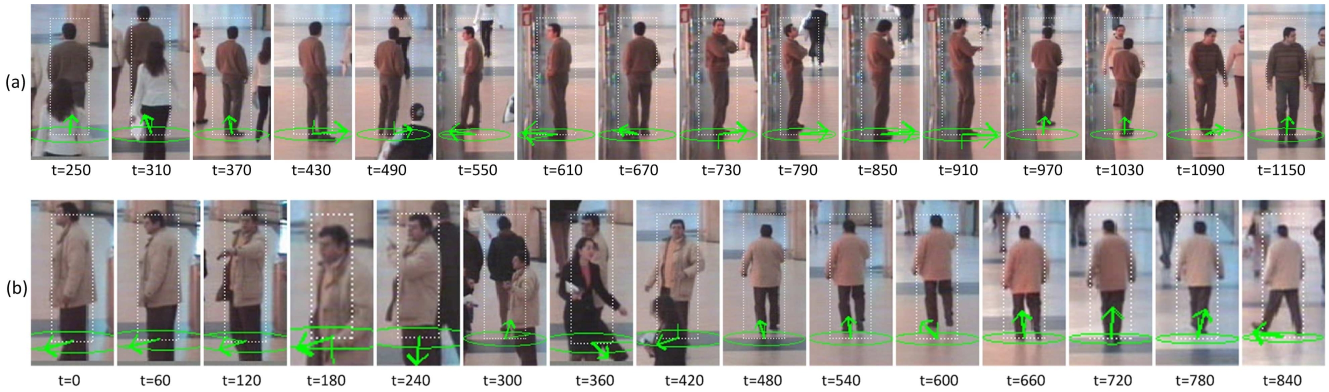


Figure 8. Results on two Caviar clips.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, ”Robust Face Recognition via Sparse Representation”, PAMI, 31(2): 210-227, 2009. 1, 2

[6] M. Andriluka, S, Roth, and B. Schiele. Monocular 3D Pose Estimation and Tracking by Detection. in CVPR, 2010. 1, 4

[7] J. Yao, and J. Odobez. Multi-Camera 3D person Tracking with Particle Filter in a Surveillance Environment. in EUSIPCO, 2008. 2

[8] N. Robertson, and I. Reid. Estimating Gaze Direction from Low-Resolution Faces in Video. in ECCV, 2006. 2

[9] N. Dalal, and B. Triggs. Histograms of Oriented Gradients for Human Detection. in CVPR, 2005. 2

[10] J. Yao, and J. Odobez. Fast Human Detection from Videos UsingCovariance Features. in ECCV-VS, 2008. 4

[11] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for largescale l_1-regularized least squares. IEEE Journal on Selected Topics in Signal Processing, 1(4), 606-617, 2007. 4