We are not Contortionists: Coupled Adaptive Learning for Head and Body Orientation Estimation in Surveillance Video

Cheng Chen
cchen@idiap.ch

Jean-Marc Odobez
odobez@idiap.ch

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

Abstract

In this paper, we deal with the estimation of body and head poses (i.e orientations) in surveillance videos, and we make three main contributions. First, we address this issue as a joint model adaptation problem in a semi-supervised framework. Second, we propose to leverage the adaptation on multiple information sources (external labeled datasets, weak labels provided by the motion direction, data structure manifold), and in particular, on the coupling at the output level of the head and body classifiers, accounting for the restriction in the configurations that the head and body pose can jointly take. Third, we propose a kernel-formulation of this principle that can be efficiently solved using a global optimization scheme. The method is applied to body and head features computed from automatically extracted body and head location tracks. Thorough experiments on several datasets demonstrate the validity of our approach, the benefit of the coupled adaptation, and that the method performs similarly or better than a state-of-the-art algorithm.

1. Introduction

A very important task in surveillance environment is the tracking and understanding of human activity. Most of the work so far has concentrated on multi-person tracking [6, 12]. From the location and trajectory information, scene structure understanding, crowd flow tracking, and trajectory abnormality detection can be conducted [8, 24].

While tracking people's location is a first step for activity understanding, there are other cues that one would need to perform a finer analysis of individual or group human behavior [15, 19]. This is the case of the body pose[1] and head pose, which both contribute to the understanding of people attention and can therefore be used in applications related to group/interaction detection, visual attraction analysis, luggage attendance monitoring, and so on [23, 5, 7]. For example, head pose was used in [3] to infer scene interest maps, and in [9] to discover interactions in an office environment.

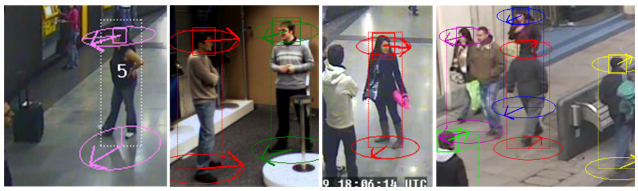

Figure 1. Sample output of our method. Body and head/gaze cues are useful for behavior analysis. They can indicate whether people pay attention to their luggage (1st image), suggest that an interaction is going on (2nd image), or indicate distraction (3rd image).

In this paper we propose an approach for body and head pose estimation in surveillance videos, as examplified in Fig. 1. This is a difficult problem, where pre-trained classifiers usually perform poorly due to low resolution, large variabilities in face and body (or clothing) appearance, combined with differences in view points and illumination. Adaptation is thus necessary, as demonstrated very recently by Benfold and Reid [5] for head pose estimation.

We formulate the body and head pose classification as a semi-supervised learning problem within a kernel framework. However, unlike [5] which only leveraged on the moving directions as weak labels to learn a scene-specific head pose classifier, we also leverage on prior knowledge provided by annotated datasets.

More importantly, our method also exploits the physical constraints that the human head can not rotate beyond some limits with respect to the body, by introducing some coupling between the head and body pose classifier output. In this way, information from head observations and from body observations, whenever available, help to improve not only the classification, but also the adaptation process by reducing the risk of classifier drift. This is particularly relevant when people remain static and no coupling with

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

[1]We use body pose to refer to the upper-body orientation in the ground plane rather than the articulated spatial configuration of the human body.

the moving direction can be exploited. To the best of our knowledge, such a coupling has never been exploited for adaptation purposes.

Our key contributions are:

- a semi-supervised learning framework for coupled adaptive classifier learning, which considers label information, manifold structure, and classifier coupling;
- a kernelized formulation of the framework that has an efficient non-iterative global optimization scheme;
- the application of our algorithm for joint head and body estimation in surveillance data, outperforming a state-of-the art head pose estimation algorithm.

The proposed learning method is different from previous semi-supervised algorithms, such as co-training (which assumes the classifiers perform the same task), multi-task learning (which assumes the features for different task lie in homogeneous space), or multi-view learning (which does not exploit the manifold structure), as detailed in Section 2. Furthermore, it is applied to a problem that has not been explored before by those techniques. Finally, note that although this paper deals with body and head pose estimation, the proposed coupled adaptive learning algorithm can be applied to any other application with coupled tasks.

Thorough experiments on several public and non-public databases demonstrate the validity of our approach and the benefit of the coupling during adaptation in comparison with traditional coupled filtering methods.

The rest of the paper is organized as follows. Related work is discussed in Section 2. Section 3 summarizes our approach, and describes the person and head tracking algorithms, as well as our head and body features. The coupled classifier adaptation method is more thoroughly described in Section 4. It is followed by experiments and discussions in Section 5, and the conclusion.

2. Related work

Coupled classifier adaptation. The coupled adaptation scheme we propose share similarities with semi-supervised methods [25] that leverage on unlabeled data to improve supervised classification. Below we review three related categories of approaches in this general framework, namely multi-task learning, co-training, and multi-view learning, and point out the differences of our work.

Multi-task learning [1, 18] jointly learns several classifiers on different but related tasks, e.g. learning jointly horse and cow classifiers. It usually assumes that both the feature and classifier parameter spaces are the same for all tasks, so that the task similarity can be modeled by imposing similarities between the classifier parameters. In our case, however, body and head features lie in different spaces, and similarity is enforced at the output.

In co-training [2, 11] two classifiers are learned on the same task using unlabeled data. Samples confidently classified by one classifier are used to update the other. This label propagation process is often iterative. Differences in our method are: first, the label dependency is directly encoded into the joint objective function which is solved using a more efficient non-iterative optimization process; second, body and head pose estimation are dependent but nonetheless different tasks. The dependency is enforced by soft coupling rather than hard constraints.

Our coupled adaptation training method is also related to multi-view learning, in which an item is assumed to be sensed by multiple views (e.g. modalities) upon each of which a classifier is trained [22, 21]. As with co-training, most multi-view learning algorithms assume that all classifiers solve the same task, i.e. the multi-view data corresponds to the same label, which differs from our softly coupled multi-task problem. Furthermore, our method is more general as it leverages not only on the inter-cue coupling (e.g. as in [22]) but also on the intra-cue manifold structure (the adaptation component of our approach).

Head and body pose estimation. Due to its potential as attention and social cue, head pose estimation in surveillance scenarios has recently become an important research topic, [17, 13, 4, 10]. For instance, as a pioneer work, [17] proposed to estimate head poses into 8 directions using visual features based on skin detection. In [13], head pose is estimated using an SVM classifier and the mean appearance model at different poses. Besides building classifiers, authors like in [17] also investigated the coupling of head pose and speed direction. However, classifier adaptation was not addressed, with the recent exception of [5] that performs scene level adaptation. Also, none of the above work exploited body pose features.

Although full body pose estimation in smart room settings has received some attention [26], very few works have addressed body pose estimation in surveillance settings. Several of them introduced body orientation as a link between the head pose and body movement cues, [17, 16], but without exploiting body pose related features. This approach is problematic when a person does not move, as the velocity becomes too noisy to provide reliable information for body pose (and ultimately head pose) estimation. This contrasts with the work in [7], which uses multi-level HOG body features and sparse representation in a temporal filtering framework to estimates body orientation. The work in [20] estimates body pose, but relies on 3D space carving in a multi-camera set-up not available in most surveillance systems. In all cases (except [20]), classifier adaptation has not been addressed.

3. Method overview, and feature extraction

Fig. 2 describes the overall scheme. Given a video, we first apply multi-person and head location tracking algorithms. Multi-person localization is conducted with a track-

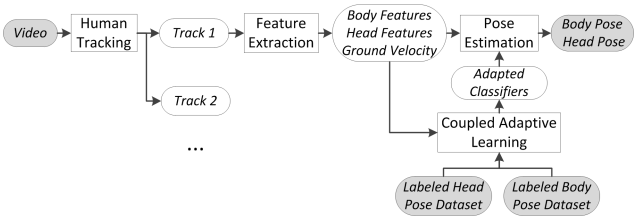

Figure 2. Workflow of our approach.

ing by detection framework relying on a Conditional Random Field method similar to [6], while head tracking is performed as described in Section 3.1. For each resulting track, the sequences of head and body features (described in 3.2) as well as ground velocities are extracted and used along with external labeled datasets to learn pose classifiers within our coupled adaptive learning scheme. The learned classifiers are then used to predict body and head poses.

3.1. Tracking head locations

Accurate head pose estimates rely on precise head image localization. This is achieved using a robust tracking-by-detection method. For each person, we first perform head detection in each frame around the upper part of the body bounding box using a pre-trained Histogram-of-Gradient (HoG) based SVM head detector. Due to noise, there might be no, wrong, or multiple responses. These detections are then filtered by finding an optimum sequence of locations. This is achieved by building a graph and finding the most probable path in it, as illustrated in Fig. 3.

Graph. Each detection response v_i, whose frame index is t_i, is a node in the graph, and detections that are close enought in time are connected. More precisely, an edge e_{ij} connecting v_i and v_j is created if and only if $1 <= t_j - t_i <= \Delta T$[2]. In addition, we introduce a source node v_{so} and a sink node v_{si}.

Tracking. Given the graph, we define a transition probability term for each edge e_{ij} according to:

$$p_{ij} = p_{\text{siz}}(v_j|v_i)\, p_{\text{loc}}(v_j|v_i)\, p_{\text{app}}(v_j|v_i)\, p_{\text{mis}}(v_j|v_i) \quad (1)$$

where the different terms are described below. The head tracking problem then consists of finding the path $H = (v_{so}, ..., v_{si})$ from v_{so} to v_{si} with maximum probability under Markov assumption, i.e. find the maximum $p(H) = \prod_{e_{ij} \in H} p_{ij}$. By defining $-\log(p_{ij})$ as the cost of edge e_{ij}, the problem is equivalent to finding a shortest path from v_{so} to v_{si}. Thanks to the different terms involved in Eq. 1, we were able to reliably and efficiently track the head locations in the presence of wrong or miss detections.

Probability terms. They favor the head tracking continuity using the following principles:

– p_{siz} favors the continuity in scale, i.e. it is higher when the scales of v_j and v_i are similar.

[2]We use ΔT large enough to cover gaps, which works well in practice.

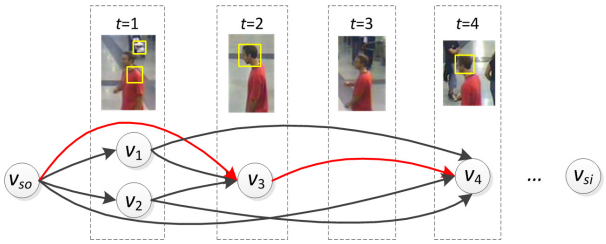

Figure 3. Head tracking from multiple, wrong and miss head detections. Example of a tracking graph, where nodes are detection responses, and lines are edges. The optimal path is drawn in red.

– p_{app} favors continuity in appearance, i.e. it is higher if the head patches represented by v_j and v_i have similar appearance. We use cross correlation as similarity measure.

– p_{loc} addresses location similarity, and contains two terms: the first performs a local template tracking of the region v_i from frame t_i to frame t_j, and compares the predicted location in t_j with v_j, The second term simply measures the distances between detections .

–p_{mis} considers the miss detection rate, where each link that skips some frames is penalized by the probability that there is no correct head detection in the skipped frames. This allows connection (with some penalty) between two detection responses not on successive frames, allowing the skip of frames where there are no detections or all detections are wrong (e.g. frame $t = 1$ and frame $t = 3$ in Fig. 3).

3.2. Body and head pose features

Given the body and head regions, we extract multi-level HoG features. For the body, we use three levels (1×3, 2×6 and 4×12). For the head, we use two levels (2×2 and 4×4). Each block is divided into 2×2 cells, and for each cell we construct HoG with nine unsigned bins. In this way, we end up with a $d_b = 2268$ dimensional body feature vector, and a $d_h = 720$ dimensional head feature vector.

4. Coupled Adaptive Classifier Learning

In this section we first provide the overview and main principles of our method, and then describe in more details the different terms involved in the model.

4.1. Approach Overview

We first present the different datasets involved in the algorithm as well as some notations. Then we introduce the task, problem formulation, and modeling strategy.

Data. Let us denote by $\mathcal{D}_b = \left\{\left(\mathbf{x}_i^b, \mathbf{y}_i^b\right), i = 1..n_b\right\}$ the prior labeled dataset for body pose, where $\mathbf{x}_i^b \in \mathbb{R}^{d_b}$ is the body feature, and $\mathbf{y}_i^b \in \{0,1\}^{d_{disc}}$ denotes the ground-truth pose label. As we formulate our problem as a discrete classification problem, we represent the ground-truth pose angle belonging to the $j^{\text{th}} (1 \leq j \leq d_{disc})$ class as a

d_{disc} dimensional binary vector, where all but the j^{th} element are zero. Currently, we use $d_{disc} = 8$ orientations for both the body and head poses, but finer quantization (and different ones for body and head) could be used as well. Similarly, we define the prior labeled dataset for head pose $\mathcal{D}_h = \left\{\left(\mathbf{x}_i^h, \mathbf{y}_i^h\right), i = 1..n_h\right\}$.

For adaptation, we also have an unlabeled target dataset $\mathcal{D}_t = \left\{\left(\tilde{\mathbf{x}}_i^b, \tilde{\mathbf{x}}_i^h, \mathbf{v}_i, u_i\right), i = 1..n_t\right\}$, where $\tilde{\mathbf{x}}_i^b$ and $\tilde{\mathbf{x}}_i^h$ are the body and head features, $\mathbf{v}_i \in \{0,1\}^8$ denotes the velocity direction expressed in the label space, and $u_i \in \{0,1\}$ is a binary flag indicating whether the velocity magnitude is large enough (we use 3 km/h as threshold). This dataset is unlabeled, but all observations are synchronized (they are extracted from the same person in the same frame).

In addition, we denote $\mathbf{z}_i^b$ as the (both labeled and unlabeled) body features in $\{\mathcal{D}_b, \mathcal{D}_t\}$, i.e. $\mathbf{z}_i^b = \mathbf{x}_i^b$ when $i \leq n_b$ and $\mathbf{z}_i^b = \tilde{\mathbf{x}}_{i-n_b}^b$ when $i > n_b$. We also define $\mathbf{z}_i^h$ similarly.

Task. Our goal is to learn body pose and head pose classifiers which are adapted to the target data, by leveraging on multiple information sources (labeled data $\mathcal{D}_b$ and $\mathcal{D}_h$, test data $\mathcal{D}_t$, coupling between classifier outputs or with the weak velocity direction labels). We denote by $f^b : \mathbb{R}^{d_b} \rightarrow \mathbb{R}^8$ the body classifier function we want to learn, and similarly f^h for the head classifier[3].

Problem formulation. We adopt a kernel-based view. Assume that there is a non-linear mapping $\phi^b : \mathbb{R}^{d_b} \rightarrow \mathcal{F}^b$ that maps the body feature to a high dimensional (possibly infinite) Reproducing Kernel Hilbert Space $\mathcal{F}^b$. According to the Representer Theorem, for any $\mathbf{x}^b$, $f^b(\mathbf{x}^b)$ is linear with regard to its inner product with the data samples in $\mathcal{F}^b$:

$$\begin{aligned} f^b\left(\mathbf{x}^b\right) &= \sum_{i=1}^{n_b+n_t} \mathbf{w}_i^b \left(\phi^b\left(\mathbf{z}_i^b\right)\right)^\mathsf{T} \phi^b\left(\mathbf{x}^b\right) \\ &= \left(\mathbf{W}^b\right)^\mathsf{T} \left[\mathbf{\Phi}^b, \tilde{\mathbf{\Phi}}^b\right]^\mathsf{T} \phi^b\left(\mathbf{x}^b\right), \text{ with} \end{aligned} \tag{2}$$

$\mathbf{\Phi}^b = \left[\phi^b(\mathbf{x}_1^b), .., \phi^b(\mathbf{x}_{n_b}^b)\right]$, $\tilde{\mathbf{\Phi}}^b = \left[\phi^b(\tilde{\mathbf{x}}_1^b), .., \phi^b(\tilde{\mathbf{x}}_{n_t}^b)\right]$, and $\mathbf{W}^b = \left[\mathbf{w}_1^b, .., \mathbf{w}_{n_b+n_t}^b\right]^\mathsf{T} \in \mathbb{R}^{(n_b+n_t)\times 8}$. Given a kernel function $k\left(\mathbf{x}_i^b, \mathbf{x}_j^b\right) = \phi^b\left(\mathbf{x}_i^b\right)^\mathsf{T} \phi^b\left(\mathbf{x}_j^b\right)$, learning f^b reduces to the learning of the weight parameters $\mathbf{W}^b$. The classifier $f^h\left(\mathbf{x}^h\right)$ has a similar form as in Eq. (2), with parameters $\mathbf{W}^h$.

Modeling Strategy. Our goal is thus to learn the set of weights $\mathbf{W} = \left[(\mathbf{W}^b)^\mathsf{T}, (\mathbf{W}^h)^\mathsf{T}\right]^\mathsf{T}$. To this end, we design an objective function $E(\mathbf{W})$ that takes into account several factors, as explained below:

• *Label information factor* E_l. The classifier functions should respect the label information encoded in $\mathcal{D}_b$ and $\mathcal{D}_h$.

• *Manifold structure factor* E_m. The classifier functions should be smooth over the manifold structure encoded by both labeled and unlabeled body (or head) features in $\{\mathcal{D}_b(\text{or}\mathcal{D}_h), \mathcal{D}_t\}$, i.e. similar features should generate similar labels.

• *Body and head pose coupling factor* E_c^{bh}. In the target data $\mathcal{D}_t$, body pose and head pose tend to be aligned due to anatomical constraints and the fact that people tend to look into the same direction as that faced by their body. Thus, on this data, the output of the head and body pose classifiers should be similar.

• *Body pose and velocity coupling factor* E_c^{vb}. When people are moving, their body tend to be oriented in the moving direction. For the target data $\mathcal{D}_t$, when the velocity is large enough, its direction can thus be used as a weak label for body pose.

• *Regularisation factor* E_r. We want to control the complexity of $\mathbf{W}$ for better generality.

Ultimately, the objective function is thus defined by:

$$E\left(\mathbf{W}\right) = E_l + \alpha E_m + \beta E_c^{bh} + \gamma E_c^{vb} + \lambda E_r, \tag{3}$$

where $\alpha, \beta, \gamma, \lambda$ are (non-negative) parameters, and the specific expressions for each factor are given below.

4.2. Objective function factors

Label factor E_l**.** For body pose, we define E_l^b as the discrepancy between the output of the classifier f^b and the label measured on the labeled dataset $\mathcal{D}_b$:

$$\begin{aligned} E_l^b &= \frac{1}{n_b} \sum_{i=1}^{n_b} \left\| \left(\mathbf{M} f^b\left(\mathbf{x}_i^b\right) - \mathbf{M}\mathbf{y}_i^b\right) \right\|_F^2 \\ &= \frac{1}{n_b} \left\| \mathbf{M}\left(\left(\mathbf{W}^b\right)^\mathsf{T} \mathbf{K}^b - \mathbf{Y}^b\right) \right\|_F^2 \end{aligned} \tag{4}$$

where $\mathbf{Y}^b = \left[\mathbf{y}_1^b, ...\mathbf{y}_{n_b}^b\right]$, $\mathbf{K}^b = \left[\mathbf{\Phi}^b, \tilde{\mathbf{\Phi}}^b\right]^\mathsf{T} \mathbf{\Phi}^b$ is the kernel matrix, and $\mathbf{M} \in \mathbb{R}^{8\times 8}$ is the label smoothing matrix[4]. A similar expression can be obtained for E_l^h by changing the superscripts $\square^b$ to $\square^h$ in Eq. (4). By defining $\mathbf{K}_l = \text{diag}\left(\mathbf{K}^b, \mathbf{K}^h\right)$, $\mathbf{Y} = \left[\mathbf{Y}^b, \mathbf{Y}^h\right]$, and $\mathbf{O} = \text{diag}\left(\mathbf{I}_{n_b}/n_b, \mathbf{I}_{n_h}/n_h\right)$, the expression for the label term $E_l = E_l^b + E_l^h$ is given by[5]:

$$E_l = \text{Tr}\left(\mathbf{M}\left(\mathbf{W}^\mathsf{T}\mathbf{K}_l - \mathbf{Y}\right)\mathbf{O}\left(\mathbf{W}^\mathsf{T}\mathbf{K}_l - \mathbf{Y}\right)^\mathsf{T}\mathbf{M}^\mathsf{T}\right) \tag{5}$$

Manifold factor E_m**.** For all body features in $\{\mathcal{D}_b, \mathcal{D}_t\}$, we construct a similarity matrix $\mathbf{S}^{b\tilde{b}} \in \{0,1\}^{(n_b+n_t)\times(n_b+n_t)}$, where $s_{ij}^{b\tilde{b}} = 1$ iff $\mathbf{z}_i^b$ is the k nearest neighbors of $\mathbf{z}_j^b$ or vice-versa. We define E_m^b as the violation of this similarity

[3] The classifiers are trained using labels in $\{0,1\}^8$ with only one non-zero component. However, in practice the classifier outputs are real-valued 8 dimensional vectors with each dimension reflecting classification score in each class. Post-processing is applied to transform the classifier output into a angular output. This will be described in Section 5.1.

[4] We use $\mathbf{M} = \begin{bmatrix} 11000001 \\ ... \\ 10000011 \end{bmatrix}$. It is to "diffuse" the label, posing less penalty on adjacent misclassifications (e.g. classifying "left" as "left-front") than complete mistakes (e.g. classifying "left" as "right").

[5] Note the property $\|\mathbf{A}\|_F^2 = \text{Tr}\left(\mathbf{A}\mathbf{A}^\mathsf{T}\right) = \text{Tr}\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)$ is used here.

at the output of f^b, i.e. we impose a large penalty if $\mathbf{z}_i^b$ and $\mathbf{z}_j^b$ are similar but their predicted poses are not:

$$\begin{aligned} E_m^b &= \frac{1}{\sum_{i\neq j} s_{ij}^{b\tilde{b}}} \sum_{i\neq j} s_{ij}^{b\tilde{b}} \left\| f^b\left(\mathbf{z}_i^b\right) - f^b\left(\mathbf{z}_j^b\right)\right\|_F^2 \\ &= 2\,\mathrm{Tr}\left(\mathbf{M}\left(\mathbf{W}^b\right)^\mathsf{T}\mathbf{K}^{b\tilde{b}}\mathbf{L}^{b\tilde{b}}\left(\mathbf{K}^{b\tilde{b}}\right)^\mathsf{T}\mathbf{W}^b\mathbf{M}^\mathsf{T}\right) \end{aligned} \tag{6}$$

where $\mathbf{K}^{b\tilde{b}} = [\mathbf{\Phi}^b, \tilde{\mathbf{\Phi}}^b]^\mathsf{T}[\mathbf{\Phi}^b, \tilde{\mathbf{\Phi}}^b]$ is the kernel matrix, and $\mathbf{L}^{b\tilde{b}}$ is the (trace normalized) Laplacian matrix of $\mathbf{S}^{b\tilde{b}}$. Defining $\mathbf{K}_m = \mathrm{diag}(\mathbf{K}^{b\tilde{b}}, \mathbf{K}^{h\tilde{h}})$, and $\mathbf{L}_m = \mathrm{diag}(\mathbf{L}^{b\tilde{b}}, \mathbf{L}^{h\tilde{h}})$, we have:

$$E_m = E_m^b + E_m^h = \mathrm{Tr}\left(\mathbf{M}\mathbf{W}^\mathsf{T}\mathbf{K}_m\mathbf{L}_m(\mathbf{K}_m)^\mathsf{T}\mathbf{W}\mathbf{M}^\mathsf{T}\right) \tag{7}$$

Body and head pose coupling factor E_c^{bh}. It is defined as the discrepancy between the body pose and head pose classifier outputs over $\mathcal{D}_t$:

$$\begin{aligned} E_c^{bh} &= \frac{1}{n_t}\sum_{i=1}^{n_t}\left\|\mathbf{M}f^b\left(\tilde{\mathbf{x}}_i^b\right) - \mathbf{M}f^h\left(\tilde{\mathbf{x}}_i^h\right)\right\|_F^2 \\ &= \frac{1}{n_t}\left\|\mathbf{M}\left(\mathbf{W}^b\right)^\mathsf{T}\mathbf{K}^{\tilde{b}} - \mathbf{M}\left(\mathbf{W}^h\right)^\mathsf{T}\mathbf{K}^{\tilde{h}}\right\|_F^2 \\ &= \mathrm{Tr}\left(\mathbf{M}\mathbf{W}^\mathsf{T}\mathbf{K}_{c1}(\mathbf{K}_{c1})^\mathsf{T}\mathbf{W}\mathbf{M}^\mathsf{T}\right) \end{aligned} \tag{8}$$

where $\mathbf{K}^{\tilde{b}} = [\mathbf{\Phi}^b, \tilde{\mathbf{\Phi}}^b]^\mathsf{T}\tilde{\mathbf{\Phi}}^b$ and $\mathbf{K}^{\tilde{h}} = [\mathbf{\Phi}^h, \tilde{\mathbf{\Phi}}^h]^\mathsf{T}\tilde{\mathbf{\Phi}}^h$ are corresponding kernel matrices, and $\mathbf{K}_{c1}^\mathsf{T} = \frac{1}{\sqrt{n_t}}[(\mathbf{K}^{\tilde{b}})^\mathsf{T}, -(\mathbf{K}^{\tilde{h}})^\mathsf{T}]$. Note that all samples from $\mathcal{D}_t$ contributes to this term[6].

Body and velocity coupling factor E_c^{vb}. It is defined as the discrepancy between the body pose and the velocity direction, provided the velocity magnitude is large enough:

$$\begin{aligned} E_c^{vb} &= \frac{1}{\sum u_i}\sum_{i=1}^{n_t} u_i\left\|\left(\mathbf{M}f^b\left(\tilde{\mathbf{x}}_i^b\right) - \mathbf{M}\mathbf{v}_i\right)\right\|_F^2 \\ &= \left\|\mathbf{M}\left(\left(\mathbf{W}^b\right)^\mathsf{T}\mathbf{K}^{\tilde{b}} - \mathbf{V}\right)\mathbf{U}^{\frac{1}{2}}\right\|_F^2 \\ &= \mathrm{Tr}\left(\mathbf{M}\left(\mathbf{W}^\mathsf{T}\mathbf{K}_{c2} - \mathbf{V}\right)\mathbf{U}\left(\mathbf{W}^\mathsf{T}\mathbf{K}_{c2} - \mathbf{V}\right)^\mathsf{T}\mathbf{M}^\mathsf{T}\right) \end{aligned}$$

where $\mathbf{V} = [\mathbf{v}_1, ...\mathbf{v}_{n_t}]$, $\mathbf{U} = \mathrm{diag}\left(u_1, ..., u_{n_t}\right)/\sum u_i$, and $\mathbf{K}_{c2}^\mathsf{T} = [(\mathbf{K}^{\tilde{b}})^\mathsf{T}, \mathbf{0}^{n_t\times(n_h+n_t)}]$. Note that due to u_i, only samples with large speed contributes to this term.

Regularization factor E_r: It is simply defined as:

$$E_r = \mathrm{Tr}\left(\mathbf{W}^\mathsf{T}\mathbf{W}\right). \tag{9}$$

4.3. Optimization

It can be shown that our objective function Eq. (3) is convex and thus we have to solve:

$$\frac{\partial E\left(\mathbf{W}\right)}{\partial \mathbf{W}} = 2\mathbf{G}\mathbf{W}\mathbf{M}^\mathsf{T}\mathbf{M} + 2\lambda\mathbf{W} - 2\mathbf{H} = \mathbf{0} \tag{10}$$

where

$$\begin{aligned} \mathbf{G} =& \mathbf{K}_l\mathbf{O}(\mathbf{K}_l)^\mathsf{T} + \alpha\mathbf{K}_m\mathbf{L}_m(\mathbf{K}_m)^\mathsf{T} \\ &+\beta\mathbf{K}_{c1}(\mathbf{K}_{c1})^\mathsf{T} + \gamma\mathbf{K}_{c2}\mathbf{U}(\mathbf{K}_{c2})^\mathsf{T} \end{aligned} \tag{11}$$

$$\mathbf{H} = \mathbf{K}_l\mathbf{O}\mathbf{Y}^\mathsf{T}\mathbf{M}^\mathsf{T}\mathbf{M} + \gamma\mathbf{K}_{c2}\mathbf{U}\mathbf{V}^\mathsf{T}\mathbf{M}^\mathsf{T}\mathbf{M} \tag{12}$$

Eq. (10) is a *Sylvester equation* with a closed-form solution.

5. Experiments

5.1. Experimental protocol

Data. We use the TUD Multiview Pedestrians dataset [14] and the Benfold dataset [3] as the labeled prior datasets $\mathcal{D}_b$ and $\mathcal{D}_h$ for all experiments. We used the tracks of four other datasets as $\mathcal{D}_t$ for evaluation: (1) the **CHIL dataset** comes from the CLEAR 2007 head pose estimation contest. It features an indoor scenario. We used the 4 available subjects for our experiments. For head pose, we used the ground-truth (GT) distributed with the data[7] and obtained from a magnetic sensor, while the body pose GT was annotated by us. (2) the **MetroStation dataset** contains several clips from a surveillance camera in a metro station. We manually annotated the GT body and head poses. (3) the **Indoor dataset** contains clips captured from an indoor surveillance camera. GT annotation was done manually for both body and head poses. (4) the **TownCentre dataset** is provided by [5]. The data comes with tracking output for body and head (that we used as input), but no pose information. Therefore, we manually annotated 15 tracks for evaluation purpose in this paper. In total, the above datasets contains over 20 minutes video with 25 persons for quantitative evaluation.

Performance measure. The performance is evaluated by the average angular error between the GT and predicted pose angles. Note that we need to transform each 8 dimensional classification output $\{o_i, i = 1..8\}$ into an angle, where o_i can be interpreted as classification score for the angle θ_i. To this end, we used the angle of the weighted average vector $\sum_{i=1}^8 o_i\vec{n}_{\theta_i}$, where $\vec{n}_{\theta_i}$ denotes the unit vector associated with θ_i

Algorithms. To evaluate, and understand the benefits of the different components of the approach, we tested several algorithms. The **Proposed (default):** algorithm corresponds to our full coupled adaptive learning method with default parameters: $\alpha = 1$, $\beta = 0.5$, $\gamma = 0.5$, $\lambda = 0.01$, and a Laplacian kernel $k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\sqrt{\delta}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right)$ with $\delta = 10$. The **Proposed (baseline)** corresponds to our approach without adaptation and coupling ($\alpha = \beta = \gamma = 0$). The **Proposed (no velocity)** is the same as the default approach, but without using the velocity coupling (i.e. $\gamma = 0$). The **Proposed (no prior data)** method does not exploit the prior labeled information (E_l is dropped during optimization), and relies only the coupling between head and body

[6] Here we assume the body and head features for all data in $\mathcal{D}_t$ are valid. In practice, for better handling of occlusion, we can introduce a binary flag to exclude data with occlusion, similar to the flag u_i in E_c^{vb}.

[7] The dataset also provides ground-truth head locations in the images. However, we do not use them. We use our head location tracking method.

Methods	CHIL	MetroStation	Indoor	TownCentre
Proposed (default)	35.3 / 36.0	29.4 / 30.0	23.6 / 23.6	17.4 / 18.4
Proposed (baseline)	50.7 / 56.9	53.8 / 40.5	59.9 / 29.4	48.1 / 44.8
Proposed (no velocity)	35.3 / 36.0	31.3 / 30.1	23.4 / 24.0	26.5 / 27.6
Proposed (no prior data)	80.7 / 85.1	63.5 / 66.7	63.9 / 68.2	18.3 / 19.4
Proposed (Benfold setting [5])	80.7 / 85.1	82.2 / 85.4	63.5 / 66.7	18.4 / 20.5
Walking direction	78.7 / 79.5	79.9 / 77.1	66.3 / 66.7	19.3 / 22.9
TF with coupling [7, 16]	44.5 / 46.7	42.2 / 40.5	36.3 / 33.8	20.1 / 24.9
Proposed + TF	37.7 / 35.2	32.8 / 31.0	24.9 / 23.9	19.0 / 25.0

Table 1. Evaluation on several datasets. Each cell contains two numbers (body pose error/head pose error). All errors are in degree.

pose classifiers. Finally, the **Proposed (Benfold setting [5])** only relies on a coupling between velocity and the head direction (i.e. without coupling with the body pose and prior data) as was done in [5].

For comparison purposes, we also report other alternatives. The **Walking direction** baseline uses the walking direction as the body and head pose output. The **Temporal Filtering (TF) with coupling** method is similar to [7] and [16]. It relies on a particle filtering framework which considers intra-cue temporal smoothness and inter-cue dependencies, (i.e. the coupling between velocity, body pose and head pose is exploited in the dynamical model), and likelihood models built from the labeled data only (without adaptation). Finally, the **Proposed + TF** approach corresponds to a refinement of our proposed method, by applying the TF step just described but using the adapted classifiers for the head and body pose likelihoods.

5.2. Results

The results are shown in Table 1. We can make several comments. First, the method we propose performs the best in all cases. Comparing with our proposed baseline that only relies on labeled data, we see that the coupled adaptive learning contributes largely to the significant improvement, demonstrating the need for adaptation to leverage the gap between training and testing data. Indeed, just introducing the coupling at the filtering level [7, 16] does improve the result compared to the baseline, but much less than through adaptation. We can also notice that adding this coupled filtering step on top of our approach (cf "proposed+TF") does not further improve the results, since the coupling has already been exploited, and the intra-cue temporal smoothness is implicit encoded in the manifold structure term E_m, which requires that similar features generate similar poses, and people appearance changes smoothly temporally.

Velocity coupling. Two couplings are exploited during adaptation: head and body pose output consistency, and velocity direction. In absence of velocity information ("no velocity" results), our method still performs much better than the baseline, and with only slight degradation with respect to using the velocity. Indeed, the level of degradation also reflects the different dataset/scene types. In TownCenter, people mainly move straight in the street and dominantly look in the moving direction[8]. In this case, the motion direction is a good prediction of head and body pose ("Walking direction" results), and can reliably be exploited: we thus obtain a significant error reduction gain of 9^o when using it. When people are static (i.e. not moving forward), e.g. while waiting or during interaction, the gain using velocity is marginal, showing that most of the improvement is due to the coupling between the head and body.

Comparison with [5]. In their setting, [5] uses only the coupling with velocity and no prior data to infer the head pose. The "Benfold setting" algorithm reproduce this situation (separately for head and body pose) using our method. As can be seen, it perform much worse than our method on the first three datasets. Even in TownCenter, where people keep moving and mainly look in their moving direction our method still provides a gain of 2^o for head pose estimation. Note that on TownCenter, [5] reports an average error of 23.9^o for the head pose (and 25.9^o when using the walking direction), but we can not make a direct comparison since their annotations are not available and probably differ from ours. Our belief is that, for head pose estimation, both the body pose and velocity provide complementary information (and at different instants). Body information may not always be availbale due to occlusion, and similarly for velocity when people don't move.

Prior dataset. The "no prior" results show that using prior data is important for adaptation and that relying only on the walking direction to provide some weak labels is not sufficient when the amount of data is smaller, or people are not moving a lot and therefore look more around. In the opposite case (TownCenter), the benefit of using the prior data is reduced, and we obtain only a 1^o gain.

Qualitative results. Figs. 4 to 7 show some results on each of the evaluated datasets (more results are shown in the supplementary material). In each image, the rectangles show

[8]The proportion of data points with reliable velocity orientation is 73% in this case, as compared with 0%, 24%, and 5% respectively for CHIL, MetroStation, and Indoor.

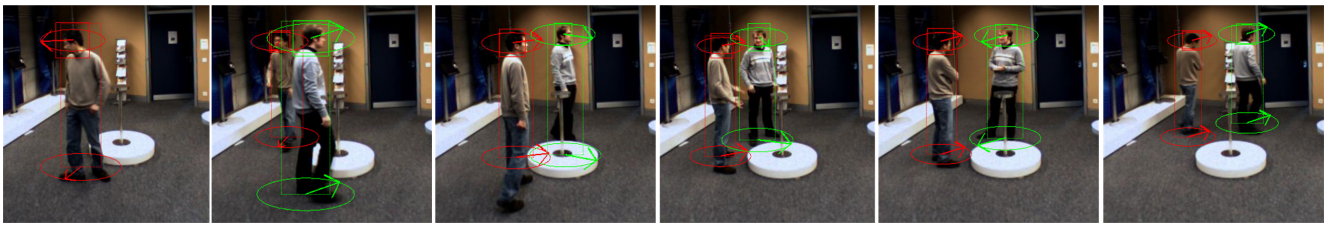

Figure 5. Output of our method (default parameter) on the indoor surveillance dataset. Frame size is 640×480.

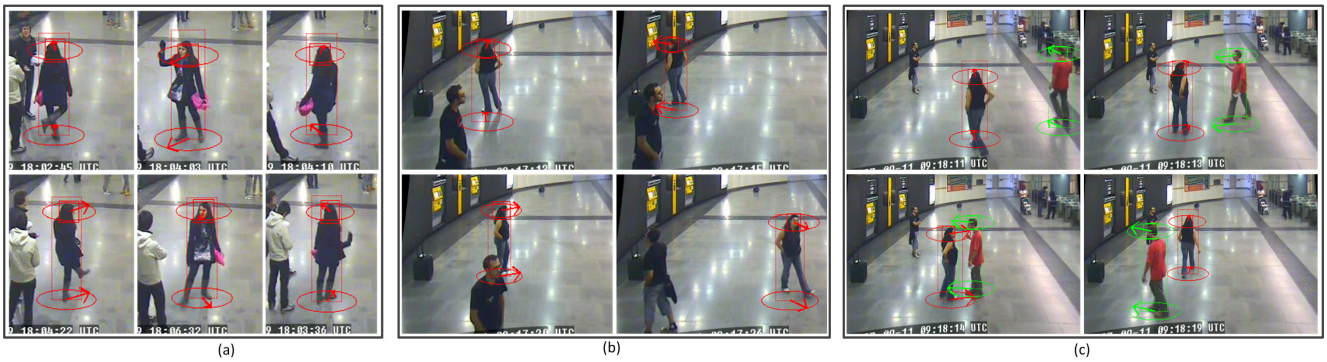

Figure 6. Output of our approach (default parameter) on the metro station dataset. Frame size is 355×288.

Figure 4. Output of our method on CHIL.

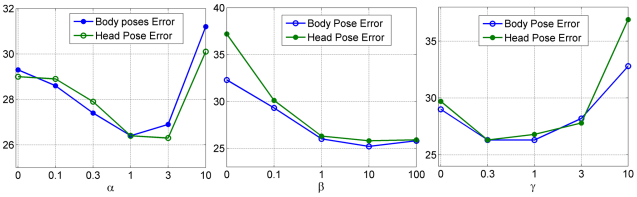

Figure 8. Performance variations with different parameters $\alpha = 1$, $\beta = 0.5$ and $\gamma = 0.5$. Varying α (left), β (middle), and γ (right).

the result of the body and head tracking algorithms. The body and head poses are shown by arrows in ellipses. Despite the difficulty of the task, our method successfully estimates the body and head poses in most cases. Still, some (sometimes large) errors can be spotted as well. Note from these results that our head/body coupling is soft and allows some discrepancy between body and head poses.

Parameter sensitivity analysis. Our learning method has three parameters α, β and γ, which control the importance of manifold structure constraint, coupling between body and head pose, and coupling between velocity and body pose. Here we analyse the performance variation with different parameter values. At each time we try different values for one parameter, leaving the other two to the default values. Fig. 8 reports the results for the three parameters. To avoid cluttering the figure, the error values are averaged

	CHIL	MetroStation	Indoor	TownCentre
No	51 / 57	54 / 41	60 / 30	48 / 44
Ind.	35 / 36	29 / 30	24 / 24	17 / 18
Global	37 / 41	34 / 32	27 / 24	19 / 17

Table 2. Compare adaptation strategy. 1st row: no adaptation. 2nd row: per track individual adaptation. 3rd row: global, using all track jointly. Numbers are rounded to integers to save space.

over all the four datasets we evaluated. We can see that a proper value for each parameter contributes to the results, which justifies the exploitation of different factors in our learning algorithm. For α, the proper value lies around 1 or 3. For β, the performance is similar as soon as its value is above 1. For γ, the proper value is around 0.3 and 1.

Adaptation strategy. So far, our method relied on an "per-track adapt (PTA)" where each single track is used as $\mathcal{D}_t$ to adapt the classifiers. In other words, the model is automatically adapted to each specific person. As an alternative, we could take a "global scene adapt (GSA)" way by using all tracks together in $\mathcal{D}_t$, to adapt the classifier at the scene level. The results of this second approach is shown in Table 2. We see that PTA performs slightly better or comparable compared to GSA. In practice, both strategies have their advantages. In PTA, we get slightly higher accuracy bacause the algorithm is concentrated on one track, which contains "purer" data. On the contrary, in GSA, we get a single estimator for multiple tracks with potentially better generalization ability.

Efficiency. Our method performs learning and adaptation in batch mode. To give an idea of the efficiency, it takes the

Figure 7. Output of our approach (default parameter) on the TownCentre dataset. Frame size is 1920×1080.

same magnitude of time to learn and adapt on a person track as the duration of the track (assuming a 25 fps video).

6. Conclusions

In this paper, we proposed a novel semi-supervised approach for coupled adaptive learning. The method was successfully applied to the joint estimation of body and head pose in surveillance videos, in which the classifier outputs were adapted to exploit multiple information sources. Experiments on several datasets demonstrated the validity of our method and its similar or better performance compared to a recent state-of-the-art head pose estimation approach.

Future work include the extension to multiple cameras, and the modeling of behaviors and interactions using the output of the model.

References

[1] A. Argyriou and T. Evgeniou. Multi-task feature learning. In *NIPS*, 2007. 2

[2] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *ICCV*, 2003. 2

[3] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *BMVC*, 2009. 1, 5

[4] K. Smith, S. Ba, J. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *TPAMI*, 30(7):1212 – 1229, 2008. 2

[5] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *ICCV*, 2011. 1, 2, 5, 6

[6] A. Heili, C. Chen, and J. Odobez. Detection-based multi-human tracking using a CRF model. In *ICCV Workshop VS*, 2011. 1, 3

[7] C. Chen, A. Heili, and J. Odobez. A joint estimation of head and body orientation cues in surveillance video. In *ICCV workshop SISM*, 2011. 1, 2, 6

[8] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008. 1

[9] C-W. Chen, R. C. Ugarte, C. Wu, and H. Aghajan. Discovering social interactions in real work environments. In *FG*, 2011. 1

[10] D. Huang, M. Storer, F. D. Torre, and H. Bischof. Supervised local subspace learning for continuous head pose estimation. In *CVPR*, 2011. 2

[11] F. Tang, S. Brennan, Q. Zhao, and H. Tao. Co-tracking using semi-supervised support vector machines. In *ICCV*, 2007. 2

[12] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. In *ICCV*, 2011. 1

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

[14] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010. 5

[15] M. Cristani, V. Murino, and A. Vinciarelli. Socially intelligent surveillance and monitoring: analysing social dimensions of physical space. In *CVPR Workshop SISM*, 2010. 1

[16] N. Krahnstoever, M-C. Chang, and W. Ge. Gaze and body pose estimation from a distance. In *AVSS*, 2011. 2, 6

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

[18] Q. Liu, X. Liao, H. Li, J. R. Stack, and L. Carin. Semi-supervised multitask learning. *TPAMI*, 31(6):1074 – 1086, 2009. 2

[19] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *TPAMI*, 34(5): 1003 – 1016, 2012. 1

[20] M. Hofmann, and D. Gavrila. Multi-view 3D human pose estimation in complex environment. *IJCV*, 2011. 2

[21] U. Brefeld, T. Gartner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *ICML*, 2006. 2

[22] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *ICML Workshop on Learning with Multiple Views*, 2005. 2

[23] X. Liu, N. O. Krahnstoever, T. Yu, and P. H. Tu. What are customers looking at? In *AVSS*, 2007. 1

[24] X. Song, X. Shao, H. Zhao, J. Cui, R. Shibasaki, and H. Zha. An online approach: learning-semantic-scene-by-tracking and tracking-by-learning-semantic-scene. In *CVPR*, 2010. 1

[25] X. Zhu. Semi-supervised learning literature survey. Technical report, Department of Computer Sciences, University of Wisconsin, Madison., 2005. 2

[26] C. Chen, Y. Yang, F. Nie and J. Odobez. 3D human pose recovery from image by efficient visual feature selection. *CVIU*, 115(3):290 – 299, 2011. 2