IMPROVING HEAD AND BODY POSE ESTIMATION THROUGH SEMI-SUPERVISED MANIFOLD ALIGNMENT

Alexandre Heili[⋆], *Jagannadan Varadarajan*[†], *Bernard Ghanem*[‡], *Narendra Ahuja*[⋆†], *Jean-Marc Odobez*[⋆]

[⋆] Idiap Research Institute, École Polytechnique Fédérale de Lausanne, Switzerland
[†]Advanced Digital Sciences Center, Singapore, [⋆†] University of Illinois at Urbana-Champaign
[‡] King Abdullah University of Science and Technology, Saudi Arabia

ABSTRACT

In this paper, we explore the use of a semi-supervised manifold alignment method for domain adaptation in the context of human body and head pose estimation in videos. We build upon an existing state-of-the-art system that leverages on external labelled datasets for the body and head features, and on the unlabelled test data with weak velocity labels to do a coupled estimation of the body and head pose. While this previous approach showed promising results, the learning of the underlying manifold structure of the features in the train and target data and the need to align them were not explored despite the fact that the pose features between two datasets may vary according to the scene, e.g. due to different camera point of view or perspective. In this paper, we propose to use a semi-supervised manifold alignment method to bring the train and target samples closer within the resulting embedded space. To this end, we consider an adaptation set from the target data and rely on (weak) labels, given for example by the velocity direction whenever they are reliable. These labels, along with the training labels are used to bias the manifold distance within each manifold and to establish correspondences for alignment.

Index Terms— head and body pose, weak labels, manifold, semi-supervised, domain adaptation, surveillance.

1. INTRODUCTION

Video-based human behavior analysis is an important problem studied by the computer vision community. An integral part of any behavior analysis system is tracking and human (body/head) pose estimation. While tracking allows us to infer gross statistics about human movement, human pose gives more fine grained information about human attention and person to person interaction. Due to this rich information content, human pose plays a vital role in several domains such as surveillance, human computer interaction and retail. For instance, head pose was used to study social interactions in [1], customer focus in [2] and to monitor human behavior in surveillance scenes in [3, 4]. Although several works exist for human pose estimation in constrained environments, inferring pose from open spaces and unsconstrained surveillance scenes is quite challenging due to several factors such as poor quality and resolution of surveillance videos, occlusion, cluttered background, appearance changes due to facial geometry, lighting changes, and clothing.

In this paper, we present an improved approach for person independent head and body pose estimation from surveillance videos that incorporates the following aspects: i) we adopt a semi-supervised manifold alignment framework for domain adaptation, where our samples come from both external labelled datasets and unlabelled surveillance clips. For unlabelled samples, we consider the motion direction as a weak label wherever possible, otherwise we use partial annotations within the target data; ii) we introduce a biased manifold embedding term in the alignment framework that allows the manifold distance between two samples to be weighted by the difference in the pose angles too. As a result, we learn a manifold where the neighborhood of a sample is constrained to be samples that are closely related in both the feature and pose angle spaces. We present experiments on two state-of-the-art datasets that validate the effectiveness of our approach.

2. RELATED WORK

Pose estimation in surveillance scenarios has recently garnered some attention. While some approaches consider body and head pose estimation as two separate problems, techniques such as [5][6] explicitly exploit the coupling between these cues, which is a result of anatomical constraints. Besides, in a tracking scenario, moving direction usually gives a strong prior on a person's orientation, which can be leveraged upon to predict pose.

Until recently, one important limitation of existing methods was the use of pre-trained classifiers, that were not adapted to the test data, in spite of obvious appearance variabilities, as well as different viewpoints and illumination. To address this issue, some authors have proposed to perform classifier adaptation. For instance, the authors of [7] leverage on weak labels given by the velocity direction in the test set

This work was done while A. Heili was an intern at ADSC, Singapore. This study is supported by the research grant for the Human Sixth Sense Program at ADSC from Singapore's Agency for Science, Technology and Research (A*STAR). Narendra Ahuja was supported in part by the Office of Naval Research under grant N00014-06-1-0101.

to learn a scene-specific head pose classifier. The method of [8] explores transfer learning approaches for head pose classification in order to transfer knowledge from the source to the target data.

Chen et. al. [6] present an interesting framework, which addresses both the coupling and classifier adaptation aspects described above, and gives state-of-the-art pose estimation performance on several datasets. In their approach, manifold information is used to constrain samples with similar features to be assigned similar pose labels. However, their feature similarity is only based on Euclidean distance between HOG features. We propose to improve their framework by leveraging on pose information as well, when available, so that samples are tightly clustered in the feature and pose spaces. Biased manifold embedding has been proposed by [9] for dimensionality reduction of embedded head features.

Manifold alignment, and especially semi-supervised alignment have been addressed previously [10]. Recently, a least square formulation of manifold alignment has been introduced by [11]. We propose to use such techniques to align the input features fed to the classifiers, by finding a common low-dimensional space over the joint training and test data, using a subset of pairwise correspondences.

3. MODEL OVERVIEW

In this section, we introduce notations and then briefly review the learning framework proposed in [6]. Improvements are described in section 4.

Notations. Let $\mathcal{D}^b = \{(x_i^b, y_i^b), i = 1 \cdots N_b\}$ denote the prior labelled dataset for body pose where, $x_i^b \in \mathbb{R}^{K_b}$ is the feature vector from the body and y_i^b is its corresponding ground truth label vector. Since we formulate our estimation problem as a multi-class classification problem, the label vector $y_i^b \in \{0,1\}^{K_l}$ is a one of K_l vector where, all but the j^{th} element ($1 \leq j \leq K_l$) are zero. A similar treatment applies to our head pose dataset which is indicated by $\mathcal{D}^h = \{(x_i^h, y_i^h), i = 1 \cdots N_h\}$. Our target dataset for adaptation is indicated by $\mathcal{D}^t = \{(\tilde{x}_i^b, \tilde{x}_i^h, v_i, u_i), i = 1 \cdots N_t\}$, where $\tilde{x}_i^b$ and $\tilde{x}_i^h$ are body and head features, respectively, $v_i \in \{0,1\}^{K_l}$ is the motion direction expressed in the label space, $u_i \in \{0,1\}$ is a binary value indicating if the object motion is fast ($\geq$ 3km/h) or not.

Problem definition. Our goal is to learn a multi-class classifier $f^b : \mathbb{R}^{K_b} \rightarrow \mathbb{R}^8$ for body pose and similarly, f^h for head pose by leveraging various information sources, *i.e.*, labelled and unlabelled data, head-body-motion coupling and the fact that samples with similar pose lie closeby in the feature manifold. This is achieved by optimizing an objective function E as follows:

$$E = E_l + \alpha E_m + \beta E_c^{bh} + \gamma E_c^{vb} + \lambda E_r \qquad (1)$$

where the different terms model the following constraints[1]:

[1]In the following, we use $z_i^b = x_i^b$ for $i = 1 \cdots N_b$ and $z_i^b = \tilde{x}_{i-N_b}^b$ for $i = N_b + 1 \cdots N_b + N_t$.

- **Training error term E_l.** The classifier function should have minimum error on labelled training samples $\mathcal{D}^b$ and $\mathcal{D}^h$. This constraint can be encoded by the following function for the body pose:

$$E_l^b = \frac{1}{N_b} \sum_{i=1}^{N_b} ||M f^b(x_i^b) - M y_i^b||_F^2 \qquad (2)$$

  where M is a label smoothing matrix. Similarly, we obtain E_l^h for the head pose, and we have $E_l = E_l^b + E_l^h$.

- **Manifold term E_m.** The classifier function should be smooth over the manifold obtained from labelled and unlabelled samples. In other words, samples close by in the HOG feature space should generate labels that are similar too. To achive this, a binary similarity matrix $S^{b\tilde{b}}$ is constructed by setting $s_{ij}^{b\tilde{b}} = 1$ if z_i^b is one of the k nearest neighbors of z_j^b and 0 if otherwise. E_m^b is then defined as the violation of this similarity in the output.

$$E_m^b = \frac{1}{\sum_{i \neq j} s_{ij}^{b\tilde{b}}} \sum_{i \neq j} s_{ij}^{b\tilde{b}} ||f^b(z_i^b) - f^b(z_j^b)||_F^2 \qquad (3)$$

  Similarly, we obtain E_m^h for the head pose and we have $E_m = E_m^b + E_m^h$.

- **Body and head coupling term E_c^{bh}.** Due to anatomical constraints, we can expect that the head pose is mostly aligned with the body pose in our datasets. E_c^{bh} encodes this by minimizing $||M f^b(z_i^b) - M f^h(z_i^h)||_F^2$ on $\mathcal{D}^t$.

- **Velocity and body coupling term E_c^{vb}.** Similarly, when people are moving, their body pose is oriented in the moving direction. Therefore, in the target data $\mathcal{D}^t$, the body pose is constrained to be mostly aligned with the velocity direction, when $u_i = 1$, i.e. when speed is reliable.

Within a kernel-based framework in which features are mapped to a high dimensional space, learning classifiers f^b and f^h reduces to the learning of weight parameters, whose complexity is controlled by the regularization factor E_r. The non-negative parameters α, β, γ and λ control the effect of the constraints. The objective function of eq. 1 is then convex and has a closed-form solution. We refer to [6] for a more detailed explanation of the model and its optimization.

4. MANIFOLD LEARNING AND ALIGNMENT

The model described above has shown good performance on several datasets, however there are a few drawbacks that we could identify. The cost term in eq. 3 is used so that the classifier predicts similar labels for samples that lie close by in the feature space. But in practice, this assumption could be violated due to several reasons: the model in [6] considers the Histogram of Gradients (HOG) feature space as a smooth manifold of the high dimensional image space and that the training and target set share a common manifold structure despite changes in point of view, illumination, perspective and object size. We claim that the underlying manifold structure

Fig. 1: Illustration of kNN for the same query image (highlighted) in original feature space (first row, 2 mistakes) and in the biased manifold (second row, neighbors have the same pose).

is not exploited well and that it is necessary to align the train and test manifolds at the pre-processing stage. To illustrate this, we retrieved the 5 nearest neighbors (NN) using Euclidean distance for a query sample as shown in Fig. 1 (top row). We see that two of the neighbors have different pose angles w.r.t the query sample indicating the fact that proximity in the HOG feature space does not guarantee proximity in the label space. In order to overcome the above limitation, we propose to first learn more effective manifolds of the high dimensional image space for the training and target sets. This will ensure that neighboring data points in each manifold have similar pose angles. In addition, to align the training and target manifolds, we establish sparse pairwise correspondences between the training set and the target set using pose labels.

We adopt a graph-based manifold learning approach proposed in [10, 12] to learn and align the train and target manifolds. Let $X = \{x_1^b, \ldots, x_{N_b}^b\}$ be our training data and $\tilde{X} = \{\tilde{x}_1^b, \ldots, \tilde{x}_{N_t}^b\}$ be the target data[2]. Let us now define the set of corresponding points for alignment between train and target data by the set of index pairs $I_c = \{(i,j) \mid x_i^b \in X\ ,\ \tilde{x}_j^b \in \tilde{X}$ and x_i^b is in correspondence with $\tilde{x}_j^b\}$. In practice, I_c is determined by using pose information of a subset of test samples and finding for each of them the training sample with the most similar pose. The task is then to learn the linear mappings F and $\tilde{F}$ from the training and test feature spaces X and $\tilde{X}$ into the same embedded space, given the similarity matrices W and $\tilde{W}$ defined on the training and test set, respectively. The dual learning and alignment problem is solved by minimizing the cost function:

$$C(F,\tilde{F}) = \mu \sum_{(i,j)\in I_c} ||F^T x_i^b - \tilde{F}^T \tilde{x}_j^b||^2 + \sum_{i,j} ||F^T(x_i^b - x_j^b)||^2 W_{ij} + \sum_{i,j} ||\tilde{F}^T(\tilde{x}_i^b - \tilde{x}_j^b)||^2 \tilde{W}_{ij} \quad (4)$$

where the first term penalizes discrepancies between F and $\tilde{F}$ on the corresponding pairs, and the second term imposes smoothness of F and $\tilde{F}$ on the respective spaces. Fig. 2 illustrates the results of the joint learning and alignment procedure.

The method requires the knowledge of the similarity matrices W and $\tilde{W}$ between the data points in X and $\tilde{X}$, respectively. KNN with Euclidean distance is typically used to compute the entries of these matrices. In our case, we propose to exploit the label information present in the two datasets to bias the distance between two samples. More precisely, our biased distance $D'(x_i^b, x_j^b)$ between two samples x_i^b and x_j^b is given by:

$$D'(x_i^b, x_j^b) = \left(\tau + \frac{\rho}{1 + e^{r-\delta}}\right) D(x_i^b, x_j^b) \quad (5)$$

where $D(x_i^b, x_j^b)$ is the Euclidean distance in the original feature space and δ is the difference in the pose angles. The bias coefficient is a sigmoid function with parameters τ, ρ and r. We use these parameters to specify the shape of the sigmoid in terms of its upper and lower saturation points, offset, and slope such that the function diminishes the original feature distance when pose differences are within $45°$ and accentuates this distance when pose differences are more that $45°$. In practice, for pose differences more than $90°$ the distance is doubled. The similarity is then computed based on the biased distance, as $W_{ij} = e^{-D'(x_i^b, x_j^b)/\sigma}$ (heat kernel parameterized by σ). Fig. 1 (bottom row) shows the positive effect of using the biased distance.

Method. Within the respective training and test manifold, we compute pairwise distances and use our bias when pose is available. For the samples without pose information, we propagate labels from neighbors within the adaptation set[3], by using a simple kNN technique and majority voting scheme, and use the estimated label to compute the bias. We then simultaneously impose smoothness within each manifold and enforce inter-manifold alignment based on sparse correspondences. Once the projections are learned, we apply them on the features so as to project them on the common manifold where they are aligned and proceed with optimizing eq. 1.

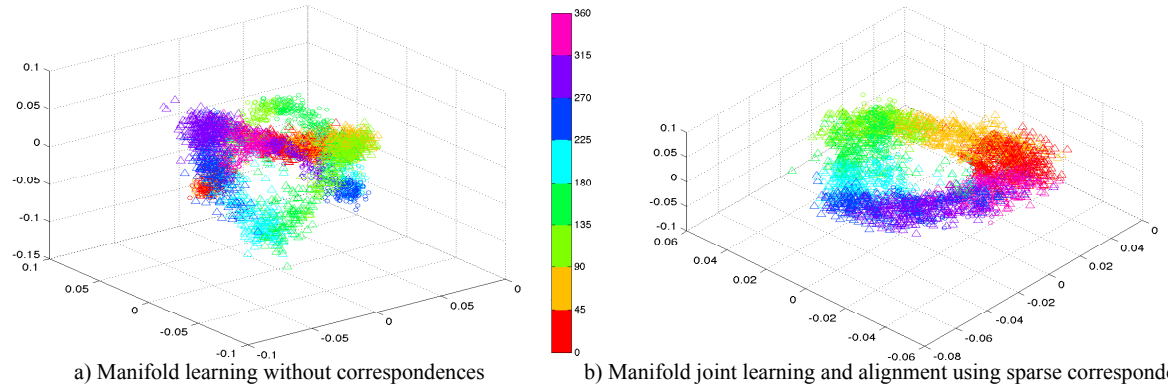
5. EXPERIMENTS

5.1. Experimental Protocol

Datasets. We show the benefits of semi-supervised manifold alignment on pose estimation using two datasets, for which the method of [6] obtains state-of-the-art performance. The **CHIL** dataset [13] contains videos of static people, rotating around a fixed point and moving the head freely. We consider 4 of these subjects for our experiments. The **TownCentre** dataset [7] is a high resolution video of a busy city street, in which we consider the tracks of 15 people. For both datasets, we use the ground truth (head and body pose annotations) provided by [6] for evaluation. Similarly to [6], we use the TUD Multiview Pedestrians dataset [14] and the Benfold dataset

[2] In the following, we detail the procedure for the body feature. The same procedure is applied for the head feature.

[3] We denote by adaptation set the part of the target data that has (weak) labels associated to it and that is used to bias the test manifold and to establish correspondences with the training manifold.



a) Manifold learning without correspondences b) Manifold joint learning and alignment using sparse correspondences

Fig. 2: Illustrations of manifold learning and alignment on CHIL data for the body pose feature. Projection of train and test samples to a 3D subspace with F and $\tilde{F}$ learned from eq. 4: a) without correspondences ($\mu = 0$); the learned manifolds are not aligned. b) using sparse correspondences; the learned manifolds are aligned. Colors represent the ground-truth pose classes for the pan angle. Triangular symbols represent test samples, circular symbols represent train samples (best viewed in color and zoomed in).

[3] as the external, labelled datasets for body and head pose features, respectively.

Performance measure. We report the mean body and head absolute pan angle errors in degrees, w.r.t the ground truth[4].

Procedure. Similarly to [6], we extract multi-level HOG features from within body and head bounding boxes obtained from detection, head localization and tracking [15]. The dimensions of each body and head feature is 2268 and 720, respectively. We first apply PCA to reduce the dimension of these features, which will make the following steps faster. We select the principal components so as to keep 90% of the data variance. In the next subsection, we show that partial pose information within the test set (obtained from annotations or weak labels) can be used to align the test features with the ones from the training set in a semi-supervised way, which improves the pose estimation results.

5.2. Results

On the CHIL dataset, we partition the data samples in 4 chunks of equal size by taking 1/4 of each track. We then do a 4-fold evaluation, using in turn each data partition as the adaptation set, and the 3 others for testing. Similarly, on the TownCentre dataset, we perform a 3-fold evaluation. The mean performance obtained from the cross-validation sets is reported in Table 1. We compare our results with the baseline of [6], which was not using any annotations within the test set. Note that in our experiments, we do not use coupling between body pose and velocity ($\gamma = 0$) on the test set, so as to simulate pose estimation on static persons and better evaluate the actual learned and adapted classifiers. This is done for both the baseline [6] and our method.

In the CHIL dataset, for each adaptation set we consider, we use the ground truth annotations of these samples to find the alignment with the training manifold. Such an approach can be used in any similar scenario where we can still have access to a few annotated samples within the target data to do the semi-supervised alignment. Table 1 shows that partial annotations of the test set (here 25%) can help to gain a significant improvement over the unsupervised baseline.

	Chen et. al. [6], $\gamma = 0$	**Ours**, $\gamma = 0$
CHIL	37.6/41.4	**21.3/22.8**
TownCentre	29.0/29.1	**24.8/23.8**

Table 1: Mean body/head pose error in degrees on CHIL and TownCentre datasets. On CHIL, our method uses partial annotations within the test set. On TownCentre, our method uses available motion estimates and does not require manual annotations. Note that on the test folds, the velocity coupling has been set to 0 to better evaluate the learned classifiers.

In most real-world cases however, as manual labelling can be a tedious task, it would be more desirable to avoid such annotations within the target data. For datasets where some people are moving with a reliable velocity, like TownCentre, we propose to use the motion direction of those adaptation samples as weak pose labels[5]. The alignment therefore becomes weakly-supervised and no manual intervention is needed. The second line of Table 1 shows that our biased, weakly-supervised manifold alignment brings an improvement of around 5 degrees on TownCentre.

6. CONCLUSION

We presented a principled approach to address classifier adaptation for body and head pose estimation in videos. Our approach leverages on external, labelled training data and some partial labelling information within the test data in the form of some annotations or weak labels from reliable speed direction, when available. The labels of the training set and the (weak) labels within the adaptation set are used to bias and align manifolds. We have shown that aligning manifolds helps improve the accuracy over an already challenging benchmark for pose estimation.

[4] Note that we use the classification scores $\{o_i, i = 1..8\}$ of each class (o_i can be interpreted as classification score for the class angle θ_i) to compute a real-valued angular output using 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.

[5] We could not use motion direction as weak labels on CHIL because people remain static around a fixed point and velocity is thus unreliable.

7. REFERENCES

[1] Chih-Wei Chen, Rodrigo Cilla Ugarte, Chen Wu, and Hamid K. Aghajan, "Discovering social interactions in real work environments," in *Face and Gesture*, 2011, pp. 933–938.

[2] Xiaoming Liu, Nils Krahnstoever, Ting Yu, and Peter H. Tu, "What are customers looking at?," in *AVSS*, pp. 405–410.

[3] Ben Benfold and Ian Reid, "Guiding visual surveillance by tracking human attention," in *BMVC*, 2009.

[4] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez, "Combined estimation of location and body pose in surveillance video," in *AVSS*, 2011, pp. 5–10.

[5] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez, "A joint estimation of head and body orientation cues in surveillance video," in *ICCV Workshops*, 2011, pp. 860–867.

[6] Cheng Chen and Jean-Marc Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *CVPR*, 2012, pp. 1544–1551.

[7] Ben Benfold and Ian Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *ICCV*, 2011, pp. 2344–2351.

[8] A Rajagopal, R Subramanian, E Ricci, R Vieriu, O Lanz, and N Sebe, "Exploring transfer learning approaches for head pose classification from multi-view surveillance images," in *IJCV*. 2013, pp. 1–22, Springer.

[9] Vineeth Nallure Balasubramanian, Jieping Ye, and Sethuraman Panchanathan, "Biased manifold embedding: A framework for person-independent head pose estimation," in *CVPR*, 2007.

[10] Jihun Ham, Daniel Lee, and Lawrence Saul, "Semisupervised alignment of manifolds," in *AISTATS*, Robert G. Cowell and Zoubin Ghahramani, Eds., 2005, pp. 120–127.

[11] Chang Wang, Bo Liu, Hoa Vu, and Sridhar Mahadevan, "Sparse manifold alignment," 2012.

[12] Chang Wang and Sridhar Mahadevan, "Manifold alignment without correspondence," in *IJCAI*, 2009, pp. 1273–1278.

[13] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, R. Travis Rose, Martial Michel, and John S. Garofolo, "The CLEAR 2007 evaluation," in *CLEAR*, 2007, pp. 3–34.

[14] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, "Monocular 3D pose estimation and tracking by detection," in *CVPR*, 2010, pp. 623–630.

[15] Alexandre Heili and Jean-Marc Odobez, "Parameter estimation and contextual adaptation for a multi-object tracking CRF model," in *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2013.