Heterogeneous Auto-Similarities of Characteristics (HASC): Exploiting Relational Information for Classification

Marco San Biagio[1] marco.sanbiagio@iit.it

Marco Crocco [1] marco.crocco@iit.it

Marco Cristani[1,2] marco.cristani@univr.it

Samuele Martelli[1] samuele.martelli@iit.it

Vittorio Murino[1,2] vittorio.murino@iit.it

[1] Istituto Italiano di Tecnologia, Pattern Analysis & Computer Vision, Via Morego 30, 16163, Genova, Italy

[2] Università di Verona, Departimento di Informatica, Strada le Grazie 15, 37134, Verona, Italy

Abstract

Capturing the essential characteristics of visual objects by considering how their features are inter-related is a recent philosophy of object classification. In this paper, we embed this principle in a novel image descriptor, dubbed Heterogeneous Auto-Similarities of Characteristics (HASC). HASC is applied to heterogeneous dense features maps, encoding linear relations by covariances and non-linear associations through information-theoretic measures such as mutual information and entropy. In this way, highly complex structural information can be expressed in a compact, scale invariant and robust manner. The effectiveness of HASC is tested on many diverse detection and classification scenarios, considering objects, textures and pedestrians, on widely known benchmarks (Caltech-101, Brodatz, Daimler Multi-Cue). In all the cases, the results obtained with standard classifiers demonstrate the superiority of HASC with respect to the most adopted local feature descriptors nowadays, such as SIFT, HOG, LBP and feature covariances. In addition, HASC sets the state-of-the-art on the Brodatz texture dataset and the Daimler Multi-Cue pedestrian dataset, without exploiting ad-hoc sophisticated classifiers.

1. Introduction

Visual object classification and recognition remains one of the most studied problems in Computer Vision and Pattern Recognition. In this domain, the design of novel feature descriptors play a crucial role, with many and heterogeneous types proposed so far. In addition to the classical ”feature-based” descriptors (SIFT [13], HOG [2], LBP histograms [20] to quote some), in the recent years a novel trend has emerged, which consists of discarding the intrinsic value of the cues, encoding instead their inter-relations. We call this class of methods relation-based. The most known descriptor following this line is the covariance of features (COV) [17], in which linear correlations between features are exploited as elementary patterns. In the literature, relation-based descriptors exhibit a consistent invariance to many aspects (scale, illumination), making them ideal for object classes with high intra-class variability (as in the case of pedestrians [17]).

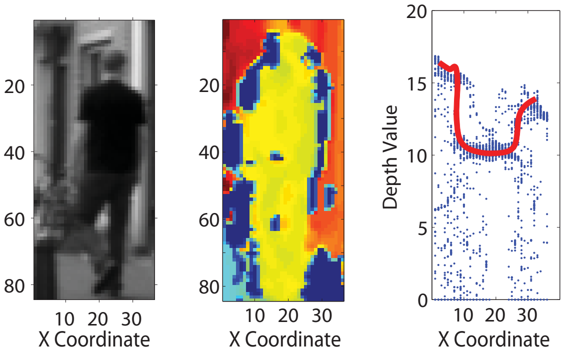

Figure 1. Example of non-linear relation between two features, depth values and x- coordinates.

In this paper, we pursue the relation-based approaches, designing a new descriptor which captures all the diverse relations that may hold between the characteristics of an object. More specifically, we claim that linear relations, in the form of covariances, are not enough to explain the complex structure of many objects. As an example, consider Fig.1, where the horizontal image coordinates x and the depth values of an image representing a pedestrian (taken from the recent Daimler Multi-Cue Occluded Pedestrian Classification Benchmark Dataset [4]) are scattered in the same plane: one can see that the depth values which model the human body remain approximately constant in the central part of the image, increasing rapidly while moving toward the external image portions, that correspond to the background scene. This demonstrates a clear example of non-linear relation between features.

To overcome this limitation, we propose a new local descriptor named Heterogeneous Auto-Similarities of Characteristics (HASC), which is able to encode at the same time linear and non-linear relations. The former are encoded by the covariance matrix of features (COV), and the latter are extracted by using information-theoretic measures, namely Entropy and Mutual Information, encoded into the here defined EMI matrix.

The entropy of a random variable measures the amount of uncertainty associated with the value of the variable itself. The mutual information (MI) of two random variables, instead, captures generic dependencies including the non-linear ones. In practice, EMI is a $d \times d$ matrix where the main diagonal contains the entropy values of the distribution of d features that characterize an object, while in the off-diagonal entries, the element i, j contains the mutual information between the $i-$th and $j-$th features.

It is worth noting that MI (and other information-theoretic measures) has been previously adopted in different computer vision tasks, i.e., to optimally align an image to a given template [19] or to detect saliency regions in an image [9], [10]. In all these approaches, these operators are employed as an objective function to be maximized in order to detect the highest similarity or saliency. We do not exploit MI for these purposes, but to encode the relational information from the data. To the best of our knowledge, EMI represents the first attempt to exploit information-theoretic measures to build a descriptor.

The modeling of linear and non-linear feature dependencies makes HASC a versatile descriptor for a large range of applications, employing heterogeneous basic features. Furthermore, we demonstrate that HASC is superior to its components (COV, EMI) considered separately, since they model complementary aspects of the same entity, and this fact will emerge many times in the paper. Through theoretical studies and synthetic experiments, we show that HASC is not a mere arbitrary juxtaposition of diverse elements, since joining together two descriptors does not guarantee an automatic improvement. In particular, if the two descriptors bring very similar information the overall performance may also decrease due to the well known curse of dimensionality phenomenon [3].

In the experiments, HASC was applied as a feature descriptor to different tasks: object recognition (Caltech-101 [6]), texture classification (Brodatz dataset [1]) and pedestrian detection (Daimler Multi-Cue [4]). In all the cases, employing simple discriminative classifiers, HASC obtains definitely higher performances than all the other considered descriptors (SIFT [13], COV [17], LBP [20], HOG [2]). In general, fed into advanced classifiers, or accompanied with other descriptors, HASC is highly competitive, especially on dealing with classes with high intra-class variance. In addition, we set the best performance on the texture classification and the multi-cue pedestrian classification tasks.

Finally, the fact that HASC has few parameters to be set, and that there is a large plateau of parameter values that ensure in general optimal performances, promotes the idea of exploiting relation-based principles for standard tasks in computer vision.

The rest of the paper is organized as follows. In Sec. 2 our approach is introduced, with a short recap on the COV descriptor (Sec. 2.1), and the definition of the EMI descriptor (Sec. 2.2): HASC is then presented in Sec. 2.3, with some toy examples aimed at highlighting the complementarity of EMI and COV. Experiments in Sec. 3 report the comparative performances of HASC, and, finally, Sec. 4 concludes the paper, with some observations and future perspectives.

2. The Proposed Approach

2.1. Covariance descriptor

Let $\mathbf{I}(x, y)$ be a color image, possibly equipped with additional channels like depth, motion flow or thermal imaging. Let $\mathbf{F}(x, y)$ be a d-dimensional feature image extracted from $\mathbf{I}(x, y)$:

$$\mathbf{F}(x, y) = \phi\left(\mathbf{I}(x, y)\right) \tag{1}$$

where the d-dimensional vector function ϕ can include any mapping such as gradient orientation and magnitude, filter responses, etc. For a given rectangular patch P in $\mathbf{F}(x, y)$, containing K pixels, let $\{\mathbf{z}_k\}_{k=1..K}$ be the set of d-dimensional feature points inside P. The covariance descriptor of the patch P can be defined as follows:

$$\mathbf{COV}_P = \frac{1}{K-1} \sum_{k=1}^{K} (\mathbf{z}_k - \bar{\mathbf{z}})(\mathbf{z}_k - \bar{\mathbf{z}})^T, \tag{2}$$

where $\bar{\mathbf{z}}$ is the average of the point set. The d diagonal entries of the $d \times d$ matrix $\mathbf{COV}_P$ are the variances of each feature, whereas the off-diagonal entries are the covariances between pairs of features.

Since covariances belong to the Riemannian manifold Sym_d^+, calculating distances among them amounts to project them into adequate tangent spaces, as pointed out in [15]; this brings to the $(d^2 + d)/2$ vectorized version:

$$\begin{aligned}\mathbf{cov}_P = vec(Proj(\mathbf{COV}_P)) = [&COV_{P11} \\ &COV_{P21}\; COV_{P22}\; COV_{P31}\; \ldots\; COV_{Pdd}].\end{aligned} \tag{3}$$

where COV_{Pij} are the elements of the projected matrix $Proj(\mathbf{COV}_P)$.

As pointed out in [17], there are several advantages of using covariance matrices as region descriptors: they have been shown to be robust to noise, to pose change and low-dimensional; nevertheless, covariance descriptors have some limitations. In particular, a single pixel outlier may drastically change the values in eq.(2), making the descriptor non-robust against impulsive noise. More significantly, the covariance among two features is able to optimally encapsulate the features of the joint PDF only if they are linked by a linear relation. As soon as their relation becomes non-linear, or their joint PDF distribution becomes multi-modal, the covariance loses its expressiveness since drastically different joint PDF may have very similar covariances. In such cases, the implicit discriminative power contained in the joint PDF is not captured by covariances.

To overcome these drawbacks, we propose a new descriptor, based on entropy and mutual information among features, described in the next section.

2.2. Entropy and mutual information descriptor

The Mutual Information (MI) of a pair of random variables A, B is defined as

$$MI(A,B) = \int_A \int_B p(a,b) \log\left(\frac{p(a,b)}{p(a)p(b)}\right) dbda \quad (4)$$

where $p(a)$, $p(b)$ and $p(a,b)$ are the PDF of A, the PDF B and their joint PDF respectively. If $A = B$ MI becomes the entropy of A

$$E(A) = MI(A,A) = -\int_A p(a)\log(p(a))da \quad (5)$$

If a finite set K of realizations pairs $\{A : a_k, B : b_k\}_{k=1..K}$ are available, MI can be estimated as a sample mean of the quantity inside the logarithm [1]:

$$MI(A,B) \approx \frac{1}{K}\sum_{k=1}^{K} log\left(\frac{p(a_k,b_k)}{p(a_k)p(b_k)}\right), \quad (6)$$

Probabilities inside the logarithm can be estimated from the K realizations by Kernel Density Estimation (KDE) method, however such procedure is computationally demanding. In this work, a definitely faster alternative procedure to estimate probabilities has been adopted by building a joint 2D normalized histogram of values of A and B. In detail, each $p(a_k, b_k)$ is estimated by taking the value of the 2D histogram bin in which the pair (a_k, b_k) falls; $p(a_k)$ and $p(b_k)$ are then estimated by summing up all the bins corresponding to a_k or b_k, respectively. Thus, the ij-th entry of the EMI matrix related to a patch P can be defined as follows:

$$\mathbf{EMI}_{P\{ij\}} = \frac{1}{K}\sum_{k=1}^{K} log\left(\frac{\tilde{p}(z_{ki}, z_{kj})}{\tilde{p}(z_{ki})\tilde{p}(z_{kj})}\right) \quad (7)$$

where $\tilde{p}(.,.)$ and $\tilde{p}(.)$ are the probabilities estimated with the histogram procedure and z_{ki} is the value of the i-th feature at pixel k. The EMI matrix depends only on two parameters: the number of bins on which the 2D histogram is calculated and the support given by the patch size, i.e. the number of pixels K.

Each diagonal entry of the EMI matrix captures the amount of uncertainty or unpredictability related to a given feature, whereas off-diagonal entries capture the mutual dependency between two different features. It is worth noting that mutual information accounts for the *strength* of mutual dependency, irrespective of the particular *kind* of dependency, be it linear or non-linear.

Therefore we choose to build the EMI descriptor by simply vectorizing the $(d^2 + d)/2$ different values of the matrix as follows:

$$\mathbf{emi}_P = vec(\mathbf{EMI}_P) = [EMI_{P11}\ EMI_{P21}\ EMI_{P22}\ EMI_{P31}\ ...\ EMI_{Pdd}]. \quad (8)$$

2.3. Combining COV and EMI: the HASC descriptor

Based on their properties, COV and EMI descriptors are largely complementary, capturing different features of the joint underlying PDFs. In particular, COV provides information about the kind of dependency. Assuming the joint PDF of two features to be bivariate Gaussian, covariance encodes the slope of the principal axes of the Gaussian, but is largely limited to linear dependencies, and its expressiveness decreases as soon as the joint distribution becomes multi-modal and/or the functional dependency becomes non-linear. On the contrary, EMI is able to encapsulate the degree of dependency among features but could not express the functional form that such dependency takes. Putting together the two descriptors in a larger feature space may boost the overall discriminative power.

This idea leads to the definition of Heterogeneous Auto-Similarities of Characteristics (HASC) descriptor, defined as the concatenation of vectorized EMI and COV:

$$\mathbf{hasc}_P = \left[\mathbf{cov}_P \quad \mathbf{emi}_P\right]. \quad (9)$$

Since HASC is defined here by components that can be related to entities in an Euclidean space, as pointed out in (9), it is also lies in an Euclidean space: therefore, its usage for machine learning algorithms is straightforward.

Concerning the computational cost of HASC, let us consider an N pixels image subdivided in R regions, possibly overlapped. The computation of R COVs (one per region) of size $d \times d$ can be efficiently addressed with integral images [16], yielding a total computational cost of

[1] MI can be seen as the expectation of the quantity inside the logarithm.

$O\left(\left(N+R\right)d\left(d-1\right)/2\right)$. On the other hand, the computation of EMI implies, for each couple of features, the estimation of 1D and 2D histograms, the logarithm and the final sum, according to Eq. 7. Calculation of 2D histograms can be efficiently addressed with the integral histograms ([14]), yielding a computational cost of $O\left(N+BR\right)$, where B denotes the number of histogram bins. The logarithm of 2D histogram can be efficiently calculated by pre-allocating a look-up table of size $N+1$ and accessing the current value ($O(B)$); for 1D histograms the cost is by far inferior as they are calculated for each feature (not for each couple). Finally the sum over the pixels of the three logarithms amount to $O(N)$ operations. In total the number of operations required for a set of EMI matrices of size $d \times d$, calculated over R regions is given by: $O\left(\left(N+BR\right)\left(d(d-1)/2\right)\right)$. To make a numeric example, $B = 64$, $N = 150 \cdot 150$, $R = 64$, $d = 11$ give a computation ratio EMI/COV equal to roughly 4. Furthermore, considering that bin values equal to zero are dropped from the logarithm calculation, computational load is even decreased.

We carried out some synthetic illustrative examples of binary classification, in order to highlight the differences in the expressive power of the COV and the EMI descriptors, as well as their complementarity which accounts for the superiority of the HASC descriptor.

In particular, we consider a dense single-pixel feature extraction operator, as would be in the case of images, that process all the pixels considered as independent entities. Given an image of K pixels, we extract K values from each of the $d = 2$ kinds of features. As a result, from the features of the image, a $d \times d$ COV and a $d \times d$ EMI descriptors are created. This was done for all the images of a class, obtaining different instances of COV/EMI, that were been subsequently fed into a linear SVM classifier, with 3000 examples for training and 1500 for testing.

In the first example, the two features x, y extracted from the examples are generated as $y = 20x + noise$ for class 1 and $y = -20x+noise$ for class 2, where $noise$ is Gaussian i.i.d. (see Fig. 2 (a) and (b): each point in the figure represents the values of the pair of features (x, y) calculated on a pixel of a given image). In this case, COV is able to differentiate the sign of the covariance, reaching an accuracy of 100%, while the EMI reaches an accuracy of about 50%.

In the second case, the relation between the two features is based on the circle equation (non-linear relation), with two different noise intensities added to differentiate the two classes (see Fig. 2 (c) and (d)). Due to the non-linear relationship between the features and different noise intensity on the two classes, classification results are opposite with respect to the previous experiments. EMI reaches an accuracy of 100% while COV is almost equivalent to a random guess.

In the third case, we create two different classes in which the relation between the two features is non-linear but can be fairly approximated with a linear one, and each class brings a Gaussian noise of different intensity. In practice, the feature relation is governed by an ellipse equation, whose inclination is related to the class the element belongs to (see Fig. 2 (e) and (f)).

In this case, the performance for COV and EMI descriptors are similar, as shown in Fig. 2 (g): COV is able to catch the different slopes of the two ellipses, while EMI captures the different degree of non-linearity, modeled by the different noises. HASC, incorporating both the models, gives 100% of accuracy.

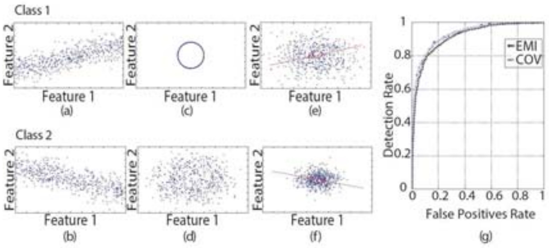

Figure 2. (a-f) Joint distribution of the two features for the synthetic examples; (g) Detection rate curves for COV and EMI.

3. Experiments

In this section, we want to show how HASC behaves in a diverse number of applications. First of all, we compare HASC against the most popular feature descriptors (HOG, SIFT, LBP, COV, SSD)[2] and the new EMI, in the object classification task, addressing the well known Caltech-101 dataset [6]. We demonstrate that employing a linear Support Vector Machine as a baseline classifier, HASC outperforms all the other descriptors. Furthermore, we carry out a detailed analysis of the impact of implementation details on the overall performances, drawing some conclusion on the descriptor robustness. Our code is publicy available at http://www.iit.it/en/datasets-and-code/code/hasc.html

This preliminary analysis served us also to discover the scenario where HASC may perform the best. In practice, this is realized in the case of discriminative non-linearities between features, where EMI can add more information with respect to COV. For this reason, we focused on the Brodatz dataset, where texture patterns offer a rich compound of relations among basis features (intensity, gradients) and on multimodal pedestrian detection, considering the Daimler Multi-Cue Occluded Pedestrian Classification Benchmark Dataset [4]. In such a case, diverse sensor modalities introduce complex links between low-level cues, considering intensity, motion and depth: EMI is able in this case to

[2] Textons have been shown to be inferior to COV in the paper [16], so this should convince us about their inferiority w.r.t. HASC, that outperforms COV.

capture them, allowing HASC to definitely outperform its two main ingredients.

On both datasets, we set the new state-of-the-art. In detail, we beat the previous best performance of [4] on Daimler Multi-Cue, by joining HASC with HOG+LBP, and the previous best performance of [12] on Brodatz. The overall underlying message is clear: HASC is able to finely encode relational class-specific information, surprisingly beating feature-based descriptions. Since relational and features-based descriptions are two sides of the same coin, joining them together appears to be a promising strategy, worth to be investigated in the future works.

3.1. Object Classification

The Caltech-101 dataset [6] represents a key benchmark for the object recognition community. It consists of 102 classes (101 object categories plus background). The significant variations in color, pose and illumination inside each of th 101 classes make this dataset very challenging. The number of images per class ranges from 31 to 800 and most of them are at medium resolution, roughly 250×280 pixels. The 15-dimensional vector of feature maps extracted from each image is defined as follows:

$$\mathbf{F}(x, y) = \begin{bmatrix} RGB & Lab & F_{med} & LBP & x & y & \mathbf{F}_V \end{bmatrix} \tag{10}$$

where R, G, B and L, a, b are respectively the three RGB and three CIELab image components, F_{med} denotes the median filter, LBP is the Local Binary Pattern [23][3], x and y are the horizontal and vertical pixel coordinates. These last two features are particularly interesting, since they allow to distill relations that hold between particular cues and their spatial position. Finally $\mathbf{F}_V$ is defined as follows:

$$\mathbf{F}_V = \begin{bmatrix} |V_x| & |V_y| & \sqrt{V_x^2 + V_y^2} & |V_{xx}| & |V_{yy}| \end{bmatrix} \tag{11}$$

where V_x, V_y, V_{xx} and V_{yy}, are the first- and second-order derivatives of the image intensity. To test our descriptor, we adopted the protocol of [21]: 30 per-class images are randomly chosen and subsequently split into 15 for training and 15 for testing; twenty different random partitions are considered and the average results with standard deviations are reported. Each image is re-scaled to 150×150 pixels, subdivided into different patches (4×4, 8×8 and 16×16) of different pixel sizes (60×60, 32×32 and 16×16), overlapping for half of their size. The aim here is to highlight the net superior expressiveness of HASC in comparison with COV and EMI taken alone, as well as other descriptors. To investigate this aspect, a baseline object model and a basic classifier are employed, in order to discard any classification strategy favoring a specific descriptor, in the same way as done in the HOG paper [2]; in other words, we are not interested in reaching top scores (actually, the performances reached are inferior w.r.t. the state-of-the-art), but to highlight the genuine differences among features. Moreover, to achieve the state-of-the-art on Caltech-101, it is nowadays mandatory to employ kernelized fusion of multiple descriptors [7], which would obscure the specific contribution of HASC.

In this setting, the final feature vector for an image was the simple concatenation of the feature vectors for each patch and a linear SVM was used as classifier (we also applied another classifier, i.e. Random Forest, but it did not provide consistent improvements probably due to the relatively high dimension of the image descriptors involved).

In Table 1, the best accuracies of each descriptor, with the related patch size, are reported. Results show that the HASC descriptor significantly outperforms all the other competitors. In particular, in all the twenty splits, HASC exceeds the performance of COV and EMI taken separately.

Table 1. Classification results for the Caltech-101 dataset

Descriptors	Results	Patch Size
HASC	**54.45** ±1.6	32 × 32 pixels
COV	51.32 ±1.3	32 × 32 pixels
EMI	47.57 ±1.1	32 × 32 pixels
LBP	41.25 ±0.5	32 × 32 pixels
HOG	38.91 ±0.6	16 × 16 pixels
SIFT	37.91 ±0.2	32 × 32 pixels
SSD	36.81 ±0.7	32 × 32 pixels

On this dataset, we investigate how the size and the number of the patches employed, and the number of bins used for the EMI computation affect the overall performances. In particular, the number of patches was constrained on their size, in a way that the whole image has to be covered by patches, that overlap for half of their sizes. The results are reported in Fig. 3 for HASC and EMI. Some observations can be drawn: first, employing HASC (or EMI) as a global descriptor, i.e. only one patch per image, is not very informative, no matter the number of bins adopted. The lower bound on the size of the patches is 8×8 pixels, since with lower sizes the performances degrade dramatically, as the statistics employed is too scarce. For EMI the best result is achieved with 12 bins and patch size of $16 - 32$; note that with increasing the patch size the best performance is obtained by increasing the number of bins as well. For HASC, the overall performance obviously increases and the dependence on the number of bins is dampened due to the contribution of COV (whose calculation does not imply bin quantization). In conclusion, HASC demonstrates notable scale invariance and robustness to bin quantization.

[3] Note that, differently from its standard use [23], here LBP is employed as a low level feature which provides just a single value, from 0 to 255, for each image pixel. Experimentally, we noted that adding this version of LBP increments systematically the performance of COV (and EMI) descriptors

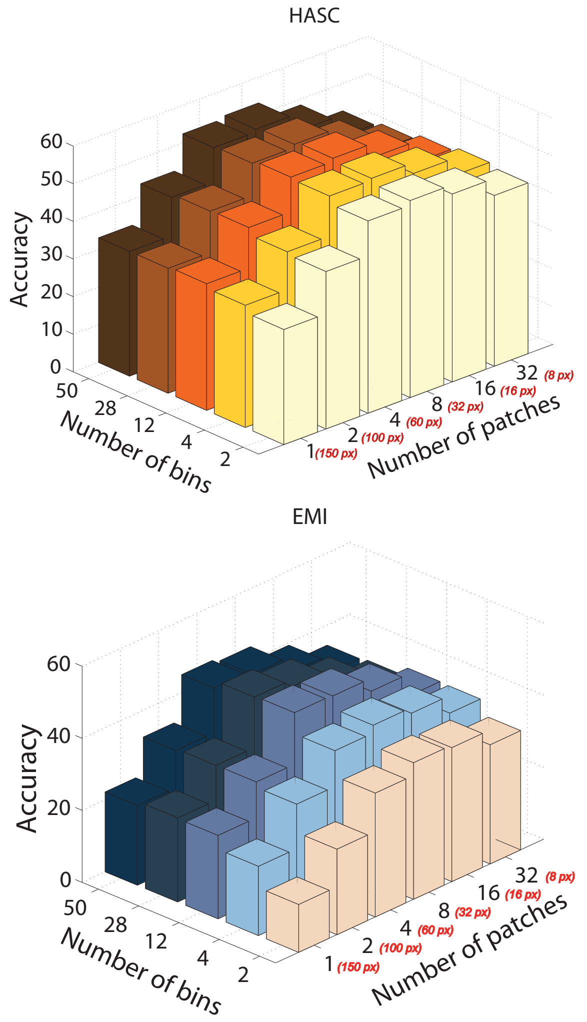

Figure 3. The accuracy of the HASC and EMI descriptor versus the number of bins and patch size. The stride (block overlap) is fixed at half of the patch size. The configuration with 8×8 patches and 28 bins performs best, with 53.72% accuracy on HASC and 45.10% on EMI.

3.2. Texture Classification

As classic application where COV achieves one of its best performances, we consider here the texture classification task on the Brodatz [1] database: covariances were originally used to encode repeated structural information, and textural imagery is one of the best benchmarks [16].

As experimental protocol, we follow [16]: we subdivide each image of the 112 classes into four sectors, obtaining 2 images per class in training and 2 images per class in testing, all of size 320×320. We extract the same features, i.e. intensity and magnitudes of first and second order derivatives:

$$\mathbf{F}(x, y) = \begin{bmatrix} V & |V_x| & |V_y| & |V_{xx}| & |V_{yy}| \end{bmatrix}. \qquad (12)$$

For each image, $s = 100$ random square patches of random sizes between 16×16 and 128×128 are extracted and the HASC descriptor is calculated on each patch. In the testing phase, the same number of patches is extracted on each image. For each patch, the distance between the extracted HASC descriptor and all the training HASCs is measured and the label is predicted according to the majority voting among the $k = 5$ nearest ones (kNN algorithm).

This classifier acts as a weak classifier, and the class of the image is determined according to the maximum votes among the s weak classifiers [16]. In Table 2 classification accuracy obtained with HASC, averaging on 10 different trials, is displayed and compared with the following methods: Lazebnik's method [11], VZ-joint [18], Hayman's method [8], Tuzel's method [16], Harris detector+Laplacian detector+SIFT descriptor+SPIN descriptor((HS+LS)(SIFT+SPIN)) [22] and L^2ECM [12] . With HASC, we reach 98.66%, which amounts to fail on 3 images, beating the best classification score of [12] (97.9) and setting state of the art performance.

Table 2. Classification accuracy for the Brodatz dataset

Descriptors	Accuracy(std)
HASC	**98.66(0.2)**
L^2ECM [12]	97.9(0.4)
Lazebnik [11]	89.8(1)
VZ-joint [18]	92.9(0.8)
Hayman [8]	95(0.8)
Tuzel [16]	97.77
HLSS [22]	95.4

A further study was conducted to assess the role played by nonlinear relations in the overall performance. To this end HASC was compared with COV and EMI taken alone, and the results portrayed in the first row of Table 3. As the performance is near to the saturation point, the improvement in respect COV is significant but limited to 2 images. Therefore, in order to investigate the expressivity of HASC, we start decreasing the number of low-level features. In particular, as done in [16], we take the intensity and magnitude of gradient and Laplacian as defined below:

$$\mathbf{F}(x, y) = \begin{bmatrix} V & \sqrt{V_x^2 + V_y^2} & \sqrt{V_{xx}^2 + V_{yy}^2} \end{bmatrix}. \qquad (13)$$

With only 3 features we reach the best result obtained in [16] with 5 features, as shown in the second row of Table 3. Moreover, the improvement brought by EMI in respect to COV is 8 images, demonstrating the complementary information held by COV and EMI.

3.3. Pedestrian Detection

For the pedestrian classification task, we consider the *Daimler Multi-Cue Occluded Pedestrian Classifica-*

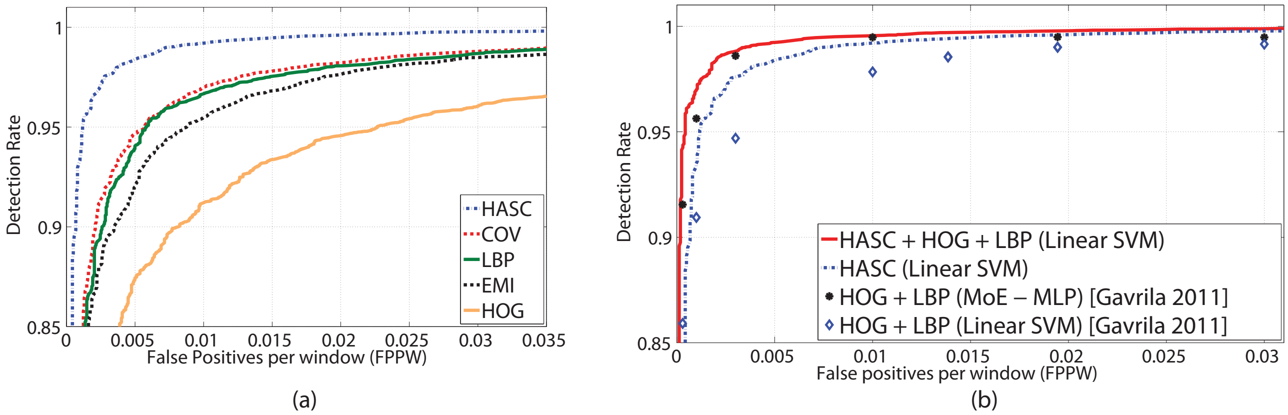

Figure 4. Pedestrian detection scenario; (a) Detection rate curves for Daimler dataset for HOG, LBP, EMI, COV and HASC; (b) Detection rate curves for Daimler dataset for HASC (linear SVM), HASC + HOG + LBP (linear SVM), HOG+LBP (linear SVM) and HOG +LBP (Multi Level Mixture of Experts and Multi Layer Perceptron).

Table 3. COV, EMI and HASC accuracy for Brodatz dataset with different numbers of features (in brackets number of images correctly classified)

	COV	EMI	HASC
5 features	97.77 (219) [16]	96.87 (217)	**98.66** (221)
3 features	94.20 (211) [16]	92.41 (207)	**97.77** (219)

tion Benchmark Dataset [4], taking into account the unoccluded part. The training part contains 52112 and 32465 positive and negative samples, respectively; the testing part has 25608 and 16235 positive and negative samples. Each image, of size 96×48, is composed of three imaging modalities: standard visible gray scale image $V(x, y)$, depth $D(x, y)$, and motion flow $M(x, y)$. As alternative descriptors, we focus on COV and EMI descriptors as well as HOG [2] and LBP [23], already applied on this dataset [5].

For each image, we have the following dense feature map $\mathbf{F}(x, y)$:

$$\mathbf{F}(x,y) = \begin{bmatrix}\mathbf{F}_V(x,y) & \mathbf{F}_D(x,y) & \mathbf{F}_M(x,y) & x & y\end{bmatrix}, \tag{14}$$

where each 96×48 map $\mathbf{F}_V(x, y)$, $\mathbf{F}_D(x, y)$ and $\mathbf{F}_M(x, y)$ denote low-level features extracted from the visible, depth, and motion flow modalities, respectively. In particular, for each modality, we extract the following low-level features (omitting the subscript for clarity):

$$\mathbf{F}_{(x,y)} = \begin{bmatrix}I & |I_x| & |I_y| & |I_{xx}| & |I_{yy}| & \sqrt{I_x^2 + I_y^2} & LBP(I)\end{bmatrix} \tag{15}$$

where I, I_x, I_y, I_{xx} and I_{yy}, are the intensity, first- and second-order derivatives of the three image modalities.

For the depth and motion flow modalities, the depth value and the module of the motion flow are considered as image intensities. For each image, EMI and COV matrices are extracted on a set of patches of different size, fusing together the different modalities in a natural way, resulting in 23×23 matrices. In particular, for each modality the following patches are extracted: 1 patch of size equal to the whole image; 3 overlapping patches corresponding to the head torso and legs regions of a pedestrian, as defined in [4]; 6, 9 and 15 overlapping patches obtained equally by subdividing the three previously defined regions into 2, 3 and 5 respectively. The global feature vector, fed to a linear SVM classifier, is given by $d + d^2$ elements of the vectorized HASC ($d = 23$) multiplied by the total number of patches (34), yielding a total of 17204 features. Results are compared with COV and EMI separately considered and extracted on the same set of patches; moreover a comparison is carried out with HOG and LBP descriptors. The object model for HOG and LBP (number, size and overlap of patches)is the same as adopted in [5].

In Fig. 4 (a) the detection rate curves for all the descriptors are reported, showing that HASC definitely outperforms COV, EMI, HOG and LBP.

A further test was conducted in order to compare HASC, alone or in combination with HOG and LBP, by the state-of-the-art results. The latter are obtained in [5] adopting HOG + LBP with two strategies. The simpler one consists in joining together all LBP and HOG descriptors and feed the final feature vector to a linear SVM. The more complex one, which set the top performance, adopt a Mixture of Experts (MoE) structure in which the scores of several nonlinear classifiers are used as features for an SVM on top. Each classifier, defined by a Multi Layer Perceptron (MLP), is related to a single feature (HOG or LBP), single visual modality, and single pedestrian pose. The results displayed in Fig. 4 (b) show that HASC definitely outperforms the combination of HOG and LBP using a linear SVM. Moreover, joining together HASC, HOG and LBP and using a linear SVM outperform the best result in [5] despite the latter is obtained with a much more complex classification scheme. This is significant, since it demonstrates how well HASC encodes exclusive aspects that the other two descriptors fail to capture. Finally, even HASC alone with a simple linear SVM outperform HOG + LBP with MLP and MoE whenever the False Positive Rate (FPR) exceeds the value of 0.02. To furtherly quantify the performance, Table 4 reports the FPR obtained fixing the Detection Rate (DR) at 90% as a common reference point. Once again the value obtained by joining together HASC + HOG + LBP improves the state-of-the-art by a factor of 1.4.

Table 4. Pedestrian detection: False positives rates on Daimler dataset for a detection rate of 90%:

	FPR
HASC + HOG + LBP (Linear SVM)	**1.85e-4**
HOG + LBP (MoE - MLP) [5]	2.6e-4
HASC (Linear SVM)	6.77e-4

4. Conclusions

In this paper, we presented a novel relation-based feature descriptor, HASC, which is capable to subsume all possible dependencies between low-level dense features of visual entities. Our proposal represents a step ahead with respect to the state-of-the-art of the relation-based strategies for object description, represented by the covariance of features (COV). While COV is limited to the modeling of linear dependencies between features, HASC can also deal with nonlinear ones. The comparative highest detection and classification scores achieved by HASC on heterogeneous tasks (object detection and classification, scene recognition, texture classification), demonstrate that non-linearities are consistently present among basic features of many visual entities, and capturing these bonds allows us to improve modeling capabilities. Future perspectives are essentially focused on embedding HASC into more challenging classifiers, in order to raise their best performances in diverse scenarios.

References

[1] B. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover Publications, New York, 1966.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 1, pages 886–893, 2005.

[3] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[4] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Proc. CVPR*, pages 990 –997, june 2010.

[5] M. Enzweiler and D. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *Image Processing, IEEE Transactions on*, 20(10):2967–2979, oct. 2011.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, apr 2007.

[7] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proc. ICCV*, pages 221–228, 2009.

[8] E. Hayman, B. Caputo, M. Fritz, and J. Eklundh. On the significance of real-world conditions for material classification. In *Proc. ECCV*, 2004.

[9] M. Jagersand. Saliency maps and attention selection in scale and spatial coordinates: an information theoretic approach. In *Proc. ICCV*, pages 195–202, jun 1995.

[10] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45:83–105, 2001.

[11] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *PAMI*, 27:1265–1278, 2005.

[12] P. Li and Q. Wang. Local log-euclidean covariance matrix (l2ecm) for image representation and its applications. In *Proc. ECCV*, volume 7574, pages 469–482, 2012.

[13] D. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, volume 2, pages 1150–1157, 1999.

[14] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *Proc. CVPR*, pages 829–836, 2005.

[15] D. Tosato, M. Spera, M. Cristani, and V. Murino. Characterizing humans on riemannian manifolds. *PAMI*, 35(8):1972–1984, 2013.

[16] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. ECCV*, volume 3952, pages 589–600. Springer, 2006.

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

[18] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *Proc. CVPR*, 2003.

[19] P. Viola and W. M. Wells, III. Alignment by maximization of mutual information. *IJCV*, 24(2):137–154, Sept. 1997.

[20] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *Proc. ICCV*, pages 32–39. IEEE, Sept. 2009.

[21] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Proc. CVPR*, pages 2126–2136, 2006.

[22] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, pages 213–238, 2007.

[23] Y. Zheng, C. Shen, R. I. Hartley, and X. Huang. Effective pedestrian detection using center-symmetric local binary/trinary patterns. *CoRR*, abs/1009.0892, 2010.