Detecting Humans With Their Pose UCLA CSD-TR 050047

Alessandro Bissacco[1], Ming-Hsuan Yang[2], and Stefano Soatto[1]

[1] Computer Science Department
University of California, Los Angeles
Los Angeles, CA 90095
{bissacco,soatto}@cs.ucla.edu
[2] Honda Research Institute
800 California Street
Mountain View, CA 94041
mhyang@ieee.org

Abstract. In this work we consider the problem of detecting humans in a single image, together with classifying their pose. Specifically, our goal is to simultaneously answer two questions: 1) is there a human body in the image and, if so, 2) does her pose match one of the pose classes present in a given set of unlabeled examples?
Starting from a set of descriptors recently proposed for human detection, we derive a probabilistic model for the statistics of these features and show how such a model can be used to answer these two questions. In particular, we show how our model is effective at detecting humans while providing an efficient representation for the tasks of pose classification and matching.

1 Introduction

Human detection and localization from a single image is an active area of research that has witnessed a surge of interest in recent years [1–5]. Simply put, given as input an image patch, we want to devise an automatic procedure that yields a positive answer whenever the patch contains a human body.

This is a hard problem because of the wide range of variability that images of humans exhibit. Given that is impractical to explicitly model the changes due to nuisance factors such as clothing, lighting conditions, viewpoint, body pose, partial and/or self-occlusions, it is natural to use a descriptive model to represent the human/non human statistics.

The problem then reduces to a binary classification task for which we can directly apply general statistical learning techniques. Consequently, the main focus of research on human detection so far has been on deriving a suitable representation ([1, 3, 5, 4]), i.e. one that is most insensitive to typical appearance variations, so that it provides good features to a standard classifier.

Building on the success in object recognition and wide-baseline matching with local descriptors based on histograms of gradient orientations [6], recently a similar representation [5] has proven to be particularly successful for human detection. The main idea of such features is to use distributions of gradient orientations in order to be insensitve to color, brightness and contrast changes and, to some extent, local deformations. However, to account for more macroscopic variations, due for example to changes in pose, a more complex statistical model is warranted. We go beyond standard classifiers operating directly on the feature set. Our approach relies on a statistical model of the feature generation process. Guided by considerations on the nature of the basic elements used to build these descriptors, we show how a special class of hierarchical Bayesian processes can be used as generative models for these features and applied to the problem of detection and pose classification.

This work can be interpreted as an attempt to bridge the gap in the literature between the two related problems of human detection and pose estimation. In human detection, since a simple yes/no answer is required, there is no need to introduce a complex model with latent variables associated to physical quantities. In pose estimation, on the other hand, the goal is to infer these quantities and therefore a full generative model is a natural approach. Between these extremes lies our approach. We estimate a probabilistic model with a set of latent variables, which do not necessarily admit a direct interpretation in terms of configurations of objects in the image. However, these quantities are instrumental to both human detection and the pose classification problem.

The main difficulty that any approach to pose classification has to face is in the representation of the pose information. Humans are highly articulated objects with many degrees of freedom, which makes defining pose classes a remarkably difficult problem. Even assuming manual labeling, how do you judge the distance between two poses? We believe that in such conditions the only avenue is an unsupervised method. We propose an approach which allows for unsupervised clustering of images of humans and provides a low dimensional representation encoding essential information on their pose. A main difference with standard clustering or dimensionality reduction techniques is that we derive a probabilistic framework, which allows principled ways to combine and compare different models, as required for tasks such as human detection, pose classification and matching.

2 Related work

The literature on human detection and pose estimation is too broad for us to review here. Therefore we will review the main approaches, emphasizing the most relevant to our work. We refer the reader to the survey [7] for further investigations.

We focus on human detection and pose estimation from a single image. We do not consider the widely studied case where temporal information or a background model are available, for which effective algorithms based on silhouettes [8–12] or motion patterns [13–15] can be applied. This is a fundamental problem with a range of sensible applications, such as image understanding and image retrieval. It makes sense to try to solve it because we know that the information is there, since humans can tell what people are doing from photographs. The question is how to represent this information, and the answer we give constitutes the main novelty of this work.

We can roughly classify the approaches to human detection according to the basic representation used: Haar wavelets [3, 1, 16], edges [4, 2], gradient orientations [5, 17], gradients and second derivatives [18, 19] and regions from image segmentation [20].

Viola et al. [1] propose an efficient coarse-to-fine approach based on Adaboost and a set of computationally efficient Haar-like features, while Papageorgiu et al. [3] use a similar representation with a support vector machine (SVM). Gavrila et al. [4] build a real-time system that matches the edge map of the image with a set of templates organized in a hierarchical structure, automatically constructed from examples. Ioffe et al. [2] assemble segments as pair of parallel edges using Adaboost, associate a likelihood to them and find the body part locations by incrementally sampling subassemblies through an efficient importance sampling scheme.

Most approaches to pose estimation are based on body part detectors, either designed ad-hoc from cues such as edges, shape, color and texture [20–24] or learned from training data [18, 19].

The optimal configuration of the part assembly is then computed using dynamic programming as first introduced in [23] and later applied in [18, 19], or by performing inference on a generative probabilistic model, using either Data Driven Markov Chain Monte Carlo [25], Belief Propagation or its non-Gaussian extensions [21, 24].

These works focus on only one of the two problems, either detection or pose estimation. Our approach is different, in that our goal is to extract more information that a simple yes/no answer, while at the same time not reaching the full level of detail of determining the precise location of all body parts. Thus we want to simultaneously perform detection and pose classification, and we want to do it in an unsupervised manner. In this aspect, our work is related to the constellation models of Weber et al. [26], although we do not have an explicit decomposition of the object in parts.

We start from the representation [5] based on gradient histograms recently applied to human detection with excellent results, and derive a probabilistic model for it. We show that with this model one can successfully detect humans and classify their poses.

The statistical tool used in this work, Latent Dirichlet Allocation (LDA) [27], has been introduced in the text analysis context and recently applied to the problem of recognition of object classes [28–30]. Contrary to most approaches (all but [29]) where the image is treated as a "bag of features" and all spatial information is lost, we encode the location and orientation of edges in the basic features so that this essential information is explicitly represented by the model.

3 A Probabilistic Model for Gradient Orientations

This section contains a brief overview of the features that we use as the basic representations of images, followed by a description of the probabilistic model that we propose and how it can be applied to the feature generation process.

3.1 Histogram of Oriented Gradients

Local descriptors based on gradient orientations are one of the most successful representations for image-based detection and matching, as was firstly demonstrated by Lowe in [31]. Among the various approaches within this class, the best performer for the case of humans as of today appears to be the feature proposed in [5]. This descriptor is obtained by computing weighted histograms of gradient orientations over a grid of spatial neighborhoods (cells), which are then grouped in overlapping regions (blocks) and normalized for brightness and contrast changes.

[5] provides a thorough experimental study of the performances of this detector for various configurations: spatial and orientation binning methodology, size and shape of the cell, number and arrangement of cells in each block and normalization schemes. We refer to the original paper for details.

Here we give a brief description of the feature for the default settings, which has been used in this work.

Assume that we are given a patch of 64×128 pixels. We divide the patch into cells of 8×8 pixel, and for each cell a gradient orientation histogram is computed. The histogram represents a quantization in 9 bins of gradient orientations in the range $0^o - 180^o$. Each pixel contributes to the neighbor bins, both in orientation and space, by an amount proportional to the gradient magnitude and linearly decreasing with the distance from the bin center.

The cells are grouped in blocks of 2×2 cells, and the contribution of each pixel is also weighted by a Gaussian kernel with $\sigma = 8$, centered in the block. Finally the cell histograms within one block are normalized to have unit L_2 norm. The final descriptor is a collection of histograms from overlapping block, with horizontal and vertical spacing of 8 pixel between neighboring blocks.

The main properties of such a representation is robustness to local deformations, illumination changes and, to a limited extent, viewpoint and pose changes due to coarsening of the histograms.

In order to handle the larger variations typical of human body images, we need to complement this representation with a model. Instead of using general machine learning techniques, we propose a probabilistic model that can accurately describe the generation process of this features.

3.2 Latent Dirichlet Allocation

In this section we review the main statistical tool used in this work, latent Dirichlet allocation. We borrow notations and terminology from the document analysis framework in which it was firstly introduced. For further details we refer the reader to [27].

We are given a collection of documents, with the following representation:

- Words w,the basic units of our data, take values in a dictionary of W unique elements, $w \in \{1, \cdots, W\}$.
- A document $\mathbf{w} = (w_1, w_2, \cdots, w_N)$ is a sequence of N words. The number of words N is a random variable, for example a Poisson process. Since the distribution of N does not affect the derivation of the model, we will not consider it in what follows.
- The corpus $\mathcal{D} = \{\mathbf{w_1}, \mathbf{w_2}, \cdots, \mathbf{w_M}\}$ is a collection of M documents.

The main goal of LDA is to find a model that assigns high probabilities to the elements of the corpus and to documents similar to them. In order to do so, it introduces a set of K latent variables, called topics. Each word in the document is assumed to be generated by one of the topics. Under this model, the generative process for each document $\mathbf{w}$ in the corpus is as follows:

1. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
2. For each word $w_n,\ n = 1, \cdots, N$:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word $w_n \sim \text{Multinomial}(\beta z_n)$.

where the hyperparameter $\alpha \in \mathcal{R}_+^K$ represents the prior on the topic distribution, $\theta \in \mathcal{R}_+^K$ are the topic proportions, and $\beta \in \mathcal{R}_+{}^{W \times K}$ are the parameters of the word distributions conditioned on topics, i.e. $\beta_{ij} = P(w = j|z_i = 1)$. In figure 1 we show a graphical representation of the LDA model.

Here we can safely assume that the topic distributions β are deterministic parameters, later for the purpose of inference we will treat them as random variables and assign them a Dirichlet prior: $\beta_{.k} \sim \text{Dirichlet}(\eta)$, where $\beta_{.k}$ denotes the k-th column of β.

Then the likelihood of a document $\mathbf{w}$ is:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{1}$$

Equation (1) shows how the documents are represented as a continuous mixture distribution. The advantage over standard mixture of discrete distributions [32], is that we allow each document to be generated by more than one topic. LDA bears close relation to probabilistic latent semantic analysis (pLSA,[33]), where documents are mixtures of topics with document-specific proportions. Unlike pLSA, however, LDA is a proper generative model and, as we will see in the next section, this allows us to combine multiple models in a principled way, as it is required in tasks such as detection and classification.

The model can be simplified if we represent documents as bags of words. Specifically, a document $\mathbf{w}$ is a collection of word counts r_j:

$$\mathbf{w} = (r_1, r_2, \cdots, r_W) \quad , \quad r_j = |i : w_i = j, i \in \{1, \cdots, N\}| \tag{2}$$

By marginalizing over the hidden variable z:

$$p(w|\theta, \beta) = \sum_z p(w|z, \beta) p(z|\theta) \tag{3}$$

we obtain the following two-level hierarchical process:

1. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
2. For each word j, $j = 1, \cdots, W$ in the dictionary, choose a word count r_j as:

$$r_j \sim p(r_j|\theta, \beta) \tag{4}$$

where the word counts are drawn from a discrete distribution conditioned on the topic proportions θ: $p(r_j|\theta, \beta) = \beta_{j.}\theta$.

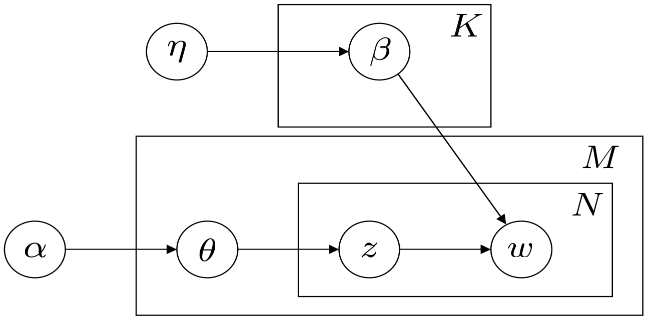


Fig. 1. Graphical representation of the Latent Dirichlet Allocation model. Each node is a random variable and each box represents sampling. The larger box generates documents in the corpus, the inner box samples words for each document, and the top box generates the topic distributions.

3.3 A Bayesian Model for Gradient Orientation Histograms

Now we can show how the described two-level Bayesian process finds a natural application in modeling the spatial distribution of gradient orientations.

Here we consider the histogram of oriented gradients [5] as the basic feature from which we build our generative model, but let us point out that the framework we introduce is very general and can be applied to any descriptor based on histograms.

In this histogram descriptor, we have that each bin represents the intensity of the gradient at a particular location, defined by a range of orientations and a local neighborhood (cell). Thus the bin height denotes the strength and number of the edges in the cell.

If we quantize histogram bins and assign a unique word to each bin, we obtain a representation for which we can directly apply the LDA framework. By analogy with document analysis, an orientation histogram computed on an image patch is a document $\mathbf{w}$ represented as a bag of words (2), where the bin heights are the word counts r_j. We assume that such a histogram is generated by a mixture of basic components (topics), where each topic z induces a discrete distribution $p(r|\beta_{.z})$ on bins representing a typical configuration of edges common to a class of elements in the dataset. By summing the contributions from each topic (3), we obtain the total count r_j for each bin, distributed according to (4).

Such feature formation process has some properties that are desirable for our application:

First, different bins can be attributed to different topics, so the underlying edge collection representing the image can be seen as a mosaic from the edge collections of the topics.

Second, even within the same bin we have contributions from multiple topics, and this models the fact that the bin height is the count of edges in a neighborhood which may include parts generated by different components.

Finally, let us point out that by assigning a unique word to each bin we model spatial information, encoded in the word identity, whereas most previous approaches [30, 28] using similar probabilistic models for object class recognition did not exploit this kind of information.

4 Probabilistic Detection and Pose Estimation

Given a set of positive and negative examples, our goal is to learn a model that can be applied for both human detection and pose estimation.

In detection, we are presented with a previously unseen image I_{new} and asked to choose between two hypotheses: either it contains a human or it is a background image.

The first step is to compute the gradient histogram representation $\mathbf{w}(I)$ for the test and training images.

Then we apply our probabilistic framework. The natural approach is to learn a model for humans and background images and use a threshold on the likelihood ratio[3] for detection:

[3] Ideally we would like to use the posterior ratio $R = P(\text{Human}|I_{new})/P(\text{Background}|I_{new})$. However notice that R is equal to (5) if we assume equal priors $P(Human) = P(Background)$.

$$L = \frac{P(\mathbf{w}(I_{new})|\text{Human})}{P(\mathbf{w}(I_{new})|\text{Background})} \tag{5}$$

For the the LDA model, the likelihoods $p(\mathbf{w}(\mathbf{I})|\alpha, \beta)$ are computed as in (1), where α, β are model parameters and can be learned from data. In practice, we can assume α known and compute an estimate of β from the training corpus. In order to do so, we can choose from two main set of inference algorithms: mean field or variational inference [27] and Gibbs sampling [34]. Mean field algorithms provide a lower bound on the likelihood, while Gibbs sampling gives statistics based on a sequential sampling scheme. As shown in figure 2, in our experiments Gibbs sampling exhibited superior performances over mean field in terms of classification accuracy. We have experimented with two variations, a direct method and Rao-Blackwellised sampling (see [35] for details). Both methods gave similar performances, here we report the results obtained using the direct method, whose main iteration is as follows:

1. For each document $\mathbf{w}_i = (r_{i,1}, \cdots, r_{i,W})$:
   (a) Sample $\theta^{(i)} \sim p(\theta|\mathbf{w}_i, \alpha, \beta)$
   (b) Sample $v_{j.}^{(i)} \sim \text{Multinomial}(\beta_{j.}\theta^{(i)}, r_{i,j})$
2. For each topic k
   (a) Sample $\beta_{.k} \sim \text{Dirichlet}(\sum_i v_{.k}^{(i)} + \eta)$

In pose classification, we start from a set of unlabeled training examples of human poses. From this data we learn the topic distribution β. This defines a probabilistic mapping to the topic variables, which can be seen as an economical representation encoding essential information on the pose. That is, from a image I_{new}, we estimate the topic proportions $\hat{\theta}(I_{new})$ as:

$$\hat{\theta}(I_{new}) = \int \theta p(\theta|w(I_{new}), \alpha, \beta) d\theta \tag{6}$$

Pose information can be recovered by matching the new image I_{new} to an image I in the training set. For matching, ideally we would like to compute the matching score S_{opt} as the posterior probability of the test image I_{new} given the training image I and the model α, β:

$$S_{opt}(I, I_{new}) = P(\mathbf{w}(I_{new})|\mathbf{w}(I), \alpha, \beta) \tag{7}$$

(7) is computationally expensive since for each pair I, I_{new} it requires computing an expectation of the form (6), so we opted for a suboptimal solution. For each training document I, in the learning step we compute the posterior topic proportions $\hat{\theta}(I)$ as in (6). Then the matching score S between the test image I_{new} and the training image I is given by the dot product between the two vectors $\hat{\theta}(I)$ and $\hat{\theta}(I_{new})$:

$$S(I, I_{new}) =< \hat{\theta}(I), \hat{\theta}(I_{new}) > \tag{8}$$

5 Experiments

In the first set of experiments we tested the efficacy of our model for the human detection task. We used the dataset provided by [5], consisting in 1822 64 × 128 images of pedestrians in various configurations and 1671 images of outdoor scenes not containing humans.

We collected negative examples by random sampling 10 patches from each of the first 1218 non-human images. These, together with 1208 positive examples and their left-right reflections, constituted our training set. We used the learned model to classify the remaining 614 positive and 4530 patches from 453 negative images.

The first step is to compute the histograms of oriented gradients from the image patches. We used the parameter settings described in section 3.2. Then, in order to apply our probabilistic model, it is necessary to quantize the histogram bins. We have verified experimentally that 8 quantization levels yields satisfactory discrimination performances while providing an economical representation.

Given the number of topics K and the prior hyperparameters α, η, we learn topic distributions β and topic proportions $\hat{\theta}(I)$ using either Gibbs sampling or Mean Field. We assume scalar priors, $\alpha_i = a, \eta_i = b$. We tested different settings for a, b and we noticed that their values are not crucial for the final results. In the experiments shown we used $a = 2/K$ and $b = 0.1$.

The number of topics K is an important parameter that should be carefully chosen based on considerations on modeling power and complexity. With a higher number of topics we can more accurately fit the data, which can be measured by the increase in the likelihood of the training set. This does not come for free: we have a larger number of parameters and an increased computational cost for learning. Eventually, an excessive topic number causes overfitting, which can be measured as the likelihood in the test dataset decreases. For the INRIA data, experimental evaluations suggested that a good tradeoff is obtained with $K = 20$.

We learn two models, one for positive and one for negative examples. For learning we run the Gibbs sampling algorithm described in section 4 for a total number of 300 samples per document, including 50 samples to compute the likelihoods (1). We also trained the model using the Mean Field approximation. For details on the implementation we refer to [35]. We compute the likelihood ratio (5) and yield positive answer if $L > 1$.

In figure 2 we show the performances of our LDA detector on the INRIA test images compared with:

- Linear SVM classifier: we used the SVMlight [36] implementation with error weight $C = 0.01$ and cost factor j equal to the ratio between positive and negative examples.
- pLSA classifier: we learned the topic distributions $p(w|z)$ for positive and negative pLSA models from the training data with standard EM [33], Then using the fold-in heuristic [33] we computed the likelihood $p(\mathbf{w}) = \prod_i \sum_z p(w_i|z)p(z)$ of test feature $\mathbf{w}$ for both models, and use the likelihood ratio (5) for detection. The only parameter here is the number of topics K, we used $K = 20$.

- K-Means classifier. We learned positive and negative models, each as a collection of $K = 20$ clusters using K-Means. The decision rule is wether the closest cluster belongs to the positive or the negative model.

From the plot we see how the results of our approach are comparable with the performances of the Linear SVM, while being far superior to the other unsupervised clustering techniques. We would like to stress that a comparison on detection performances with state-of-the discriminative classifiers[4] would be unfair, since our model targets pose classification which is a problem sensibly harder than binary detection. If a fair comparison needs to be made, then we should divide the dataset in classes and compare our model with a multiclass classifier. But then we would face the difficult problem of how to label human poses.

The particularly poor performances of the pLSA model can be explained by the fact that it is not a proper generative model, and the pseudo-likelihood provided by the fold-in heuristic for different models are not directly comparable.

Let us mention that the performances of all these classifiers could be improved by using boosting, we do not purse this avenue since we do not focus on constructing the best possible human detector.

In figure 3 we show the distributions of sample topics from the 20-topic INRIA positive model and a 8-topic model learned on a less challenging dataset, the Mobo sequences, which we will describe shortly. We see how topics roughly correspond to pose classes, this is clear in particular from the results on the Mobo dataset. It is difficult to attribute a semantic meaning to poses given the high number of degrees of freedom of the human body. Thereby our approach can be seen as an unsupervised method to discover pose classes, which are represented by the topics and the associated distributions. From figure 3 we can also see how topics concentrate most of their gradient distribution on different parts of the image, showing the mosaic effect induced by this model.

For the second round of experiments, we used the Mobo dataset [37] provided by the Carnegie Mellon group. It consists of sequences of subjects performing different motion patterns, each sequence taken from 6 different views. In the experiments we used 22 sequences of fast walking motion, picking the first 100 frames from each sequence.

In the first experiment we trained the model with all the views and set the number of topics equal to the number of views, $K = 6$. As expected, each topic distribution represents a view and by assigning every image I to the topic k with highest proportion $k = \arg\max_k \hat{\theta}_k(I)$ we correctly associate all the images from the same view to the same topic.

To obtain a more challenging setup, we restricted to a single view and tested the classification performances of our approach in matching poses. We learned a model with $K = 8$ topics from 16 training sequences, and used the remaining 6 for testing. In figure 3 we show sample topics distributions from this model. In figure 4, for each test sequence we display a sample frame and the associated top ten matches from the training data according to the score (8). We can see how

[4] For example, [5] shows that sensible improvements over the Linear SVM can be obtained using support vector machines with Radial Basis function kernel.

the pose is matched against change of appearance and motion style, specifically a test subject pose is matched to similar poses of different subjects in the training set. This shows how the topic representation factors out most of the appearance variations and retains only essential information on the pose.

In figure 5, we plot the similarity matrix $S(I_{new}, I)$ between test I_{new} and training I frames, which shows the correlation between the respective motions as measured by the model. We can see the parallel diagonal lines corresponding to the different motion cycles, similar the ones obtained in motion autocorrelation studies [13]. Let us point out that the computation of this matrix requires only dot products between low dimensional unit vectors $\hat{\theta}$, so our approach provides also an efficient method for matching poses to a large dataset. In order to give a quantitative evaluation of the pose matching performances, we labeled the dataset by mapping the set of walking poses to the interval $[0, 1]$. We manually mapped to 0 all the frames at the beginning of the double support phase, when the swinging foot touches the ground, and to 1 the frames where the legs are approximately parallel. We labeled the remaining frames automatically using linear interpolation between keyframes. The average interval between keyframes is 8.1 frames, this motivates our choice of the number of topics $K = 8$.

For each test frame, we compute the pose error as the difference between the associated pose value and the average pose of the best top 10 matches in the training dataset. We obtain an average error of 0.14, corresponding to 1.2 frames. In figure 6 we show the average pose error when matching test frames to a single train sequence. Although the different appearance affects the matching performances, overall the results shows how our approach can be successfully applied to automatically match poses of different subjects.

6 Conclusions

In this work we introduced a novel approach to human detection, pose classification and matching from single image. Starting from a representation robust to a limited range of variations in the appearance of humans in images, we derived a generative probabilistic model which allows for automatic discovery of pose information. The model can successfully perform detection and provides a low dimensional representation of the pose. It automatically clusters the images using representative distributions and allows for an efficient approach to pose matching. We verified experimentally the efficacy of our approach by applying it to human detection, pose clustering and pose matching on two large image datasets.

References

1. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. Proc. ICCV (2003) pp. 734–741
2. Ioffe, S., Forsyth, D.A.: Probabilistic methods for finding people. International Journal of Computer Vision **vol 43, no. 1** (2001) pp. 45–68
3. Papageorgiou, C., Poggio, T.: A trainable system for object detection. International Journal of Computer Vision **vol 38, no 1** (2000) pp 15–33
4. Gavrila, D.M., Philomin, V.: Real-time object detection for smart vehicles. Proc. ICCV **vol 1** (1999) pp. 87–93
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. Proc. CVPR **vol. 1** (2005) pp. 886–893

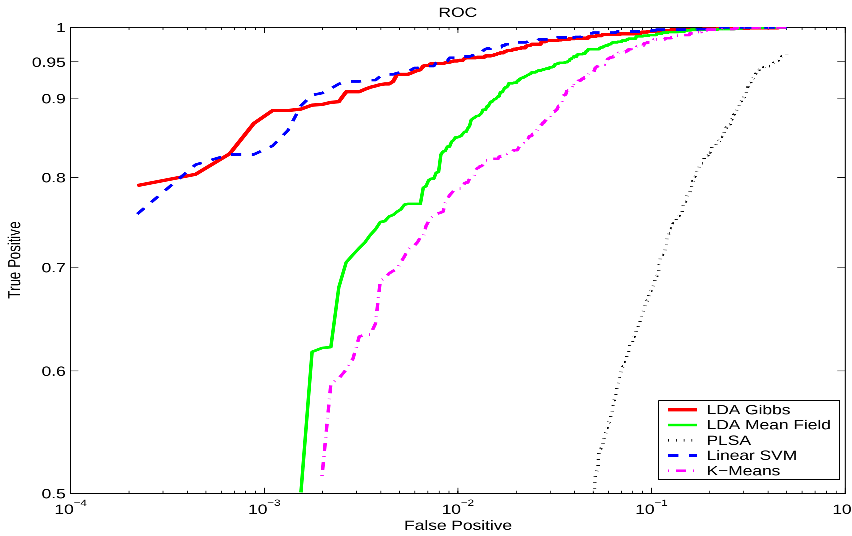


Fig. 2. Receiver Operating Characteristics on INRIA test dataset of five detectors using histogram of oriented gradients as features: LDA detectors with 20 topics learned by Gibbs Sampling and Mean Field, pLSA with 20 topics, Linear SVM and K-Means with 20 clusters. We can see how the Gibbs LDA outperform the other unsupervised clustering techniques and scores comparably with the Linear SVM, which is specifically optimized for the simpler binary classification problem.

6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **vol.60, no 2** (2004) pp. 91–110
7. Gavrila, D.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding **vol 73, no 1** (1999) pp 82–98
8. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. Proc. CVPR **vol 2** (2003) pp. 459–466
9. Agarwal, A., Triggs, B.: 3d human pose from silhouettes by relevance vector regression. Proc. CVPR **vol 2** (2004) pp. 882–888
10. Rosales, R., Sclaroff, S.: Inferring body without tracking body parts. Proc. CVPR **vol 2** (2000) pp 506–511
11. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 780–785
12. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. Proc. ICCV **vol 2** (2003) pp. 750–757
13. Cutler, R., Davis, L.: Robust real-time periodic motion detection, analysis, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **vol 22, no 8** (2000) 781–796
14. Sidenbladh, H., Black, M.J.: Learning the statistics of people in images and video. International Journal of Computer Visio **Vol. 54, No 1-3** (2003) pp. 183–209
15. Dimitrijevic, M., Lepetit, V., Fua, P.: Human body pose recognition using spatio-temporal templates. ICCV workshop on Modeling People and Human Interaction (2005)
16. Zehnder, P., Koller-Meier, E., Gool, L.V.: A hierarchical system for recognition, tracking and pose estimation. Machine Learning for Multimodal Interaction: First International Workshop (2004)
17. Shashua, A., Gdalyahu, Y., Hayon, G.: Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. Proc. of the IEEE Intelligent Vehicles Symposium (2004)
18. Ronfard, R., Schmid, C., Triggs, B.: Learning to parse pictures of people. Proc. ECCV **vol 4** (2002) pp. 700–714
19. Mikolajczyk, K., Schimd, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. Proc. ECCV **vol 1** (2004) pp. 69–81
20. Mori, G., Ren, X., Efros, A.A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. Proc. CVPR **vol 2** (2004) pp. 326–333

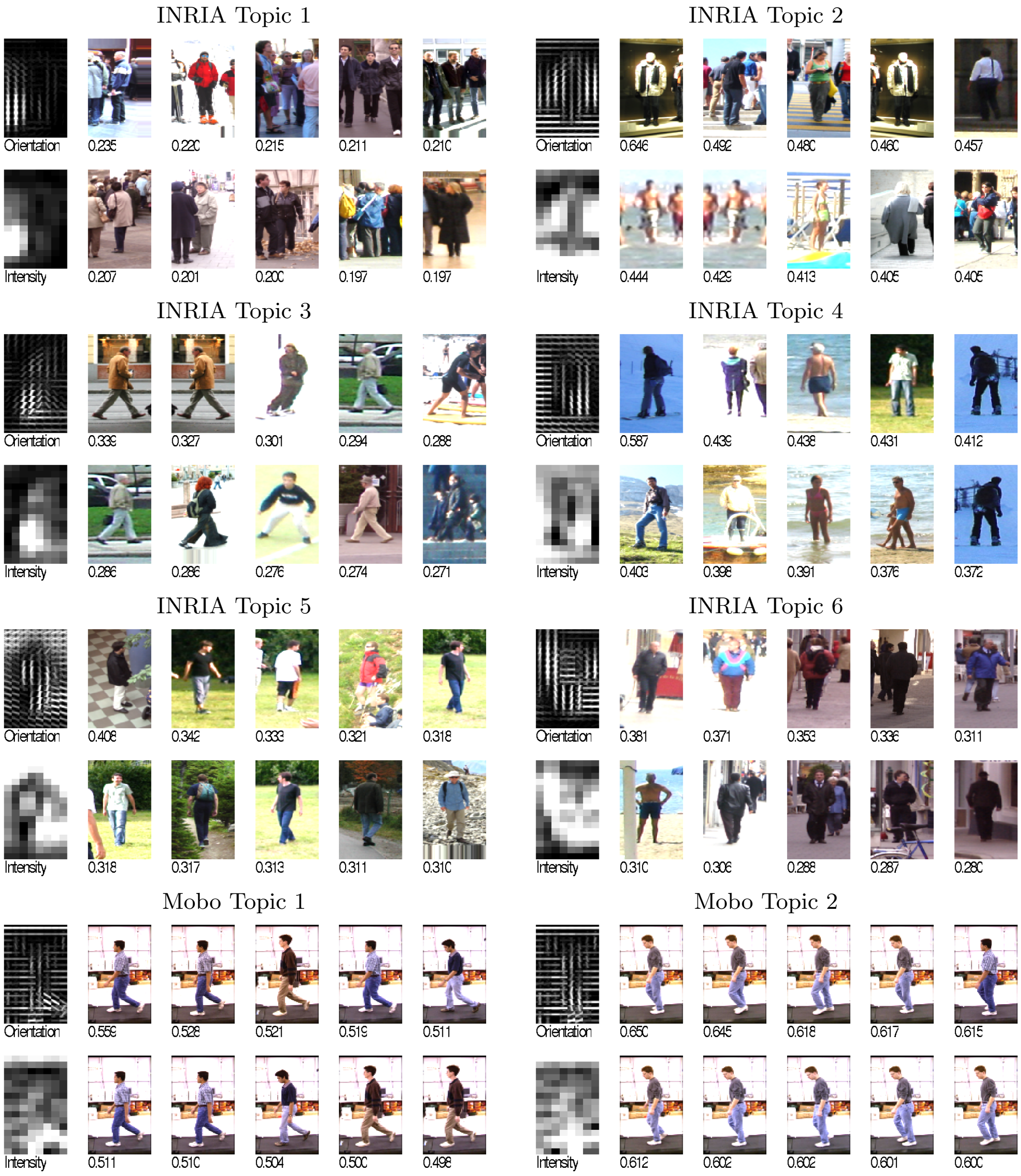


Fig. 3. Topics distributions and clusters. The first three rows show sample topics (6 out of 20) for the model learned from the positive INRIA dataset and used in the detector of figure 1. Last row shows samples from the 8-topic LDA model from one view of the Mobo sequences. For each topic k, we show 12 images in 2 rows. The first column shows the distribution of local orientations associated with topic k: (top) visualization of the orientations and (bottom) average gradient intensities for each cell. The right 5 columns show the top ten images in the dataset with highest topic proportion $\hat{\theta}_k$, shown below each image. We can see how topics are roughly related to pose classes, for example INRIA topic 1 clusters images with multiple people, topic 3 has mainly side views, topic 4 has images of people with legs apart, and so on. Let us remark the difficulty of classifying and giving a semantic meaning to the wide range of poses in this dataset. For the easier Mobo dataset, the relation between topic and pose is more clear, as we can see from the topics in the last row. We can also see how most of the energy of the topic distributions are concentrated in different parts of the image patch, this accounts for the additive nature of our mixture model.

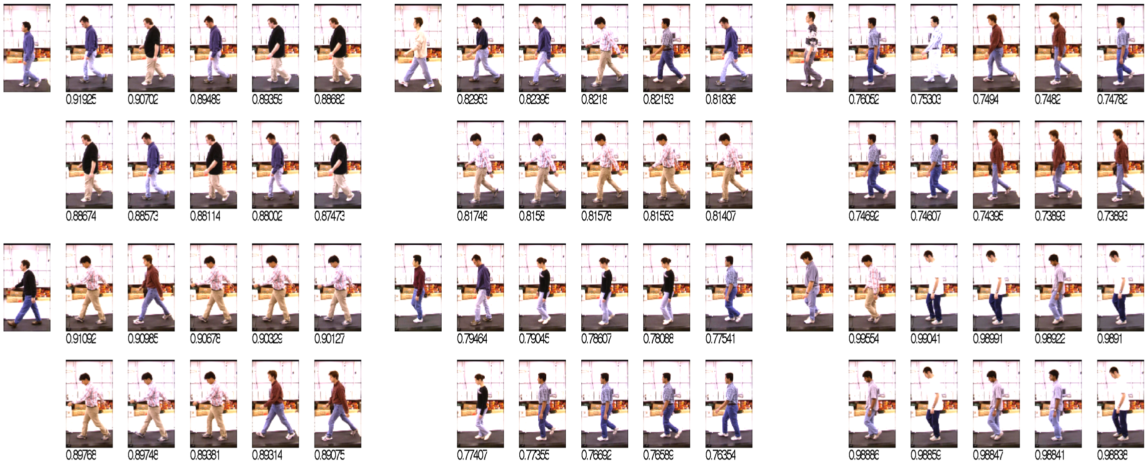


Fig. 4. Pose matching examples. On the left one sample frame from test sequences, on the right the top 10 matches in the training set based on the similarity score (8), reported below the image. We can see how our approach allows to match poses even despite large changes in appearance, and the same pose is correctly matched across different subjects.

21. Hua, G., Yang, M., Wu, Y.: Learning to estimate human pose with data driven belief propagation. Proc. CVPR **vol 2** (2005) pp. 747–754
22. Ramanan, D., Forsyth, D.A., Zisserman, A.: Strike a pose: Tracking people by finding stylized poses. Proc. CVPR (2005)
23. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient matching of pictorial structures. Proc. CVPR **vol 2** (2000) pp. 2066–2074
24. Sigal, L., Isard, M., Sigelman, B.H., Black, M.: Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In Proc. NIPS (2003) 1539–1546
25. Lee, M.W., Cohen, I.: Proposal driven mcmc for estimating human body pose in static images. Proc. CVPR (2004)
26. Weber, M., Welling, M., Perona, P.: Toward automatic discovery of object categories. Proc. CVPR **vol 2** (2000) pp. 2101–2108
27. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent drichlet allocation. Journal of Machine Learning Research **vol 3** (2003) pp. 993–1022
28. Fei-Fei, L.: A bayesian hierarchical model for learning natural scene categories. Proc. CVPR (2005)
29. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. Proc. ICCV (2005)
30. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. Proc. ICCV (2005)
31. Lowe, D.G.: Object recognition from local scale-invariant features. Proc. ICCV (1999) pp. 1150–1157
32. Nigam, K., McCallum, A.K., Thurn, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Machine Learning (2000) pp. 1–34
33. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning **vol 42, no 1-2** (2001) pp. 177–196
34. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Science **vol 101** (2004) pp. 5228–5235
35. Buntine, W., Jakulin, A.: Discrete principal component analysis. HIIT Technical Report (2005)
36. Joachims, T.: Learning to Classify Text Using Support Vector Machines. Kluwer (2002)
37. Gross, R., Shi, J.: The cmu motion of body dataset. Technical report, CMU (2001)

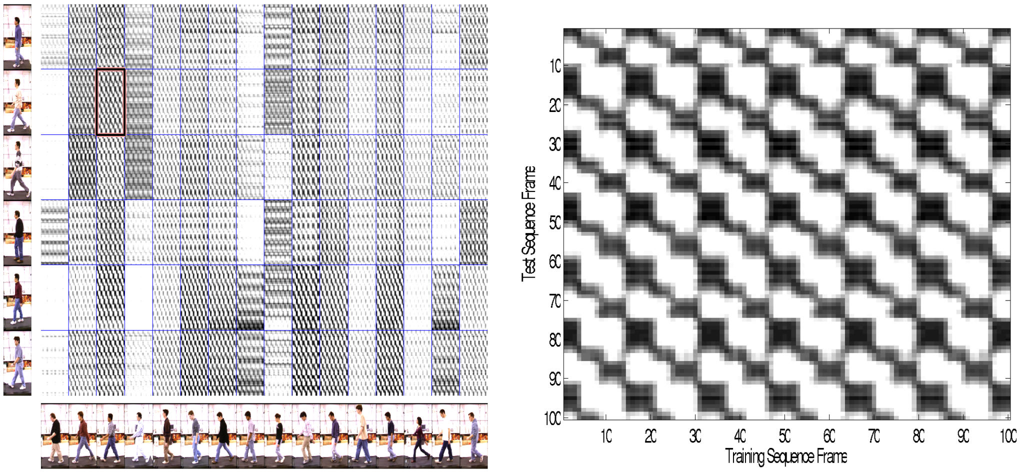
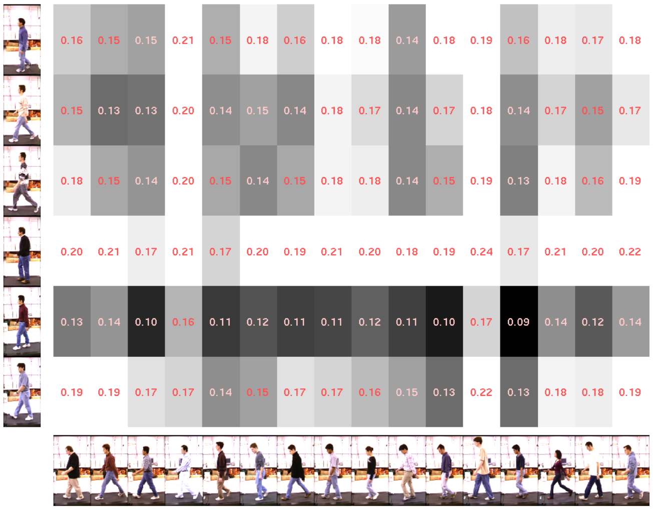


Fig. 5. Similarity measure (8) between entire test and training dataset. Every block is a 100 frame-long sequence, each rows is a test frame and each columns is a training frame. Dark indicates large similarity, light denotes small values. We can see that our measure perform consistently across subjects. On the right, a zoom-in of the box highlighted of the left, showing the matching between a single test and train sequence. We can clearly see the diagonal lines typical of walking gait motion patterns.



0.16	0.15	0.15	0.21	0.15	0.18	0.16	0.18	0.18	0.14	0.18	0.19	0.16	0.18	0.17	0.18
0.15	0.13	0.13	0.20	0.14	0.15	0.14	0.18	0.17	0.14	0.17	0.18	0.14	0.17	0.15	0.17
0.18	0.15	0.14	0.20	0.15	0.14	0.15	0.18	0.18	0.14	0.15	0.19	0.13	0.18	0.16	0.19
0.20	0.21	0.17	0.21	0.17	0.20	0.19	0.21	0.20	0.18	0.19	0.24	0.17	0.21	0.20	0.22
0.13	0.14	0.10	0.16	0.11	0.12	0.11	0.11	0.12	0.11	0.10	0.17	0.09	0.14	0.12	0.14
0.19	0.19	0.17	0.17	0.14	0.15	0.17	0.17	0.16	0.15	0.13	0.22	0.13	0.18	0.18	0.19

Fig. 6. Pose matching for test and training sequence pairs. Each row is a test sequence, each column a training sequence. We display the average pose error in matching the test frames to the frames of the specific training sequence. The highest error corresponds to about 2 frames, while the mean error is 0.32 and amounts to approximately 1.3 frames.