Encoding Classes of Unaligned Objects Using Structural Similarity Cross-Covariance Tensors

Marco San Biagio[1], Samuele Martelli[1], Marco Crocco[1], Marco Cristani[1,2], and Vittorio Murino[1,2]

[1] Istituto Italiano di Tecnologia - Pattern Analysis & Computer Vision
Via Morego 30, 16163, Genova, Italy
[2] Università degli Studi di Verona - Dipartimento di Informatica
Strada le Grazie 15, 37134, Verona, Italy
{marco.sanbiagio,samuele.martelli,marco.crocco,
marco.cristani,vittorio.murino}@iit.it

Abstract. Encoding an object essence in terms of self-similarities between its parts is becoming a popular strategy in Computer Vision. In this paper, a new similarity-based descriptor, dubbed Structural Similarity Cross-Covariance Tensor is proposed, aimed to encode relations among different regions of an image in terms of cross-covariance matrices. The latter are calculated between low-level feature vectors extracted from pairs of regions. The new descriptor retains the advantages of the widely used covariance matrix descriptors [1], extending their expressiveness from local similarities inside a region to structural similarities across multiple regions. The new descriptor, applied on top of HOG, is tested on object and scene classification tasks with three datasets. The proposed method always outclasses baseline HOG and yields significant improvement over a recently proposed self-similarity descriptor in the two most challenging datasets.

Keywords: object recognition, scene classification, covariance.

1 Introduction

In pattern recognition, the representation of an entity can be addressed following two complementary paradigms: feature-based and similarity-based. In the first case the characteristics of the entity, or of parts of it, are encoded by descriptors concerning for example shape and color. Most descriptors (e.g. SIFT [4], LBP histograms [5], HOG [6]) are enclosed in this class. In the latter case the focus is on a similarity measure allowing to relate new entity to a set reference ones.

Whenever an entity can be structurally represented by its parts, the similarity philosophy can be applied to the internal relationship among parts, each one represented in terms of features. In other words, a self-similarity descriptor can be constructed on top of feature descriptors related to different entity parts, joining the advantages of the two approaches. An example of this strategy, applied to the pedestrian detection task, can be found in [7]: each image is subdivided in

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 133–140, 2013.

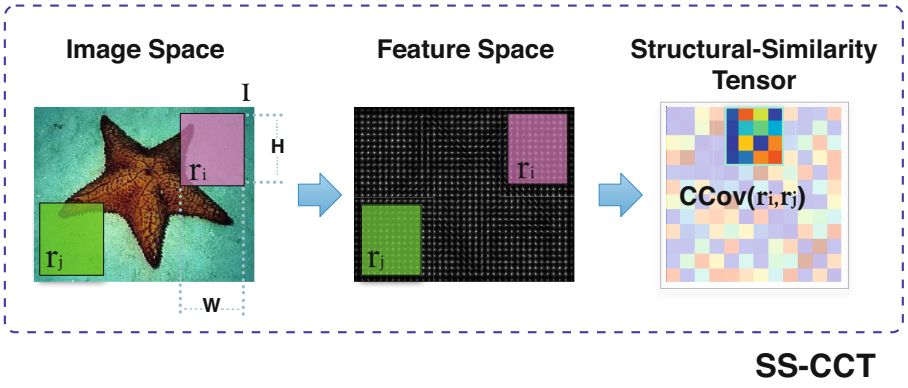

Fig. 1. Building process of the SS-CCT: each region is described by a set of local feature descriptors; the pairwise similarity among two regions is encoded by a cross-covariance matrix of the feature descriptors.

regions from which HOG are extracted; similarity among these regions are then encoded by Euclidean pairwise distances among HOG descriptors. This approach is effective and computationally efficient but has some drawbacks, which are shared by all the similarity-based approaches relying on point-wise distances. In particular, if entities to be detected are not aligned, i.e. the entity's parts do not occupy the same image regions across the images, point-wise distance approaches are not statistically robust, as the single distance may undergo too much variability in the same entity class. Moreover, all the information on the similarity among two descriptors (i.e. two vectors) collapses in a single scalar value, potentially obscuring discriminative relations between single elements of the descriptor (e.g. the single bins of an HOG).

In order to overcome these limitations a different self-similarity approach is here proposed: the key idea is to provide a rich and, at the same time, statistically robust notion of similarity among different regions of an image, exploiting covariance measures among couples of low-level features across different regions.

Covariances of low-level features, in the form of covariance matrices, bear several advantages when used as single region descriptors, as pointed out in [1,2,3]. The representation provides a natural way of fusing multiple features that might be correlated. The single pixel noise is largely filtered out with the average operation intrinsic to the covariance calculation. In comparison to other statistical descriptors, such as multi-dimensional histograms, covariances are intrinsically low-dimensional as their size is only $O(N^2)$, with N being the number of features. Since covariance matrix is invariant with respect to pixels position inside the region, the descriptor has also some degree of robustness against pose change and object rotation.

Till now covariances of low-level features have been employed essentially as a *single region* descriptors [1,2,3]. What we propose here is to employ covariances as a measure of similarity *across different regions*. Thus, covariance matrices

have to be generalized with the *Cross-Covariance matrices*, which capture the covariance among two generally different feature vectors, in our case related to two different regions. In particular, a *Structural Similarity Cross-Covariance Tensor (SS-CCT)* is here proposed, which encodes all the pairwise similarities among regions by means of Cross-Covariance matrices, each one encoding all the pairwise relationships between the single features extracted in a given couple of regions. Any region descriptor can be ideally adopted (e.g. *HOG* [6], *SIFT* [4], *LBP* [5]).

As a proof of concept and for computational reasons, the proposed method is applied to the well-known *HOG* feature descriptor, implemented according to [6], and tested on two different classification tasks: objects and scenes. The classification results show significant performance improvements with respect to both the simple feature-based descriptors and the point-wise similarity based approach in [7].

The remaining of the paper is organized as follows: in Section 2 the SS-CCT descriptor is introduced; in Section 3 some information on the object model is provided; in Section 4 the SS-CCT performance on Caltech 101 [8], Caltech-256 [9] and SenseCam ([10]) datasets is displayed and compared with two literature methods; finally, in Section 4 some conclusions are drawn.

2 Proposed Method

Given an image I, we define R regions each one of size $W \times H$ pixels (see Fig. 1). Each region is divided into M patches and, for each patch, a given feature descriptor is applied, obtaining M feature vectors of size N.

The global *Feature Level* descriptor (FL) of the image I is obtained stacking together the feature vectors for all the regions and all the patches as follows:

$$FL = [\mathbf{z}_{1,1}^T \dots \ \mathbf{z}_{r,m}^T \ \dots \mathbf{z}_{R,M}^T] \tag{1}$$

where $\mathbf{z}_{r,m}$ is the feature vector obtained applying the descriptor to the patch m in the region r.

The proposed *Similarity Level* structural descriptor is built on top of FL, encoding the similarity among each couple of regions. In order to achieve a statistically robust and highly invariant description of this similarity, we calculate the covariance among each couple of features, using the patches of the two regions as spatial samples (Fig. 1).

In detail, given two regions r_1 and r_2, we calculate the $N \times N$ cross-covariance matrix $\mathbf{Ccov}_{r_1,r_2}$ among the feature vectors $\mathbf{z}_{r,m}$ in the following way:

$$\mathbf{Ccov}_{r_1,r_2} = \frac{1}{M-1} \sum_{m=1}^{M} (\mathbf{z}_{r_1,m} - \bar{\mathbf{z}}_{r_1})(\mathbf{z}_{r_2,m} - \bar{\mathbf{z}}_{r_2})^T, \tag{2}$$

where $\bar{\mathbf{z}}_{r_1}$ and $\bar{\mathbf{z}}_{r_2}$ are the mean of the feature vectors inside regions r_1 and r_2, respectively. In practice the i,j-th element of $\mathbf{Ccov}_{r_1,r_2}$ is the spatial covariance of feature i in region r_1 and feature j in region r_2. Notice that Cross-Covariance

matrices do not share the same properties of covariance matrices. In particular, $\mathbf{Ccov}_{r_1,r_2}$ are *not* symmetric and, consequently, *not* semi-definite positive. Therefore cross-covariance matrices do not live on the Riemannian manifold defined by the set of semi-definite positive matrices [1], and the only known modality to use these descriptors in a machine learning framework is to vectorize them.

Calculating Eq. (2) across all the possible region pairs, we define a block matrix **CcovBlock** of size $NR \times NR$ as follows:

$$\mathbf{CcovBlock}(I) = \begin{bmatrix} \mathbf{Ccov}_{1,1} & \cdots & \mathbf{Ccov}_{1,R} \\ \vdots & \ddots & \vdots \\ \mathbf{Ccov}_{R,1} & \cdots & \mathbf{Ccov}_{R,R} \end{bmatrix}. \tag{3}$$

It can be noticed from Eq. (3) that this matrix is block-symmetric, i.e. $\mathbf{Ccov}_{r_1,r_2} = \mathbf{Ccov}_{r_2,r_1}$. Therefore the final structural descriptor, named *Structural-Similarity Cross Covariance Tensor (SS-CCT)*, is built vectorizing $\mathbf{CcovBlock}(I)$ in the following manner:

$$\begin{aligned} \textit{SS-CCT} = [&\mathrm{Vec}(\mathbf{Ccov}_{1,1})\ \mathrm{Vec}(\mathbf{Ccov}_{1,2})\ \ldots \\ &\mathrm{Vec}(\mathbf{Ccov}_{1,R})\ \ \mathrm{Vec}(\mathbf{Ccov}_{2,2})\ \ldots\ \mathrm{Vec}(\mathbf{Ccov}_{R,R})]. \end{aligned} \tag{4}$$

where Vec is the standard vectorization operator.

The length of the *SS-CCT* descriptor is therefore $(R+1)(R/2)N^2$. The final descriptor is obtained joining together the *Feature Level* (Eq. 1) and the *Similarity Level* (Eq. 4) descriptors, with a final length equal to $(R+1)(R/2)N^2+RMN$.

3 Object Model

The adopted object model is dependent on the size of the images considered and on the general characteristics of the dataset. In general, given an image I, containing the object of interest, we calculate the low-level descriptor on a uniformly sampled set of MR patches, of size $w \times w$, whose overlap is $w/2$ in both x and y dimensions. For every patch, we encoded the appearance of an object through the use of *Histograms of Oriented Gradients* descriptor, as defined in [6]. We adopted this descriptor since it is relatively fast to compute and still considered one of the most expressive one.

After that, we defined a set of R regions, subdividing the MR patches in R corresponding subsets of size M. The region size is defined considering the following criteria: 1) each region should contain a number of patches sufficient to yield a significant statistics in the cross-covariance matrix calculus; 2) the patch size should be sufficiently large so as to retain the descriptor expressiveness; 3) finally, the region size should match the size of significant parts of the objects to be detected or classified.

In this way, we calculate the SS-CCT descriptor evaluating the cross-covariance between all the couples of regions as formalized in Eq. 3 and Eq. 4. The final descriptor, here dubbed *SS-CCT(HOG)*, is given by the concatenation of SS-CCT and the HOG descriptors.

4 Experiments

In this section, we report experimental results obtained on two different tasks, using three datasets: object classification (Caltech-101 [8] and Caltech-256 [9]), and scene classification (SenseCam Dataset [10]). In all the experiments, we employ a multiclass one-vs-all linear Support Vector Machine classifier [11].

The comparisons are carried out with the HOG baseline descriptor [6] and the *Self-Similarity Tensor* described in [7]. The latter, named SST(HOG), is built joining together the HOG descriptor and the pairwise Euclidean distances between all the patches, sharing the mixed feature-based and similarity-based philosophy of SS-CCT.

4.1 Object Classification

In the object classification community, Caltech-101 [8] dataset represents an important benchmark. It consists of 102 classes (101 object categories plus background) with a number of images per class ranging from 31 to 800. Despite its importance, Caltech-101 has some cues, notably the presence of strongly aligned object classes, which significantly ease the classification process. To overcome such limitation, the larger Caltech-256 dataset was subsequently introduced. It consists of 256 classes (256 + Clutter class) with a minimum of 80 images per class and a total number of images equal to 30607. In Caltech-256 objects position inside the image is significantly varying for a lot of classes, as can be seen observing the average images for the 256 classes in Fig. 2, so making the classification task more challenging with respect to Caltech-101.

To test our descriptor, the object model introduced in Sec. 3 is adopted. The HOG descriptor is calculated on dense patches of size 32×32 with an overlap of 16 pixels. The number of regions R is set to 9, 3 along both the horizontal and vertical image direction. For Caltech-101 we considered 15 images per class for training and 15 images per class for testing, repeating the experiments with five different splits according to a standard procedure [12]. The same was done for Caltech-256 except for the number of training images which ranged from 5 to 30 with a step of 5.

Experimental results on the Caltech-101 are displayed in Tab. 1. As can be seen both SS-CCT(HOG) and SST(HOG) outperform the baseline HOG with a 6% increment in the overall accuracy. On the other hand, SS-CCT(HOG) and SST(HOG) yield roughly the same performance: this is easily explainable considering that in Caltech-101 images are strongly aligned, reducing the need for robustness against position variation.

Results on the Caltech-256 in terms of accuracy vs the number of training images per class, are displayed in Fig. 3. As figure shows, our method outperforms both HOG and SST(HOG) in all the cases and the gap between our method and the others increases with the increase of the training set size. Differently from the Caltech-101 case, the higher complexity of the dataset highlights the superiority of our method with respect to SST(HOG).

Table 1. Classification results on the Caltech-101 dataset

	HOG	SST(HOG)	SS-CCT(HOG)
Accuracy %	41.3	47.6	**47.77**

Fig. 2. Average of the images of the Caltech-256 dataset

4.2 Scene Classification

In the second experiment, the proposed framework is tested on the SenseCam Dataset [10]. This dataset consists of images acquired with a SenseCam, a wearable camera which automatically shoots a photo every 20 secs. It consists of 912 images labeled according to 32 classes (e.g. Bathroom Home, Car, Garage Home, Biking...). The images are divided into 479 images for training and 433 for testing. The dataset is challenging because most images present dramatic viewing angle, translational camera motions and large variations in illumination and scale: Fig. 4 shows four images belonging to two classes extracted from the dataset.

The HOG descriptor has been calculated on dense patches of size 32×32 with an overlap of 16 pixels. The number of regions was set to $15 : 5$ along the x axis and 3 along the y axis. Experimental results are displayed in Tab. 2.

Our method outperforms both HOG and SST(HOG) with a difference in accuracy of about 8% and 3% respectively, so confirming its effectiveness in classifying images containing objects with an high degree of position variability.

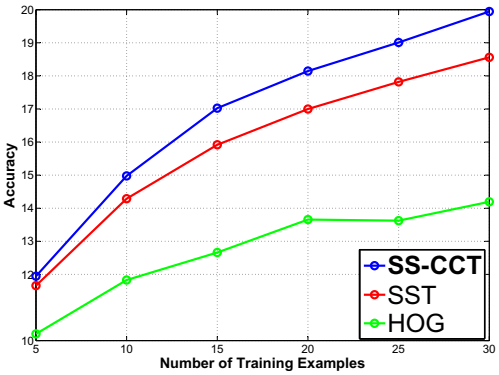

Fig. 3. Results obtained on the Caltech-256 dataset

Table 2. Classification results for the SenseCam dataset

	HOG	SST(HOG)	SS-CCT(HOG)
Accuracy %	36.72	41.10	**44.12**

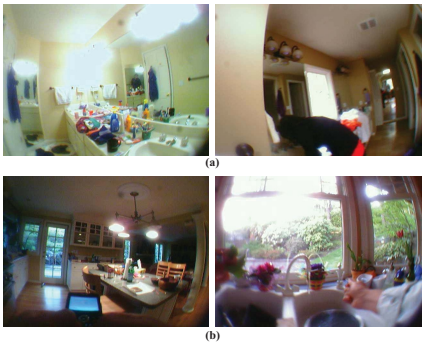

Fig. 4. Four images extracted from the SenseCam Dataset: (a) Bathroom Home and (b) Kitchen

5 Conclusions and Future Works

This paper proposes a novel similarity-based descriptor for image classification purposes. The idea is to encode similarities among different image regions by means of cross-covariance matrices calculated on low level feature vectors, obtaining a robust and compact representation of structural (dis)similarities of a given entity. The final descriptor, obtained joining together the low-level features (HOG in our case) and their structural similarities, has proven to outperform baseline HOG, on all the datasets tested, and a recent literature similarity-based

method [7], on the two most challenging datasets. This is a seminal work, and, despite the encouraging results obtained, needs further study for setting the best object model (number, shape and displacement of the parts) and the best features in a given context. This will allow the comparison with popular state-of-the-art approaches for detection and classification.

References

1. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Trans. PAMI, 1713–1727 (2008)
2. Tosato, D., Spera, M., Cristani, M., Murino, V.: Block Characterizing humans on riemannian manifolds. IEEE Trans. PAMI, 2–15 (2013)
3. San Biagio, M., Crocco, M., Cristani, M., Martelli, S., Murino, V.: Heterogeneous Auto-Similarities of Characteristics (HASC): Exploiting relational information for classification. In: Proc. ICCV (2013)
4. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. ICCV, vol. 2, pp. 1150–1157 (1999)
5. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: Proc. ICCV, pp. 32–39 (2009)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR, vol. 1, pp. 886–893 (2005)
7. Martelli, S., Cristani, M., Bazzani, L., Tosato, D., Murino, V.: Joining feature-based and similarity-based pattern description paradigms for object detection. In: Proc. ICPR (2012)
8. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. CVIU 106(1), 59–70 (2007)
9. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007)
10. Perina, A., Jojic, N.: Spring lattice counting grids: Scene recognition using deformable positional constraints. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 837–851. Springer, Heidelberg (2012)
11. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27 (2011), `http://www.csie.ntu.edu.tw/cjlin/libsvm`
12. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proc. ICCV, pp. 606–613 (2009)