Tensor Representations for Object Classification and Detection

Tosato, Diego; Cristani, Marco; Murino, Vittorio; Gong, Shaogang; Xiang, Tao

ISSN 2043-0167

Tensor Representations for Object Classification and Detection

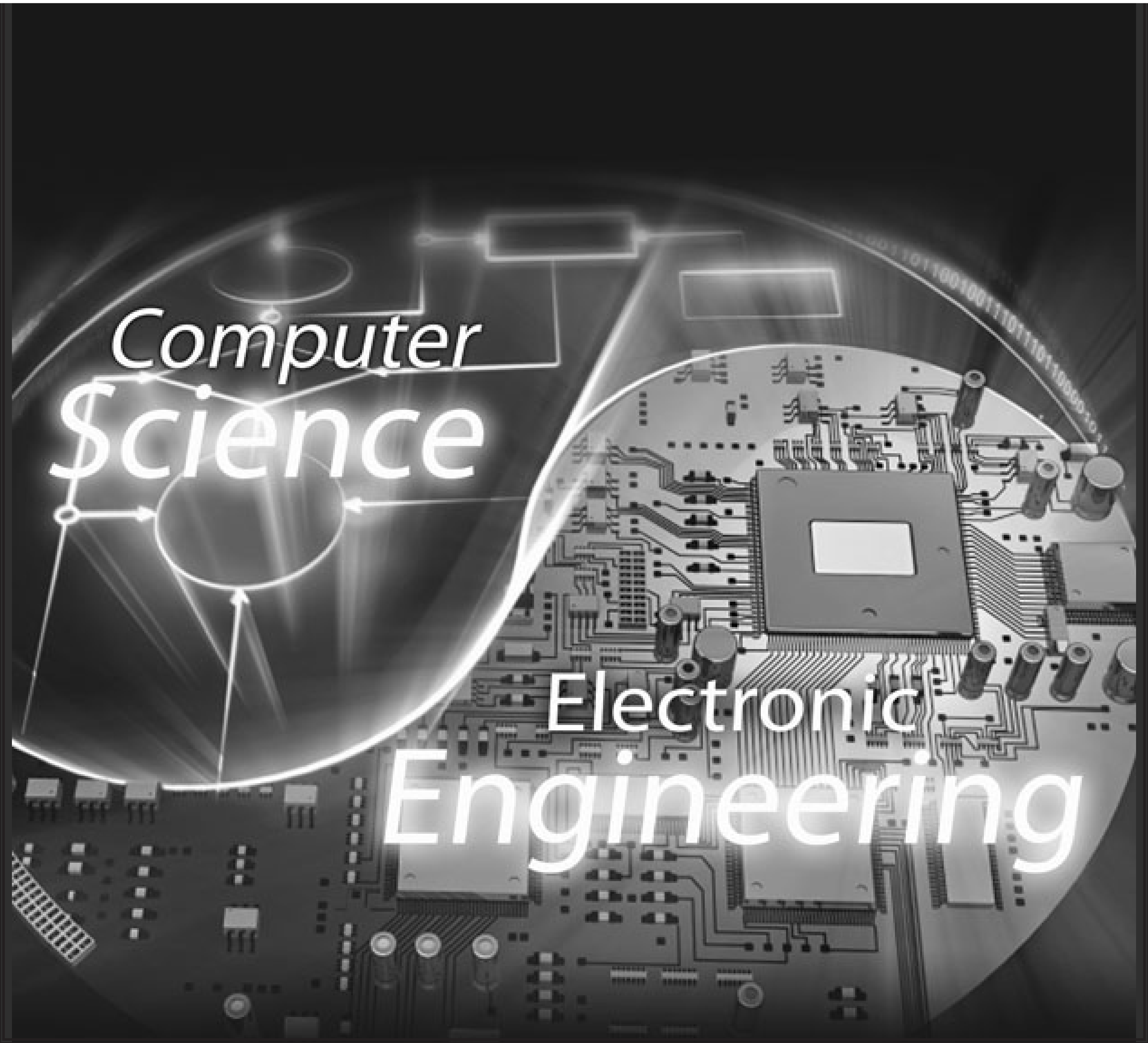
Diego Tosato, Marco Cristani, Vittorio Murino, Shaogang Gong and Tao Xiang

Queen Mary
University of London

EECSRR-11-04

October 2011

School of Electronic Engineering
and Computer Science



Tensor Representations for Object Classification and Detection

Diego Tosato[1], Marco Cristani[1,3], Vittorio Murino[1,3], Shaogang Gong[2], and Tao Xiang[2]

[1]Dipartimento di Informatica, University of Verona, Italy
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[3]Istituto Italiano di Tecnologia (IIT), Genova, Italy

Abstract

A novel method is proposed for robust detection and categorization of people from a wide-area distance by appearance and action. We adopt the tensor representation that is able to describe the interactions of multiple factors inherent to image formation and separately to encode the higher order statistics of each of these factors. Drawing inspiration by some successful works using covariance tensors [27, 28, 25], we propose two new kind of tensors that we have called EMI (Entropy and Mutual Information) tensor and SST (Self Similarity Tensor) that outperform the covariance representation on different classification and detection tasks. Then we present a complete framework for pedestrian detection based on the SST combined with Hausdorff distance that is able to manage object's description of variable lengths. We combine different object models with tensor representation. For all those the basic ingredient is the patch which replaces the concept of part because, in a video surveillance context, it guarantees better performances in terms of robustness. Finally we test the proposed tensor representations and the classification and detection approaches on different object classification (LabelMe [24] and Pascal VOC 2009 [12]), object recognition (Cifar 10 and Cifar 100 [19]), and pedestrian detection (DaimlerChrysler [23] and Caltech Pedestrian [9]) datasets.

Chapter 1

Introduction

A key problem in object recognition is finding a suitable object representation. For historical and computational reasons, vector descriptions that encode particular statistical properties of the data have been broadly applied. However, employing tensor (matrix) representation we are able to describe the interactions of multiple factors inherent to image formation and separately encode the higher-order statistics of each of these factors. Successful works that inspire what we are going to present is the covariance tensor (matrix) [27, 28] that has demonstrated to lead to state-of-the-art results for several classification and detections tasks. More generally, structure tensors and deformation tensors are used in image understanding, especially for segmentation, grouping, motion analysis and texture segmentation [5], and can also be utilized in regularization approaches for medical image registration [1, 15].

Mathematically speaking, a covariance tensor corresponds to an SPD (Symmetric Positive Definite) matrix and the value of its determinant is a direct measure of the dispersion of the associated Gaussian multivariate random variable. Fixing the SPD structure, but changing the information contained, we want to figure out if it is possible to build a tensor representation able to outperform the covariance matrix. In this report we have studied novel kinds of objects' tensor representation. To be more precise, we propose (1) two different kinds of tensor representation for objects' description that we have called EMI (Entropy and Mutual Information) tensor and SST (Self Similarity Tensor). EMI tensor is composed mixing entropy and mutual information and shows its potentiality in general object classification problems where it outperforms covariance representation. Differently, SST measures the self-similarity of an object composed by parts and it is suited for the object detection task. (2) We propose a framework for the task of the pedestrian detection in urban scenarios where pedestrians filmed by surveillance cameras can be at very low resolution. In a nutshell, the idea we propose is to replace the definition of person described as a set of fixed parts with a set of non-fixed patches (different number and position), which share a certain space location in the image. Patches are then pruned and only *reliable* patches survive that process. Finally, to decide if an image contains a person, we learn a binary SVM (Support Vector Machine) for which the kernel is built combining SST representation and Hausdorff distance in order to manage a human representation with a variable number of patches.

This report is organized as follow. Chapter 2 describes the EMI and SST. For each of them different sets of image features object models are utilized. Once the tensor representation is introduced, we show some experimental results on state-of-the-art dataset of object classification, recognition, and detection. Chapter 3 a complete framework based on covariance and SST and covariance tensors for pedestrian detection is described and some experiments highlighting the robustness to the occlusions of the proposed approach are reported. Then we test our framework on the [9] Caltech Pedestrian dataset. Finally, in Chapter 4, we draw our conclusions and we outline the future works.

Chapter 2

Multiple Features Tensor Representation for Object Description

In this section, we present an experimental study on objects' representation using the tensor description in which multiple features are combined together. In particular, we focus our attention on the Sym_d^+ ($d \times d$ symmetric positive definite) matrices. Through the covariance matrices [27], tensor representation has become popular, and it is applied on different computer vision problems like texture classification, [27], clinical imaging analysis and smoothing [15], pedestrian detection [28], visual object tracking [29], head orientation classification [25], and person re-identification [3].

Our goal is to understand if it is possible to exploit the Sym_d^+ matrix representation to build a more powerful object descriptor, processing features' information (like color, shape, etc.) to obtain better classification accuracy results. For this reason we introduce the Entropy-Mutual Information (EMI) tensor in Sec. 2.1 that shares the Sym_d^+ structure, but processes the information combining the more robust histogram representation and the entropy and mutual information measures. We apply EMI tensor to general object classification problems, and finer human body parts classifications finding that EMI tensor leads to considerably better performance than the covariance's representation.

In Sec. 2.2 we introduce another tensor representation for object (i.e. pedestrian) detection that we called SST (Self-Similarity Tensor). Differently from the tensors mentioned above it is designed to capture the structural information of an object. SST is built on a robust regular grid structure which suits well for the pedestrian detection task even in very low resolution conditions because the pedestrian's structure is similar in all the images. As for EMI tensor, we show that SST outperforms the covariance matrix representation. Moreover in the next Chapter we show how to use SST to build a kernel matrix for pedestrian detection where pedestrians can have a variable representation, namely they can be described with different number of parts.

2.1 Entropy-Mutual Information (EMI) Tensor

Similarly to covariance matrices, *Entropy-Mutual Information* (EMI) tensor is a dense region descriptor. In fact, given an image I of $W \times H$ pixels and a set of d feature maps $\Phi(I)$ of $W \times H \times d$ pixels:

$$\Phi(I) = [F_{1_{W\times H}}(I)),\ F_{2_{W\times H}}(I), \ldots, F_{d_{W\times H}}(I)], \tag{2.1}$$

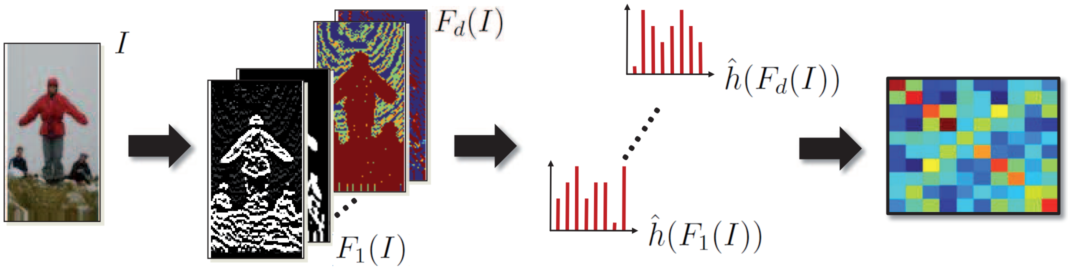


Figure 2.1: EMI descriptor. The d-dimensional feature map set $\Phi(I)$ is constructed from input image I.

where $F_1, \ldots, F_d$ are image features as shown in Fig. 2.1. Then, we use $\Phi(I)$ to build d histograms of n bins:

$$H(\Phi(I)) = [h(F_1(I))_{1\times n}, h(F_2(I))_{1\times n}, \ldots, h(F_d(I))_{1\times n}], \tag{2.2}$$

in which h is the operator used to build a histogram. In order to obtain a probability distribution from each feature, we normalize each row of $H(\Phi(I))$ such as $\sum_{i=1}^{n} h(F_j(I))_n = 1$ and $j \in \{1, \ldots, d\}$. We call the normalized version $H(\Phi(I))$ as $\hat{H}(\Phi(I))$:

$$\hat{H}(\Phi(I)) = [\hat{h}(F_1(I))_{1\times n}, \hat{h}(F_2(I))_{1\times n}, \ldots, \hat{h}(F_d(I))_{1\times n}]. \tag{2.3}$$

Using Eq. (2.3), we are ready to define the EMI tensor as follows:

$$\text{EMI}(I) = \begin{bmatrix} \text{E}(\hat{H}_1(\Phi(I))) & \cdots & \text{MI}(\hat{H}_{1d}(\Phi(I))) \\ \vdots & \ddots & \vdots \\ \text{MI}(\hat{H}_{d1}(\Phi(I))) & \cdots & \text{E}(\hat{H}_d(\Phi(I))) \end{bmatrix}, \tag{2.4}$$

where $\text{E}(\hat{H}_i(\Phi(I)))$ is the entropy operator defined as

$$\text{E}(\hat{H}_i(\Phi(I))) = \sum_{j=1}^{n} \hat{h}(F_i(I))_j \log(\hat{h}(F_i(I))_j) \quad i \in \{1, \ldots, d\}, \tag{2.5}$$

and $\text{MI}(\hat{H}_{d1}(\Phi(I)))$ is the mutual-information operator

$$\text{MI}(\hat{H}_{ij}(\Phi(I))) = \sum_{l=1}^{n}\sum_{k=1}^{n} \hat{h}(F_i, F_j(I))_{lk} \log(\frac{\hat{h}(F_i, F_j(I))_{lk}}{\hat{h}(F_i(I))_l \hat{h}(F_j(I))_k}) \quad i, j \in \{1, \ldots, d\}. \tag{2.6}$$

We represent the joint probability in Eq. (2.6) as $\hat{h}(F_i, F_j(I))$. Since EMI matrix belongs to Symmetric positive-definite matrices (or Sym_d^+ matrices) of real numbers, it is called tensor [2]. For classification purpose, we build a minimal representation EMI. Since it has only $d(d+1)/2$ independent coefficients, which are the upper triangular or lower triangular part of the matrix, we decide to consider only the upper triangular part vectorizing it. The resulting vector belongs to $\mathbb{R}^{\frac{d(d+1)}{2}}$ and the standard machine learning framework can be used with this representation.

2.1.1 Object Models for Object Classification

We decide to represent an object using different strategies for robust object classification. Firstly, we tried to use a single tensor for the entire object image. That representation has its pros and cons: it gives a compact and global picture of the object, and, since we use a tensor representation of fixed ($d \times d$) dimension, it is independent to the images size and resolution. Unfortunately it cannot manage occlusions, and it loses object's details which are useful to discriminate similar objects of different classes. However, using a single tensor we can obtain a clear picture of EMI's tensor potential regardless the object representation. The results of a comparison between covariance (COV) tensors and EMI tensors are reported in the first part of Sec. 2.1.2.

In order to improve the classification accuracy a more complex object's representation is adopted. To build a sufficiently general but discriminative descriptor we follow the idea proposed in [4] where a pyramidal patch based representation is used. In particular, each image is divided into a sequence of increasingly finer spatial grids by repeatedly doubling the number of divisions in each axis direction. The cell counts at each level of resolution are the bin counts for the histogram representing that level. We decide to adopt a 3 level pyramid and since EMI is based on multiple features we call that pyramid Multiple Features spatial pyramid, which is depicted in Fig 2.2.

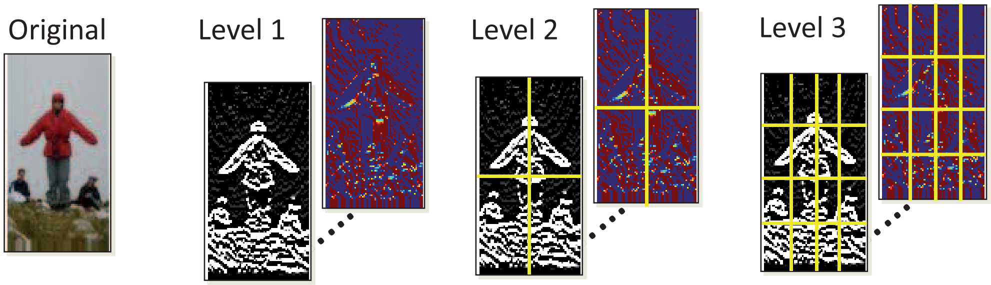


Figure 2.2: Spatial pyramid representation. An image on the left and grids for levels 1 to 3.

In order to make a fair comparison we decide to adopt the same structure as the one depicted in Fig 2.2 for COV tensors. In that case the histogram representation is not built because covariance is computed directly on the values of the pixels.

2.1.2 A Comparative Experimental Study

In this section a comparative study on different public available datasets for the object classification task is described.

PASCAL VOC 2009. This dataset [12] consists of a few 17895 high resolution images annotated with bounding boxes for objects of twenty categories (e.g., car, bus, airplane, ...). The goal of this challenge is to classify objects in realistic scenes (i.e. not pre-segmented objects). It is fundamentally a supervised learning problem where a training set of labeled images is provided. We choose that dataset to compare EMI to COV tensor representation accuracy. Recalling Eq. 2.1, we instantiate the same feature set for both COV and EMI:

$$\Phi(I, x, y) = \left[F_1(Y) \ldots F_8(Y) \; Y \; C_b \; C_r \; G_{|\,|}(Y) \; G_O(Y)\right], \tag{2.7}$$

where $F_1(Y) \ldots F_8(Y)$ is the filter bank consisting of scaled symmetric DOOG (Difference Of Offset Gaussian) [7], applied only to the luminance channel of the perceptually uniform CIELab color space. Y, C_b, and C_r are the three color channels obtained transforming the original RGB image. $G_{|\,|}(Y)$ and $G_O(Y)$ are the gradient magnitude and orientation calculated on the Y channel map, respectively.

To test the potentiality of the tensor representations, we use global representation of the objects using only one 13×13 ($d = 13$) tensor. After the vectorization (see Sec. 2.1), it produces a compact 96-dimensional vector representation. Random Forests [17] is used as a supervised learning toolbox. For this method, which builds an ensemble of tree classifiers, four parameters must be defined: (1) for each node, the feature to split a node is selected among a random subset of all the d_v features. The number of candidate feature is fixed to $\sqrt{d_v}$; (2) to guarantee good generalization performances of the classifier the number of samples per leaf is fixed to at least τ; (3) each tree is trained on a randomly drawn bootstrap sub-sample of the data, and here it is fixed using approximately $2/3$ of the examples. (4) the number of trees is fixed to $T = 100$ to reduce the amount of memory necessary to instantiate the classifier, since the implementation we have adopted [17] is not optimized.

For evaluation purposes we train the system using a 5-fold cross validation procedure. At each iteration, 100 examples per class are used in the training phase and all the remaining examples are utilized for testing. In Fig. 2.3 we report the best confusion matrices for EMI and COV. In this experiment EMI clearly outperforms COV representation with an average accuracy of 48% against the 45% provided by the COV tensor.

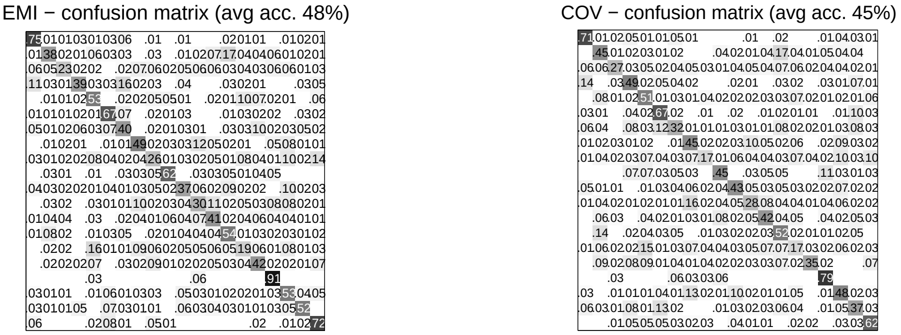


Figure 2.3: Confusion Matrices (CMs) for the PASCAL VOC 2009 [12] dataset. On the left the CM given using the EMI tensor, while on the right the CM associated with the COV tensor.

We want also to test tensor's representation in function of the images' resolution. Using bilinear resampling function provided by Dollar toolbox [7], we have downsampled all the Pascal's images. As you can see in Fig. 2.4, we made two different kinds of downsampling: in the first case we do not preserve the image's size, while in the second case we do. This is due to the fact that we want to study the behavior of the tensor in function of both image's size and resolution.

LabelMe. We used the annotated LabelMe [24] dataset to test the ability of the tensors representation to discriminate among more fine categories compared with the previous case. LabelMe is a database and an online annotation tool that allows the sharing of images and annotations. It is designed for object class's recognition and it contains various object classes. We extract from this dataset only 4 different object classes, all belonging to the same object as you can see in Fig. 2.5. The classes are 4 human body parts: arm, head, leg and torso. Images are reflected building a dataset of 16288 examples. Also in this case, as for Pascal VOC 2009, a 5-fold cross-validation procedure has been used. During each training phase 2000 randomly selected examples per class populate the training set and all the remaining are used for testing purposes. Each example is described with the feature set of Eq. 2.7 and, again, one tensor is used to describe an object's image. In Fig. 2.5 the CMs of EMI and COV tensors are shown. It is clear that EMI outperforms COV also in this finer classification task. Moreover, since the classes are highly overlapped we can claim that EMI manages better the presence of noise in images. This is probably due to the fact that it uses the histogram intermediate representation that improves the description robustness compared to COV tensors.

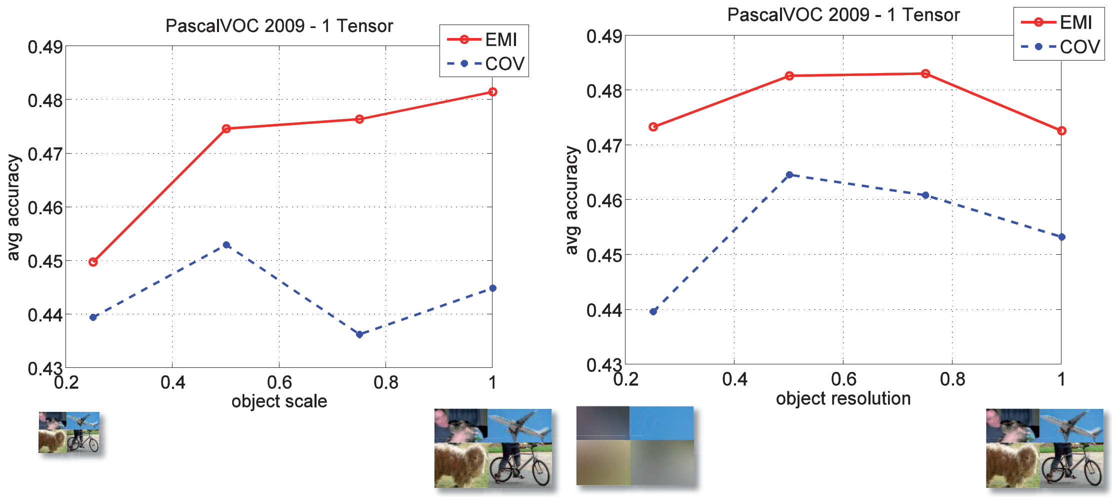
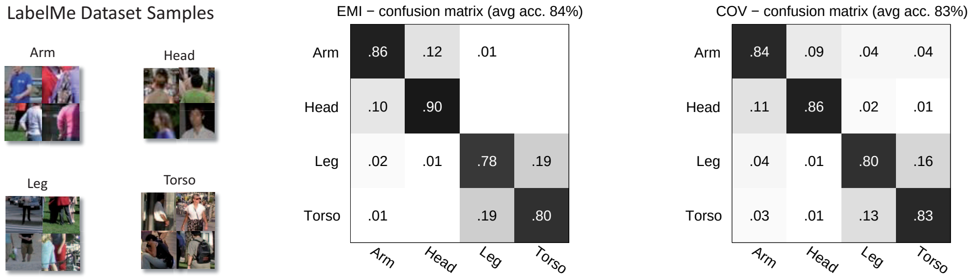


Figure 2.4: Classification performances of EMI and COV tensors on PASCAL VOC 2009 in terms of mean classification accuracy varying objects' scale and resolution.



EMI − confusion matrix (avg acc. 84%)

	Arm	Head	Leg	Torso
Arm	.86	.12	.01	
Head	.10	.90		
Leg	.02	.01	.78	.19
Torso	.01		.19	.80

COV − confusion matrix (avg acc. 83%)

	Arm	Head	Leg	Torso
Arm	.84	.09	.04	.04
Head	.11	.86	.02	.01
Leg	.04	.01	.80	.16
Torso	.03	.01	.13	.83

Figure 2.5: Some examples and Confusion Matrices (CMs) for the LabelMe [24] dataset. On the left the CM given using the EMI tensor, while on the right the CM associated with the COV tensor.

CIFAR10. The CIFAR10 dataset [19] is a hand-labeled subset of a larger dataset of 80 million tiny images. These images were downloaded from the Internet and down-sampled to $32x32$ pixels. The CIFAR10 subset has 10 object categories, namely airplane, car, bird, cat, deer, dog, frog, horse, ship, and truck (see Fig. 2.6). The training set has 5000 examples per class, the test set has 1000 examples per class. The low resolution and variability make recognition very difficult and a traditional method based on features extracted at interest points does not work. We learn RF as for the previous experiments in order to compare COV and EMI tensors. Unlike the previous case here the cross-validation procedure is not applied since the training and the testing set are already given. Since the recognition task on this dataset is hard, we decide to enhance the feature description using the pyramidal descriptor described in Sec. 2.1.1 that adds to the single (top layer) descriptor utilized before 2 sub-layers. For each patch of that pyramidal structure a tensor is extracted, vectorized and concatenated. The dimension of the final object descriptor is clearly larger if compared to using just one tensor for object description. Therefore we adopt PCA (Principal Component Analysis) to automatically reduce the dimensionality of the final object description. We chose a method developed by [30] because it automatically establishes the optimal feature descriptor dimensionality fixing to 96% the data's energy that should be preserved after the linear projection. That procedure is used both for EMI and COV tensors. In Tab. 2.1 we report a comparison using both EMI and COV tensors on CIFAR's images resized at a resolution of 128×128. Different features sets already implemented in the Dollar's toolbox [7] have been applied. The first filter bank has been already presented in Eq. (2.7). It is composed

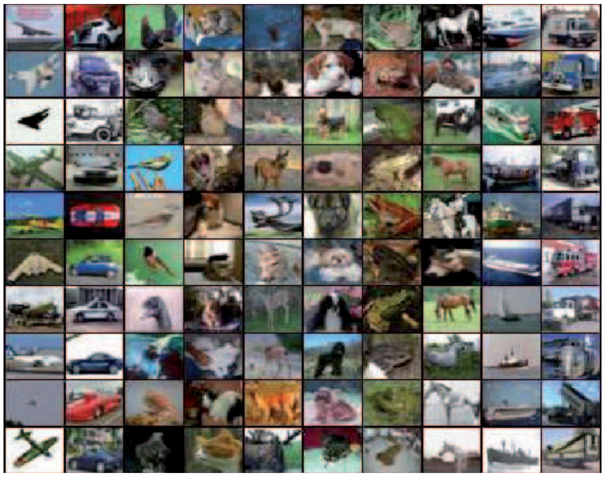


Figure 2.6: Example of images in the CIFAR10 dataset.

by a set of 8 DOOG filters and other Gradient and color features. We called that feature set DOOG in Tab. 2.1. Replacing the filters' set with a different filter bank from Serge Belongie [22] composed by 40 filters we build a much more informative filter representation that we called Belongie in Tab. 2.1. We observe that using the pyramidal EMI representation combined with Belongie's filters set we obtain the best performance outperforming the covariance representation. To consolidate that result we try to apply the comparison between EMI and COV on a much more difficult dataset in the next experiment.

Tensor Representation	Filters' Set	Avg Accuracy
EMI	Belongie	$\mathbf{52}\%$
EMI	DOOG	49%
COV	Belongie	40%
COV	DOOG	38%

Table 2.1: Test recognition accuracy on the CIFAR10 dataset produced by different pyramidal tensor representation.

CIFAR100. CIFAR100 dataset [19], as CIFAR10, is a hand-labeled subset of a larger dataset of 80 million tiny images. Also in this case images were downloaded from internet and down-sampled to $32x32$ pixels. CIFAR100 is made of 100 categories of objects. Its training set has 100 examples per class and its testing set has 100 examples per class. We use the same experimental setting as CIFAR10 that we have described above. In Tab. 2.2 the experimental results are reported. As for CIFAR10 the best average accuracy is obtained using pyramidal EMI tensor and Belongie's filters set, which confirms the superiority of EMI on COV tensor representation.

Tensor Representation	Filters' Set	Avg Accuracy
EMI	Belongie	$\mathbf{26}\%$
EMI	DOOG	32%
COV	Belongie	19%
COV	DOOG	18%

Table 2.2: Test recognition accuracy on the CIFAR100 dataset produced by different pyramidal tensor representation.

2.2 Self-Similarity Tensor (SST)

We have investigated also a different tensor that we have called *Self-Similarity Tensor* (SST) which can be used to robustly describe the structure of an object. SST is similar in spirit to structure tensors, which are powerful tools which can be used in such computer vision tasks as edge or corner detection [26] and spatio-temporal recognition [20]. The main idea that motivates the introduction of SST is to build an object representation to tackle different problems of interest in the surveillance context where a robust representation is necessary. We propose to build an object descriptor with a pyramidal layout which gives a coarse-to-fine object representation. It is probable that the top layer of this representation is more suitable for the detection task while the layers below for a finer classification or recognition.

From a mathematical point of view, the intuition is that given a patch-based representation of an object, it can be possible to find a compact and useful object description capturing the relationships between patches in a SPD matrix of distances among the patches (or parts). Then, SST can be vectorized and used as an object descriptor. More precisely, given an image I of $W \times H$ pixels and a set $\Lambda(I)$ of $W \times H \times m$ pixels of m image patches described by any kind of feature description (like HOG, COV, etc.):

$$\Lambda(I) = [f_{1\times n}(P_1(I)),\ f_{1\times n}(P_2(I)), \ldots, f_{1\times n}(P_m(I))], \tag{2.8}$$

in which f is a function producing an n-dimensional vector descriptor and P extracts a patch from the image I. Using Eq. (2.8), we define the SST as follows:

$$\text{SST}(I) = \begin{bmatrix} d(f(P_1(I)), f(P_1(I))) & \cdots & d(f(P_1(I)), f(P_m(I))) \\ \vdots & \ddots & \vdots \\ d(f(P_m(I)), f(P_1(I))) & \cdots & d(f(P_m(I)), f(P_m(I))) \end{bmatrix}, \tag{2.9}$$

Since the basic ingredient of STT are the covariance matrices that are proven to give superior performance in low resolution images [28], we decide to use different kinds of metrics to measure the distances among these matrices. In the simpler case we adopt the Euclidean distance ignoring the geometry of covariance tensors. On the contrary, in the second case we consider their geometry using a Riemannian metric. To be more precise, given a pair of COV tensors $X and Y$ the following distance is utilized:

$$d^2(X, Y) =< \log_X(Y), \log_X(Y) >_X = \text{trace}(\log^2(X^{-1/2} Y X^{-1/2})), \tag{2.10}$$

where $\log_X(Y) = X^{-1/2} \log(X^{-1/2} Y X^{-1/2}))X^{-1/2}$ is the Riemannian logarithm map and $\log$ is the ordinary matrix logarithm (see [25] for further details).

2.2.1 Object Model for Pedestrian Detection

We combine SST with a patch-based structure as described in Sec. 2.1, since that structure guarantees both a high level of robustness and generality to describe different classes of objects. Moreover, focusing our attention on small pedestrians, it is difficult to automatically extract a set of meaningful parts because of the low object resolution. However, a main issue still remains how to decide the patch size or rather the grid layout. Our hypothesis is that a rougher grid layout is suitable for a task like object detection in which the object model must be invariant (or at least less sensitive) to object details. Adopting a finer one, we necessarily have to change the task into an object classification task in which a high level of details is necessary to discriminate among classes. To this end in the next experimental section we confirm that hypothesis on a state-of-the-art pedestrian dataset.

2.2.2 Experiments

We present an experimental study to use that representation for *small* pedestrian detection task in real scenarios. To that end the DaimlerChrysler dataset [23] is chosen because it contains very small pedestrians. **DaimlerChrysler.** The DaimlerChrysler dataset [23] contains 4000 pedestrian (24000 with reflections and small shifts) and 25000 nonpedestrian images. The dataset was organized into three training and two test sets, each of them having 4800 positive and 5000 negative examples. The small size of the pedestrian windows (18×36 pixels), combined with a carefully arranged negative set, makes detection on the DaimlerChrysler data set extremely challenging. For this dataset we want to compare SST representation against COV representation. First of all to make that comparison as fair as possible, we utilize the same feature representation for both the tensors. In particular, we grid each image extracting 8 patches using the covariance of gradient-based information for each patch. The color information is not considered since it is not available for this dataset. More formally, the feature set is:

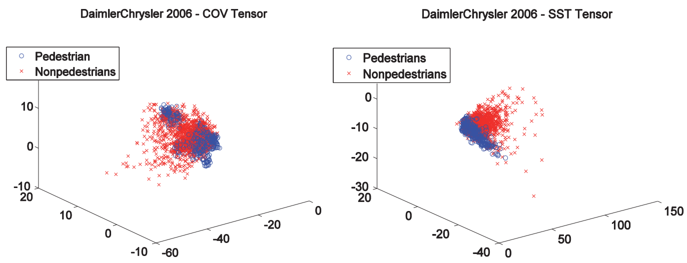


Figure 2.7: DaimlerChrysler feature space visualization via PCA using COV and SST.

$$\Phi(I, x, y) = \left[G_{|\,|}(I)\ G_O(I)\ D_x(I)\ D_y(I)\ D_{xx}(I)\ D_{yy}(I)\right], \tag{2.11}$$

where $G_{|\,|}(Y)$ and $G_O(Y)$ are the gradient magnitude and orientation, and $D_x(I), D_y(I), \ldots$ are intensity derivatives. Then we build a covariance matrix for each image patch using the feature set above. Covariances are vectorized and used as feature descriptors. Then SST is built computing the distance between each pair of descriptors as formalized in Eq. (2.9) where d is the Euclidean distance. On the contrary COV is built concatenating all the vectorized covariance matrices. In Fig. 2.7 we use PCA (Principal Component Analysis) to visualize the distribution of the negative and positive sets using the two different representations. We observe that SST offers a more linearly separable feature space with respect to COV. Hence, we expect that the detection performances of SST are reasonably better than COV. Now, we show in Fig. 2.8 another experiment in order to evaluate the behavior of SST at different patches' resolution. For this figure we build a pyramidal SST dividing an image into a sequence of increasingly finer spatial grids by repeatedly doubling the number of divisions in each axis direction. Hence, we show the feature space for different layers of the spatial pyramid. For each level of that pyramid an SST is computed and the feature space associated with each pyramid layer is visualized. We observe that a rough grid is more suitable for the detection task, while a finer grid subdivision can be used for a different classification task in which a high level of details is necessary (e.g. pose classification). To verify our assertion we compare the performances of the different pyramid layers for the pedestrian detection task. In Fig. 2.9(a), we plot the DET (Detection Error Tradeoff) curve on a log-log scale, whose y-axis corresponds to the miss rate, and the x-axis corresponds to false positives per window (FPPW). We notice that the first (top) layer is the most indicated for the detection task due to the fact that its rough image subdivision captures only the essential information to characterize an

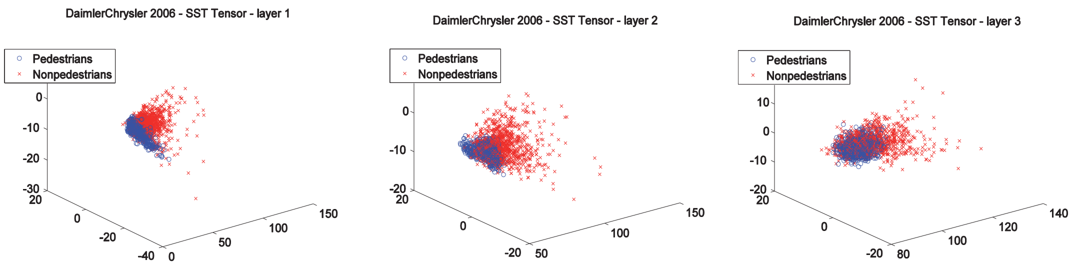
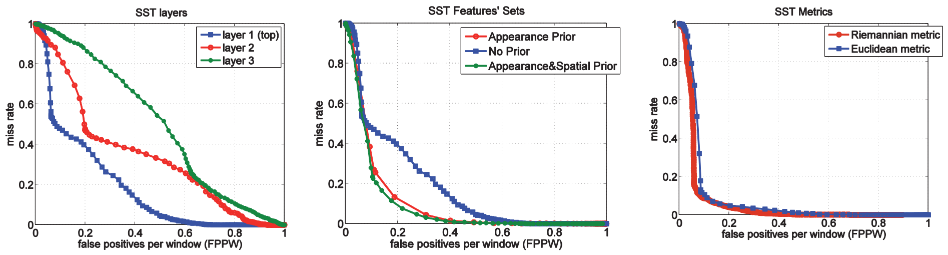


Figure 2.8: DaimlerChrysler SST's feature space visualization via PCA at different patches' size.

object avoiding object's details unnecessary for the detection task.



(a)

(b)

(c)

Figure 2.9: DET curve on the DaimlerChrysler dataset using the SST's tensor. (a) depicts the detection performances associated with different levels of the spatial pyramid. (b) shows how adding the spatial layout and an appearance prior to the feature descriptors the detection performances can be increased. (c) compares two different metrics that can be used to build the SST.

Chapter 3

Kernel's building via SST for Robust Pedestrian Detection

The capability to detect people in images of crowded scenes is fundamental for a large variety of applications, such as video surveillance or automatic driver-assistance systems. If people detection is performed in a non-problematic scenario such as one where people are not occluded, with a limited range of scales and pose variations there are already a lot of effective frameworks [6, 27, 16, 8, 14] usable to solve this task. On the other hand, if the scenario is problematic, among these systems only few are really useful. We highlight three of them which are able to manage different difficult problems which typically are present jointly in images of crowded scenes. [23] effectively deals with small scale pedestrians, [21] manages the presence of occlusions and [14] covers extreme changes of pose or occlusions of pedestrians. Since anyone of the previous frameworks is able to give a solution to all the previous problems, in this paper we want to propose a unified framework capable to jointly cope with the mentioned issues. Therefore, the goal is to detect as many people as possible even when it is not possible to infer the human body layout. A typical example of the scenario in which we want to work is depicted in Fig. 3.1. Here, it is very difficult to define a part-based model able to describe each person.

We propose to replace the definition of a person as a set of fixed parts as a set of non-fixed combination of human patches which share a defined space location in the image. Initially, an image is divided into a set of multi-scale overlapping patches on which a binary patch classifier is learned in order to highlight the patches belongings to people. Then we assign the human patches, if it is possible, to the different people in the image.

The ideas below our approach are: $1)$ a person is represented as a variable set of patches depending on a probabilistic evaluation of the patches' visibility, or rather if a human is occluded the patches containing the occlusion are automatically removed from the model. $2)$ since the number of patches is variable a classifier based on a set distance is used to discriminate between human and nonhuman image ROIs (Regions of Interest).

3.1 The Approach

The proposed approach is a five-phase process. (1) A set of features is calculated on a set of overlapping patches for each image. In the training stage, it assumes to have a set of ROI (Region Of Interest) containing fully-visible people at the same scale. Then the training set is populated by other problematic examples

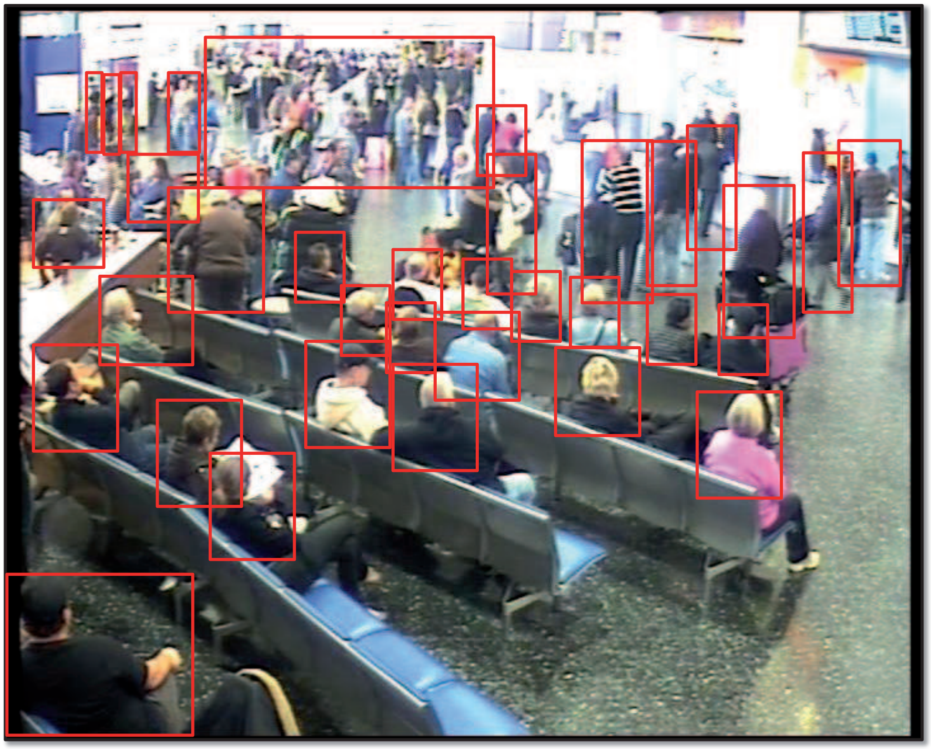


Figure 3.1: An example of video surveillance scenario from the i-LIDS dataset.

where occlusions are present. (2) From the training ROIs a set of feature is extracted and (3) extracting a fixed number of patches computed on a regular grid, their tensor descriptor is computed. (4) A robust binary patch classifier is used to detect the foreground (human) patches. (5) The survived patches are organized as sets and using a classifier based on a set distance we finally detect the presence of a human in the original ROI. The set distance is necessary since the number of foreground patches is variable. In Fig. 3.2 a the entire approach pipeline is depicted.

3.1.1 Person Representation

For the pedestrian detection task the most reliable source of information is related to the image gradient. As shown in [10], that information is strictly dependent on the image resolution. In particular for low resolution pedestrians (less than 30 pixels tall) Haar Wavelet features [23] are a simple and effective choice, while for medium and high resolution pedestrians it is preferable to use directly the gradient information or its orientation as done by HOG (Histogram of Oriented Gradients) [6]. To be able to manage people at different resolutions, combinations of the previous features are used [8]. This combination is typically a straightforward concatenation among some of the previous features. This leads to two problems: $1)$ using different features the normalization is not an easy task and it becomes more difficult proportionally to the number of the features involved. $2)$ The dimension of the final vector representation can be extremely high leading to the curse of dimensionality problem. A more proper way to combine different features and automatically solving both the previous problems is using covariance tensors as feature descriptors [27]. Due to the use of integral representation, these descriptors are fast to compute, making it suitable for

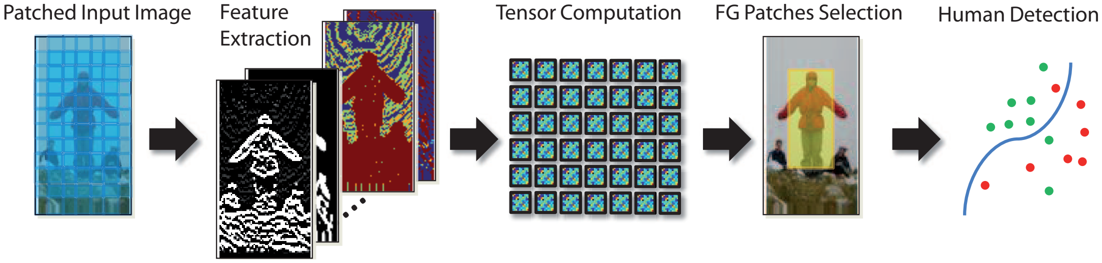


Figure 3.2: The proposed approach pipeline.

detection tasks. We have shown that there are others tensor representations (see Chap. 2) able to outperform the covariance, but their calculation time is still to long for object detection purposes.

Regular Grid Human Body Layout. As introduced at the beginning of this Chapter, we believe that in order to find a good representation for a person in a crowd scene where small pedestrians are presents, it is necessary to make a step further to the definition of human body part widely used for the current pedestrian detectors [14]. That because $1)$ a configuration of body parts changes in function of the object resolution. Even if multiple models are instantiated (one for each object's resolution), their management could be tricky and computationally expensive. $2)$ Defining a part automatically involve the part alignment problem. Since the part configuration can vary slightly with highly non-rigid object (as a human) or in case of occlusion, the research of the correct position and scale could lead to very poor results. $3)$ Parts are extremely unusable descriptors in crowded situations where it is hard to correctly assign parts to different overlapped human bodies.

We propose to divide an image I in overlapping patches on a regular grid. Each patch is described by a COV tensor. More formally, a set of patches $\{P_i\}_{i=1,\dots,N}$ of 4×4 pixels is sampled from I as shown in Fig. 3.2. We want to stress that differently from many successful people detector [28, 8], here the patch dimension p is not optimized in order to obtain the best performance on a benchmark dataset. This should be led to a more general detector in which the concept of fixed human parts is replaced by one that describe it as variable human patches.

3.1.2 Combinations of Features

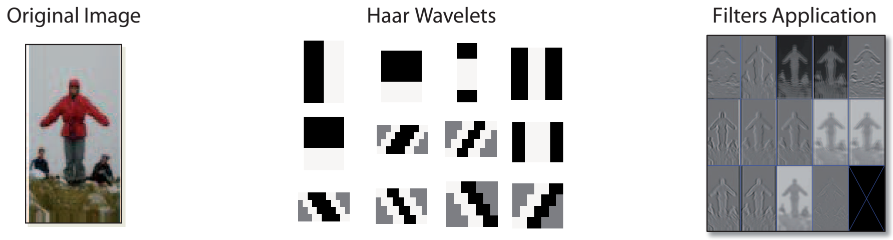


Figure 3.3: Haar wavelet bank for small pedestrians detection. Some of those wavelets are applied using kernel of different size

Each patch P_i is represented by a covariance matrix of d image features

$$\Phi = [H_1 \ H_2 \dots H_{10} \ G \ O] \, , \tag{3.1}$$

where d is equal to 12. $H_1, \dots, H_{10}$ represent the results of the application of a set of Haar Wavelets depicted in Fig. 3.3. G and O which are the gradient module and gradient orientation are computed the first two Haar Wavelets (see Fig. 3.3). The descriptor encodes information of the variances of the defined features and their correlations with each other, which are useful to detect both high and low resolution people. In order to build quickly a set of covariance matrices, given a set of feature Φ, in [27] is proposed a good solution based on the integral representation which is adopted in this paper.

Given a set of $d \times d$ covariance descriptors $\{C_i\}_{i=1,\dots,N}$ where $C_i \in Sym_d^+$ (the group of the symmetric positive definite matrices), they are one-to-one with their relative patches $P_1, \dots, P_N$. A very important preprocessing operation is the normalization of these descriptors to enhance the robustness to also include illumination variations in I. Unlike the local normalization proposed in [28], we propose to use a global normalization which is much more robust in presence of occlusions and noise. The normalized version of a covariance matrix C_i is denoted as $\hat{C}_i$ and is computed by dividing the columns and rows of C_i with the square root of the maximum variance of the image features Φ (Eq. (3.1)):

$$\hat{C}_i = \text{diag}(V)^{-\frac{1}{2}} C_i \, \text{diag}(V)^{-\frac{1}{2}}, \tag{3.2}$$

where $\text{diag}(V)$ is a diagonal matrix in which at the diagonal entries there are the maximum variance of the image features. This is equivalent to first globally normalizing the feature vectors to have zero mean and unit standard deviation and then computing the covariance descriptor.

Covariance matrices are an interesting way to combine information not only for the previous motivation, in fact their particular geometry provides an implicit framework to represent multi-modal distributions. So, if we are particularly interested in focusing our attention on a sub-set of covariances (i.e. people patches), exploiting their geometry a set of tools is naturally provided to find an highly discriminative Euclidean space to analyze them as described in the next section.

Covariance Tensors. Since covariance matrices do not live on a Vector Space but they can live on a Riemannian Manifold [2] it is necessary to map them on a particular tangent space of this Manifold where the covariances can be treated as vectors. More formally, given a normalized covariance matrix $\hat{C}_i$ it can be projected applying the following equation which represents the logarithmic mapping

$$\mathbf{c}_i = M^{\frac{1}{2}} \log_{Id}(M^{-\frac{1}{2}} \hat{C}_i M^{-\frac{1}{2}}) M^{\frac{1}{2}}, \tag{3.3}$$

where $M \in Sym_d^+$ is the Karcher mean point computed considering only the covariances belonging to people image examples and it is computed [18]. The $\log_{Id}(A)$ map is equal to $U \log(D) U^T$, where UDU^T is the eigenvalue decomposition of A. Please note that $\log_{Id}$ and $\log$ are different operators. The first one is a standard operator of the Riemannian geometry and the second one is the usual logarithm of a scalar value (for further details see [25]).

Since $\mathbf{c}_i \in Sym_d$, it contains only $d(d+1)/2$ independent coefficients which can be the upper triangular part of the matrix. As in [28], an orthonormal coordinate system for the tangent space is defined as follows:

$$\text{vec}(\mathbf{c}_i) = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \mathbf{x}_{d(d+1)/2}], \tag{3.4}$$

where $\mathbf{x} = M^{-\frac{1}{2}} \mathbf{c}_i M^{-\frac{1}{2}}$.

Having $d = 12$, a tangent vector is a 78 dimensional. Since not all the features are informative, linear PCA (Principal Component Analysis) is applied. According to [30] we preserve the 96% of the energy selecting the principal components, which number is automatically selected. We denote with $\tilde{\mathbf{c}}_i$ the principal components vector after the projection

$$\tilde{\mathbf{c}}_i = T\mathbf{c}_i,\ T \in \mathbb{R}^{d(d+1)/2 \times d_p} \tag{3.5}$$

where T is leaned during the training phase and d_p is automatically selected. As done above for the patch dimension, also here our goal is not to find the best feature set Φ to obtain the best performance on a benchmark dataset. We have collected a reasonable feature set that can be used to describe pedestrian at different scales and we have decided to use PCA [30] to select automatically the most informative subset of the original covariance C_i.

We add to $\tilde{\mathbf{c}}_i$ a further dimension containing a rough spatial information position in order to avoid patch configuration clearly infeasible. Dividing the ROI in 3 equal horizontal layers we assign 1 to the *top* body part, 0 to the *middle* and -1 to the *bottom*.

3.1.3 Patch Classification

We collect a large number of human and nonhuman patches and we learn a binary classifier using RF. We define $\Pr(\tilde{\mathbf{c}}_i)$ that represents the probability of a patch to belong to a human. That probability is computed as

$$\Pr(\tilde{\mathbf{c}}_i) = \frac{1}{T_n} \sum_{t=1}^{T_n} g_t(\tilde{\mathbf{c}}_i), \tag{3.6}$$

where T_n is the cardinality of a set of decision trees and $g_t(\tilde{\mathbf{c}}_i)$ is a decision function given by the t-th tree. Hence, $\Pr(\tilde{\mathbf{c}}_i)$ is computed as the mean of the decision responses coming from all the decision trees. Finally, if $\Pr(\tilde{\mathbf{c}}_i) > .5$ we decide that $\tilde{\mathbf{c}}_i$ is associated with a human patch. Clearly, we do not expect that this classifier is accurate, since extracting small patches the human and nonhuman classes has a large overlap. This is actually the reason why we have chosen RF as classifier, in fact it is able to manage very noisy data. However, we want to find a rough subdivision that removes patches that certainly do not belong to a human.

3.1.4 Object Detection based on Hausdorff distance

After the previous pruning phase we await to have a reliable set of patches for each example in the training set. Then, we build another classifier able to manage a variable representation of the same object to label a ROI as a pedestrian. First of all we decide to treat the feature descriptors of the survived patches independently, so we do not concatenate the descriptors in a unique vector because removing some patches we lose the order among the patches. Moreover, standard machine learning techniques cannot manage representation of different dimensionality. A popular distance among two sets of points that work regardless the number of descriptors in each set is the Hausdorff distance. It has been already used for object recognition in quite recent works [11, 13], but in these case object description were image coordinates. Since we work in $\mathbb{R}^n$ ($n = d(d+1)/2$) , we generalize the usual Hausdorff distance using the Euclidean norm of $\mathbb{R}^n$. Therefore to compute the Hausdorff distance of a pair of descriptors' sets $\tilde{C}_1, \tilde{C}_2$ we do as follow:

$$d_H(\tilde{C}_1, \tilde{C}_2) = \max[\max_{\tilde{\mathbf{c}}_i \in \tilde{C}_1}(\min_{\tilde{\mathbf{c}}_j \in \tilde{C}_2}(||\tilde{\mathbf{c}}_i, \tilde{\mathbf{c}}_j||)), \max_{\tilde{\mathbf{c}}_j \in \tilde{C}_2}(\min_{\tilde{\mathbf{c}}_i \in \tilde{C}_1}(||\tilde{\mathbf{c}}_j, \tilde{\mathbf{c}}_i||))] \quad \tilde{\mathbf{c}}_i, \tilde{\mathbf{c}}_j \in \mathbb{R}^n. \tag{3.7}$$

We choose the Euclidean norm for computational convenience, but any norm of $\mathbb{R}^n$ can be used to into Eq. (3.7). Than we embed d_H into an SST (see Sec. 2.2) computed on the training set that we call D. After that we build a kernel matrix exploiting D. Since D cannot satisfy the Mercer inequality itself to build a valid kernel that can be used in a SVM (Support Vector Machine) we apply the following non-linear transformation to D:

$$D^{+} = \exp(-\frac{1}{\mu}D), \tag{3.8}$$

where μ is the mean value of D. Applying that transformation we satisfy the Mercer inequality, hence the D^{+} is a valid kernel. In Fig. 3.4 an example of the kernel matrix based on d_H is shown.

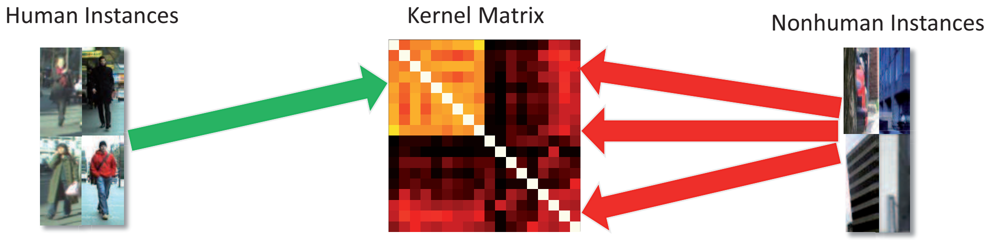


Figure 3.4: An example of Kernel matrix based on the proposed Hausdorff distance.

Once the kernel is built a binary SVM is learned for the final pedestrian detection task.

3.2 Experiments

In the first experiment we want to show the probabilistic output of the patch classifier described in Sec. 3.1.3 in presence of different types of synthetic occlusions. The goal of this experiment is to try to find a reliable set of patches that can be used to describe a human. In Fig. 3.5 we show the result of the application of different kinds of occlusions. We notice that, despite the grid of image patches is quite rough, the patches classifier provides useful information on which is the actual object's ROI for each occluded image. Ones can except that the segmentation should be finer, but for detection purposes we have to minimize the computational burden, therefore a rough image segmentation is enough for this first pruning phase. In the next experiment, regarding again the output of the patch classifier (Sec. 3.1.3), we show the probabilistic map produced the patch classifier in function of the image resolution. It is interesting to observe that the final probabilistic map is still reliable even when the original occluded image is heavily downsampled. That means two things: $1)$ the patch classifier can provide reliable information also in presence of heavy noise and low resolution images, $2)$ the feature set we have built (see Eq. (3.1)) is effective, so it captures discriminating information in very low resolution images. We train the proposed framework in the INRIA data set [6]. It contains 1774 pedestrian examples (3548 with reflections) and 1671 nonpedestrian images. The pedestrian annotations were scaled into a fixed size of 64×128 ROI, which includes a margin of 16 pixels around the pedestrians in the training images. The data set is partitioned into two, where 2416 pedestrian annotations and 1218 nonpedestrian images, from which we extract 100000 nonpedestrian ROIs of 64×128 pixels, are selected as the training set. The remaining images compose the testing set. Since that dataset does not include low resolution pedestrians, we decide to test our framework on the images of the Caltech data set [10] which contains several images with very low resolution pedestrians in crowded scenarios. In Fig. 3.7 some qualitative results are depicted. The proposed method achieves good performance in the pedestrian detection where the pedestrians are small. The number of false alarm is low, but many pedestrians are lost.

Discussion There are two main issues that must be tackled in order to improve the performance of the proposed detection approach. The first issue regards the efficiency: in fact, the usage of kernel methods in detection problems is very limited due to its computational burden. Since the patch detector permits to a considerable number of false positives to reach the kernel based classifier, it is difficult to build a light kernel

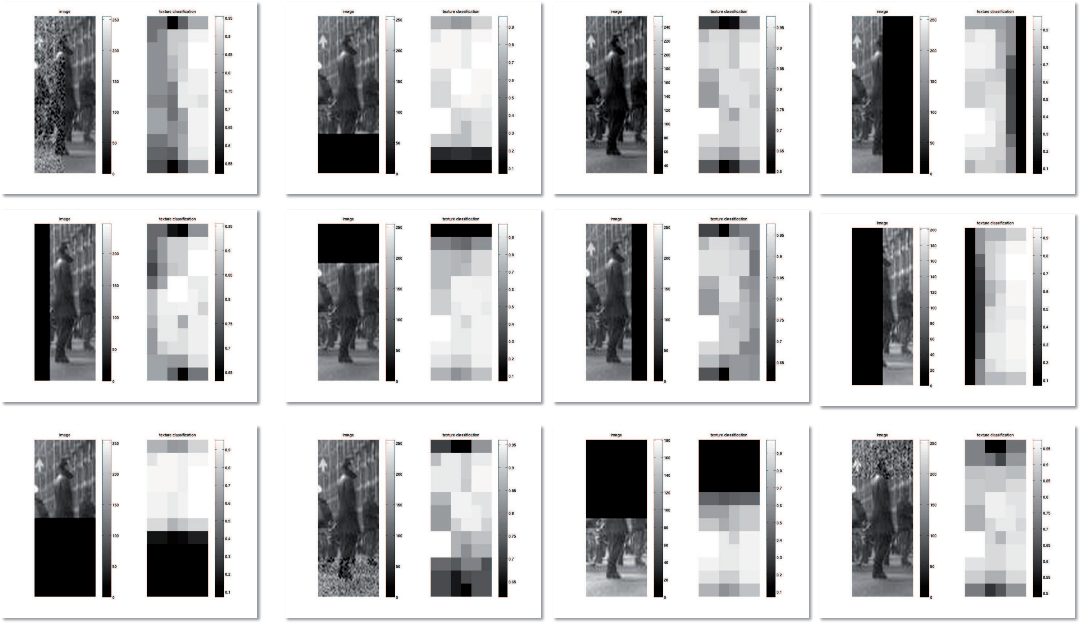


Figure 3.5: Patch classification in presence of different type of synthetic occlusions. For each picture we show on the left the occluded image and on the right the image patches classification that produces a probabilistic map. We randomly use different levels of occlusion: from soft (25% of the image size) to hard (50% of the image size). We also try various kinds of noise: full occlusion and salt& pepper noise.

that permits a fast detection. Thus, it is necessary to improve the performance of the patch classifier using contextual an spatial information during the pruning phase.

Another issue concerns the Hausdorff distance. That distance assumes that the information contained into the descriptor vectors is geometrical, namely vectors should contain coordinates of a $1, 2, \ldots, N$ dimensional space. In our case the descriptors contain different kind of information. That leads to an unclear meaning of that distance from the geometrical point of view. However, we have shown that the proposed distance is effective of the pedestrian detection task (see Fig. 3.4).

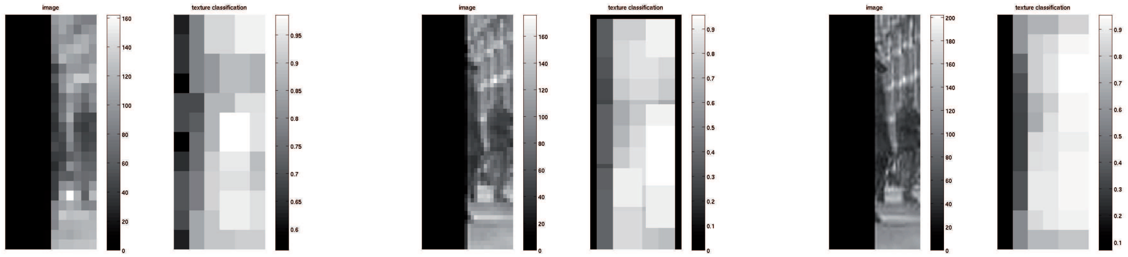


Figure 3.6: Patch classification at different image resolution. On the right we show the full resolution image that is downsampled one time to obtain the central image and two times for the left image. Each image present two maps: on the left the occluded image and on the right the image patches classification that produce a probabilistic map.

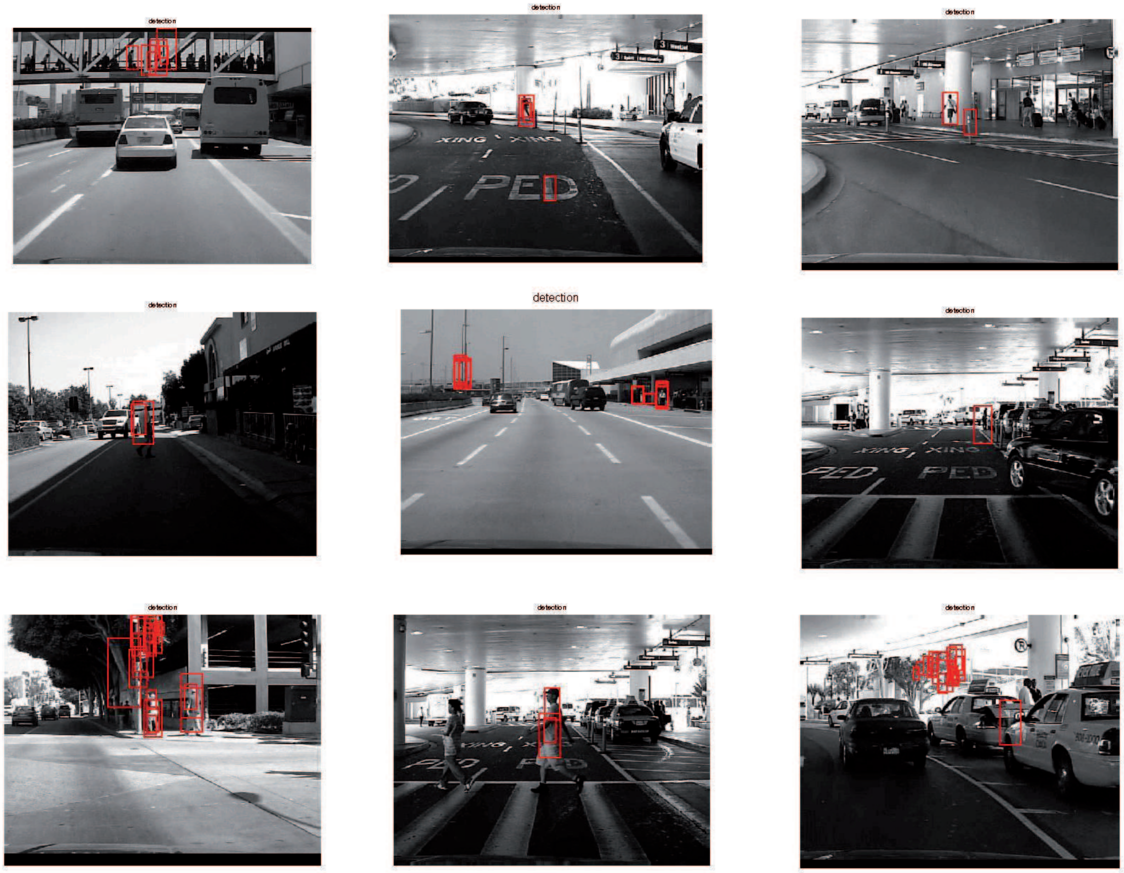


Figure 3.7: Detection examples. The classifier is trained on the INRIA data set [6]. Red boxes all the detection results without filtering or maximum suppression. In the first two rows there are good detection examples considering medium and low resolution pedestrians. In the last row we show problematic detection images.

Chapter 4

Conclusion and Future Works

We have presented a study of how to represent objects using tensors for classification and detection purposes. We have decided to use tensor representation because of the successful performances achieved using covariance tensors to represent objects. Thus, we have proposed two different tensor representations that we have called EMI and SST that lead to better performances compared to COV tensors. EMI tensor is composed mixing entropy and mutual information and shows its potentiality in general object classification problems where it outperforms COV representation. SST measures the self-similarity of an object composed by parts; it uses that structural information to discriminate an object. In the low resolution pedestrian detection task where pedestrian shape is almost the same, it outperforms the COV tensor on the DaimlerChrysler dataset.

We have proposed different object models that can be associated with tensor representation. In the straightforward case, a single tensor to represent an object is utilized. Then we have adopted a pyramidal structure in which each layer contains a regular grid of overlapping patches decried by tensors. In the latter case we have kept the regular grid structure, which is effective in surveillance scenarios, but we organize the patches in a set, therefore we overcome the necessity to have a fixed structure to describe an object improving the ability of managing occlusions or particular object poses. Combining the SST and the Hausdorff distance we have built a kernel for an SVM for pedestrian classification, able to manage a variable object representation. Since kernel methods are computationally expensive, we will look for a lighter version of the proposed Hausdorff distance that can guarantee a good level of performances in terms of detection computational time. Furthermore, we investigate a possible distance more suitable to compare feature descriptors in the same spirit as the Hausdorff distance. We will extend the experimentation on EMI and SST in order to find the most discriminative representation of object detection and classification. There are also two other aspects of tensor representation that must be investigated: first, a tensor normalization for SST and EMI tensor to achieve better classification and detection performances. Second, as for covariance tensor, can be interesting to figure out if it is possible to equip the proposed tensors with a metric different from the Euclidean that can simplify the learning phase.

Finally, another aspect that we will investigate is the combination of the tensors' representation. To be more precise, we want to see if combining tensors can lead to better classification performances.

Acknowledgements

This research was partially funded by the EU FP7 project SAMURAI with grant no. 217899.

Bibliography

[1] N. Archip, O. Clatz, S. Whalen, D. Kacher, A. Fedorov, A. Kot, N. Chrisochoides, F. Jolesz, A. Golby, P.M. Black, et al. Non-rigid alignment of pre-operative mri, fmri, and dt-mri with intra-operative mri for enhanced visualization and navigation in image-guided neurosurgery. *Neuroimage*, 35(2):609–624, 2007.

[2] V. Arsigny. *Processing Data in Lie Groups: An Algebraic Approach. Application to NonLinear Registration and Diffusion Tensor MRI.* PhD thesis, Ecole polytechnique, 2006.

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proc. AVSS*, 2010.

[4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proc. ACM Multimedia*, pages 401–408. ACM, 2007.

[5] T. Brox, J. Weickert, B. Burgeth, and P. Mrázek. Nonlinear structure tensors. *Image and Vision Computing*, 24(1):41–55, 2006.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 1, page 886, 2005.

[7] P. Dollar. Piotr dollar toolbox howpublished = ”http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html”, 2010.

[8] P. Dollar, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *Proc. BMVC*, 2009.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. CVPR*, pages 304–311. IEEE, 2009.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. PAMI*, 99(PrePrints):PrePrints, 2011.

[11] M.P. Dubuisson and A.K. Jain. A modified hausdorff distance for object matching. In *Proc. ICPR*, volume 1, pages 566–568. IEEE, 1994.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2009 (voc2009) howpublished = ”http://www.pascal-network.org/challenges/voc/voc2009/workshop/index.html”.

[13] P.F. Felzenszwalb. Learning models for object recognition. In *Proc. CVPR*, volume 1, pages I–1056. IEEE, 2001.

[14] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. PAMI*, 32(9):1627–1645, 2010.

[15] P. Fillard, X. Pennec, V. Arsigny, and N. Ayache. Clinical dt-mri estimation, smoothing, and fiber tracking with log-euclidean metrics. *IEEE Trans. MI*, 26:1472–1482, 2007.

[16] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *Proc. CVPR*. IEEE, 2009.

[17] A. Jaiantilal. `http://code.google.com/p/randomforest-matlab/`, 2009.

[18] H. Karcher. Riemannian Center of Mass and Mollifier Smoothing. *Comm. Pure and Applied Math.*, 30:509–541, 1997.

[19] A. Krizhevsky and GE Hinton. *Learning multiple layers of features from tiny images*. PhD thesis, Master's thesis, Department of Computer Science, University of Toronto, 2009.

[20] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *CMUI*, 108(3):207–229, 2007.

[21] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. CVPR*, pages 878–885, 2005.

[22] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 43(1):7–27, 2001.

[23] S. Munder and D.M. Gavrila. An experimental study on pedestrian classification. *IEEE Trans. PAMI*, 28:1863–1868, 2006.

[24] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1):157–173, 2008.

[25] D. Tosato, M. Farenzena, M. Cristani, M. Spera, and V. Murino. Multiclass classification on riemannian manifolds for video surveillance. In *Proc. ECCV*, pages 378–391. Springer, 2010.

[26] B. Triggs. Detecting keypoints with stable position, orientation, and scale under illumination changes. *Proc. ECCV*, I:100–113, 2004.

[27] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. ECCV*, 2006.

[28] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEE Trans. PAMI*, 30:1713–1727, 2008.

[29] J. Yao and J.M. Odobez. Fast human detection from videos using covariance features. In *Proc. IWVS*, 2008.

[30] W.S. Zheng, J.H. Lai, and P.C. Yuen. Penalized preimage learning in kernel principal component analysis. *IEEE Trans. NN*, 21(4):551–570, 2010.