A Joint Estimation of Head and Body Orientation Cues 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

The automatic analysis and understanding of behavior and interactions is a crucial task in the design of socially intelligent video surveillance systems. Such an analysis often relies on the extraction of people behavioral cues, amongst which body pose and head pose are probably the most important ones. In this paper, we propose an approach that jointly estimates these two cues from surveillance video. Given a human track, our algorithm works in two steps. First, a per-frame analysis is conducted, in which the head is localized, head and body features are extracted, and their likelihoods under different poses is evaluated. These likelihoods are then fused within a temporal filtering framework that jointly estimate the body position, body pose and head pose by taking advantage of the soft couplings between body position (movement direction), body pose and head pose. Quantitative as well as qualitative experiments show the benefit of several aspects of our approach and in particular the benefit of the joint estimation framework for tracking the behavior cues. Further analysis of behavior and interaction could then be conducted based on the output of our system.

1. Introduction

In surveillance systems, detecting and tracking people is probably the most important task. There has thus been extensive work on tracking the location of single person [4, 17] or multiple persons [5, 3, 1]. This enables location- or trajectory- based analysis of people activities, for instance people counting, scene structure understanding and trajectory abnormality detection, and even some social situation understanding like the identification of groups [9] and social networks [18].

However, to make the surveillance system really "intelligent", we want to know not only "where the people are", but also "what they are doing". In other words, position and trajectory are not enough if we want to make the system aware of the ongoing behaviors and interactions. To move beyond this location-based understanding, our aim in this paper is to propose and study an algorithm for the extraction of behavioral cues, namely body and head poses (orientations), which characterise people's activity and interactions more precisely. Indeed, when observing a single person, his body and head poses indicate which part of the space he is facing and looking at, which could be useful for instance to assess if he his paying attention to his (dropped) luggage. Also, a large discrepancy between his movement direction, body pose, or head pose might indicate an interesting attentional behavior. This can be due either to an intentional pose shift towards an object or region of interest, or to a distraction by something in the scene, which could be useful for abnormality detection. When considering multiple persons, body and head pose would be particularly useful in group/interaction analysis since they provide direct evidence of interaction: people tend to face to and look at each other when they are interacting [8], as illustrated in Figure 1.

Figure 1. Behavior cues such as body pose and head pose are very informative for behavior and interaction analysis.

The workflow of our approach is summarised in Fig. 2. First, we employ a multi-person tracker to generate continuous tracks, where each track contains a noisy bounding box sequence in the image for a person identity. Then, for each bounding box, we perform some static analysis, namely body pose feature extraction, head localization and

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

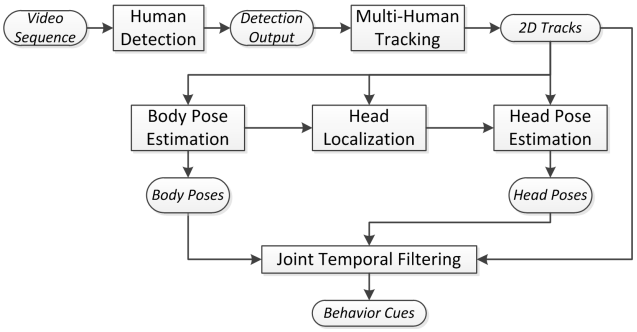

Figure 2. Workflow of our approach.

head pose feature extraction, along with the body and head pose likelihood evaluation for all potential (discrete) poses. This analysis generates noisy observations on body pose and head pose. Finally, based on the observations, we perform a joint estimation of all the cues in a particle filtering framework. The joint estimation takes into account the smoothness of cues over time (which is ensured by the temporal filtering itself), and the dependency between the cues. More precisely, we propose to use soft coupling between body position (movement), body pose and head pose. The coupling is also dependent on the speed: when the person is moving fast, the body orientation is more sharply aligned to the movement direction, and vice versa).

Note that we rely on a separate stage to localise the head in the human bounding box. This ensures that correct head patches are used for pose classification, and is clearly better than some other works like [16] assuming a fixed top-center position on the body as head region, as head pose classification is known to be very sensitive to alignment errors.

Several works have addressed body pose or head pose estimation in surveillance videos [2, 11]. However, despite the obvious link between those two cues, they were mostly treated as completely separate cues. For example, [14] estimated the body pose of tracked people in videos (discretized in eight directions), but the dependency between pose and velocity is not exploited. [13] perform face detection and head pose estimation using a network of far-field cameras, without exploiting the dependency between different cues. The authors in [7] uses 3D distance and head pose to classify pairwise interactions in a work environment, but the 3D position tracking and the head pose classification are completely separate, probably because in their setting, people are static most of the time and there is no coupling between position and head orientation (but coupling between body and head pose exists). A similar approach is used in [8].

The coupling between 3D position and body/head pose has been exploited in previous work [16, 12], but the problem when people are static or with slow speed is not solved. For example, in [16], the coupling between head pose and body movement is constant regardless of the magnitute of speed. This has problems when the person is moving slowly, as the speed direction is highly noisy in this cases. [12] exploited a loose coupling at low speed, but they did not have an explicite observation model for body pose estimation, resulting in a similar problem when the person is moving slowly. To our knowledge, our work is the first to address both body pose and head pose, as well as the inter-cue dependency, and it works fine when the person is moving fast, slowly or is static.

This work is based on and extends the previous work [6], which only deals with body pose. In the current paper, we make two contributions:

- a head localization method which reliably localizes the head of a person from a human detection bounding box, and a head pose classifier estimating the head pose from the localization output;
- a framework for the joint estimation of body position, body pose and head pose, relying on the soft coupling between these cues.

The remaining of the paper is organized as follows. Section 2 introduces the per-frame analysis: body pose feature and likelihood models, head localization and head pose feature and likelihood model. In Section 3 we present the joint estimation by temporal filtering. Experiments are shown in Section 4 and conclusions are given in Section 5.

2. Head and Body Pose Representation

As shown in Fig. 2, for each frame of a human track, we extract several features characterizing the body pose and the head pose. In this Section, we describe these features, along with the head detector that is used to localize the head, and the likelihood models associated with the pose features.

2.1. Body pose representation

For this step, we rely on the previous work described in [6], which is summarized here. The body pose angle (the orientation of the torso) in the image plane is discretized into eight directions (N, NE, E, SE, S, SW, W, NW). Given the human detection output, a multi-level HoG (Histogram of Oriented Gradients) feature is extracted from the image, and corresponds to our body pose observations $\mathbf{z}^{\mathrm{b}}$. This feature vector is then decomposed into a linear combination of the training samples using a sparse representation technique. The (normalized) sum of the weights of the samples belonging to a pose angle class k is then used to defined the likelihood $p^{\mathrm{b}}(\mathbf{z}^{\mathrm{b}}|k)$ of the observation for each pose class.

2.2. Head localization

Prior to extracting head pose observations, we first estimate the head localization and size from the human body bounding box. Here, routines such as face detection or

skin color detection can not be exploited since they will fail when the head is not faced towards the camera. To design a robust pose-independent head localization algorithm, we rely on a HoG based feature and Adaboost classifier.

Features: A given head patches is evenly divided into a 6×6 grid. The features are then defined on the multi-size rectangular blocks associated with the grid boundaries. On each potential block, we extrcated a HoG feature vector. More specifically, the gradient orientation of the pixels within the block are quantized into 9 unsigned bins, and each pixel votes to the corresponding directions using the (cropped) gradient magnitude as the weight.

Given the initial grid, there are 441 possible blocks ranging from size 1×1 to 6×6, and each block is associated with a 9-dimensional feature vector. One possibility would be to concatenate all these sub-features to get a 441×9 dimensional holistic feature vector. However, we note that for head detection, not all sub-features are of the same importance. For example, due to the variation of head pose, the inner texture in the head patch can be quite different and thus not very discriminative. On the other hand, the gradients near the contour of the head remains roughly unchanged (they resemble the overall ellipse shape of the head regardless of the head pose). This inspires us to use boosting technique to select the relevent features.

Weak classifier: Following [10], a first possibility is to define weak classifiers on a block basis, using a 9d to 1d transformation trained in a discriminant fashion (e.g. use Fisher Linear Discriminant - FLD) on the feature vectors [10], and use Adaboost to select relevant weak classifers from the 441 candidates. In this paper we take a different approach, keeping the block approach, but using the primitive HoG differential classifiers as weak classifiers, defined as follows

$$h_{l,d_1,d_2,\sigma}(\mathbf{p}) = \text{sign}\left(f_l^{d_1}(\mathbf{p}) - f_l^{d_2}(\mathbf{p}) - \sigma\right) \quad (1)$$

where $\mathbf{p}$ is an image patch, $l \in \{1, ..., 441\}$ is a block index, $f_l^d(\mathbf{p})$ is the d^{th} dimension of the feature vector in the l^{th} block of image patch $\mathbf{p}$, and σ is a threshold. In other words, a weak classifier h first selects a block l, and then compare two gradient directions in that block against a threshold. Compared to [10], the idea is to have a more sparse weak classifier by comparing only two directions in a block, rather than to learn a full FLD between all directions. When the amount of data is not huge, we expect such weak classifier to be more robust and lead to a better generalization. For example, at the top center of the head patch, we will expect the predominant gradient direction to be somewhat vertical. Due to data variability and noise, the learned FLD relation involving the vertical direction and all the other directions might not be very accurate, and non zero values might lead to noisy classifier values at test time. On the other hand, simply requiring that the vertical gradient direction should be stronger than the horizontal direction might be a looser but more noise-robust choice. This is confirmed by our experiments reported in Section 4.

Training. Our strong classifier is trained using Adaboost on a training dataset $\{(\mathbf{p}_i, y_i)\}$, where $y_i \in \{+1, -1\}$ is the label for positive (head) or negative (non-head) samples. Adaboost learns a strong classifier H by selecting T weak classifiers (and optimizing for the variance parameter):

$$H(\mathbf{p}) = \text{sign}\left(\sum_{j=1}^{T} \alpha_j h_t(\mathbf{p})\right), \quad (2)$$

where α_j are the learned weights associated to the weak classifiers.

Testing and Head localization. Given the human body bounding box, we test possible head localization configurations (with different locations, sizes and aspect ratios of the head bounding box). However, if we directly use the binary classifier of Eq. (2), we might get many hits around the true head location. Instead, we use the real-valued score of the detector (i.e. H without the sign function) to build a confidence map on the possible head locations, and perform non-maximum supression to find local maxima as localized heads. The detector score of those local maxima is used to accept or reject the detection, but we always assume the presence of at least one head. Finally, to select the single head location used for further processing, we apply a separate and simple temporal filtering on the head location candidates by enforcing a head location smoothness over time.

2.3. Head pose estimation

We represent head pose as pan $\tilde{\alpha}$ and tilt $\tilde{\beta}$ angles in the image plane[1]. Considering the resolution of surveillance video, we discretize the pan into 12 angles with 30° interval [2], and we discretize tilt angle into 3 classes: up ($\tilde{\beta} > 30^\circ$), middle ($-30^\circ < \tilde{\beta} < 30^\circ$) and down ($\tilde{\beta} < -30^\circ$). Therefore, overall there is a set of 36 pose $(\tilde{\alpha}_m, \tilde{\beta}_m)$.

Defining the head pose feature $\mathbf{z}^{\mathbf{h}}$. We use both texture and color features for head pose estimation. As texture feature, we use again a multi-level HoG descriptor. The head patch is divided into non-overlapping blocks at two levels: 2×2 and 4×4 blocks. Each block in turn consists of 4 cells. The gradient orientation is quantized into 9 unsigned bins, and the 4×9 entries of a block are normalized to 1. In this way, for each head patch we end up with a 720 dimensional feature vector. For color feature, we use the histogram-based skin color detector proposed in [15] to detect the skin region in the head patch. Then, the head patch

[1] That is, the pose is defined with respect to the viewing direction, which means that $(\tilde{\alpha}, \tilde{\beta}) = (0, 0)$ corresponds to a person looking straight at the camera, whatever his image position.

[2] Unlike some work where the head pose is only estimated in frontal and profile views, we allow the 360° full pan range to include back view.

is resized into a 20×20 binary skin mask as our 400 dimensional color feature.

Head pose likelihood model p^{h}. Learning the likelihood is conducted assuming training data with known head poses. For each class m, we calculate the mean texture feature $\mathbf{r}_m^{\text{text}}$ and mean color feature vectors $\mathbf{r}_m^{\text{col}}$. Then, the likelihood of a head patch observation $\mathbf{z}^{\mathrm{h}} = \left(\mathbf{z}^{\text{text}}, \mathbf{z}^{\text{col}}\right)$ for a given pose class m is expressed as:

$$p^{\mathrm{h}}\left(\mathbf{z}^{\mathrm{h}}|m\right) = p^{\text{text}}\left(\mathbf{z}^{\text{text}}|m\right) p^{\text{col}}\left(\mathbf{z}^{\text{col}}|m\right) \tag{3}$$

where each component likelihood is in turn expressed as:

$$p^{F}\left(\mathbf{z}^{F}|m\right) = \exp\left(-\lambda^{F} d^{F}\left(\mathbf{z}^{F}, \mathbf{r}_m^{F}\right)\right), \tag{4}$$

where $F = \{\text{text}, \text{col}\}$ is the feature type, λ^F is a parameter, and $d^F()$ is the distance between the observed feature and the mean feature. For the texture feature, we use the $L2$ distance. For the color feature, we use the $L1$ distance.

2.4. Summary

Given the human detection bounding box output associated with a human track at time t, we have the following observations:

- $\mathbf{z}_t^{\text{loc}} = [u_t, v_t]$, which denotes the bottom-center position of the body in the image plane;
- $\mathbf{z}_t^{\mathrm{b}}$, the body pose feature described in Section 2.1;
- $\mathbf{z}_t^{\mathrm{h}}$, the feature described in Section 2.3.

In addition, we have defined $p^{\mathrm{b}}()$ and $p^{\mathrm{h}}()$, the functions providing the likelihood of the corresponding observed features for a given body pose class k or head pose class m. Note that the body and head pose classes are defined in the image plane.

3. Joint Estimation of Behavior Cues

Up to now, for each human detection output, we have extracted 2D location, body and head pose features. Using the defined likelihood models, for each frame, we could estimate the body and head pose cues. However, such estimates would be quite noisy. For example, the bounding box jumps in the image due to the uncertainty of the human detector, and the body/head pose estimation can be wrong due to poorly localized bounding boxes or partial occlusion.

In this section, we perform the estimation over time and in the 3D space of all the behavior cues. To improve the accuracy, we use temporal filtering to exploit the intra-cue temporal smoothness, and the estimation is conducted jointly to also exploit the inter-cue dependencies.

Particle filtering framework. 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:

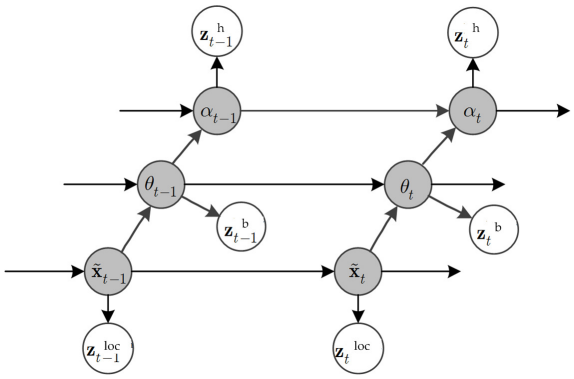

Figure 3. Dynamical model. State variables are shaded, and observation variables are unshaded.

$$p(\mathbf{s}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{s}_t) \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) p(\mathbf{s}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{s}_{t-1}. \tag{5}$$

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. Fig. 3 provides the graphical model of our approach, highlighting the dependency assumptions between variables.

State space: The state vector is defined as $\mathbf{s}_t = [\mathbf{x}_t, \dot{\mathbf{x}}_t, \theta_t, \alpha_t]^T$, 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, $\theta_t (0^\circ \leq \theta_t < 360^\circ)$ is the body orientation angle on the ground plane, $\alpha_t (0^\circ \leq \alpha_t < 360^\circ)$ is the 3D head pan angle. Note that all the elements in the state vector are defined with regard to the 3D world coordinate frame.

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

$$\begin{aligned} p\left(\mathbf{s}_t|\mathbf{s}_{t-1}\right) &= p\left(\mathbf{x}_t, \dot{\mathbf{x}}_t|\mathbf{x}_{t-1}, \dot{\mathbf{x}}_{t-1}\right) \\ &\times p\left(\theta_t|\theta_{t-1}, \dot{\mathbf{x}}_t\right) p\left(\alpha_t|\alpha_{t-1}, \theta_t\right). \end{aligned} \tag{6}$$

Location dynamics: The first term of Eq. (6) describes the

position and velocity evolution, and for this we use a linear dynamical model:

$$p\left(\mathbf{x}_t, \dot{\mathbf{x}}_t | \mathbf{x}_{t-1}, \dot{\mathbf{x}}_{t-1}\right) = \mathcal{N}\left(\tilde{\mathbf{x}}_t; \mathbf{H}\tilde{\mathbf{x}}_{t-1}, \mathbf{Q}_t\right), \quad (7)$$

where $\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$ is the Gaussian probability distribution function (pdf) with mean μ and variance $\mathbf{\Sigma}$, $\tilde{\mathbf{x}}_t = [\mathbf{x_t}, \dot{\mathbf{x}}_t]^T$ 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 $\mathbf{Q_t}$ is the system variance.

Body pose dynamics and coupling with motion direction: The second term of Eq. (6) describes the evolution of body pose over time. It is in turn decomposed as:

$$p\left(\theta_t | \theta_{t-1}, \dot{\mathbf{x}}_t\right) = \mathcal{V}\left(\theta_t; \theta_{t-1}, \kappa_0\right) \mathcal{V}\left(\theta_t; \text{ang}\left(\dot{\mathbf{x}}_t\right), \kappa_{\dot{\mathbf{x}}_t}\right), \quad (8)$$

where $\text{ang}()$ is the angle of the velocity vector (in ground plane), κ_0 is the system concentration parameter for body pose, and $\mathcal{V}(\theta; \mu, \kappa) \propto \exp^{\kappa\cos(\theta-\mu)}$ denotes the pdf function of the von Mises distribution parameterized by mean orientation μ and concentration parameter κ.

The Eq. (8) sets two constraints on the dynamics of body pose. The first term expresses 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 body pose dependency concentration, $\kappa_{\dot{\mathbf{x}}_t}$, is dependent on the magnitude of velocity and is defined as:

$$\kappa_{\dot{\mathbf{x}}_t} = \begin{cases} 0, & \text{if } \|\dot{\mathbf{x}}_t\| < \tau \\ \kappa_1 \left(\|\dot{\mathbf{x}}_t\| - \tau\right)^2, & \text{otherwise} \end{cases}. \quad (9)$$

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.

Head pose dynamics and coupling with body pose: The third term of Eq. (6) describes the evolution of the head pose over time. It is decomposed as:

$$p\left(\alpha_t | \alpha_{t-1}, \theta_t\right) = \mathcal{V}\left(\alpha_t; \alpha_{t-1}, \kappa_1\right) \mathcal{V}\left(\alpha_t; \theta_t, \kappa_2\right) \quad (10)$$

Similarly to Eq. (8), the Eq. (10) sets two constraints on the dynamics of head pose. The first term ensures the temporal smoothness of the head pan evolution, wheras the second term models the soft coupling between the head and body orientations. However, in this case, the concentration parameter κ_2 is constant (and lower than κ_1 since the coupling is looser than between the body orientation and motion direction).

Observation model: Recalling Section 2.4, at each time t the observation feature is defined as $\mathbf{z}_t = \left(\mathbf{z}_t^{\text{loc}}, \mathbf{z}_t^{\text{b}}, \mathbf{z}_t^{\text{h}}\right)$. Under observation conditionnal independence assumptions, the observation likelihood is given by:

$$p\left(\mathbf{z}_t | \mathbf{s}_t\right) = p\left(\mathbf{z}_t^{\text{loc}} | \mathbf{s}_t\right) p\left(\mathbf{z}_t^{\text{b}} | \mathbf{s}_t\right) p\left(\mathbf{z}_t^{\text{h}} | \mathbf{s}_t\right) \quad (11)$$

where each term is defined as follows. The position likelihood is calculated as:

$$p\left(\mathbf{z}_t^{\text{loc}} | \mathbf{s}_t\right) = \mathcal{N}\left([u_t, v_t]; \mathbf{C}\left(\mathbf{x}_t\right), \mathbf{\Sigma}_{\text{loc}}\right) \quad (12)$$

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. This term simply expresses that the detected location should be close to the (projected) estimated state.

For the pose observations, we can rely on the likelihood models introduced in Section 2. Since these likelihood models are defined for pose values expressed in the local image frame coordinate system, we first transform the body pose angle θ_t and head pose angle α_t from the 3D world coordinate frame to the local image coordinate (note that this depends on the person's position). Then, body and head observation likelihoods are simply defined as the data likelihood given the closest body or head pose class, i.e.:

$$p\left(\mathbf{z}_t^{\text{h}} | \mathbf{s}_t\right) = p^{\text{h}}\left(\mathbf{z}_t^{\text{h}} | m^{\text{clo}}(\mathbf{s}_t)\right) \text{ and} \quad (13)$$

$$p\left(\mathbf{z}_t^{\text{b}} | \mathbf{s}_t\right) = p^{\text{b}}\left(\mathbf{z}_t^{\text{b}} | k^{\text{clo}}(\mathbf{s}_t)\right) \quad (14)$$

where $m^{\text{clo}}(\mathbf{s}_t)$ returns the head pose class label whose orientation angle is the closest to the (transformed) state orientation $\mathbf{s}_t$ (and similarly for $k^{\text{clo}}(\mathbf{s}_t)$), and $p^{\text{h}}()$ and $p^{\text{b}}()$ are the head and body likelihoods defined in Section 2.

4. Experiments

In this section, we present quantitative evaluation on head localization and headpose estimation, as well as qualitative results of the joint tracking on surveillance videos.

4.1. Head localization task

To collect ground-truth head localization, we manually labeled the head bounding boxes from 1000 positive human detection outputs obtained on a dataset of metro station surveillance videos. The labeled head boxes are used as positive head patches. The negative patches were derived by automatically generating patches inside the body bounding boxes and avoiding the true head locations. We trained a classifier as described in Section 2.2, using 500 weak classifier to construct the strong one.

Evaluation. To quantitatively test the performance, we also annotated head locations on some other human detection outputs to be used as testing data (we used another videos recorded at another time from the same place but a different camera view). We use the IOU (Intersection over Union)

	IOU top	IOU all	n_candidates
Our method	0.60	0.64	≈ 2.3
HoG FLD [10]	0.58	0.65	≈ 3.0

Table 1. Evaluation of head localization.

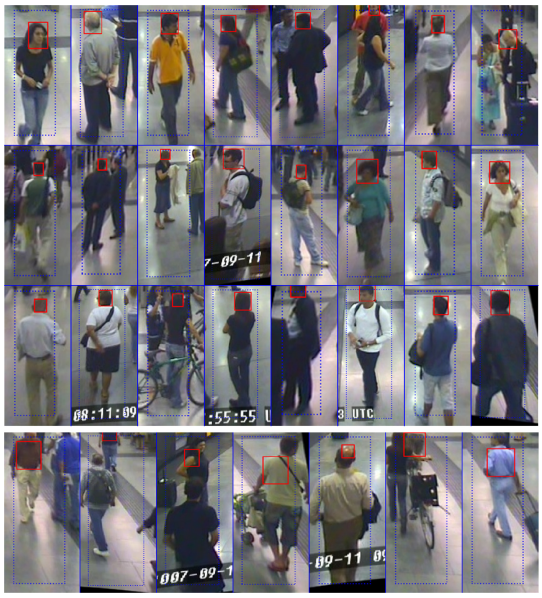

Figure 4. Head localization results. Dashed blue box is the human detection output. Solid red box is the head localization outputs (first candidate). Failures are shown in the last row

between the ground-truth head box $\mathbf{r}_{\text{gt}}$ and the detected head box $\mathbf{r}$ as performance measure:

$$\text{IOU}\left(\mathbf{r}_{\text{gt}}, \mathbf{r}\right) = \frac{\text{area}\left(\mathbf{r}_{\text{gt}} \bigcap \mathbf{r}\right)}{\text{area}\left(\mathbf{r}_{\text{gt}} \bigcup \mathbf{r}\right)}. \tag{15}$$

Note that the IOU measure is symmetric and relatively strict. For example, if the common part of two rectangles occupies 80% of each rectangle, the IOU is just 66.7%.

The evaluation result is shown in Table 1. Remember that as described in Section 2.2, for each human bounding box we may get multiple local maxima and hence multiple head location candidates. Table 1 shows the mean IOU measure when using only the first candidate (ranked by the classifier's score) and when using all candidates (in this case, only the best IOU from all candidates is kept as result), as well as the average number of candidates per human detection output. We also compare our method (HoG differential feature) with the HoG FLD feature [10]. It can be seen that our method generates better results by achieving comparable accuracy with fewer numbers of candidates (i.e. better efficiency). Fig. 4 shows some examples. Our method succesfully localizes most heads, but fails on some examples as illustrated in the last row of Fig. 4. Failures are usually due to poorly localized human bounding box, or head-like textures.

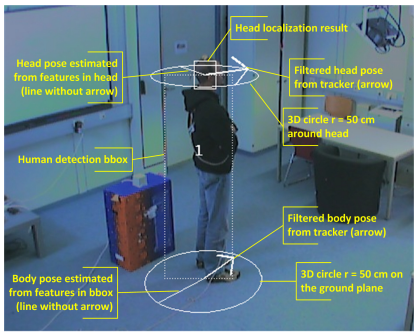

Figure 5. Legend for the illustration.

4.2. Joint tracking: qualitative results

We tested our joint tracking approach on surveillance videos acquired in a metro station (with head pose models learned from the CHIL data, see Section 4.3). Fig. 5 is the legend of the illustration, and Figs. 6-8 show the results on two clips. To save space, the images are cropped and only the region around the active person is shown. We show the human detection bounding boxes as dash rectangles, and the head localization outputs as small solid rectangles. To provide a 3D perception sense, we display two 3D horizontal circles of radius 50cm centered on the bottom-center of the person and the head position, respectively. The body poses and head poses (in 3D space) are shown using radial lines within the circles. More precisely, the body/head poses estimated directly from the feature are shown using radial lines without arrows[3], and the body/head poses returned by the temporal filtering are shown using thicker lines and arrows.

Fig. 6 shows the result on a clip with interaction between two persons. The two persons walk, meet, discuss and then separate. We show the results separately for the woman and the man in the two rows in Fig. 6. At $t = 20, 220, 300$ there is no result for the man either because he is outside the camera view range or the tracking is lost due to occlusion. Note how our joint filtering approach successfully manages to extract accurate body and head poses from noisy observations, even when people are almost static.

Fig. 7 illustrates the same video clip as Fig. 6 in a top-down bird view. Here, the 3D body and head poses can be more easily interpreted and their importance for interaction analysis becomes obvious. Each person is represented by two circles and three arrows. The arrows indicate (from outer to inner) the tracked body speed, body pose, and head pose. Four representative frames are shown. At $t = 60$, both persons are walking with a notable speed, and according to our model, the speed direction provides a good prior for the body pose (and head pose indirectly). At $t = 140$, the persons are talking, with body and head oriented towards each other. In this case, the speed magnitude

[3]The body/head pose classes with the highest likelihood are shown.

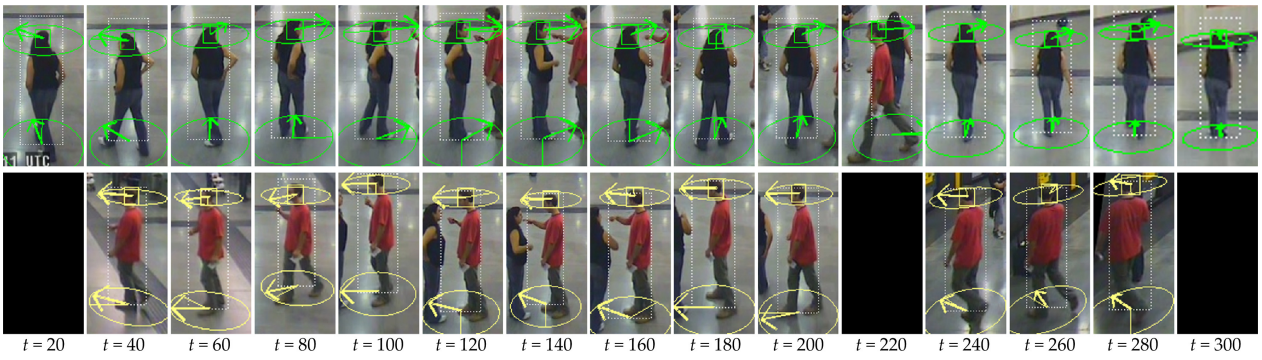

Figure 6. Results on a metro station surveillance video with human interaction. Image resolution is 486×363.

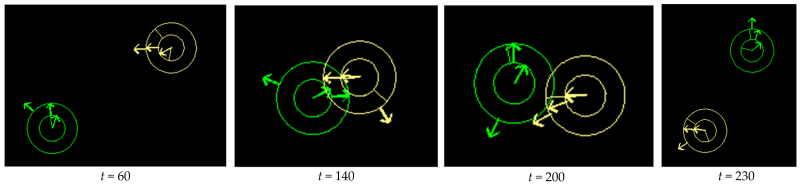

Figure 7. Top-down view illustration (same clip as in Fig. 6).

is very small, resulting in a noisy speed direction which is ignored by our method for the body and pose estimation. At $t = 200$, although the distance between the persons are close, we can still infer that the interaction just stopped because they are not facing each other. At $t = 230$, the two persons have separated.

Fig. 8 shows the result on another clip. In this case, a woman is walking and turning around near a luggage (suitcase) on the ground. Although the person alternates static and slow motion with frequent and fast orientation changes, our method succesfully estimates the body pose and head pose, from which we can easily tell whether the person is attending the luggage or not. Note that in this example, at around $t = 150$, our head localization output is incorrect, but it is automatically corrected shortly after.

4.3. Joint tracking: quantitative evaluation

We use the CHIL dataset of CLEAR 2007 head pose estimation contest. It contains annotated data for 10 persons (id 6-15) where people in the videos are turning their body and head orientation. For each frame, the ground-truth head poses are provided by a magnetic field location and orientation tracker. We used the person id 6-11 for training the head pose model, and 12-15 for testing. For body pose evaluation, we manually labeled the body orientation of 100 randomly selected frames using a 3D interface.

As performance measure, we use pose accuracy defined as the average error angle between the predicted and ground-truth pose angles in the 3D space. To evaluate the effectiveness of our joint tracking approach, we compare our method with the results obtained on a per-frame basis ('observation'), and with a baseline where body and head pose are filtered separately without exploiting the soft coupling between them (i.e. we have $\kappa_2 = 0$ in Eq. (10)). The comparison is dipicted in Table 2. It can be seen that our joint filtering method significantly outperforms the separate filtering approach, and that the accuracy is quite high given that only one camera is used.

	observation	separate filtering	joint filtering
body pose	41.7	32.9	21.9
head pose	30.3	25.8	17.6

Table 2. Evaluation on the joint tracking approach. All numbers are in degree.

Fig. 9 illustrates the results on a sample test clip. Here it is straightforward to see the advantage of our approach. By exploiting the soft coupling between body pose and head pose, we can get better accuracy for both. For example, at $t = 600$ and $t = 1200$, the incorrectly estimated head pose is corrected by the body pose. At $t = 1000$, $t = 2000$ and $t = 2600$, the body pose is corrected by the head pose. On the other hand, our soft coupling remains loose enough to still allows some discrepency between body pose and head pose, which is useful when the head is turning away from the body orientation (e.g. $t = 1400$ and $t = 2600$).

5. Conclusions

We have presented a approach for the joint tracking of pose behavioral cues in surveillance videos. Given the tracks generated by a multi-person tracker, we first localize the head and extract body and head pose features. These features are used to jointly estimate the body position, body pose and head pose in 3D space using a particle filtering ap-

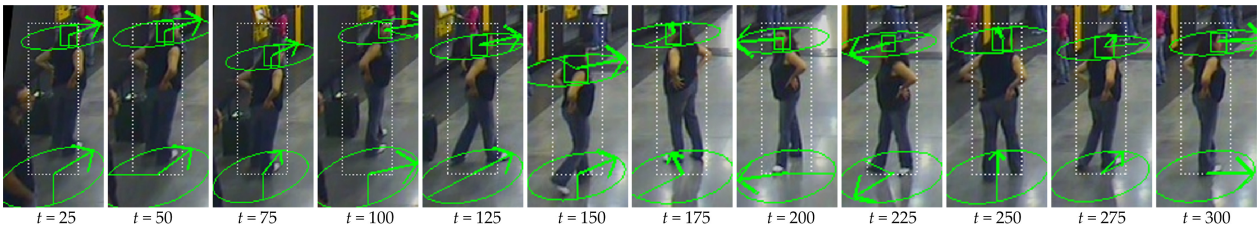

Figure 8. Results on a metro station surveillance video with luggage attendence. Image resolution is 486×363.

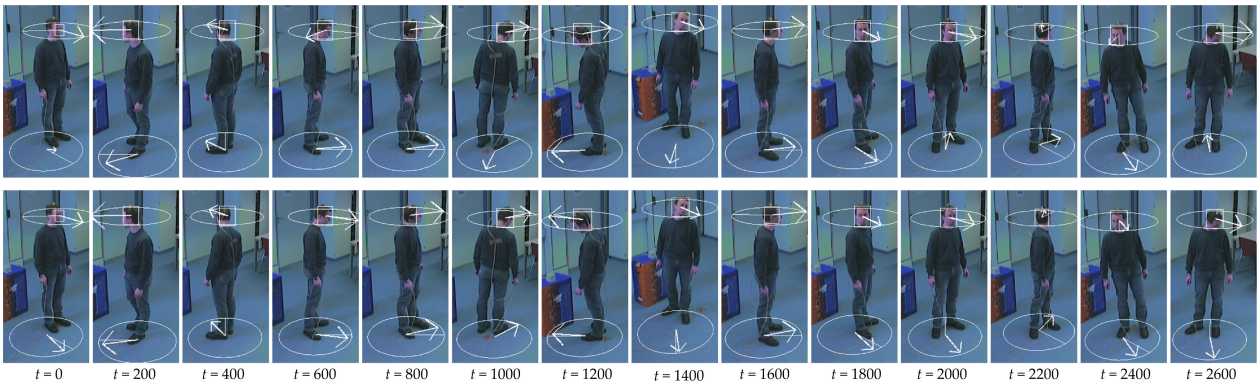

Figure 9. Comparison on a CHIL sequence. Image resolution is 640×480. First row: without soft coupling. Second row: our approach.

proach that exploits the conditional coupling between body position (movement direction) and body pose, and the soft coupling between body pose and head pose. Qualitative and quantitative experiments are provided.

In the future, we would like to investigate the issues of wrong estimates due to occlusion (in particular for the body), and exploit multi-camera environments to resolve ambiguities. We will also investigate the modeling of human interaction based on the output of our methods.

References

[1] A. Andriyenko and K. Schindler. multi-target tracking by continuous energy minimization. In *CVPR*, 2011. 1

[2] B. Benfold and I. D. Reid. Guiding visual surveillance by tracking human attention. In *BMVC*, 2006. 2

[3] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011. 1

[4] B. Liu, J. Huang, C. Kulikowski, and L. Yang. Robust tracking using local sparse appearance model and k-selection. In *CVPR*, 2011. 1

[5] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a CRF model. In *CVPR*, 2011. 1

[6] C. Chen, A. Heili, and J. Odobez. Combined estimation of location and body pose in surveillance video. In *AVSS*, 2011. 2

[7] C.-W. Chen, R. C. Ugarte, C. Wu, and H. Aghajan. Discovering social interactions in real work environments. In *FG*, 2011. 2

[8] M. Farenzena, A. Tavano, L. Bazzani, D. Tosato, G.Paggetti, G. Menegaz, V. Murino, and M.Cristani. Social interactions by visual focus of attention in a three-dimensional environment. In *Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis (PRAI*HBA)*, Dec. 2009. 1, 2

[9] W. Ge, R. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2009. 1

[10] I. Laptev. Improving object detection with boosted histograms. *Image and Vision Computing*, 27(5):535–544, 2008. 3, 6

[11] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *BMVC*, 2009. 2

[12] J. Yao and J. Odobez. Multi-camera 3D person tracking with particle filter in a surveillance environment. In *EUSIPCO*, 2008. 2

[13] K. Sankaranarayana, M.-C. Chang, and N. Krahnstoever. Tracking gaze direction from far-field surveillance cameras. In *IEEE Workshop on Applications of Computer Vision and Applications*, 2011. 2

[14] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010. 2

[15] M. J. Jones and J. M. Rehg. Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision*, 46(1):81–96, 2002. 3

[16] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, 2006. 2

[17] T. Lee and S. Soatto. learning and matching multiscale template descriptors for real-time detection, localization and tracking. In *CVPR*, 2011. 1

[18] T. Yu, S. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *CVPR*, 2009. 1