

PART-BASED HUMAN DETECTION ON RIEMANNIAN MANIFOLDS

*D. Tosato*¹, *M. Farenzena*¹, *M. Cristani*^{1,2}, *V. Murino*^{1,2}

¹Dipartimento di Informatica, University of Verona, Italy

²Istituto Italiano di Tecnologia (IIT), Genova, Italy

ABSTRACT

In this paper we propose a novel part-based framework for pedestrian detection. We model a human as a hierarchy of fixed overlapped parts, each of which described by covariances of features. Each part is modeled by a boosted classifier, learnt using Logitboost on Riemannian manifolds. All the classifiers are then linked to form a high-level classifier, through weighted summation, whose weights are estimated during the learning. The final classifier is simple, light and robust. The experimental results show that we outperform the state-of-the-art human detection performances on the INRIA person dataset.

Index Terms— Riemannian Manifolds classification, human detection.

1. INTRODUCTION

Robust object detection is important for many applications. In particular, in the context of video surveillance, pedestrians are a very important, and very challenging, class of objects to detect. Among the recent approaches proposed in literature, part-based models [1, 2, 3] seem to provide the best performances, as shown in [4]. This is because these models are intrinsically robust to partial, inter-object occlusions. Following the same direction, we propose here a new part-based model for pedestrian detection. The parts are hierarchically structured, and a-priori fixed, as in [2]. Each part is described by a covariance matrix descriptor, that encodes information of the variances of a set of defined features inside a region (patch), along with their correlations and the spatial layout. This descriptor is robust to illumination and scale variations. It has been proposed and successfully used for pedestrian detection in [5], where the human model is an ensemble of many overlapping patches, each described by a covariance matrix. In that work, the model is properly learnt by LogitBoost on Riemannian manifolds of symmetric positive definite matrices, which is the space where covariance matrices live. In particular, Logitboost is used for a greedy estimation of the most discriminative patches and to classify on them, i.e. for features selection and classification at the same time. The same reasoning, using boosting for feature selection and classification, has been applied to other approaches in the literature, as

for example in [2, 3].

In this paper, we claim that, i) injecting a-priori knowledge about the human structure by suggesting the parts where to focus on and ii) thanks to an adequate learning of such parts by boosting via polynomial fitting, the feature selection phase is not necessary. The resulting framework is light (the computational cost of the learning phase is dramatically reduced with respect to [5]), and it outperforms the state-of-the-art methods on the INRIA person dataset.

The rest of the paper is organized as follows. Section 2 presents the architecture of our classification system on Riemannian manifolds. Practical details and experimental results are explicated in Section 3. Finally, we conclude the paper and outline the future work.

2. SYSTEM ARCHITECTURE

Inspired by [2], we divide the human body into parts, according to their semantic meaning (head, torso, etc.). These parts are then organized in a hierarchy of three levels, for a total number of eleven parts (see Figure 1).

A covariance descriptor is associated to each part, and it is estimated as follows. As in [5], for each pixel (x, y) inside the region we gather a bunch of information about that pixel into a features vector:

$$\left[x \ y \ |I_x| \ |I_y| \ \sqrt{I_x^2 + I_y^2} \ |I_{xx}| \ |I_{yy}| \ \arctan \frac{I_x}{I_y} \right]^T, \quad (1)$$

where I_x, I_{xx} , etc are intensity derivatives and the last term is the edge orientation. From these vectors their covariance matrix can be estimated. This operation is done efficiently using integral images [6].

Given the descriptors, the system is composed by two phases: first, each body part is learnt separately, using LogitBoost; second, the parts classifiers are combined together. These two phases are detailed in the following.

2.1. Phase 1: boosting the part models

Let $\{\mathbf{X}_{ip}, y_i\}_{i=1, \dots, N}$ be the set of training examples (covariance matrix descriptors), of a fixed human part p . These examples are points in the Riemannian Manifold \mathcal{M} of symmetric positive definite matrices. Learning a classifier on \mathcal{M}

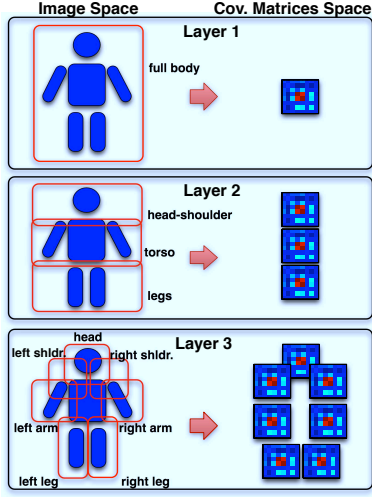


Fig. 1. Part-based human model. The human body is hierarchically divided into 11 parts, and each part is described by a covariance matrix descriptor.

implies projecting all points into the local tangent space $T_{\mathbf{X}}$ of a point $\mathbf{X} \in \mathcal{M}$. $T_{\mathbf{X}}$ is a Euclidean space, so that a classical classification algorithm can be employed on the projected points. In [5], the authors empirically show that a good choice of \mathbf{X} is the Karcher mean μ_p of $\{\mathbf{X}_{ip}\}_{i=1,\dots,N}$, i.e. the point that minimizes the sum of squared Riemannian distances.

The framework proposed in [5] is a greedy algorithm, where at each boosting iteration the most discriminative patch inside the detection window, i.e. the patch on which a single weak classifier gives the best classification performance, is selected. This implies having several covariance descriptors sets, corresponding to each of the possible patches, projecting them into their tangent spaces T_{μ_p} and choosing the one where positive and negative examples are better separated.

We follow a different direction, which is simpler, less computationally expensive, and gives good performances at the same time. We tell the classifier which are the most interesting (discriminative) areas for the human body and we let it concentrate on classification rather than features selection. This means that we build a strong classifier for each of the body parts, so that the final human detector is the composition of a few strong classifiers, instead of many weak classifiers.

In practice, for each part, we estimate μ_p and we project all examples in T_{μ_p} . The mapping of points on the Riemannian manifold to T_{μ_p} and vice-versa is done using the \log_{μ_p} and \exp_{μ_p} operators, respectively, as in [5]. This mapping is done once, because all the following reasoning are done on T_{μ_p} directly. Let us call the projected training examples as $\{\Sigma_{ip}, y_i\}_{i=1,\dots,N}$, with $\Sigma_{ip} \in T_{\mu_p}$ and labels $y_i \in \{1 = \text{human part}, 0 = \text{other}\}$.

Using the binary LogitBoost algorithm [7], we estimate a response function $F_p(\Sigma_{ip}) : T_{\mu_p} \mapsto \{0, 1\}$ that divides

the tangent space in 2 parts, according to the training set of labeled items. This function, the strong classifier, is defined as a sum of weak classifiers. The probability of Σ_{ip} being in class 1 is represented by

$$P(\Sigma_{ip}) = \frac{e^{F_p(\Sigma_{ip})}}{e^{F_p(\Sigma_{ip})} + e^{-F_p(\Sigma_{ip})}}, \quad F_p(\Sigma_{ip}) = \sum_{l=1}^{N_l} f_l(\Sigma_{ip}), \quad (2)$$

where $\{f_l\}_{l=1,\dots,N_l}$ is the iteratively selected set of weak learners (WLs). Each WL is estimated by solving a weighted least-square regression problem:

$$f_l = \sum_{i=1}^N w_{ip} |z_{ip} - f(\Sigma_{ip})|^2, \quad (3)$$

where z_{ip} and w_{ip} denote the response values and the weights, respectively, in the following forms:

$$z_{ip} = \frac{y_{ip} - P(\Sigma_{ip})}{P(\Sigma_{ip}) - (1 - P(\Sigma_{ip}))}, \quad (4)$$

$$w_{ip} = P(\Sigma_{ip}) - (1 - P(\Sigma_{ip})). \quad (5)$$

As regressors, we employ second degree polynomial functions with no mixed terms. This is because we experimentally found that this class of regressors is a good compromise between classification accuracy and computational complexity. In fact, the multidimensional polynomial fitting can be formalized as a linear problem [8], and the complexity grows linearly with the number of terms used to solve the linear system. The second degree polynomial functions double the complexity wrt the linear case, but the classification performances are clearly increased (see Fig. 2).

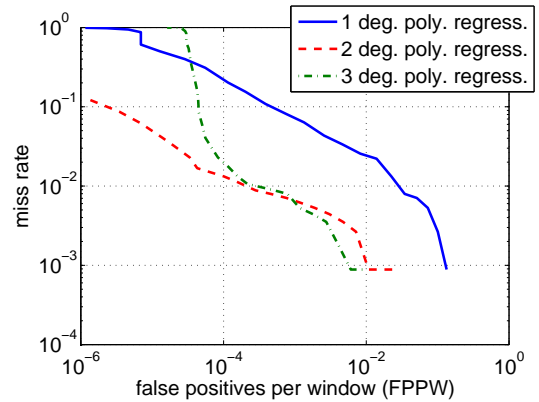


Fig. 2. Performances in terms of DET curve employing different regressor models – linear, 2nd degree polynomial, 3rd degree polynomial – in our framework. The best performances are obtained with the 2nd degree polynomial.

Part rejection cascade. We combine the LogitBoost classifier of each body part with a *rejection cascade structure* of K

levels. Using a cascade makes the part detectors more robust to false positives. We sample $N_n = 10^4$ negative examples for each cascade level and we join them to the N_p positive examples to form a training set of $N = N_p + N_n$ elements. We rewrite F_p in F_p^k to emphasize the dependence of the classifier on the current level.

We classify all the negative examples with the cascade of the previous $k - 1$ classifiers, where $k \in \{2, \dots, K\}$. The examples that are correctly classified (i.e. classified as negative) are removed from the training set, nevertheless keeping at least 1000 examples.

Learning on cascade level k stops if a combined condition is satisfied; we impose that the learning process correctly classifies at least 99.8% of the positive examples, and that it rejects at least 35% of the negatives. To verify this condition we sort the data set according to descending probabilities (Eq. (2)). Then, we check that $F_p^k(\Sigma_{ip}) > 0$ for at least the 99.8% of positives and $F_p^k(\Sigma_{ip}) < 0$ for at least the 35% of negatives. The F_p^k value of the $(0.35N_n)$ th element with the smallest probability, denoted as $thrd_p^k$, is used for testing: a point Σ_{ip} is classified as positive if $F_p^k(\Sigma_{ip}) - thrd_p^k > 0$.

2.2. Phase 2: Combination of parts classifier

When the robust part classifiers are learnt, we combine their strong responses into a unique human detection as follows:

$$F_{\text{comb}}(I_W) = \sum_{p=1}^{11} w_p \cdot F_p^*(\Sigma_p), \quad (6)$$

where I_W is the detection window and Σ_p is the covariance matrix descriptor estimated on the body part p , projected into T_{μ_p} and F_p^* is the classification response produced by the rejection cascade. I_W will be classified as positive if $F_{\text{comb}}(I_W) > \tau$.

Since the location of the human body parts is fixed by construction and the variability of human postures is high, it is reasonable that some parts detectors are more reliable than others. This is why a weight w_p is associated to each part classifier. Given a set of positive images, we instantiate a validation dataset, where we count the number of correct detections per part. The normalized resulting values become the w_p s. w_p is proportional to the ability of F_p^* to correctly classify its respective body part, and it says which part is more suitable for the detection of human bodies.

3. EXPERIMENTAL RESULTS

We evaluate our approach considering the INRIA Person dataset. The training set has 2416 human images scaled into a fixed detection window of 64×128 pixels and 1218 person-free images. The testing set contains 1132 human images and 453 person-free images. The dataset is not well-suited for learning part-based classifiers (even if [3] uses it for the same

purpose), because the data is not aligned, and different poses are present. This fact, and the excellent results gained by our approach, witness the capability of the part-based classifier to absorb even strong pose variations.

We implement our approach with Matlab on an Intel Xeon 2.83 Ghz processor with 4.00 Gbytes of RAM. The training of the classifiers takes 15 minutes to generate a part-based classifier, for all the 11 parts, with at most 5 weak classifiers per level. The state-of-the-art method in [5] needs more than two weeks in the same hardware setting.

In Fig. 3, we compare our framework with [5] and the methods in [9, 10, 11, 12, 13], whose statistics are extracted from [4]. The performances are evaluated by adopting the Detection Error Tradeoff (DET) curve, that expresses the proportion of true detections against the proportion of false positives on a log-log scale. As visible in the figure, we outperform all

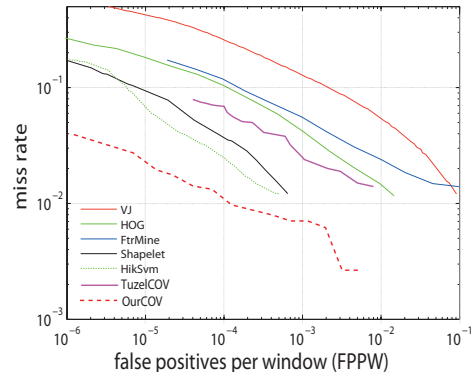


Fig. 3. Comparison with the state-of-the methods on INRIA Person dataset. The curves for other approaches are generated from [4] and [5].

the other methods reaching the best performances, both considering the FPPW (False Positive Per Window) rate and the miss rate. Moreover, this holds in a boosting framework with very few weak learners (for example, [5] has %50 more weak learners).

We show in Tab. 1 the ability of our system to detect human body parts. The table is built by considering the cascade level $k = 5$. Considering that in the INRIA Person dataset several people are not aligned, we show here that our part detector is in general able to detect single parts with high accuracy. In specific, the more reliable region is the torso, meaning that our part descriptor is particularly suited for that body portion, capturing all the pedestrian intra-class variability. Such variability pops out considering the variance image of the INRIA training dataset (see Fig. 4), where in each pixel we portray the associated per-pixel variance. It is evident that, even if that portion is characterized by the highest variability, it is the one that our framework better models.

The weights used to build the final detection response are

Human body part	Accu.% ($k = 5$)
full body	99.4%
head-shoulder	93.8%
torso	97.2%
legs	92.8%
left shoulder	82.7%
head	83.6%
right shoulder	87.4%
left arm	88.6%
right arm	88.4%
left leg	81.8%
right leg	84.9%

Table 1. Per-part detection accuracy. The detection ability of the part detectors in the cascade level $k = 5$ is shown.

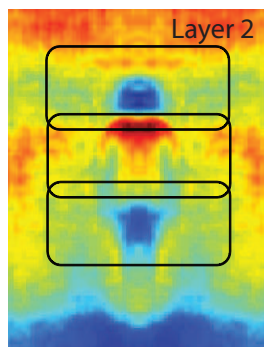


Fig. 4. Capturing the intra-class variation: the central part, even if characterized by the highest variance, is the one better modeled by the part-classifier.

proportional to the number contained in Tab. 1.

4. CONCLUSION AND FUTURE WORK

We describe a part-based human detection method based on covariance features that describes single body parts. We define a weighting system that combines the boosted responses of part detectors, and that shows the upper part of the body as the best part usable to capture human beings. Our approach is light and robust and it outperforms the state-of-the-art systems on the INRIA person dataset.

Acknowledgment

This research is funded by the EU-Project FP7 SAMURAI, grant FP7-SEC-2007-01 No. 217899.

5. REFERENCES

- [1] K. Mikołajczyk, C. Schmid, and A. Zisserman, “Human detection based on a probabilistic assembly of ro-

bust part detectors,” *Lecture notes in computer science*, pp. 69–82, 2004.

- [2] B. Wu and R. Nevatia, “Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses,” *International Journal of Computer Vision*, vol. 82, no. 2, pp. 185–204, 2009.
- [3] Y. Chen, C. Chen, K. Chang, “Multi-Class Multi-Instance Boosting for Part-Based Human Detection,” in *In Proc. of the International Conference On Computer Vision Workshops*. pp. 1177–1184, 2009.
- [4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: A benchmark,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.
- [5] O. Tuzel, F. Porikli, and P. Meer, “Pedestrian detection via classification on Riemannian manifolds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1713–1727, 2008.
- [6] O. Tuzel, F. Porikli, and P. Meer, “Region covariance: A fast descriptor for detection and classification,” *Lecture Notes in Computer Science*, vol. 3952, pp. 589, 2006.
- [7] J. Friedman, T. Hastie, and R. Tibshirani, “Special invited paper. additive logistic regression: A statistical view of boosting,” *Annals of statistics*, pp. 337–374, 2000.
- [8] T.B. Deng, “Linear approach to the least-squares multidimensional polynomial fitting,” in *Information, Communications and Signal Processing*, vol. 3, 1997.
- [9] P. Viola and M. Jones, “Robust real-time object detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2002.
- [10] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005.
- [11] P. Dollar, Z. Tu, H. Tao, and S. Belongie, “Feature mining for image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
- [12] P. Sabzmeydani and G. Mori, “Detecting pedestrians by learning shapelet features,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
- [13] S. Maji, A.C. Berg, and J. Malik, “Classification using intersection kernel support vector machines is efficient,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.