

# Normalized Autobinomial Markov Channels For Pedestrian Detection

Cosmin Țoca  
cosmin.toca@gmail.com

Mihai Ciuc  
mihai.ciuc@upb.ro

Carmen Pătrașcu  
cpatrascu@ceospacetech.pub.ro

University Politehnica of Bucharest,  
Faculty of Electronics, Telecommunications  
and Information Technology,  
Applied Electronics and Information Engineering  
Department,  
RO

## Abstract

This paper brings significant contributions to the field of pedestrian detection by learning probabilistic dependencies and contextual information that draw special attention to the human body characteristics and silhouette shapes and play down other irrelevant features. More precisely, we introduce the NAMC (Normalized Autobinomial Markov Channels) and study the efficiency of different configurations of cliques, providing a detailed experimental evaluation. Our proposed features outperform most of the solutions that have laid the foundations of pedestrian detection [1, 2]. Moreover, if we combine our novel features with gradient-based descriptors [3] and apply an efficient local decorrelation algorithm [4] to each channel, our results outperform the majority of the state-of-the-art solutions currently present in the Caltech Pedestrian Detection Benchmark [5]. We focus on a thorough analysis of the proposed feature model using the INRIA Pedestrian Dataset [6] as a benchmark to evaluate various parameter settings.

## 1 Introduction

Pedestrian detection represents one of the most important components of engineering devices that use automated vision to help decision systems take quick and accurate actions. Nowadays, such systems are defined and customized to be useful for different needs, such as monitoring and aided surveillance, or increasing safety features in automotive industry.

Given the large spectrum of applications that use pedestrian detection, demand has increased in recent years for the development of feasible solutions which can be integrated in devices such as smartphones, tablets, as well as professional and consumer cameras.

There are two main factors that contribute to the performances of object detection algorithms. The first one is the learning algorithm they use. Examples include AdaBoost [8], Boosted Trees [9, 10, 11], linear SVMs [12, 13] or latent SVMs [14, 15, 16], discriminative deep models [17, 18], Random Forests [19], and LogitBoost [20]. The second factor is the feature design and representation method. Well known examples include Haar [8], HOG [10, 11, 12, 13, 14] and channel features [1, 2, 3, 15, 16, 21].

In our algorithm, we use a standard boosted cascade of trees [22] and focus on finding probabilistic features that highlight the human body characteristics regardless of contextual

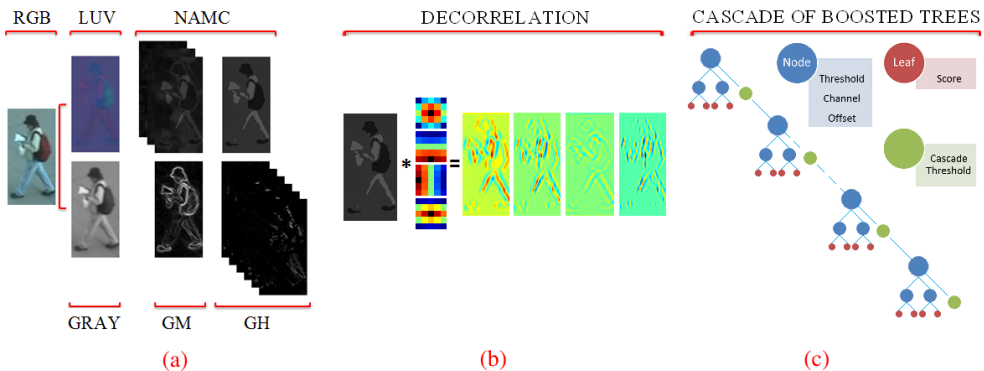


Figure 1: **a) Feature representation.** Several registered channels have been computed using different transformations of the input RGB image. The initial image has been converted to CIE-LUV color space to compute the NAMC (normalized autobinomial Markov channels), and to GRAY to obtain the GM (gradient magnitude) and GHs (gradient histograms). **b) Local decorrelation.** Multiple decorrelation filters have been learned separately for each channel feature and then applied to the corresponding channels. **c) Cascade of boosted trees.** The classification is based on a cascade of boosted trees, where each node examines a feature value in a certain channel and each leaf returns a score.

information in images. Information in images is often spatially correlated, which means that adjacent values are likely to have similar values. For this reason, our method employs only close neighbouring pixels for the probabilistic description of our target object. We propose to learn the local dependencies in the image channels and also between pairs of pixels coming from different channels and convert them into a probability space that fully respects the Markov property for a given local specification. A statistical model that respects all these requirements and best fits the multidimensional characteristics of the data can be found in the scientific literature as MRF (Markov Random Fields) theory.

Markov random fields are widely used probabilistic models providing a basis for modelling contextual constraints in visual processing and interpretation, since they allow the integration of prior knowledge of images and scenes. The foundations of the theory of Markov random fields was laid in the early '70s [29, 35, 39], but due to their generic nature, they have found widespread use across image processing, and in particular for image restoration [20, 40], denoising [8, 25], segmentation [23, 26] and texture classification [9, 7, 4].

While much of this work could apply to other areas of computer vision as well, we focus on generic Markov random fields models and apply them in pedestrian detection.

Recent years have brought significant results in pedestrian detection, and several approaches using boosted trees and channel features have achieved state-of-the-art results on major benchmarks.

Our paper is based on Dollár's *et al.* [15] work, which involves learning a cascade of boosted trees from a feature space that consists of channel features such as image channels, gradient magnitude and gradient histograms. The main purpose of Dollár *et al.* [15] paper has been related to feature scaling. The author has proposed a method to approximate channel features for various scales by extrapolating from nearby ones, rather than explicitly computing them. Starting with the assumption that spatially decorrelated features could produce better results, Woonhyun *et al.* [30] have proposed a method for learning decorrelation filters matching each channel feature.

We propose a novel set of channel features based on autobinomial Markov-Gibbs random fields (see Fig. 1). Moreover, we include optimizations for fast feature calculation, and test a random selection method for the neighbourhood system. In the same way as Woonhyun *et al.* [30], we get specialized decorrelation filters matching our newly introduced channels. We use Dollár’s [16] implementation of the learning algorithm and his scaling method to speed up the scanning process.

The remainder of this paper is organized as follows. Sec. 2 begins with an introduction of channel features, gives a detailed overview of our normalized autobinomial Markov channels with implementation details, and briefly describes channel features such as gradient magnitude and gradient histograms, as well as algorithms for feature pyramids and local decorrelation used in our autobinomial model. In Sec. 3 we argue the choice of boosting method. A detailed experimental evaluation of the quality of the results is performed in Sec. 4. Conclusions and ideas for future work are given in Sec. 5.

## 2 Channel Features

The decision to use a particular learning algorithm is highly dependent on the application. This choice can be influenced by several factors, such as the need of having a quick rejection or otherwise a good performance measured at the end of the scanning process. When choosing which features to use in a solution, both the computational cost (that can affect the speed of the algorithm), and the memory space (required to store all channel features at different scales) are highly important. The performance of the whole detector is directly influenced by the performance of each feature apart. One of the most important aspects related to features is their ability of separation between positive and negative classes.

A channel feature can be seen as an image, meaning that it is a function of the same spatial variables as the initial image. However, pixel values hold knowledge about image features (e.g. weighted averages, gradient magnitudes, gradient orientations) instead of luminance intensity or color.

In case of pedestrian detection applications, extracting only one type of feature may not be enough to get relevant information from the image data. As an alternative, several different features are extracted from the image, resulting in more channel features at each image point.

### 2.1 Normalized Autobinomial Markov Channels

Let us consider that  $X = \{X_\xi\}_{\forall \xi \in \Omega}$  is a random process defined on a probability space  $(\Omega, \mathcal{F}, \mathcal{P})$  where  $\Omega$  is a finite set that refers to the pixel’s locations in the image configuration, with elements denoted by  $\xi$  and called sites. Let  $\mathcal{F}$  be a finite set, called the phase space, that refers to all possible values of the pixels in a spectral plane, and  $\mathcal{P}$  the assignment of probabilities for each location.

A neighbourhood system on  $\Omega$  is a family  $\mathcal{N} = \{\mathcal{N}_\xi\}_{\forall \xi \in \Omega}$  of subsets of  $\Omega$  such that for all  $\xi \in \Omega$  a site is not neighbouring to itself  $\xi \notin \mathcal{N}_\xi$ , and the neighbouring relationship is mutual  $\eta \in \mathcal{N}_\xi \Leftrightarrow \xi \in \mathcal{N}_\eta$ .

**Markov Property:** The Markov Property says that given the set of neighbours  $\mathcal{N}_\xi$ , the variable  $\xi$  is conditionally independent of all variables in the random field except the neighbours, so the probability  $\mathcal{P}(X_\xi = \gamma_\xi \mid X_{\Omega \setminus \xi} = \gamma_{\Omega \setminus \xi}) = \mathcal{P}(X_\xi = \gamma_\xi \mid X_{\mathcal{N}_\xi} = \gamma_{\mathcal{N}_\xi})$  for

all  $\xi \in \Omega$ ,  $\gamma \in \mathcal{F}$ . This means that the distribution of the phase at a site  $\xi$  is directly influenced only by the phases of the neighbouring sites  $\mathcal{N}_\xi$ .

The local characteristic of a Markov random field at site  $\xi$  is defined as the probability to assign to each pixel  $\xi$  a certain value given the values of all other pixels in the neighbouring system. In other words  $\pi^\xi(\gamma) = \mathcal{P}(X_\xi = \gamma_\xi \mid X_{\mathcal{N}_\xi} = \gamma_{\mathcal{N}_\xi})$ , where  $\pi^\xi : \mathcal{F} \rightarrow [0, 1]$ .

**Energy function:** An unpublished manuscript written by Hammersley and Clifford [24] describes how to interpret the local property of a Markov random field in terms of energy and potential. The proof is found in literature as the Markov-Gibbs Equivalence or Hammersley-Clifford Theorem and defines the probability at a site  $\xi$  as being:

$$\pi^\xi(\gamma) = \frac{e^{-\sum_{C \ni \xi} V_C(\gamma)}}{\sum_{\varphi \in \mathcal{F}} e^{-\sum_{C \ni \xi} V_C(\varphi, \gamma_{\Omega \setminus \xi})}}, \quad (1)$$

where  $V_C$  is the potential function, and  $\varphi \in \mathcal{F}$ .

To get the probability that at a site  $\xi$  the state is  $\gamma$ , we need to define a potential function  $V_C(\gamma)$  in the neighbouring system, here denoted by a collection of cliques  $\mathcal{C}$ . To be able to do this, we refer to the auto-binomial model that was introduced by Besag [4] to describe certain types of spatial processes, examining some stochastic models that occur in the texture of various physical materials.

**Potential function:** The potential at a certain state is given by:

$$V_C(\gamma) = \begin{cases} -\ln \left( \frac{\Gamma}{\gamma_\xi} \right) + \gamma_\xi & \text{if } C = \{\xi\} \\ \frac{\gamma_\xi \cdot \gamma_\eta}{\nu} & \text{if } C = \{\xi, \eta\} \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where  $\nu$  is a normalization constant equal to the number of sites in the neighbourhood system. Detailed explanations of the Markov random fields and Gibbs distribution are found in Pierre Bremaud's book [5].

The set of sites is  $\Omega = \mathbb{Z}_\omega^{m \times n \times d}$ , and the phase space is  $\mathcal{F} = \{0, \dots, \Gamma\}$ , where  $m, n, d \in \mathbb{N}$  and  $\Gamma$  is the maximum pixel value in a certain color plane. In the context of this paper, a site  $\xi$  is a pixel, and a phase  $\gamma \in \mathcal{F}$  is a shade of gray.

The neighbourhood system is  $\mathcal{N}_\xi = \{\eta \in \Omega \mid \eta \neq \xi, d^2(\xi, \eta) \leq \delta\}$ , where  $\delta$  is a fixed positive integer and  $d^2(\xi, \eta)$  is the squared Euclidean distance between  $\xi$  and  $\eta$ .

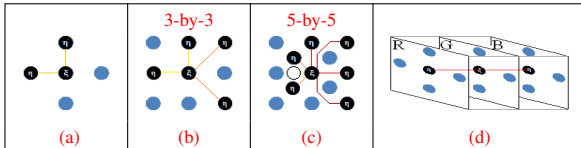
In this model, the only cliques participating in the energy function are singletons and pairs of mutual neighbours, the set of cliques appearing in the energy function being a disjoint sum of collections of cliques  $\mathcal{C} = \sum_{k=1}^{\omega(\delta)} C_k$ .

**Feature calculation:** If we replace Eq. 2 in Eq. 1, we get the probability assigned to a local system as being:

$$\pi^\xi(\gamma) = Z^{-1} \left( \frac{\Gamma}{\gamma_\xi} \right) \sigma^{\gamma_\xi} (1 - \sigma)^{\Gamma - \gamma_\xi}, \quad (3)$$

where  $Z = \left( \frac{\Gamma}{\Gamma/2} \right)$  is a normalization constant, and  $\sigma = \sigma(\mathcal{N}_\xi) = (e^{\langle \alpha, \beta \rangle}) / (1 + e^{\langle \alpha, \beta \rangle})$ .

Here  $\langle \alpha, \beta \rangle$  is the scalar product of two vectors  $\alpha$  and  $\beta$  of sizes  $\omega(\delta)$ , and elements in  $\alpha$  normalize the absolute difference of gray levels between pairs of cliques found in the same



$\delta$	1	2	4	1
$\omega(\delta)$	3	5	7	3

Table 1: **Examples of neighbourhood systems.** The maximum distance between two sites  $\{\xi, \eta\}$  is denoted by  $\delta$  and  $\omega(\delta)$  represents the number of possible cliques for each local specification, including only singletons and pairs of mutual neighbours.

neighbourhood of a certain pixel  $\beta_k = \beta_k(\gamma_{\mathcal{N}_\xi}) = |\gamma_\eta - \gamma_{\eta'}|$ , for  $\eta \neq \eta'$ , and  $\eta, \eta' \in \mathcal{N}_\xi$ , where  $\{\eta, \xi\}$  and  $\{\eta', \xi\}$  are two pairs in  $\mathcal{C}_k$  containing  $\xi$ .

**Implementation details:** Each pixel  $\xi$  in the input color plane has been converted using Eq. 3 to a probability considering a fixed neighbourhood  $\mathcal{N}_\xi$ , the same for each color plane (see Tab. 1.c). To obtain one channel feature that characterizes correlations between color planes, a simple neighbourhood system has been used (see Tab. 1.d). For corners and boundaries we only consider the available neighbours.

To avoid an increase of the computational costs, the product of the first two terms in Eq. 3 is pre-computed and quantized on  $\Gamma$  values, one for each possible pixel value, and therefore used as a look-up table. The same can be done for  $\sigma$ .

Regardless of the choice of the sites inside the neighbourhood systems, the normalized autobinomial Markov channels are strongly invariant if the distance  $\delta$  is kept constant. This fact allows us to randomly choose which site becomes a part of the neighbourhood.

## 2.2 Local Characteristics of Gradient Magnitude and Gradient Histograms

Histogram of Oriented Gradient descriptors were first introduced by Dalal and Triggs [10], who used their algorithm for pedestrian detection in static images. The fundamental idea behind these descriptors is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. Following Piotr Dollár *et al.* [15], we adapt the next definitions to our assumptions and notations.

**Gradient magnitude:** The gradient magnitude at site  $\xi$ , denoted by  $\mathcal{M}(\gamma_\xi)$  and defined as  $\mathcal{M}(\gamma_\xi) = \sqrt{\frac{\partial}{\partial x} \mathcal{N}_\xi^2 + \frac{\partial}{\partial y} \mathcal{N}_\xi^2}$ , captures undirected edge strength.

Therefore, the orientations have been obtained from previously computed gradients as being  $\mathcal{O}(\xi) = \arctan\left(\frac{\partial}{\partial y} \mathcal{N}_\xi / \frac{\partial}{\partial x} \mathcal{N}_\xi\right)$ , values that may be quantized in a fixed number of bins, where each bin  $\theta$  accumulates a fixed range of orientations. The orientations are spread from 0 to 180 degrees, the half circle being divided into six equal sectors, each one accumulating angles from a range of 30 degrees. To account for variations in contrast and illumination, the gradient strengths must be locally normalized, which requires grouping the cells together into larger and spatially connected blocks.

As shown in Piotr Dollár *et al.* [15], the normalization of the gradient magnitude  $\mathcal{M}$  is made by dividing each pixel by the average of its  $11 \times 11$  neighbourhood, this being computed by convolving  $\mathcal{M}$  with an L1 normalized  $11 \times 11$  triangle filter. The author shows that using the normalized gradient improves the results in the context of object detection.

**Gradient histograms:** The gradient histograms are weighted histograms where bin indexes can be determined by gradient angles and then weighted by gradient magnitudes. The channel features are given by  $\mathcal{H}_\theta(\xi) = \delta \cdot \mathcal{M}(\xi)$ , where  $\delta = 1$  if  $\mathcal{O}(\xi) = \theta$ , and  $\delta = 0$  otherwise. As shown in Fig. 1, the gradient magnitude channel is computed from luminance, leading to a set of six gradient histogram channels, one for each of the six orientations.

The inherent difference between gradient histogram channels and histogram of oriented gradients is that the first employs spatial distribution of orientations at each pixel location, whereas the latter, accumulating the votes into orientation bins over local spatial regions, discards spatial information. Keeping spatial gradient information allows an approach for predicting the behavior of gradients in resampled images without resorting to analytical derivations that may be difficult and computationally expensive. The idea is exploited in Dollár’s *et al.* [24, 25] works, showing that it is possible to create high fidelity approximations of multiscale gradient histograms using gradients computed at a single scale, and briefly described in Sec. 2.4, as a general theory applicable to various feature types.

## 2.3 Local Decorrelation

Each node in our boosted cascade of decisional trees checks one feature at a time, which means that the data is orthogonally split. However, for highly correlated data it may be more effective to use oblique splits which involve checking two or more features at a time. This methodology requires decisional trees that are not limited to binary trees and, consequently, are much more computational expensive. Breiman’s [5] foundational work on random forests experimented with oblique trees. To achieve similar advantages while avoiding the computational expenses of oblique trees, Woonhyun *et al.* [30] proposed to decorrelate features prior to applying orthogonal trees. They have shown that orthogonal trees trained on decorrelated data may be equally or more effective than oblique trees.

Decorrelation is a common pre-processing technique used for classification and clustering. Inspired from Hariharan *et al.* [22] who proposed an efficient scheme for estimating covariances between Histograms of Oriented Gradients with the goal of replacing linear SVMs with LDA and thus allowing for fast training, Woonhyun *et al.* [30] have shown that significant improvements are brought if the channel features are decorrelated after they were effectively computed.

To obtain the decorrelated channel features, for each patch  $p$  of a fixed size in the channel features, the covariance matrices  $\Sigma$  should be computed. Woonhyun *et al.* [30] proposed to compute a single fixed covariance matrix  $\Sigma$  for each channel, shared across all local channel neighbourhoods, under the assumption that the statistics of natural images are transitionally invariant. The covariance between two features should depend only on their relative offset. For every patch we can create a decorrelated representation of the features by computing  $Q^T p$ , where  $Q A Q^T$  is the eigendecomposition of  $\Sigma$ .

The authors have shown that using 4 eigenvectors per channel for patches of size 5 is sufficient, and have argued that the projection  $Q^T p$  can be computed by a series of 4 convolutions between a channel image and each learned  $5 \times 5$  filter reshaped from its corresponding eigenvector (column of  $Q$  matrix). This is possible because the covariance matrix  $\Sigma$  is shared across all patches per channel and hence the derived  $Q$  is spatially invariant.

Fig. 1 shows the results of the convolution between an image and the four decorrelation filters that have been learned for normalized autobinomial Markov channels. For each channel, a set of 4 decorrelation filters of size  $5 \times 5$  pixels were learned from training sets and applied in both training and scanning processes. To fairly compare the performance of

using locally decorrelated channels over the original ones in the feature selection process, each decorrelated channel has to be downsampled by 2 which means that the total number of features is 4 times smaller and so equal to the initial feature space.

## 2.4 Scaling Features

To speed up the scanning process, we compute the channels only for the native resolution  $\Omega_b$  and approximate them to other twenty four scales  $\Omega_s$  by using Piotr Dollár's *et al.* [14] feature pyramids. A feature pyramid is a multi-scale representation of an image  $\Omega_b$  where channels are computed at every scale  $s$ .

As explained in the previously mentioned work, the channels for different scales may be extrapolated from nearby scales by weighting the values from the original scale with an empirically determined ratio  $\mu$ , each channel type having its own corresponding ratio. The factor  $\mu$  represents the mean of a distribution of overall divisions between the original channel's values and the ones computed explicitly for a scale  $s$ .

Note that here for each of the two scales the channel features were explicitly computed, and therefore, to be able to make any sum between the correspondent locations they have been brought at the same scale. By upsampling we do not lose information, but when an image is downsampled many of the high frequencies are cut.

The best scenario is to bring both images at the least common multiple, but this alternative is not used in practice. The ratio is exponentially dependent of  $s$  being defined as  $\mu_s = \hat{\mu}_0 e^{-\hat{r}s}$ , where  $\mu_0$  is an artefact of the bilinear interpolation.

To estimate  $\hat{r}$  and  $\hat{\mu}_0$  the author used a least squares fit of the exponential function to the twenty four means computed explicitly over a set of images, one for each scale.

## 3 Boosted Trees

Over the past few years the algorithms for boosting trees have evolved from the application of boosting methods to regression trees. The general idea is to compute a sequence of simple trees as being weak classifiers, where each successive tree is built for the prediction residuals of the preceding tree. Fig. 1 allows a broad view of a cascade of boosted trees and lists its major components.

This method builds binary trees by partitioning the data into two sets at each split node which implies that at each step of the boosting, a best partitioning of the data is determined, and the deviations of the observed values from the respective means are computed and the scores of each leaf are set. The next tree in the cascade will then be fitted to those residuals, to find another partition that will further reduce the residual variance for the data, given the preceding sequence of trees.

**Learning algorithm:** The learning algorithm we use is represented by such a cascade of boosted trees, where each binary tree has a maximum depth of two levels, leading to a maximum of three decisional nodes and four leaves.

In the training process, for each decisional node, both a channel identifier and the offset in that channel that has the minimum separation error between positive and negative samples are assigned. Each leaf returns a positive or negative value which is added to the overall sum when scanning a certain window. When the sum drops below the cascade threshold, the window is rejected. Particularly, the training process consists of a couple of stages gradually increasing the number of weak classifiers and bootstrapping negatives after each stage.

Learning the classifiers in more stages helps to bootstrap negative samples between each training stage. This means that after each stage is trained, more negative samples are added to the initial set by testing the previously learned cascade of trees on new negative samples and keeping those that were misclassified.

## 4 Experiments and Performance Evaluation

In this section we discuss the effectiveness of using the normalized autobinomial Markov channels. We evaluate the performance of different configurations, such as various color spaces (GRAY, RGB, HSV, CIE-LUV), lower and higher distance conditions (that influence the number of the neighbourhood systems, one channel being acquired for each neighbourhood system), features computed on a certain color plane or between channels, with and without smoothing.

The current practice is to use the INRIA Pedestrian Dataset for training and testing, even if it is amongst the oldest and has comparatively few images. However, it benefits from high quality annotations of pedestrians in diverse settings and scenarios.

	NAMC			Color Channels 3	GM 1	GH 6	Smooth	LD ×4	Log-Avg. MR [ $4 \cdot 10^{-3} - 10^{-1}$ ] FP
	Gray 1	Color 3	Multispectral 1						
<b>NAMC+</b>		✓	✓	✓	✓	✓	✓	✓	<b>15.57%</b>
NAMC-LDCF-4C		✓	✓	✓	✓	✓	✓	✓	15.84%
NAMC-LDCF-1C			✓	✓	✓	✓	✓	✓	17.08%
NAMC-ACF-8C	✓			✓	✓	✓	✓		18.82%
NAMC-ACF-4C		✓	✓	✓	✓	✓	✓		19.8%
<b>NAMC</b>			✓				✓	✓	<b>50.56%</b>
NAMC-4C-LUV		✓	✓						68.8%
NAMC-4C-RGB		✓	✓						96.21%
NAMC-4C-HSV		✓	✓						96.42%
ACF [15]				✓	✓	✓	✓		22.38%
LDCF [60]				✓	✓	✓	✓	✓	17.45%

Table 2: **Performance Evaluation.** From left to right each column refers to the feature type, and the corresponding number of channels. For those that use color channels, we refer to LUV color space, therefore GM comes from gradient magnitude, while GH are the gradient histograms. A smoothing may be applied across the channels, and channel features may be LD (locally decorrelated) or not. The MRs (miss rates) have been obtained by averaging only the results in the [ $4 \cdot 10^{-3} - 10^{-1}$ ] interval of FP (false positive) rates.

We have chosen to evaluate pre-trained detectors but also re-trained some of them, mainly those that we have been interested to improve [15, 60] by using our proposed features, or those that employ additional processing steps proven to increase performance [60].

Tab. 2 exemplifies the types of the features used for each test, as well as the post-processing algorithms, employed in each individual case, while Fig. 2 plots the results we achieved for different configurations.

If we only use our NAMC features with 4 channels (one for each color plane plus one computed between color planes) we outperform Viola and Jones’s [68] Haar-like features (VJ), Dalal and Triggs’s [10] HOGs, Felzenszwalb’s *et al.* [18] DPMs (trained on PASCAL), Dollár’s *et al.* [10] channels (FtrMine), while our best performance is obtained for NAMC+ that improves LDCF [60] with  $\sim 1.8\%$ . Fig. 3 shows the state-of-the art solutions on INRIA, but also includes others, in order to fill a wide range of feature types used so far.



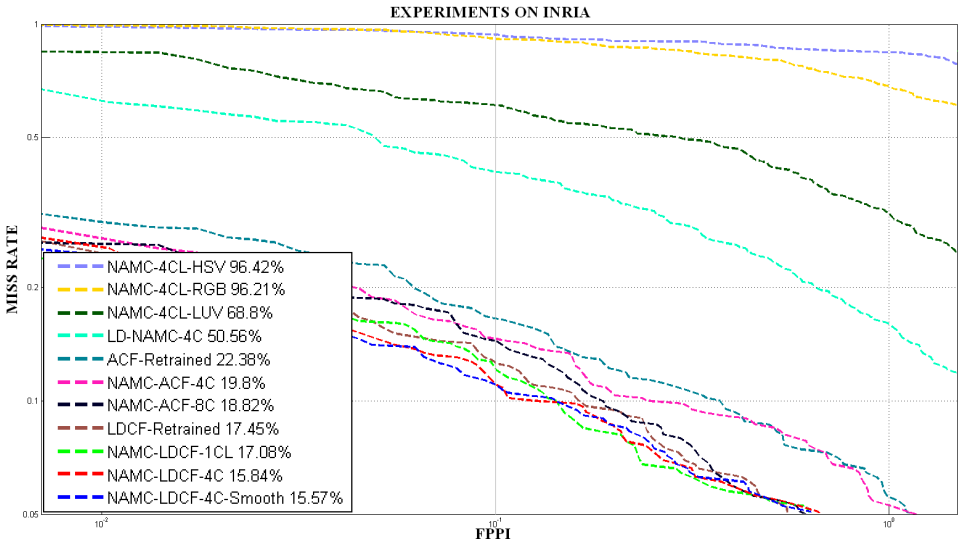


Figure 2: **Experiments on INRIA Pedestrian Dataset.** The normalized autobinomial Markov channels on LUV performs much better than for other color spaces, while by decorrelating our features we have got improvements up to  $\sim 18\%$ . It is better to use many normalized autobinomial Markov channels on Y, but if we extract fewer from different color spaces we get close results. An inexpensive configuration of the normalized autobinomial Markov channels used with gradient based features or/and a post-processing local decorrelation brings improvements up to  $\sim 2.6\%$  against ACF (Aggregated Channel Features) [15] and  $\sim 1.8\%$  against LDCF (Locally Decorrelated Channel Features) [6].

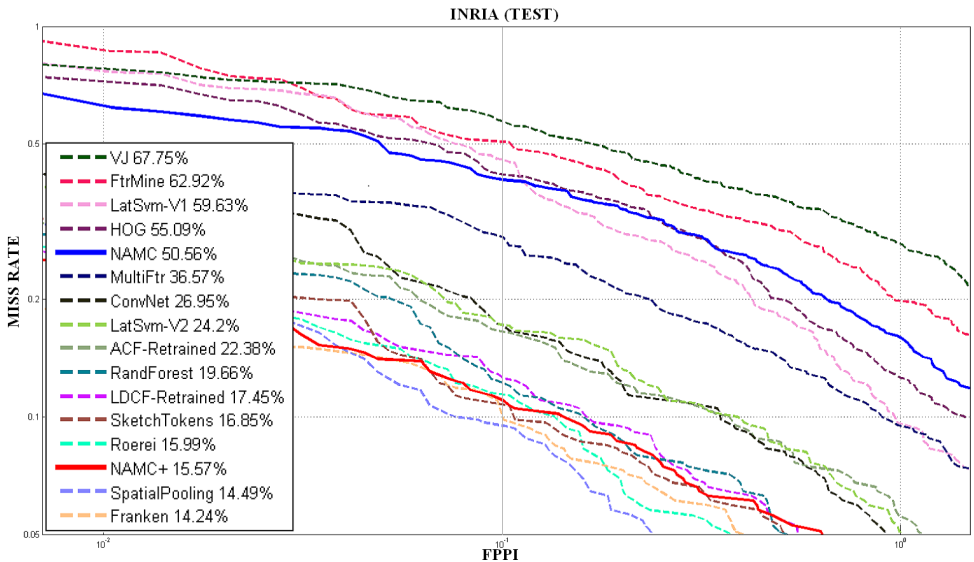


Figure 3: **Results on INRIA Pedestrian Dataset.** The log-average MRs (miss rates) have been computed considering a reasonable FP, lower than  $10^{-1}$  FPPI (false positives per image). Our **NAMC+** ranks in top three, preceded by Franken [28] and SpatialPooling [33], while our **NAMC** outperforms most of the solutions that have laid the foundations of pedestrian detection.

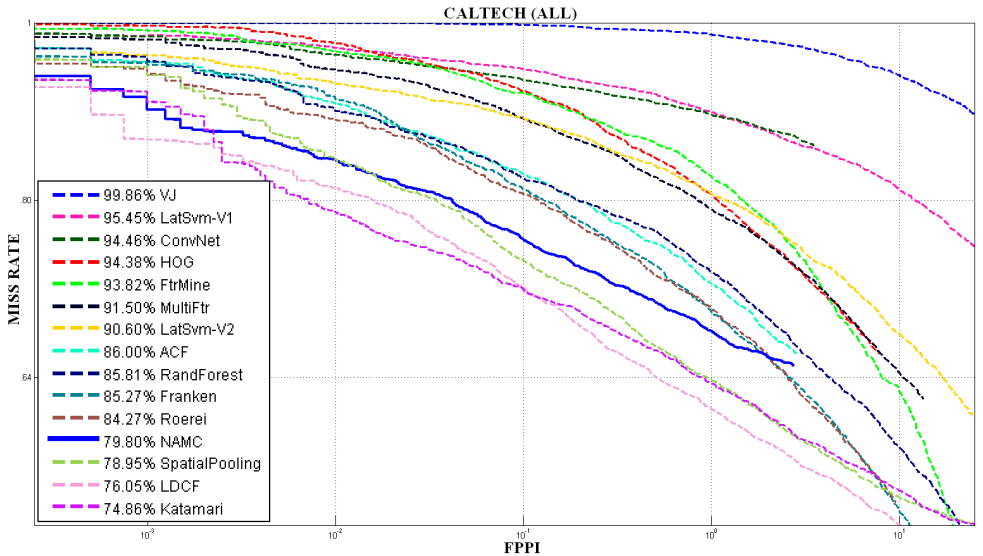


Figure 4: **Results on CALTECH Pedestrian Dataset.** The test includes all marked pedestrians, with a height of the bounding box of at least 16 pixels and less than 80% occlusion.

In order to ensure our results cover a large range of real case scenarios, we have employed the Caltech Pedestrian Dataset [14], currently ranked as the largest and most complete dataset in the field of pedestrian detection. The obtained results shown in Fig. 4 demonstrate the efficiency of our approach against the tested state-of-the-art solutions.

The main advantage of using the normalized autobinomial Markov channels as feature descriptor comes from the property of randomly selecting pixels to be part of the neighbourhood system, having a significant contribution for pedestrian detection in noisy scenarios. Another benefit is given by the fact that it shows several possibilities of optimizations by turning many of the computations into memory accesses, as briefly shown in Sec. 2.1.

## 5 Conclusions

Our work has been motivated by the inherent characteristic of natural images, namely the fact that adjacent pixel values are correlated. This characteristic allowed us to use probabilistic methods that, by their nature, are suitable for such problems. More precisely, we proposed the use of Markov random fields for pedestrian detection.

We have shown that combining the normalized autobinomial Markov channels together with gradient-based features in a cascade of boosted trees and using a method that substitutes the need for oblique splits, we outperform the majority of the existing features and methods for pedestrian detection.

Further work will consist in an extensive evaluation of robustness and functional limitations of the proposed methodology, as well as assessing the computational costs against the actual state of the art solutions.

### Acknowledgement

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398.

## References

- [1] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *CVPR*, 2013.
- [2] Rodrigo Benenson, Mohamed Omran, Jan Hendrik Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? *CoRR*, 1411.4304, 2014.
- [3] M. Berthod, Z. Kato, S. Yu, and J. Zerubia. Bayesian image classification using markov random fields. *Image and Vision Computing*, 14(4):285 – 295, 1996.
- [4] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 1974.
- [5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [6] P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Texts in Applied Mathematics. Springer, 1999.
- [7] R. Chellappa and S. Chatterjee. Classification of textures using gaussian markov random fields. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(4): 959–963, 1985.
- [8] J. Chen and C.K. Tang. *Spatio-Temporal Markov Random Field for Video Denoising*. 2007.
- [9] F.S. Cohen, Z. Fan, and M.A. Patel. Classification of rotated and scaled textured images using gaussian markov random field models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):192–202, 1991.
- [10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893, June 2005.
- [11] P. Dollar, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.
- [12] P. Dollar, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *Proceedings of the British Machine Vision Conference*, pages 91.1–91.11. BMVA Press, 2009.
- [13] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, June 2009.
- [14] P. Dollar, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *Proceedings of the British Machine Vision Conference*, pages 68.1–68.11. BMVA Press, 2010.
- [15] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1532–1545, Aug 2014.
- [16] Piotr Dollár. Piotr’s Computer Vision Matlab Toolbox (PMT).

- [17] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34, 2012.
- [18] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.
- [19] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, Sept 2010.
- [20] S Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- [21] J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. *Unpublished*, 1971.
- [22] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Computer Vision-ECCV 2012*, volume 7575, pages 459–472. Springer Berlin Heidelberg, 2012.
- [23] K. Held, E.R. Kops, B.J. Krause, W.M.I.I.I. Wells, R. Kikinis, and H.-W. Muller-Gartner. Markov random field segmentation of brain mr images. *Medical Imaging, IEEE Transactions on*, 16(6):878–886, 1997.
- [24] D. Levi, S. Silberstein, and A. Bar-Hillel. Fast multiple-part based object detection using kd-ferns. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 947–954, June 2013.
- [25] M. Malfait and D. Roose. Wavelet-based image denoising using a markov random field a priori model. *Image Processing, IEEE Transactions on*, 6(4):549–565, 1997.
- [26] B.S. Manjunath and R. Chellappa. Unsupervised texture segmentation using markov random field models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):478–482, 1991.
- [27] J. Marin, D. Vazquez, A.M. Lopez, J. Amores, and B. Leibe. Random forests of local experts for pedestrian detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2592–2599, Dec 2013.
- [28] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool. Handling occlusions with franken-classifiers. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1505–1512. IEEE, 2013.
- [29] J. Moussouris. Gibbs and markov random systems with constraints. *Journal of Statistical Physics*, 10(1):11–33, 1974.
- [30] W. Nam, P. Dollar, and J.H. Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems 27*, pages 424–432. Curran Associates, Inc., 2014.

- [31] Wanli Ouyang and Xiaogang Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3258–3265, June 2012.
- [32] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2056–2063, 2013.
- [33] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *Computer Vision-ECCV 2014*, volume 8692, pages 546–561. Springer International Publishing, 2014.
- [34] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *Computer Vision-ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 241–254. 2010.
- [35] F. Preston. Gibbs states on countable sets. *Cambridge University Press*, 1974.
- [36] Z. Tu. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1589–1596. IEEE, 2005.
- [37] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(10): 1713–1727, 2008.
- [38] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 734–741. IEEE, 2003.
- [39] W. Woods. Two-dimensional discrete markovian fields. *IEEE Transactions on Information Theory*, 18(2):721–741, 1972.
- [40] J. Zhang. The mean field theory in em procedures for blind markov random field image restoration. *Image Processing, IEEE Transactions on*, 2(1):27–40, 1993.