Region Moments: Fast invariant descriptors for detecting small image structures

Gianfranco Doretto Yi Yao
Visualization and Computer Vision Lab, GE Global Research, Niskayuna, NY 12309
doretto@research.ge.com yaoy@ge.com

Abstract

This paper presents region moments, a class of appearance descriptors based on image moments applied to a pool of image features. A careful design of the moments and the image features, makes the descriptors scale and rotation invariant, and therefore suitable for vehicle detection from aerial video, where targets appear at different scales and orientations. Region moments are linearly related to the image features. Thus, comparing descriptors by computing costly geodesic distances and non-linear classifiers can be avoided, because Euclidean geometry and linear classifiers are still effective. The descriptor computation is made efficient by designing a fast procedure based on the integral representation. An extensive comparison between region moments and the region covariance descriptors, reports theoretical, qualitative, and quantitative differences among them, with a clear advantage of the region moments, when used for detecting small image structures, such as vehicles in aerial video. The proposed descriptors hold the promise to become an effective building block in other applications.

1. Introduction

The design of *appearance descriptors* has received a considerable amount of attention during the last decade. This is because many practical applications require the computation of descriptors for supporting tasks that range from image feature matching [10, 12, 24] to object matching [5, 21], or that range from object detection [2, 25, 22] to object category recognition [23, 3, 11, 14, 17]. Perhaps the most important property of a descriptor is its distinctiveness, but another critical issue is its robustness with respect to noise and image transformations. The computational complexity of a descriptor is also a major property. This may become an issue during the training phase (e.g. feature selection), as well as at runtime. In fact, almost every task entails the computation of large amounts of statistics (i.e. appearance descriptors) from data that correspond to several image regions. Even though there have been some efforts to reduce the amount of computations [9], the need for fast appearance descriptors is still very high.

Detecting small image structures (e.g. vehicles in aerial video), based solely on their appearance, is the testbed application considered by this work. Detection is done by classifying an appearance descriptor, and here a class of new descriptors, called *region moments*, is introduced. They are required to be computationally efficient, as well as invariant to scale and rotation of the image region containing the image structure. This is needed because the image structure of interest can appear at different sizes and orientations.

Region moments are descriptors based on an extension of the idea of image moments [4, 19]. Rather than moments of raw image values, *moments of image features* are computed (Section 2). This enables the design of a map from the image space onto the space of *image features*, where every pixel is assigned to an image feature vector, containing the desired *isotropic* or *anisotropic* local description of the image. This gives a lot of freedom to experiment with the most appropriate and robust features. Since rotation invariance is a primary concern, a new pool of isotropic features is proposed (Section 5).

More importantly, moments represent a statistical description of the image data which is a projection onto the space of polynomials with two variables. Consequently, *the map from the space of image features onto the space of moments of image features is linear*. Therefore, Euclidean norms as well as linear classifiers can be used effectively, and efficiently. This is in contrast with other statistical representations, such as [21], which map the image features onto a manifold with non-trivial curvature that requires the use of appropriate, and costly to compute, distances for comparing descriptors or designing classifiers.

Three region moment descriptors are introduced and analyzed (Sections 3 and 4). The first one computes the central moments of the image features of a given region. The second one is the set of the Hu's invariants [7] computed over every image feature, and the third one computes the radial moments of the image features. The descriptors can be rotation invariant, depending on the isotropic properties of the image features.

The issue of computational efficiency is addressed by designing procedures that extend the use of the integral image representation [25], to the computation of moments of image features (Section 6). For a given image, this enables the computation of new region moments in constant time, modulo the pre-computation of a number of quantities. In previous work, [16], the integral image was extended to enable the fast computation of histograms, whereas [26] added the fast computation of correlograms.

The framework is applied to the problem of vehicle detection in aerial video. The performance of the region moments are thoroughly compared, and Section 7 reports that they outperform the region covariance (RC) descriptor [21], as they are more efficient while used in conjunction with a classifier, such as a support vector machine (SVM) [18].

Related work This work builds on the idea of image moments, and moment invariants. In [7] Hu exploited the theory of algebraic invariants to derive seven statistics based on image moments, which are invariant to rotation. Since then, several invariant statistics have been derived (see [4, 19] for a concise survey). In [13] a number of generalized moment invariants are introduced, which combine the information from different color channels, and which are invariant to affine changes of illumination and image domain transformations. [12] compares them against other descriptors (e.g. SIFT [10], and shape context [1]) for the purpose of image feature matching.

This framework relates to the RC descriptor [21], which was used with a set of anisotropic features for object matching and localization, and for texture classification. This work differs from [21] in many respects. Invariance to rotation is a requirement, for which an appropriate descriptor design and a set of image features are proposed. The linear mapping of the image features enables the use of Euclidean norm and linear classifiers. The region moment computation procedure (for constant time descriptor computation), has a pre-computation phase with cost that scales linearly with the dimensionality of the image features, and not quadratically, like in [21]. Finally, this work focusses on detection of small image structures.

2. Moments of image features

In this section we introduce some basic notation, and what we mean by *image features* and *moments* of image features.

Image features Let I be either a monochromatic, or an RGB image, defined for every $(x, y) \in \Omega \subset \mathbb{R}^2$. With Φ we indicate a generic filtering operation that maps I, and the pixel (x, y), to the d dimensional *image feature* vector

$$\mathbf{f}(x, y) = [f_1(x, y), \cdots, f_d(x, y)]^T \doteq \Phi(I, x, y) \ . \quad (1)$$

Let us suppose to operate a rotation of the image I around the pixel (x, y). If $f_i(x, y)$ has not changed, we say it is an *isotropic* feature. Otherwise, f_i is *anisotropic*. For instance, $I_x \doteq \partial I/\partial x$ is an anisotropic feature, whereas the magnitude of the image gradient $|\nabla I|$ is an isotropic feature.

Moments With $\mathbf{m}_{pq}$, where p and q are non-negative integers, we indicate the *moment*[1] of the image features $\mathbf{f}$ of *order* $n = p + q$, which we define as

$$\mathbf{m}_{pq} \doteq \int_\Omega x^p y^q \mathbf{f}(x, y) \, \mathrm{d}x\mathrm{d}y \ . \quad (2)$$

We also introduce the notation $\mathbf{m}_n \doteq [\mathbf{m}_{0n}^T, \cdots, \mathbf{m}_{n0}^T]^T$, which refers to the ordered stack of the $n + 1$ moments of order n.

Central moments Similarly, with $\boldsymbol{\mu}_{pq}$ we indicate the *central* moments of the image features $\mathbf{f}$ of *order* $n = p + q \geq 2$, with components $\mu_{pq,i}$ defined as

$$\mu_{pq,i} \doteq \int_\Omega (x - \frac{m_{10,i}}{m_{00,i}})^p (y - \frac{m_{01,i}}{m_{00,i}})^q f_i(x, y) \, \mathrm{d}x\mathrm{d}y \ . \quad (3)$$

As before, the notation $\boldsymbol{\mu}_n \doteq [\boldsymbol{\mu}_{0n}^T, \cdots, \boldsymbol{\mu}_{n0}^T]^T$ refers to the ordered stack of the $n + 1$ central moments of order n.

Central moment invariants From $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$, according to [7], it is possible to extract seven different statistics (known as the Hu's invariants) that are invariant to translations and rotations of the features $\mathbf{f}$. We indicate them as $\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_7 \in \mathbb{R}^d$, where $\boldsymbol{\tau}_j$ collects the j-th Hu's invariants corresponding to the d image features $\mathbf{f}$.

Radial moments Finally, we define the *radial moment* of the image features $\mathbf{f}$ of *order* $n \geq 1$ as

$$\mathfrak{m}_n \doteq \int_\Omega \left[(x - \bar{x}_\Omega)^2 + (y - \bar{y}_\Omega)^2 \right]^n \mathbf{f}(x, y) \, \mathrm{d}x\mathrm{d}y \ , \quad (4)$$

where $(\bar{x}_\Omega, \bar{y}_\Omega) \in \mathbb{R}^2$ represents the geometric center of the domain Ω. These moments have the obvious fundamental property of being invariant to rotations of the features $\mathbf{f}$.

3. Region moment descriptors

Let $D \subset \Omega$ be a simply connected domain. With $|D|$ we indicate the area of the domain. An *appearance descriptor* is an operator that maps the image features $\{\mathbf{f}(x, y) | (x, y) \in D\}$ to a *descriptor* vector $\mathbf{y} \in \mathbb{R}^J$. In this section we define three *region moment* descriptors, which are appearance descriptors based on the moments of image features introduced in Section 2.

Central Moment (CM) descriptor From the set of central moments of the image features $\{\boldsymbol{\mu}_i\}$, computed over the domain D, we define the *central moment descriptor* of *order* $n \geq 2$ as the vector

$$\mathbf{y}_{CM} \doteq \left[\frac{\boldsymbol{\mu}_2^T}{|D|^2}, \cdots, \frac{\boldsymbol{\mu}_n^T}{|D|^{\frac{n+2}{2}}} \right]^T \ , \quad (5)$$

with $J_{CM} = \frac{n^2+3n-4}{2} d$ components.

[1]Note that all the definitions of moments regard the image features as probability densities. Since they are not, we recall that as long as the $\{f_i\}$ are piecewise continuous, and have bounded support, then moments of all orders exist, and are uniquely determined [15]. In all practical cases these conditions are widely satisfied.

Central Moment Invariant (CMI) descriptor From the set of central moment invariants of the image features $\{\boldsymbol{\tau}_i\}$, computed over the domain D, we define the *central moment invariant descriptor* as

$$\mathbf{y}_{CMI} \doteq \left[\frac{\boldsymbol{\tau}_1^T}{|D|^2}, \frac{\boldsymbol{\tau}_2^T}{|D|^4}, \frac{\boldsymbol{\tau}_3^T}{|D|^5}, \frac{\boldsymbol{\tau}_4^T}{|D|^5}, \frac{\boldsymbol{\tau}_5^T}{|D|^{10}}, \frac{\boldsymbol{\tau}_6^T}{|D|^7}, \frac{\boldsymbol{\tau}_7^T}{|D|^{10}} \right]^T , \tag{6}$$

which is a vector with $J_{CMI} = 7d$ components.

Radial Moment (RM) descriptor From the set of radial moments of the image features $\{\mathbf{m}_i\}$, computed over the domain D, we define the *radial moment descriptor* of *order* $n \geq 1$ as

$$\mathbf{y}_{RM} \doteq \left[\frac{\mathbf{m}_1^T}{|D|^2}, \cdots, \frac{\mathbf{m}_n}{|D|^{n+1}} \right]^T , \tag{7}$$

which is a vector with $J_{RM} = nd$ components.

Region Covariance (RC) descriptor We report also the definition of the *region covariance* descriptor, originally introduced in [21], as we are going to compare the aforementioned descriptors against this one. In particular, we define the covariance matrix of the image features $\mathbf{f}$ of the region D, as

$$\mathbf{k}_\mathbf{f} \doteq \frac{1}{|D|} \int_D (\mathbf{f}(x,y) - \bar{\mathbf{f}})(\mathbf{f}(x,y) - \bar{\mathbf{f}})^T \mathrm{d}x\mathrm{d}y , \tag{8}$$

where $\bar{\mathbf{f}} \doteq \mathbf{m}_{00}/|D|$ is the average image feature vector over D. Given that $\mathbf{k}_\mathbf{f}$ is a symmetric positive definite matrix, the region covariance descriptor is given by the vector $\mathbf{y}_{RC}$ containing the $J_{RC} = (d^2 + d)/2$ independent entries of $\mathbf{k}_\mathbf{f}$.

4. Descriptor properties

Here we compare some of the characteristics of the descriptors, and defer the discussion on computational and quantitative comparison to Section 6 and Section 7, respectively.

Dimensionality reduction All the four descriptors can provide an effective, and very low dimensional representation of the image, especially RM (see Section 7). In contrast, as [21] points out, even the widely used representation of image features with a joint histogram would give rise to a descriptor with dimensionality $O(d^\alpha)$, with $\alpha \sim O(10)$.

General meaning of the descriptors In CM, CMI, and RM the image features are seen as probability distributions, and the descriptors summarize how the features are distributed over the domain of computation. In RC the image features are seen as a random vector, and the descriptor is its covariance. If the features contain spatial information, then RC correlates it with the other features.

Feature interaction All the descriptors provide a suitable way to fuse the information of multiple features. While RC does that by computing the correlations between each pair of image features, CM, CMI, and RM enable that by setting some image features to be the combination of others. Note that the dimensionality of these descriptors scales linearly with d, rather then quadratically, like in RC.

Invariance to translations Let us suppose that at a given instant, D contains an object over a uniform background, and at a subsequent instant, the object has translated but is still inside D. All the descriptors are invariant to such a translation with the exception of RM, which is computed relative to the geometric center of D. This tells that RM might be a better descriptor for localizing a specific object.

Invariance to scale Provided that the features at every pixel encode information that is scale invariant, it is easy to prove that the normalization with respect to $|D|$ operated on equations (5), (6), (7), and (8), make all the descriptors scale invariant. Notice, however, that even without fully scale invariant features, the normalization provides a very high degree of invariance with respect to scale.

Invariance to rotations It is easy to realize the following facts: (a) if the image features $\{f_i\}$ are all isotropic, CMI, RM, and RC are rotation invariant; (b) if some of the $\{f_i\}$ are anisotropic, CMI, RM, and RC are not rotation invariant; (c) regardless of the nature of the $\{f_i\}$, CM is not rotation invariant. Those properties are very important because depending on the application, rotation invariance (variance) may or may not be a requirement. This paper is concerned more with rotation invariance.

Invariance to illumination All the descriptors delegate the illumination invariance issue to the design of the image features. RC is invariant to illumination changes causing additive changes to the features, because the mean features are removed from the descriptor computation.

5. Image feature sets

In this section we are going to describe the filtering operations with which we produce the set of image features. We experiment with two sets of filters: Φ_A, and Φ_I. Φ_A is a set of anisotropic features that we borrow from [21]. We propose Φ_I, which is a set of isotropic features.

Anisotropic features The set of filters Φ_A is such that the image features are computed as

$$\begin{aligned} \mathbf{f}(x,y) \doteq [&x,\, y,\, R(x,y),\, G(x,y),\, B(x,y)\,, \\ &|I_x(x,y)|,\, |I_y(x,y)|,\, |I_{xx}(x,y)|,\, |I_{yy}(x,y)|]^T , \end{aligned} \tag{9}$$

where R, G, and B are the RGB color channels of the image I, and the other quantities are the magnitude of the first and second order partial derivatives of the image intensity. This is neither an illumination nor a scale invariant set of features. Finally, x and y are present only when the RC descriptor is computed.

Isotropic features We define the set of filters Φ_I such that the image features are computed as

$$\mathbf{f}(x,y) \doteq \left[\rho_D^2,\, R(x,y),\, G(x,y),\, B(x,y),\right. \\ \left. |\nabla I(x,y)|, \kappa_\rho(x,y),\, \kappa_\theta(x,y),\, s(x,y)\right]^T \;, \quad (10)$$

where we have $\rho_D^2 \doteq (x-\bar{x}_D)^2+(y-\bar{y}_D)^2$, the RGB color channels, and the magnitude of the gradient of the image intensity. As before, ρ_D is present only when the RC descriptor is computed. The last three features are shape parameters, and give a local description of the 3D shape generated by the image intensity, i.e. $\{(x,y,z) \in \mathbb{R}^3 | I(x,y) + z = 0\}$. κ_ρ and κ_θ are the representation in polar coordinates of the principal curvatures κ_1 and κ_2, whereas s is the shape index [8], which is also scale invariant. They can be computed with the fast procedure in [20], and they are defined as follows

$$\kappa_\rho = \sqrt{\kappa_1^2 + \kappa_2^2} \;, \quad (11)$$

$$\kappa_\theta = \arctan\left(\frac{\kappa_2}{\kappa_1}\right) \;, \quad (12)$$

$$s = \frac{2}{\pi}\arctan\left(\frac{\kappa_2+\kappa_1}{\kappa_2-\kappa_1}\right) \;. \quad (13)$$

Section 7 shows that the addition of the shape features significantly improves the performance.

6. Fast algorithms via integral representation

Given an image I, defined over Ω, its integral representation, commonly known as *integral image* [25], is given by $\int_{u\le x, v\le y} I(u,v)\,\mathrm{d}u\mathrm{d}v$. This representation allows to compute the integral of the image over any rectangular domain D with an $O(1)$ number of operations. Note that the area of $D = [x', x''] \times [y', y'']$ can be easily computed as $|D| = A(x',y') - A(x'',y') + A(x'',y'') - A(x',y'')$, where $A(x,y) \doteq \int_{u\le x, v\le y} \mathrm{d}u\mathrm{d}v$, which can be conveniently precomputed. In this section we exploit this representation to derive fast algorithms for computing the CM, CMI, and RM descriptors.

CM descriptor In order to compute $\mathbf{y}_{CM}$ we need the set of moments $\{\boldsymbol{\mu}_{pq} | p+q = 2, \cdots, n\}$. The binomial theorem allows to relate moments to central moments as follows

$$\mu_{pq,i} = \sum_{j=0}^{p}\sum_{k=0}^{q} \binom{p}{j}\binom{q}{k}\left(-\frac{m_{10,i}}{m_{00,i}}\right)^{p-j}\left(-\frac{m_{01,i}}{m_{00,i}}\right)^{q-k} m_{jk,i} \;. \quad (14)$$

Therefore, the central moments can be computed from the set of moments $\{\mathbf{m}_{pq} | p+q = 0, \cdots, n\}$. In turn, the moments can take advantage of the integral representation by defining the following function

$$\mathbf{M}_{pq}(x,y) \doteq \int_{u\le x\;,\;v\le y} u^p v^q \mathbf{f}(u,v)\,\mathrm{d}u\mathrm{d}v \;, \quad (15)$$

and the moments over D can be computed as follows

$$\mathbf{m}_{pq} = \mathbf{M}_{pq}(x',y') - \mathbf{M}_{pq}(x'',y') + \\ \mathbf{M}_{pq}(x'',y'') - \mathbf{M}_{pq}(x',y'') \;. \quad (16)$$

Note that the quantity $u^p v^q$ in (15) can be conveniently precomputed. Finally, the entire procedure involves the initial formation of $\frac{n^2+3n+2}{2}d$ integral representations for (15), each of which can be computed with one pass over the domain Ω [25], and then each descriptor is obtained with constant computational complexity of $O(n^2 d)$ operations.

CMI descriptor In order to compute $\mathbf{y}_{CMI}$ we need the central moments $\boldsymbol{\mu}_2$, and $\boldsymbol{\mu}_3$. They can be computed with the procedure described above, from which one can compute the set of invariants $\{\boldsymbol{\tau}_i\}$. The entire procedure involves the initial formation of $10d$ integral images, and then each descriptor is obtained with constant computational complexity of $O(d)$ operations.

RM descriptor In order to compute $\mathbf{y}_{RM}$ we need $\mathbf{m}_1, \cdots, \mathbf{m}_n$. The binomial theorem allows to relate moments to radial moments as

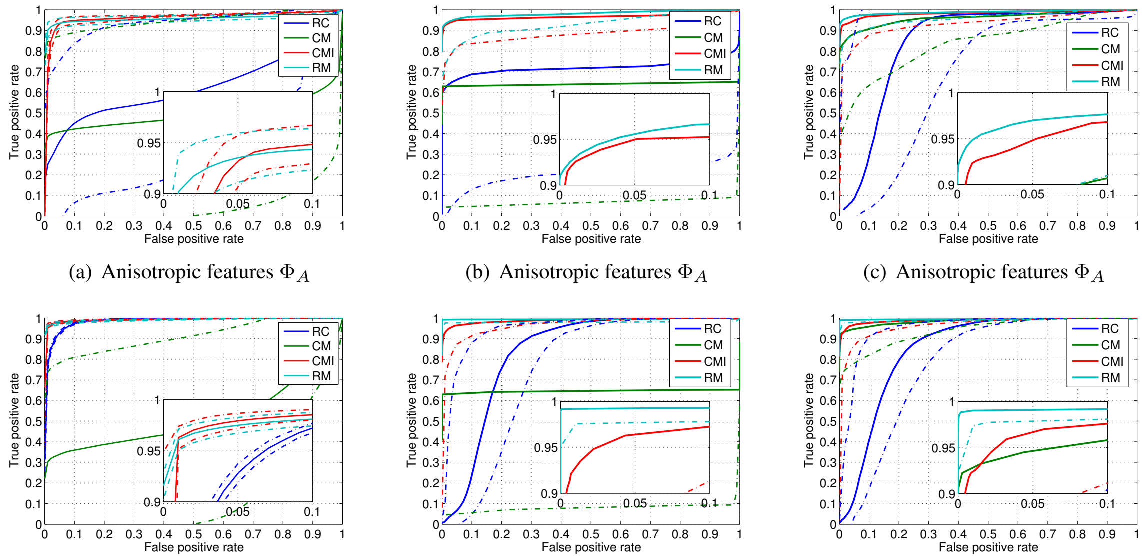
$$\mathbf{m}_{pq} = \sum_{i=0}^{n}\sum_{j=0}^{2n-2i}\sum_{k=0}^{2i}\binom{n}{i}\binom{2n-2i}{j}\binom{2i}{k}\bar{x}_D^{2n-2i-j}\bar{y}_D^{2i-k}\mathbf{m}_{jk} \;, \quad (17)$$

where $\bar{x}_D = -\frac{x'+x''}{2}$, and $\bar{y}_D = -\frac{y'+y''}{2}$. Therefore, the radial moments can be computed from $\{\mathbf{m}_{pq} | p+q = 0, \cdots, 2n\}$, which in turn can be computed as explained above. The entire procedure involves the initial formation of $(2n^2 + 2n + 1)d$ integral representations for (15), and then each descriptor is obtained with constant computational complexity of $O(nd)$ operations.

Comparison with RC Table 1 summarizes the main computational characteristics of the four descriptors. We have experienced that $n = 2$ or 3 is enough to obtain a powerful descriptor. With n fixed, the number of integral representations to be computed is $O(d^2)$ for the RC descriptor, whereas it is $O(d)$ for all the others. The same holds for the number of operations for computing one descriptor after the precomputation phase, which is $O(d^2)$ for the RC descriptor, and $O(d)$ for all the others. Note also that, with n being so small, the RM descriptor is very short.

Table 1. **Computational complexity.** Summary of the descriptor lengths, number of integral representations to be computed (passes), and order of the number of operations per descriptor, as functions of the number of image features d, and of the descriptor order n.

Descriptor	Length	Passes	Operations
RC	$\frac{d^2+d}{2}$	$\frac{d^2+3d}{2}$	$O(d^2)$
CM	$\frac{n^2+3n-4}{2}d$	$\frac{n^2+3n+2}{2}d$	$O(n^2 d)$
CMI	$7d$	$10d$	$O(d)$
RM	nd	$(2n^2+2n+1)d$	$O(nd)$



(a) Anisotropic features Φ_A (b) Anisotropic features Φ_A (c) Anisotropic features Φ_A

(d) Isotropic features Φ_I (e) Isotropic features Φ_I (f) Isotropic features Φ_I

Figure 1. **ROC curves.** Comparison between descriptors. (a), (d) Testing with synthetic camera rotations. (b), (e) Testing with real camera rotations. (c), (f) Training with synthetic camera rotations. Solid lines indicate overall performance. Dash-dotted lines indicate upper and lower one-standard-deviation performance limits.

7. Experiments

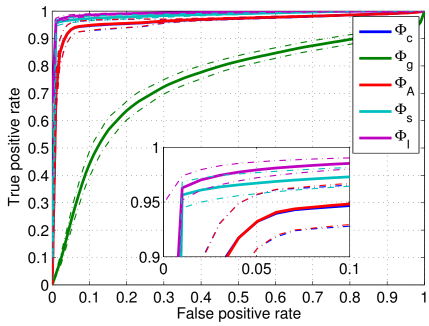
We analyze the performance of the RC descriptor, the CM descriptor of order 3, the CMI descriptor, and the RM descriptor of order 2, in combination with anisotropic, Φ_A, and isotropic, Φ_I, image features. The testbed application is vehicle detection from aerial video. We adopt a sliding window approach where at every pixel a descriptor over a squared image area is computed at 5 different scales, and feeds an SVM classifier [18], which outputs a score. Morphological filtering and connected components are then computed over the thresholded scores, in order to support the arbitration between different vehicle scales, and the establishing of a detection. Figure 3 and 4, which we later describe more in detail, show the output of this commonly used pipeline.

Classifier In all the experiments where the CM, CMI, and RM descriptors are used we deploy a linear SVM classifier. In all the experiments where the RC descriptor is used we deploy an SVM with kernel $\exp\left(-\gamma d(\mathbf{y}, \mathbf{y}')^2\right)$, where $d(\mathbf{y}, \mathbf{y}')$ is the metric for positive definite covariance matrices used in [21], that takes into account the curvature of the descriptor manifold. In fact, as proved by some of our experiments, which we do not report for brevity, the simple linear SVM would be suboptimal, given the known structure of the space of covariance matrices. The constant γ is optimized by a grid search. The interested reader can consult [6] for a framework that could be used for speeding up the computation of the exponential kernel.

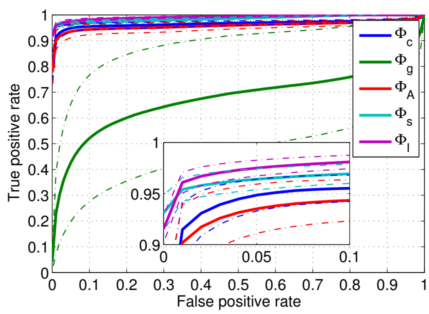
Dataset We use a dataset made of sequences of highway traffic collected from a camera mounted at various rotation angles. In particular, the sequences were collected with the camera rotated at $0°$, $20°$, $45°$, $65°$, and $90°$. For each video, we manually annotated the ground-truth vehicle locations, and sizes, by recording the rectangular regions of interest (ROIs) around each vehicle moving along the highway, resulting in a total of about 47000 different ROIs, corresponding to 590 different vehicles. The sizes of the ROIs vary from about 30×20 to 120×40 pixels, depending on the type of the vehicle. Note that the small sizes of these ROIs make it too difficult to detect points of interest for the computation of the corresponding descriptors such as SIFT. Therefore, we have chosen a region based descriptor, RC, as the baseline method for performance comparison.

Quantitative evaluation For every descriptor we provide the receiver operating characteristic (ROC) curve of the classifier output, corresponding to the positive and negative labeled samples, which are nothing but ROIs containing or not containing a vehicle, respectively. Since for every sample the camera rotation is known, it is possible to draw multiple ROC curves, corresponding to each rotation angle. For every descriptor of choice, we provide the ROC curve evaluated over all camera rotations, and we also add the one-standard-deviation upper and lower limits that express how much variation can be expected due to rotation.

Testing with synthetic camera rotations From the sequence rotated at $0°$ we use the positive and negative samples with 5-fold cross-validation. Moreover, validation is conducted by adding to the testing group synthetically rotated samples, with rotations uniformly sampled in the range between $0°$ and $180°$, with an interval of $20°$.



(a) CMI descriptor



(b) RM descriptor

Figure 2. **ROC curves.** Comparison between various sets of anisotropic and isotropic image features.

Figure 1(a) shows the performance for descriptors with anisotropic features, Φ_A, whereas Figure 1(d) shows the case with isotropic ones, Φ_I. Regardless of the features, the CM descriptor is very sensitive to rotation, which is in line with its theoretical properties. The RC descriptor is very sensitive to rotation when it uses anisotropic features, and becomes invariant with isotropic ones, with a significant performance increase. Again, this is in line with its properties. The CMI and RM descriptors are the best performers, regardless of the features. They improve by switching from Φ_A to Φ_I, and they show limited sensitivity to rotation variation, even though they are not rotation invariant when Φ_A is used. This may seem strange at first glance, but can be explained with the help of Figure 2.

Figure 2(a) shows the performance of the CMI descriptor for the following pools of features: $\Phi_c \doteq [R, G, B]^T$, $\Phi_g \doteq [|I_x|, |I_y|, |I_{xx}|, |I_{yy}|]^T$, $\Phi_A = [\Phi_c^T, \Phi_g^T]^T$, $\Phi_s \doteq [|\nabla I|, \kappa_\rho, \kappa_\theta, s]^T$, and $\Phi_I = [\Phi_c^T, \Phi_s^T]^T$. Figure 2(b) shows the same, only for the RM descriptor. Φ_c is isotropic, which makes CMI and RM rotation invariant, and that explains their good performance. In contrast, Φ_g is anisotropic, which explains the poorer performance, and greater sensitivity to rotation. When Φ_c is augmented with Φ_g to form Φ_A, neither a significant performance improvement nor a deterioration is observed. This explains the relatively good performance of CMI and RM in Figure 1(a), which is due to the dominance of the R, G, and B image features. When we

Table 2. **Area under ROC.** Comparison between the RC, CM, CMI, and RM descriptors. (a) From Figures 1(a), 1(d). (b) From Figures 1(b), 1(e). (c) From Figures 1(c), 1(f).

Descriptor	Length	$A_{ROC} \pm \sigma_{A_{ROC}}$		
		(a)	(b)	(c)
Anisotropic image features				
RC	45	0.556 ± 0.298	0.719 ± 0.456	0.840 ± 0.157
CM	49	0.505 ± 0.455	0.640 ± 0.506	0.960 ± 0.106
CMI	49	0.974 ± 0.012	0.966 ± 0.074	0.974 ± 0.043
RM	14	0.981 ± 0.006	0.982 ± 0.046	0.988 ± 0.024
Isotropic image features				
RC	36	0.981 ± 0.001	0.830 ± 0.099	0.851 ± 0.082
CM	49	0.505 ± 0.388	0.647 ± 0.501	0.986 ± 0.040
CMI	49	0.993 ± 0.002	0.979 ± 0.022	0.981 ± 0.024
RM	14	0.983 ± 0.006	0.994 ± 0.013	0.995 ± 0.006

decompose Φ_I, we notice that Φ_s is isotropic, thus CMI and RM are rotation invariant, and with performance that even exceeds the case of Φ_c. Therefore, when Φ_c is augmented with Φ_s to form Φ_I, we obtain the best performance. This analysis highlights how important the addition of the pool of features Φ_s is, which has proved to be much more effective than the addition of Φ_g.

Testing with real camera rotations From the sequence rotated at $0°$ we use the positive and negative samples for training, and we use the samples from the sequences rotated at $20°$, $45°$, $65°$, and $90°$, for testing. This is a more challenging scenario because validation is done with real rotated image data, which include new sources of noise that will affect the descriptor computation.

Figure 1(b) shows the performance for descriptors with anisotropic features, Φ_A, whereas Figure 1(e) shows the case with isotropic ones, Φ_I. The descriptors generally behave exactly as they did for the case of synthetically rotated testing samples (Figures 1(a) and 1(d)), with an increase in sensitivity to camera rotation. In particular, there is a clear deterioration of the average performance of the RC descriptor, even with isotropic features. Finally, CMI with isotropic features seems to undergo only a slight performance drop, whereas the RM appears to be the most immune to this synthetic-to-real-data switch.

Training with synthetic camera rotations From the sequence rotated at $0°$ we use the positive and negative samples, together with the positive and negative synthetically rotated samples, for training, and we use the samples from the sequences rotated at $20°$, $45°$, $65°$, and $90°$, for testing.

Figure 1(c) shows the performance for descriptors with anisotropic features, Φ_A, whereas Figure 1(f) shows the case with isotropic ones, Φ_I. Compared to Figure 1(b), with Φ_A every descriptor improves the performance, notably RC, and CM even more. This was expected, given their sensitivity to camera rotation, which is now smoothened by a richer training dataset. Compared to Figure 1(e), with Φ_I CM improves significantly. Since RC, CMI, and RM are supposed to be rotation invariant, their performance undergo only a minor change. Overall, even with a training that includes every possible rotation, the CM and RC descriptors are outperformed by the CMI descriptor, and even more by the RM descriptor.

Discussion As a comprehensive performance parameter, Table 2 provides the area under the ROC curve, A_{ROC}, with the corresponding standard deviation due to camera rotation, $\sigma_{A_{ROC}}$. A higher A_{ROC} suggests a higher true positive rate at the same false positive rate, and therefore better performance. The table confirms that the CMI, and (even more) the RM descriptors with isotropic features are the best performers.

Figure 4 shows the results of the full vehicle detector, as described at the beginning of this section. The classifier was trained only with samples from the sequence rotated at $0°$. Due to space constraints, we show results only for the RC, the CMI, and the RM descriptors with isotropic features. We use boxes to visualize the correct detections (green), the misses (yellow), and the false alarms (red). The images convey the superior robustness of the RM descriptor with respect to camera rotation. Note that only the vehicles on the highway are detected, and this is because, in the training sequence, the area outside of the highway was used to produce negative samples. We also noted that the use of linear SVMs provides a 10-fold speed-up improvement, during both training and testing, when compared against the SVM used for the RC descriptor, which requires the computation of the geodesic distance between covariance matrices.

Although Figure 4 shows images from videos coming from static cameras, the detector does not exploit motion information, and can be used when the camera is moving, without the need for video registration. Figure 3 shows this case, where vehicles are detected at multiple scales and from different viewpoints.

Finally, in this work we have treated invariances in two ways: (a) by incorporating them in the descriptor design, and (b) by incorporating them into the SVM by generating artificially transformed samples. This time we found (a) to be more effective then (b). We recall that there is also another way to treat invariances, and that involves engineering kernel functions which lead to invariant SVMs [18], and which will be the subject of future investigations.

8. Conclusions

This paper has introduced region moments, a class of appearance descriptors based on the idea of image moments applied to a collection of image features. A careful design of moments and image features allows to obtain scale and rotation invariant descriptors, which are deployable in unconstrained scenarios, such as vehicle detection in aerial video, where targets of interest can appear in different sizes and orientations. Image moments define a linear relation between image features and region moments. This allowed building a simple and yet very effective appearance based vehicle detector by training an efficient linear SVM classifier. Leveraging the integral representation has led to a fast procedure for computing region moments that makes them very attractive because the cost scales linearly in the number of image features. The extensive comparison, between the region covariance and different region moments, evaluates the detection robustness of the descriptors against rotation variations, for different choices of image features and training procedures. It was found that the region moment, with radial moments and isotropic image features, significantly outperforms the region covariance for vehicle detection. Region moments have been used also for detecting specific vehicles in aerial video [27], and hold the promise to become an effective building block in other computer vision applications.

Figure 3. **Vehicle detection in aerial video.** Top row: Image frames with ground-truth vehicle regions. Bottom row: Corresponding frames with vehicle regions detected by the RM descriptor and isotropic features Φ_I. Blue, green, red, cyan, and magenta rectangles indicate that the vehicle regions are detected at the scale of 20×20, 30×30, 40×40, 60×60, or 80×80 pixels, respectively.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24:509–522, 2002.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, June 20–25, 2005.

[3] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, pages 524–531, June 20–25, 2005.

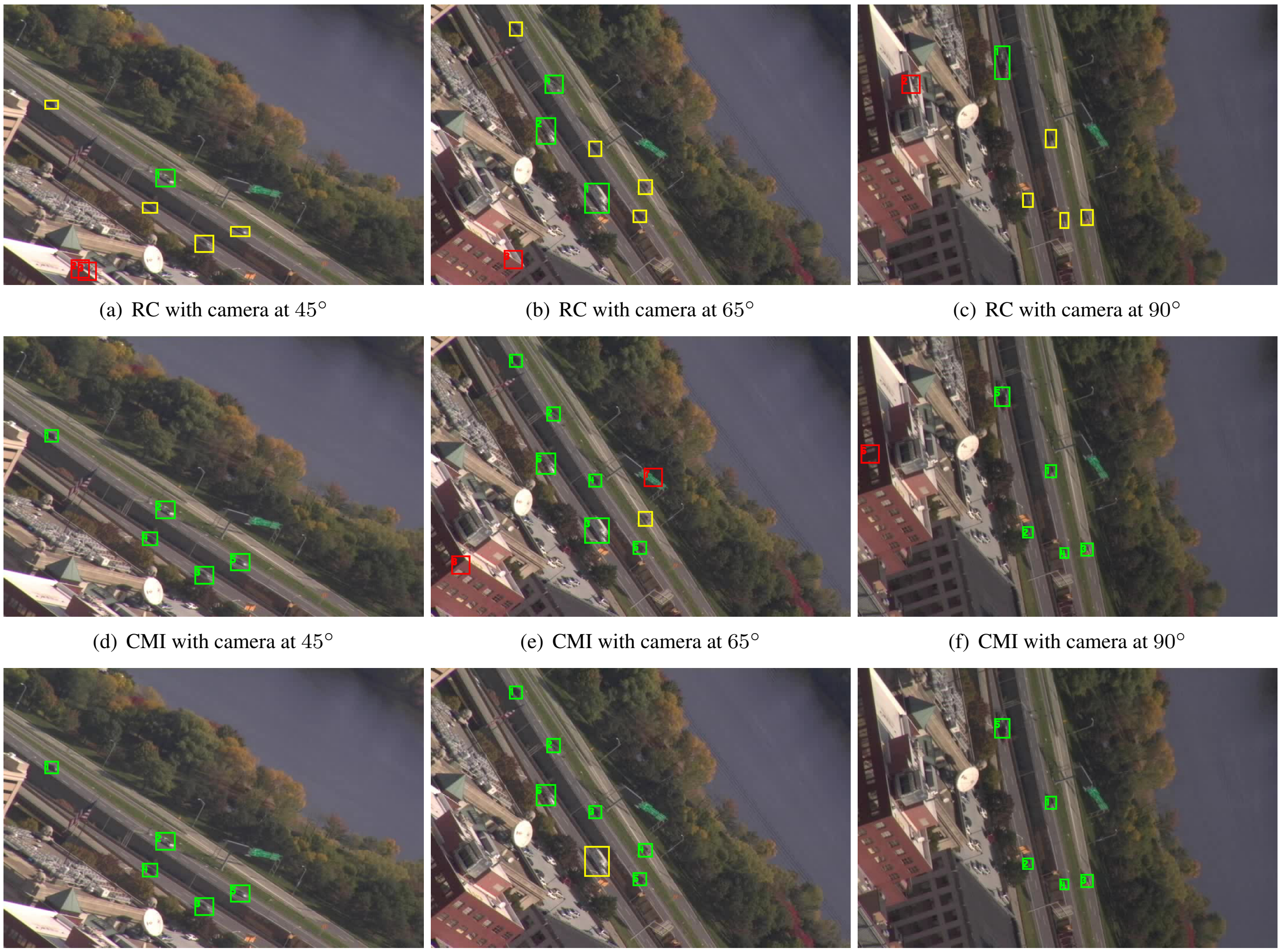
[4] J. Flusser. Moment invariant in image analysis. *Proceedings of WASET*, 11:196–201, Feb. 2006.

[5] Y. Guo, S. Hsu, Y. Shan, H. Sawhney, and R. Kumar. Vehicle fingerprinting for reacquisition & tracking in videos. In *CVPR*, volume 2, pages 761–768, June 20–25, 2005.

[6] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao. Sigma set: a small second order statistical region descriptor. In *CVPR*, 2009.

[7] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. on Information Theory*, 8:179–187, 1962.

[8] J. Koenderink and A. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–565, 1992.



(a) RC with camera at 45° (b) RC with camera at 65° (c) RC with camera at 90°

(d) CMI with camera at 45° (e) CMI with camera at 65° (f) CMI with camera at 90°

(g) RM with camera at 45° (h) RM with camera at 65° (i) RM with camera at 90°

Figure 4. **Region moments comparison.** The RC, CMI, and RM descriptors with isotropic features Φ_I, deployed for detection purposes with cameras at various rotation angles. The visualization includes the correct detections (green boxes), the misses (yellow boxes), and the false alarms (red boxes).

[9] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, pages 1–8, June 2008.

[10] D. Lowe. Distinctive image features from scale-invariant key points. *IJCV*, 60:91–110, 2004.

[11] X. Ma and W. E. L. Grimson. Edge-based rich representation for vehicle classification. In *CVPR*, volume 2, pages 1185–1192, Oct. 17–21, 2005.

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE TPAMI*, 27:1615–1630, 2005.

[13] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons. Moment invariants for recognition under changing viewpoint and illumination. *CVIU*, 94(1–3):3–27, 2004.

[14] O. C. Ozcanli, A. Tamrakar, B. B. Kimia, and J. L. Mundy. Augmenting shape with appearance in vehicle category recognition. In *CVPR*, volume 1, pages 935–942, 2006.

[15] A. Papoulis. *Probability, Ramdon Variables, and Stochastic Processes*. McGraw-Hill, 1991.

[16] F. Porikli. Integral histogram: a fast way to extract histograms in cartesian spaces. In *CVPR*, volume 1, pages 829–836, June 20–25, 2005.

[17] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR*, volume 2, pages 2033–2040, 2006.

[18] B. Schölkopf and A. Smola. *Learning with kernels: SVM, regularization, optimization, and beyond*. The MIT press, 2002.

[19] H. Shu, L. Luo, and J.-L. Coatrieux. Moment-based approaches in imaging part 1: basic features. *IEEE Engineering in Medicine and Biology Magazine*, 26(5):70–74, 2007.

[20] E. M. Stokely and S. J. Wu. Surface parameterization and curvature measurement of arbitrary 3-d objects: five practical methods. *IEEE TPAMI*, 14(8):833–840, 1992.

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

[22] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, pages 1–8, June 2007.

[23] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *IJCV*, 62:61–81, 2005.

[24] A. Vedaldi and S. Soatto. Local features, all grown up. In *CVPR*, volume 2, pages 1753–1760, 2006.

[25] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57:137–154, 2004.

[26] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *ICCV*, pages 1–8, 2007.

[27] Y. Yao and G. Doretto. Boosting for transfer learning with multiple sources. In *CVPR*, 2010.