Efficient Planar Affine Canonicalization

Alberto Ruiz[a], Pedro E. López de Teruel[b,], Lorenzo Fernández Maimó[b]

[a]*DIS, University of Murcia, Spain*
[b]*DITEC, University of Murcia, Spain*

Abstract

This paper presents a fast and accurate affine canonicalization method for planar shapes. This method improves on previous ones based on iterative optimization that produce multiple canonical versions. Canonicalization provides a common reference frame for shape comparison without the loss of discrimination ability often caused by invariant features. It also gives for free the alignment transformation between any pair of shapes. The proposed method is based on the properties of the joint angular distribution of marginal skewness and kurtosis, the so-called *SK signature*, which can be efficiently computed in closed form from the raw image moments. The experiments demonstrate that the method is robust to the non-affine distortions caused by natural perspective image conditions. Thus, it can be used as an automatic preprocessing step to add affine invariance in statistical pattern recognition applications.

Keywords: shape recognition, affine canonicalization, ICA, kurtosis

1. Introduction

Visual recognition of planar shapes is a fundamental problem in pattern recognition and computer vision, with applications in many diverse fields including autonomous robot navigation, surveillance, document understanding, localization, and augmented reality. The proliferation of low-cost mobile devices equipped with high-quality cameras (e.g., smartphones and drones) increasingly demands simpler and more accurate shape recognition methods.

A common approach to solving this problem is based on standard classifiers using a suitable set of invariant features [1, 2, 3, 4, 5, 6]. These methods are fast and do not

require costly learning stages. However, simple techniques to achieve invariance may introduce additional perceptual aliasing, reducing discrimination ability. Aggregation methods based on point distributions [7] or shape contexts [8] have similar drawbacks.

Shape discrimination can be improved by alignment, obtaining an explicit model-target transformation [9, 10, 11, 12, 13]. This allows comparing registered shapes directly in the measurement domain by means of a simple Euclidean metric or the more powerful Hausdorff distance [14, 15]. The inferred transformations are also useful to discard inconsistent matching hypotheses and provide pose estimates for self-localization and navigation applications [16]. Several ideas have been proposed for homography estimation from planar contours [17, 18, 19, 20, 21], including recursive probabilistic filters [22, 23], statistical theory of shape [24, 7, 14], and differential geometry [25]. Another interesting approach to alignment is based on estimating a non-parametric probability model for the transformations of a set of training instances with respect to a "congealed" version determined by minimization of pixelwise entropy [26, 27]. In general, alignment techniques are computationally expensive for multiclass shape recognition, especially when the parameters of the transformation cannot be expressed in closed form (due, for example, to the lack of explicit corresponding landmarks) and iterative approximations are needed for registration of all possible target-model pairs.

The extraordinary computing power of recent graphic processing units (GPU) has produced a considerable interest in machine learning techniques that use massive amounts of training data (natural or synthetic). In particular, deep convolutional neural networks have proved remarkably successful in many challenging vision applications [28], including image alignment [29, 30]. In a promising step towards automatic canonicalization, generic spatial transformer neural modules allow the networks to learn how to transform feature maps to minimize the training error [31]. However, this kind of deep models have some disadvantages such as long learning times, ad-hoc selection of the network architecture, heuristic tuning of hyperparameters, and difficult interpretation of the learned models.

Hierarchical probabilistic generative models have also been recently proposed [32]. This approach admits a wide range of transformations, requires very few training sam-

ples, and supports transfer learning between categories. Moreover, the underlying perceptual model has some cognitive plausibility. These advantages, though, come at the cost of very long inference times.

In contrast with the above general approaches, we are interested here in the specific problem of efficient planar shape recognition from natural images captured by ordinary cameras. Many computer vision methods assume weak perspective projection, modeled by simple affine transformations. This assumption usually holds in practice when object depth is small compared to the distance to the camera. In any case, it is not a severe limitation as full perspective shape recognition without explicit correspondences can be easily achieved by iterative refinement of a good affine initial solution [16, 33]. Therefore, we will explore affine alignment methods that are robust to moderate departures from weak perspective caused by out-of plane rotation. We will focus on efficient one-shot learning (using a small number of training samples for each class, ideally just one) and closed-form algorithms for the whole data processing pipeline.

The rest of the paper is organized as follows. Section 2 reviews the canonicalization approach to registration. Section 3 introduces the so called *SK signature* and describes its applications to shape recognition and alignment. A closed-form, efficient canonicalization algorithm based on this signature is developed in Section 4. The stability and range of application of the proposed method is experimentally evaluated in Section 5. The paper concludes with a summary of contributions and future research directions.

2. Canonicalization

Optimal registration is computationally expensive for classification applications, requiring a separate optimization process for each model. A faster alternative is provided by *canonicalization*, which allows precomputation of a good approximation to all possible alignment transformations and evaluation of shape similarity in a common reference frame. A traditional alignment transformation works with two input images, while canonicalization needs just one.

Invariance to a group of transformations can be achieved without any additional loss of class separability by using canonical representatives. Different shapes corre-

spond to the classes of equivalence induced on the set of planar regions by the transformations in the group. Each shape is represented by a particular canonical element in the class, characterized by certain conventional geometric properties. The alignment transformation T_{ab} between any two elements a and b can be immediately obtained as $T_{ab} = C_a^{-1} C_b$ from the respective canonicalization transformations C_a and C_b. This process is much faster than computing a different transformation from scratch for every model-target pair[1]. Furthermore, shape similarity can be directly evaluated in the canonical frame, making on-line classification very efficient as only one 'warping' transformation of the target shape is required. This approach is still suboptimal because the abstract canonical frame does not have any physically meaningful metrics, and the canonicalization transformations are not optimized to reduce registration residuals. However, as demonstrated in the experimental section, it provides excellent approximations for most practical purposes.

Canonicalization for the planar affine group (6 d.o.f.) is generally thought to be a simple task: we first apply a whitening transformation[2] and then fix one single remaining rotational degree of freedom [34, 35]. In other words, we must find a characteristic, or "intrinsic" orientation of the (whitened) shape. In principle this can be easily done by considering geometric properties like big concavities, bitangents, most distant points, or Fourier phases, among many other ideas [36]. Unfortunately, most of these proposals only work well for special sets of shapes. Moreover, they do not always provide a unique orientation even for clearly asymmetric figures, and are unstable under noise or small non-affine distortions.

A popular method to detect a dominant orientation is based on the mode of gradient directions. This method is commonly used by keypoint detectors like SIFT to normalize salient image patches in order to compute invariant feature descriptors [37]. This has proved very successful for highly textured images, but in flat or binary regions typ-

[1]Certain highly symmetric shapes can be taken to the canonical version via different and equally valid alternative transformations; this ambiguity should be managed in an application-dependent fashion.

[2]This denotes an affine transformation that produces uncorrelated variables with zero mean and unit variance. This preprocessing transformation is widely used in data analysis and can be easily computed in closed-form from the first and second-order moments (see Appendix I).

ically arising in shape recognition the aggregated gradient is a purely local property of the boundary. This feature disregards the relative location and structure of the internal points and it is sensitive to noise. Rounded shapes do not have strongly dominant gradient orientations, and those with straight edges may have multiple histogram maxima even though their structure is rich enough to induce a single distinguished orientation.

A promising canonicalization approach using global image information is based on Independent Component Analysis (ICA) [38]. In contrast with the maximum variance projections of a data set obtained by the principal components, ICA looks for a linear transformation such that the new variables are as much statistically independent as possible. This new representation may provide a useful reinterpretation of the data set in terms of meaningful components. For example, a common application is blind separation of mixed signals. Computational ICA techniques typically start from whitened data and then iteratively optimize an orthogonal transformation to get new variables as different from Gaussian distributions as possible. The key idea is that any linear combination of (non Gaussian) random variables has more entropy, and therefore is "more Gaussian", than the original variables. Practical optimization costs are marginal kurtosis and relative entropy. Efficient implementations include FastICA [39] and RobustICA [40].

In the context of shape recognition, ICA has been applied in order to compute affine invariant descriptors and alignment homographies [41, 42]. These methods eventually work with Fourier or Zernike rotation invariant features, which partially defeat the advantages of canonicalization. Other ICA methods [43, 44] work with the contour coordinates as separate 1D signals instead of the whole set of 2D points (the joint distribution) in a general figure, which may include separate fragments and holes.

The above proposals use standard ICA implementations and produce unnecessary multiple orientations. While general multidimensional ICA require expensive iterative local optimization, the 2D case arising from the shape orientation problem is the simplest one, with just one degree of freedom. In offline applications it could even be solved by exhaustive search of all rotation angles. In this paper we present a fast and simple closed-form solution for affine canonicalization, based on the concepts developed in the next section.

3. The Skewness-Kurtosis signature

We will informally use the term 'shape' to refer to a planar binary region, although most of the following results are equally valid for gray level images.

A region R is defined by its indicator function $I_R(x, y) = 1$ if $(x, y) \in R$ and $I_R(x, y) = 0$ otherwise. The sum of function f over R is

$$E_R\{f\} = \iint_{\mathbb{R}^2} I_R(x, y) f(x, y) dx dy, \tag{1}$$

and the moments of R are

$$m_{pq} = E_R\{x^p y^q\}. \tag{2}$$

Let λ_1^2 and λ_2^2 be the eigenvalues of the covariance matrix of R. Degenerate regions ($\lambda_2/\lambda_1 \ll 1$) cannot be whitened in a numerically stable way, but in that case the orientation is trivially given by the principal direction[3]

The symbol μ_{pq} denotes moments of whitened regions, which verify $m_{10} = m_{01} = m_{11} = 0$ and $m_{00} = m_{20} = m_{02} = 1$.

Except in certain symmetric cases that require special treatment, the marginal moments of any variable, say x, μ_{p0}, $p > 2$, depend on shape orientation and can be used to specify a canonical affine invariant reference frame. We will focus on the simplest and most interesting cases: $p = 3$ (skewness) and $p = 4$ (kurtosis). Higher order moments are also related to symmetry and uniformity—more precisely, weight of tails—and do not provide useful additional information for our purposes.

3.1. Marginal moments

Consider the marginal skewness and kurtosis of the whitened shape for every clockwise rotation angle θ around the origin:

[3] It is advisable to use a soft threshold on λ_2/λ_1 such that the borderline shapes are treated in a special way: they are initially whitened, but if the subsequent processing stages produce a low-confidence result we reconsider the shape as one-dimensional.

$$S(\theta) = E_R\{(x\cos\theta + y\sin\theta)^3\},$$
$$K(\theta) = E_R\{(x\cos\theta + y\sin\theta)^4\}. \tag{3}$$

Both functions are fully determined by their values in $\theta \in [0, \pi)$, and verify the following periodicity conditions:

$$S(\theta + \pi) = -S(\theta),$$
$$K(\theta + \pi) = K(\theta). \tag{4}$$

Similar to the geometric representation of a covariance matrix as an ellipse of uncertainty, a measure of uniformity and asymmetry along each direction can be represented by the following characteristic 2D curves:

$$\mathcal{S}(\theta) = (S(\theta)\cos\theta, S(\theta)\sin\theta),$$
$$\mathcal{K}(\theta) = (K(\theta)\cos\theta, K(\theta)\sin\theta). \tag{5}$$

Fig. 1 illustrates the aspect of these functions for an example image. They are strongly sensitive to orientation although the distribution of image intensities is isotropic up to second order.

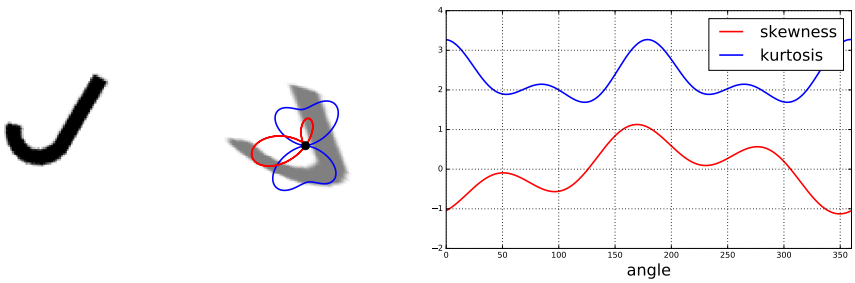

Figure 1: Left: original image. Center: whitened image and superposed polar representation $\mathcal{S}$ (red) and $\mathcal{K}$ (blue). Right: corresponding angular functions S and K.

In the polar plot in Fig. 1 we actually show $\max \mathcal{K} - \mathcal{K}$ to emphasize the most subgaussian projections. As shown below, for 2D distributions with finite support, subgaussianity corresponds better to a subjective concept of distinguished or intrinsic orientation. Kurtosis consistently detects the most uniform projections and it is very

robust to non-affine perturbations. There are, however, up to two local minima together with upside-down ambiguity, denoted by $K_1 = K(\theta_1) = K(\theta_1 + \pi) <= K_2 = K(\theta_2) = K(\theta_2 + \pi)$, totaling four candidate directions in perfectly orientable regions.

Because of antiperiodicity, S does not suffer from the 'upside-down' ambiguity of K. If $\max(S) > 0$ we have at most three local maxima $S_1 = S(\alpha_1) \geq S_2 = S(\alpha_2) \geq S_3 = S(\alpha_3)$, and hence there are up to three distinguished directions. If the local maxima are not all equal, they can be used to define a single orientation. For example, we could choose α_1 if $S_1 > S_2$ or α_3 if $S_2 > S_3$.

Skewness has actually been used to estimate the alignment rotation for registration of point samples [45]. Together with his well-known invariants, Hu [1] also proposed using the sign of μ_{30} to remove the 180° ambiguity of the principal axis in an early attempt to achieve metric canonicalization. Unfortunately, some clearly orientable figures produce very weak S responses and, as shown in Section 5.1, the orientation induced by $\mathcal{S}$ is unstable, so skewness cannot be the only basis of a general canonicalization method. Instead, it is an excellent method to remove the ambiguity of the candidate orientations provided by kurtosis.

To achieve this, we define a global orientation detector $\overline{\mathcal{S}}$ based on the 'average' location of the curve $\mathcal{S}$, which is more stable than any local α_k. It can be easily proved (see Section 4.2) that $\overline{\mathcal{S}}$ reduces to the following simple expression:

$$\overline{\mathcal{S}} \equiv \frac{1}{2\pi} \int_0^{2\pi} \mathcal{S}(\theta) d\theta = \frac{3}{8}(\mu_{30} + \mu_{12}, \mu_{21} + \mu_{03}). \tag{6}$$

For a whitened region, if $||\overline{\mathcal{S}}||$ is large, $\overline{\mathcal{S}}$ defines a unique affine invariant orientation. Interestingly, $||\overline{\mathcal{S}}||^2$ is essentially the fourth traditional orthogonal invariant proposed by Hu [1]. It was derived from algebraic considerations and it can now be given an intuitive geometrical interpretation. This simple idea has apparently not been proposed before.

3.2. Orientability

In addition to $\overline{\mathcal{S}}$ we define the following descriptors (Fig. 2):

$$
\begin{aligned}
r &\equiv \max K - \min K, \\
t &\equiv K_2 - K_1, \\
s &\equiv ||\overline{\mathcal{S}}||.
\end{aligned}
\tag{7}
$$

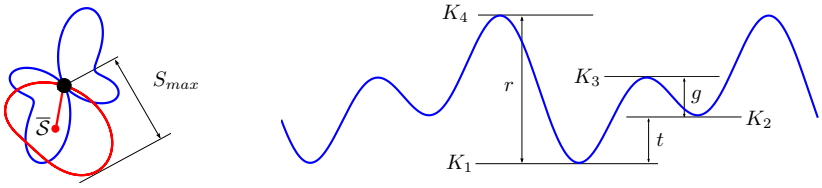

Figure 2: Orientability properties of the SK signature. S_{max} and g will be discussed later.

The 'orientability' of any figure on the basis of the third and fourth moments can be derived from r, t, and s. Figure 3 shows $\mathcal{S}$ and $\mathcal{K}$ for several basic shapes together with the corresponding values of r, t, and s.

If r is small, the figure is essentially not orientable (top row). If r is large and s is small, the figure is semi-orientable (second row). In this case, if t is small there are four alternative orientations, or else there are just two. Finally, if r and s are both large we can define a single orientation (bottom row). Again, if t is small we must choose from the four extrema, or else just from the two dominant ones. This contrasts with previous ICA-based efforts [46] that produce eight candidate orientations.

In practice we must use reasonable thresholds for these conditions to be robust to noise and non-affine distortions. A general procedure to do this will be described in Section 3.4. We first provide a general impression of the stability of the characteristic curves. Figure 4 illustrates the response of $\mathcal{S}$ and $\mathcal{K}$ to salt and pepper noise and image dilations in a few test images. Dilation produces a slight attenuation of the responses but orientation is not affected. Noise has a similar effect.

The characteristic curves are also reasonably resistant to non-affine deformations. Fig. 5 shows the signatures of a letter imaged from different view points with moderate out-of plane rotation. An experimental study of robustness will be presented in Section 5.

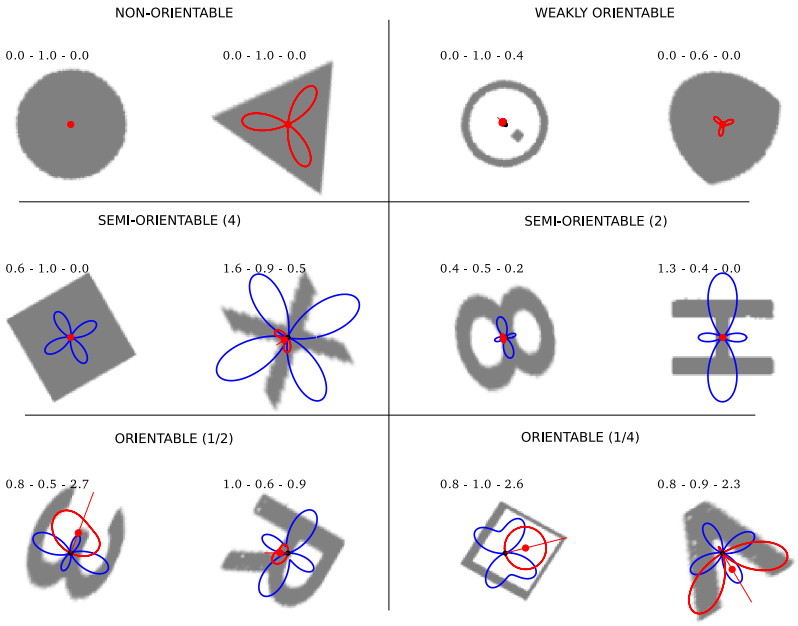

Figure 3: Geometric representation of marginal kurtosis $\mathcal{K}(\theta)$ (blue) and skewness $\mathcal{S}(\theta)$, as well as $\overline{\mathcal{S}}$ (red) for whitened regions with different 'orientability' properties. The values of r, $1-t/r$, and $10s$ are displayed above each case.

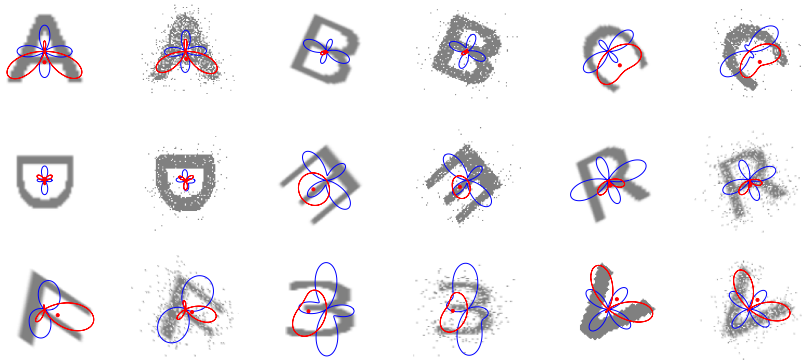

Figure 4: Response of $\mathcal{S}$ and $\mathcal{K}$ to dilation and noise.

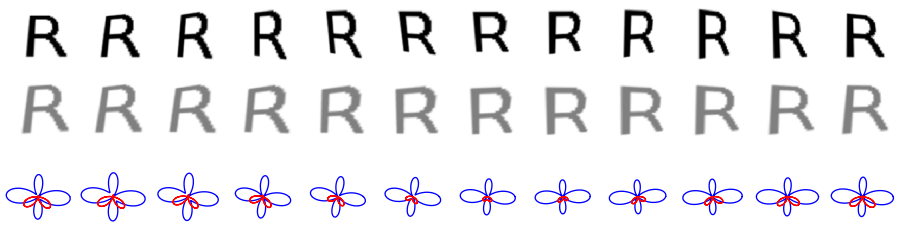

Figure 5: Stability of the SK signature for non-affine deformations. The first row shows a set of views with out-of-plane rotation caused by a 30° tilt angle, assuming that the original view has a field of view (FOV) equal to 40^{o}. The second and third rowshow the corresponding whitened shapes and SK signatures, respectively.

3.3. The joint SK signature

The extrema of S and K, together with $\overline{\mathcal{S}}$, induce virtual reference points and directions that can be transferred to the original observation frame and are covariant with affine transformations (Fig. 6).

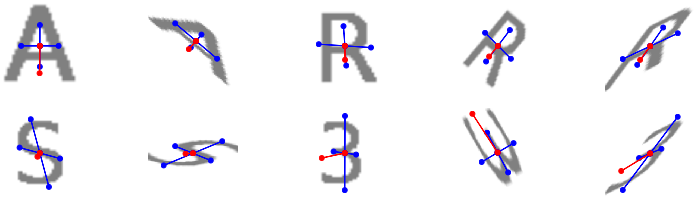

Figure 6: Reference points and directions induced by $\mathcal{K}$ and $\overline{\mathcal{S}}$ back-projected to the input frames.

It is also possible to combine the two curves into an invariant object. Let us define the joint skewness-kurtosis signature of a region as the 2D parametric curve:

$$SK(\theta) \equiv (S(\theta), K(\theta)). \tag{8}$$

In contrast with the separate functions $S(\theta)$ and $K(\theta)$, the joint SK signature as a curve independent of the parameter θ in the abstract (S, K) plane is an affine invariant property of the region. Fig. 7 illustrates the signatures of different shapes. Even though S and K are only a tool for efficient canonicalization, the joint SK signature is actually

highly discriminant and useful to define simple feature vectors for the classification or fast rejection of bad matches, as demonstrated in Section 5.5.

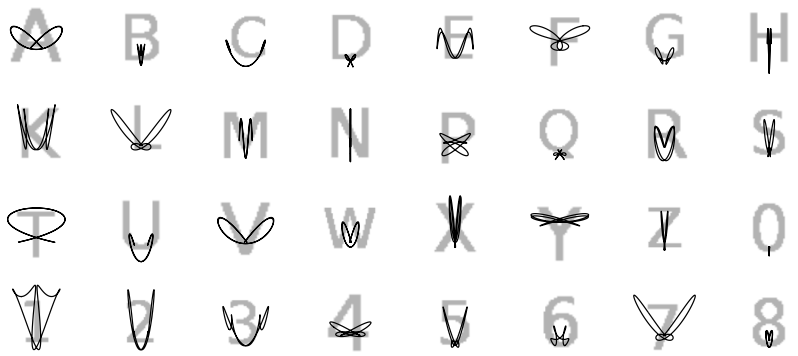

Figure 7: The joint $\mathcal{S}K$ signatures of a set of alphanumeric characters in a standard sans-serif font.

3.4. Canonicalization procedure

Our main goal is to select a canonical direction from the orientability properties r, t, and s described above. A single orientation must be determined when possible, but in order to correctly deal with symmetries, noise, and non-affine distortions, in some weakly orientable cases we must provide a set θ^* of candidate orientations. This can be accomplished by the following selection algorithm. Its sensitivity is controlled by three thresholds ϵ_r, ϵ_t, and ϵ_s, one angular tolerance δ, and one offset o.

i. If $r < \epsilon_r$ the shape is not orientable. Otherwise, let θ_1, θ_2, $\theta_3 = \theta_1 + \pi$, and $\theta_4 = \theta_2 + \pi$ be the four local minima of K.

ii. If $1 - t/r < \epsilon_t$ then $\Theta = \{\theta_1, \theta_3\}$. Otherwise, $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. (We discard directions weaker than a proportion ϵ_t of the strongest one.)

iii. If $s < \epsilon_s$ then $\theta^* = \Theta$ and stop. Otherwise,

iv. Set up overlapping decision regions for the elements in Θ as depicted in Fig. 8, and include in θ^* all the directions associated with the regions containing $\overline{\mathcal{S}}$.

In step (iv) we select the orientation more directly indicated by $\overline{\mathcal{S}}$, or located at its right side, with the precaution of establishing ambiguity margins of width δ that

generate two candidates. These common sectors are unavoidable to guarantee that changes in $\arg \overline{\mathcal{S}} < \delta$ do not result in an abrupt change in the selected orientation.

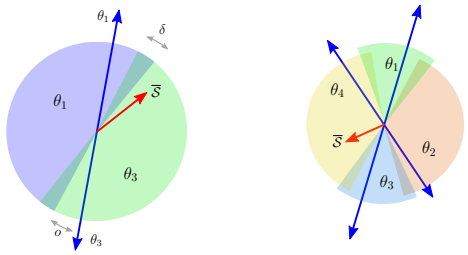

Figure 8: Decision regions for the key step of the selection algorithm.

Fig. 9 illustrates the results of the algorithm on the alphanumeric characters with reasonably safe default thresholds: $\epsilon_r = 0.2$, $\epsilon_t = 0.75$, $\epsilon_s = 0.04$, $\delta = 20°$. The offset $o = 10°$ has been chosen to minimize the number of orientable cases producing two directions (see letter `K`). Letter `O` does not pass the ϵ_r test, while number zero is less rounded and, hence, semi-orientable. The well-defined orientation found for the letter `X` in this font ($10s = 0.55$) was initially considered to be a programming bug, but careful inspection of the image revealed that this letter design is actually asymmetric. Only five weakly orientable shapes (`BDPQ8`) produce more than a single orientation.

This reduced number of canonical candidates—compared to using all four alternatives per model—has important practical consequences. Due to the quadratic dependence of the number of required model-target comparisons on the average number of alternative orientations, classification time becomes much faster with negligible performance degradation (see Section 5).

Figure 10 shows the stability of the canonical version to deformations caused by a large tilt with large focal length, so that perspective distortion is not very far from the affine assumption of weak perspective.

Finally, Figure 11 illustrates shape alignment. If there is a good match, image registration could be further improved using, for example, the Lucas-Kanade method [9]. The inverse compositional variant can be applied very efficiently since the Jacobian of the transformation can be precomputed for each model. In Section 5.4 we study the

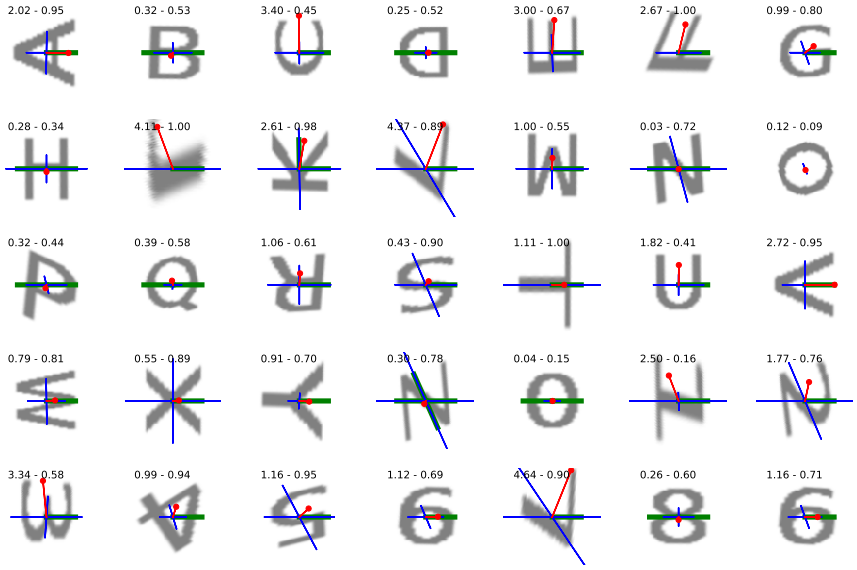

Figure 9: Directions selected by the proposed method (green) from the kurtosis extremes (blue) and $\overline{\mathcal{S}}$ (red). On top of each shape we show $10s$ and $1 - t/r$.

Figure 10: Canonical models from perspective views with a 70° tilt and large focal length (FOV=10^o). Some deformations look extreme but they are actually close to the affine group.

consequences of evaluating shape similarity in the canonical frame.

The third and fourth order moments have been previously used for shape alignment [45] and canonicalization [46] but, to the best of our knowledge, the remarkable joint properties of K and $\overline{\mathcal{S}}$ have not been fully exploited before. In Section 4 we will develop a simple closed-form algorithm to efficiently compute them from the raw image moments.

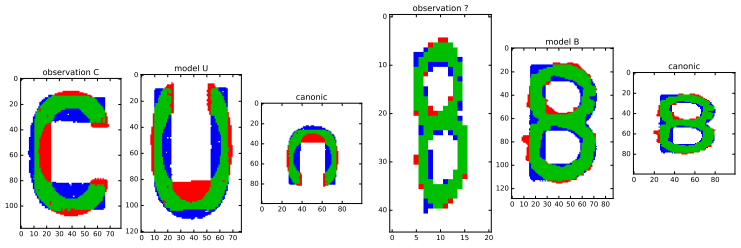

Figure 11: Two examples of canonical registration in the observation, model, and canonical reference frames. Coincidence is shown as white and green areas, and the alignment errors are red and blue. Note the difficulty of B-8 discrimination in a low resolution image.

4. Computation of the SK signature

We now turn to the computation of the SK signature in (3). Let us define

$$t_{pq}(\theta) = \cos^p \theta \sin^q \theta. \tag{9}$$

Using this auxiliary function, $S(\theta)$ and $K(\theta)$ can be written in terms of the third and fourth order moments of the whitened shape as follows:

$$\begin{aligned} S(\theta) &= \mu_{30} t_{30}(\theta) + 3\mu_{21} t_{21}(\theta) + 3\mu_{12} t_{12}(\theta) + \mu_{03} t_{03}(\theta), \\ K(\theta) &= \mu_{40} t_{40}(\theta) + 4\mu_{31} t_{31}(\theta) + 6\mu_{22} t_{22}(\theta) + 4\mu_{13} t_{13}(\theta) + \mu_{04} t_{04}(\theta). \end{aligned} \tag{10}$$

Any moment μ_{pq} can be efficiently computed from the raw moments m_{pq} of the raster image as described in Appendix I. Alternately, if the figure is represented compactly by a contour, the moments can be directly computed from powers of node coordinates by using Green's Theorem (see Appendix II).

Once the nine required moments have been computed by either method, the SK signature can be numerically obtained by (10) on a discretized domain Ψ with the desired angular resolution. This operation can be efficiently performed using precomputed values of $t_{pq}(\theta_k)$. In this case $S(\theta_k)$ and $K(\theta_k)$ are just linear combinations or fixed vectors of dim $|\Psi|$. Their extrema and distinguishable angles are immediately obtained from the discretized signature.

4.1. Extremes of K

The above numeric solution is accurate enough for many applications and takes a small proportion of the total computing effort, which for raster images is dominated by the computation of moments and image warping. Even though this might be sufficient, we also provide an exact and faster closed-form solution. It illustrates some general properties of the SK signature and may significantly reduce computing time in applications using contour-based image representations.

The local extrema of marginal kurtosis satisfy $K'(\theta) = 0$. Since the trigonometric terms $t_{pq}(\theta)$ verify the following relation,

$$\begin{aligned} t'_{pq}(\theta) = \frac{d}{d\theta} \cos^p \theta \sin^q \theta = -p \cos^{p-1} \theta \sin^{q+1} \theta + q \cos^{p+1} \theta \sin^{q-1} \theta = \\ = q\, t_{p+1,q-1}(\theta) - p\, t_{p-1,q+1}(\theta), \end{aligned} \tag{11}$$

the derivative $K'(\theta)$ is also a linear combination of the same trigonometric terms,

$$\begin{aligned} K'(\theta) = 4\mu_{31} t_{40}(\theta) + 4(3\mu_{22} - \mu_{40}) t_{31}(\theta) + 12(\mu_{13} - \mu_{31}) t_{22}(\theta) + \\ + 4(\mu_{04} - 3\mu_{22}) t_{13}(\theta) + (-4)\mu_{13} t_{04}(\theta), \end{aligned} \tag{12}$$

that can be expressed as

$$K'(\theta) = 4 \sum_{k=0}^{4} d_k\, t_{4-k,k}(\theta), \tag{13}$$

where

$$\begin{aligned} d_0 &= \mu_{31} \\ d_1 &= 3\mu_{22} - \mu_{40} \\ d_2 &= 3\mu_{13} - 3\mu_{31} \\ d_3 &= \mu_{04} - 3\mu_{22} \\ d_4 &= -\mu_{13}. \end{aligned} \tag{14}$$

The roots of $K'(\theta)$ do not change if we divide it by $\cos^4(\theta)$ to get a polynomial in $\tan\theta$:

$$\frac{t_{4-k,k}}{\cos^4\theta} = \frac{\cos^{4-k}\theta \sin^k\theta}{\cos^4\theta} = \cos^{-k}\theta \sin^k\theta = \tan^k\theta. \tag{15}$$

Therefore the solutions of $K'(\theta) = 0$ satisfy $\tan\theta = x$ where x is a root of

$$\sum_{k=0}^{4} d_k x^k = 0. \tag{16}$$

This equation can be easily solved by the polynomial root finding routines available in all standard scientific packages. There are up to four real solutions: two of them are maxima of $K(\theta)$, corresponding to the most kurtotic projections; the other two are minima, for the most uniform (more precisely, subgaussian or *platykurtic*) projections. If there are only two real solutions then $K(\theta)$ does not have the local extremes K_2 and K_3 in Fig. 2. Each solution x of (16) gives two angles θ and $\theta + \pi$ from $\tan\theta = x$, which is consistent with the expected upside-down ambiguity of marginal kurtosis. The angles $\theta = \pm\pi/2$ cannot be found by this method but this case is easily detectable as $d_4 = 0$, producing a lower degree polynomial that finds the three remaining solutions.

Although it is not strictly needed by the proposed canonicalization method, a similar solution can be derived for the three extrema of $S(\theta)$ (or any other moment).

4.2. Closed form expression for $\overline{\mathcal{S}}$

We now proceed to prove (6). Using the auxiliary function (9) and the expansion (10) the components of $\mathcal{S}$ in (8) can be written as

$$\begin{aligned} S(\theta)\cos(\theta) &= \mu_{30}t_{40}(\theta) + 3\mu_{21}t_{31}(\theta) + 3\mu_{12}t_{22}(\theta) + \mu_{03}t_{13}(\theta), \\ S(\theta)\sin(\theta) &= \mu_{30}t_{31}(\theta) + 3\mu_{21}t_{22}(\theta) + 3\mu_{12}t_{13}(\theta) + \mu_{03}t_{04}(\theta). \end{aligned} \tag{17}$$

We only need the following integral for selected p and q:

$$I_{pq} \equiv \int_0^{2\pi} t_{pq}(\theta) d\theta. \tag{18}$$

This integral cancels out for odd powers, so we immediately get $I_{13} = I_{31} = 0$. Integrating by parts we obtain the remaining even powers

$$I_{22} = \frac{\pi}{4}, \quad I_{40} = I_{04} = \frac{3\pi}{4}, \tag{19}$$

which completes the proof.

4.3. Computational complexity

The most expensive step is computing the image moments from a raster image. We need the 15 raw moments up to order 4: m_{pq} for $p \geq 0$, $q \geq 0$, and $p + q \leq 4$. The six ones up to order 2 are required for whitening, hence a common cost for any affine invariant method. The 9 normalized higher order ones μ_{pq} can be derived in closed form from the raw ones as described in Appendix I. For a $w \times h$ image we need $\sim 9wh$ additional operations to compute the full canonicalization transformation, without the need of going through an intermediate whitened image.

In contrast, a histogram of gradient orientations requires two convolutions for the gradient and element-by-element operations for magnitude, angle, and histogram accumulation. Using a simple 3×3 Sobel mask the cost is $\sim (6 + 6 + 3)wh$ operations. The gradients must be computed on a whitened image patch, requiring an additional auxiliary warping operation.

5. Experiments

Under ideal affine transformations shape alignment will be perfect and recognition mistakes will be caused by application-dependent circumstances such as class variability, allowed shape deformations (e.g., dilations), or image noise. We are more interested in natural imaging conditions, so we have experimentally measured robustness to the non-affine perspective effects caused by tilted views.

5.1. Stability of $\overline{\mathcal{S}}$ vs K

Skewness alone could in theory be used to find a single distinguishable orientation for a wide range of shapes. Unfortunately, as anticipated in Section 3, it is not very robust to non-affine deformations, which relegates it to the disambiguation role for the more stable directions generated by kurtosis. In this section we experimentally study this phenomenon.

We define the angular error Δa as the absolute difference between the intrinsic orientation on a reference figure and that obtained from a tilted view. Fig. 12 shows

the mean angular errors for increasing tilt angles measured on the illustrative set of shapes in Fig. 7.

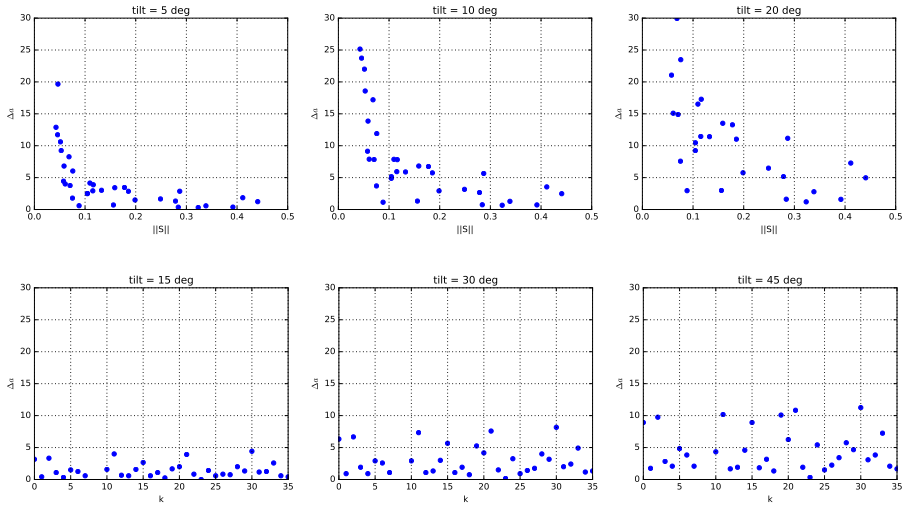

Figure 12: Stability of orientation for the alphanumeric shapes as measured by Δa. The ordinate of each blue dot is the mean angular error of a figure at the given tilt angle for 12 regularly spaced tilt directions (FOV = $40°$). Top row: orientation based on the α angle of $||\overline{\mathcal{S}}||$. Bottom row: based on θ_1 angle of K_1. Note that the tilt angles in the bottom row for kurtosis cover a much wider range. In the top row we can also observe the dependence of Δa on $||\overline{\mathcal{S}}||$. In the bottom row the abscissa is just the position of the image in the set.

We observe that the aggregated skew-based angle α is very sensitive to tilt: figures with small $||\overline{\mathcal{S}}||$ have high errors even for small angles, and the more asymmetric figures, with $||\overline{\mathcal{S}}|| > 0.1$, have mean errors $> 10^o$ for tilt = 20^o. In contrast, the minima of kurtosis have significantly lower angular errors at higher tilt angles.

5.2. Shape recognition

In this section we evaluate the quality of the SK signature for shape recognition using invariant features and canonicalization. For this experiment we use the set of official traffic plate Spanish symbols, comprised by 32 capital letters and digits depicted in Fig. 13. Only one instance of the affine equivalent shapes `OO` and `69` are included; `I` is also removed because it is nearly degenerate in this font and requires special treatment.

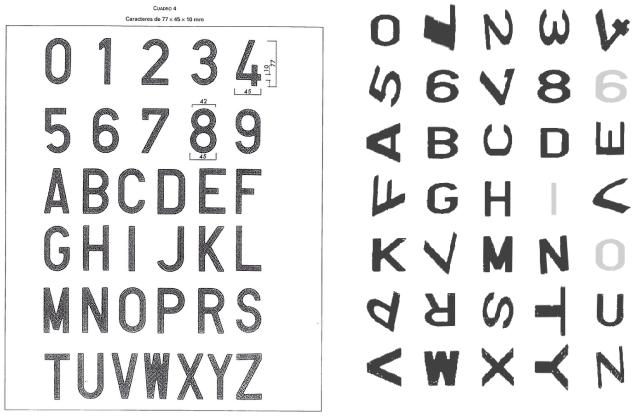

Figure 13: Official Spanish traffic plate symbols and their canonical versions.

This small dataset is nevertheless not easy to deal with because of the small differences between some pairs of shapes when transformed to the common canonical frame. Fig. 14 shows all pairwise Hausdorff distances for an image resolution of 100×100 pixels including six standard deviations. The most similar symbol pairs are the following:

```
8B 4.38, NZ 4.39, 2Z 5.00, DO 5.00, 3E 5.10, 5S 5.10, 68 5.66
```

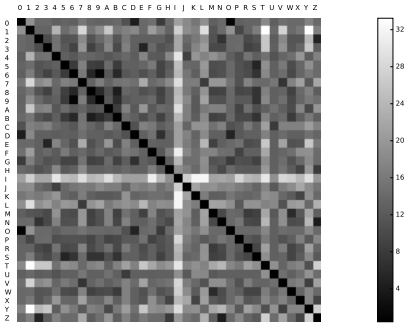

Figure 14: Pairwise Hausdorff distances in the canonical frame for the traffic plate dataset.

We use a single frontal prototype for each class, extracted from the low resolution image shown in Fig. 13. The test samples are a set of 12 synthetic perspective views for

each model (obtained as in Fig. 5) with increasingly difficult tilt angles ($15°$, $30°$, $45°$, and $60°$) for an effective FOV of $30°$. This corresponds to a relatively short distance to the camera, which may produce strong perspective distortion (the full FOV of the standard webcams and cameras found in mobile devices is usually not much larger than $60°$).

We have studied recognition methods based on the following attribute vectors and similarity functions:

1. 'I-Fourier': Rotation invariant Fourier contour descriptor (10 lowest frequencies) of the whitened shape. Used as baseline comparison.
2. 'I-SK': Simple rotation invariant feature vector extracted from the SK signature

$$\left(\frac{K_{min} + K_{max}}{2}, r, S_{max}, s, \frac{t}{r}, \frac{g}{r} \right), \tag{20}$$

   where g is the 'prominence' of the second local minimum of K (Fig. 2).
3. 'C-XOR': Total area of the symmetric difference of canonical images.
4. 'C-Hausdorff': Hausdorff distance of canonical images. The Hausdorff distance is efficiently computed from a precomputed distance transform [47] of the canonical shape.

In all cases we classify by minimum distance to the single model and reject the classification if it exceeds a given threshold. Methods 1-2 use Euclidean distance. In methods 3-4 we compare all canonical candidate pairs (1, 2 or 4) and return the best match. We have studied two strategies for selection of multiple canonical versions: a) the recommended default thresholds shown in Fig. 9; and b) the most conservative case, which always generates 4 canonical representatives for every shape. The experiments have been replicated to construct ROC curves showing the classification error rate (relative to the size of the test set) for increasing rejection rates.

Fig. 15 shows the classification results of the faster selection strategy. For the test set (32 models $\times$ 12 tilt orientations = 384 elements) it produces 574, 507, 524, and 483 canonical candidates (an average of ~ 1.4 representatives per shape), and for the 32 models it produces 47 candidates (~ 1.5 per model), totaling $\sim 25K$ target-model comparisons.

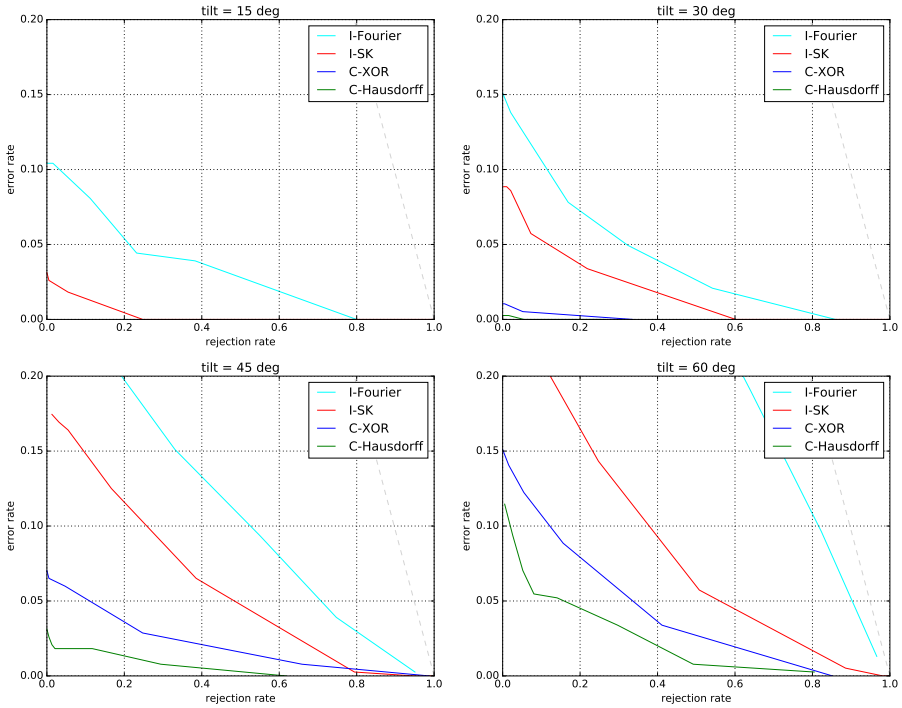

Figure 15: Robustness to perspective distortions for the traffic plates dataset using efficient selection (the errors of C-XOR and C-Hausdorff at 15^o are 0% at all rejection levels). The dashed line indicates the baseline performance of a random classifier.

These results confirm that the simple I-SK feature vector is reasonably accurate, as expected from the variability of the curves shown in Fig. 7. In general, though, rotation invariant distances like I-SK and Fourier are only appropriate for small tilt angles. They use exact affine invariant descriptors which discard information needed for shape discrimination. In contrast, the alignment approach on the canonical frame is significantly better, with Hausdorff distance outperforming symmetric difference.

Fig. 16 shows the results of the same experiment with the safest but slower selection strategy. In this case we have an average of 1460 target canonical candidates (out of the $1536 = 4 \times 384$, as a few cases have only one local minimum of K), and 122 model canonical candidates (out of $128 = 4 \times 32$), requiring $\sim 178K$ target-model comparisons. We obtain a small reduction in the error rates at the cost of more than

$7\times$ computing time[4]. For low distortion levels the individual selection method can be up to four times faster, as a safe alternative is taking one arbitrary canonical version for the models and all four for the observations. These gains would be even greater if compared with the eight orientations considered in [46].

We conclude that the selection of a unique canonical model (or a minimum set of candidates) based on the analysis of the SK signature is a significant improvement on earlier methods based on fixed sets of multiple candidate orientations.

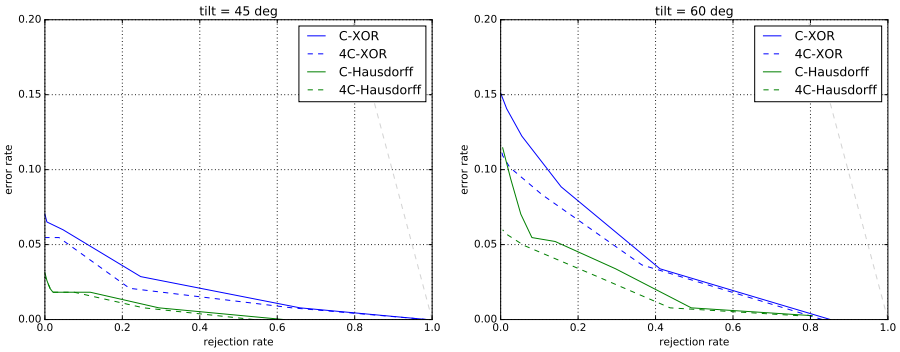

Figure 16: Error reduction using full selection (for tilt angles $< 30^\circ$ we obtain essentially perfect classification).

5.3. Validity range of the weak perspective assumption

Nonlinear alignment, required for example for full perspective transformations or deformable models, is based on iterative refinement. A good starting point is essential, but it is equally important that a reduced set of possible model matchings be selected to avoid the wasted optimization cost of eventually mismatched shapes. Hence the importance of an efficient and robust affine canonicalization.

As previously illustrated in Fig. 10, for small FOV (or equivalently large focal length) the affine assumption may be valid even for large tilt angles. It is interesting

[4]Duplicate canonical models produced by symmetric shapes (e.g. `HNSX8`) are not removed in both strategies. If this were done the absolute computing times would slightly decrease but the speedup would be similar.

to study the breaking point of the proposed method with respect to the effective field of view of the image. Fig. 17 shows the error rate (at 0% rejection) of several classification methods in 30^o and 45^o tilted views, for increasing FOVs. Notably, Hausdorff distance on the canonical frame remains below 3% error for FOVs below 30^o in the harder 45^o tilted view case.

This wide operating range demonstrates that the method can be safely embedded as a module for fast initialization from scratch in more advanced shape recognition systems.

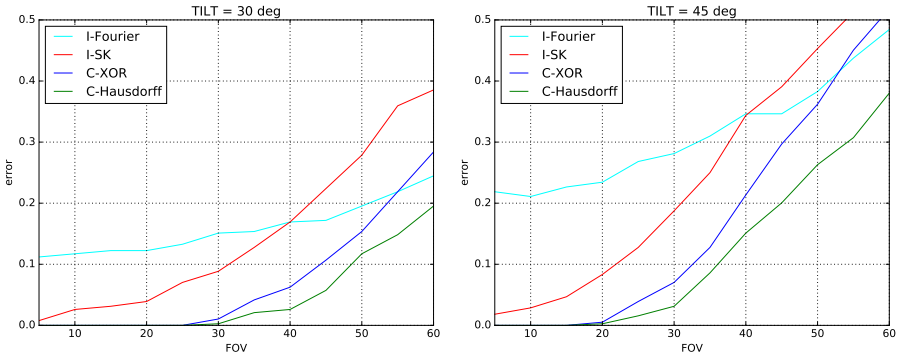

Figure 17: Graceful degradation of classification accuracy (traffic plate dataset) for increasing perspective distortion at fixed tilt angles.

5.4. Canonical vs physical error alignment

In principle, shape similarity should be measured on the physical sensor frame, since the abstract canonical frame does not have any meaningful metrics (Fig. 11). This requires costly separate warping transformations of all models into the original target frame.

We have compared the classification accuracy of Hausdorff distance evaluated in the canonical frame and in the observation frame. We have also studied an efficient modification of Hausdorff distance described in Appendix III, which combines distances in the observation and model frames. Fig. 18 shows the ROC curves of the above character recognition problem for several tilt angles.

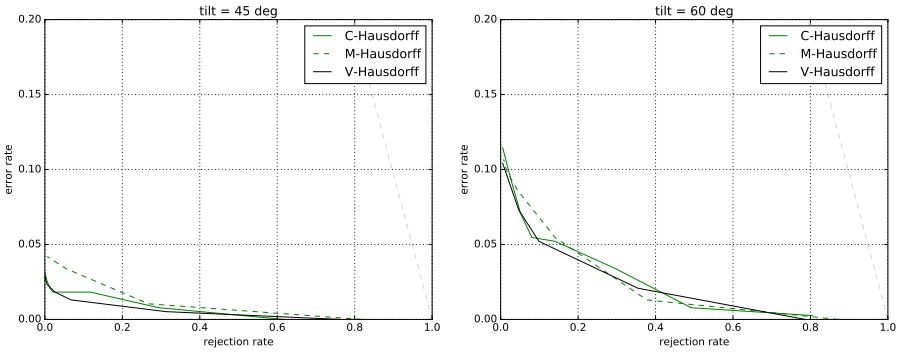

Figure 18: Accuracy of Hausdorff distance in different reference frames. C: canonical frame, V: target input frame, M: combined frames. For tilt angles $< 30^o$ the three methods get essentially perfect classification.

Perhaps surprisingly, no significant differences are observed between the three methods. A possible explanation is that the pairwise alignment transformations are a by-product of individual canonicalization, not optimized to reduce any error. This may have a regularization effect that protects against overfitting. We conclude, then, that shape similarity can be safely evaluated in the abstract canonical frame but further studies are required to determine the practical relevance of this issue.

5.5. Fast rejection of wrong matching candidates.

The experiments in Section 5.2 suggest that simple morphological features of the SK signature (20) can be used to discard a large number of wrong matchings without the need of explicitly computing registration error. For this idea to be sound it is essential that no false negatives be produced. Fig. 19 shows the joint distribution of the I-SK distance and the Hausdorff distance in the canonical frame for all pairs of synthetic perspective views of the traffic plate character dataset, generated at 30^o FOV and 30^o tilt. Clearly, the I-SK distance provides a safe bound on Hausdorff distance. For example, if we decide to reject candidate matchings with a Hausdorff distance > 4 pixels, we must consider only models with I-SK distance < 0.7.

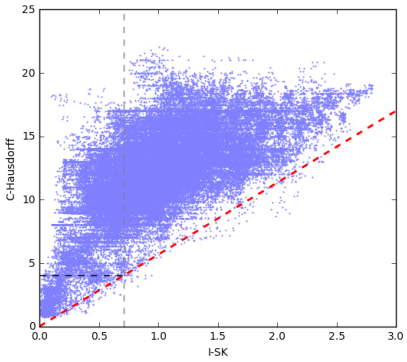

Figure 19: Joint distribution of the I-SK distance and the Hausdorff distance (blue), and safe rejection threshold (red).

5.6. Real shapes

We have tested the method on a plate recognition task with very low resolution real input images. Fig. 20 shows a small 259×457 rectangular region cropped from a 13 Mpixels photo taken with a consumer smartphone (focal length = 4.6mm). This region includes only 0.9% of the original image area, so the plate characters appear extremely pixelated. We use again the official symbol templates shown in Fig. 13.

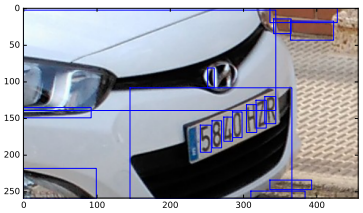

Figure 20: Bounding boxes of dark regions detected as potential plate symbols.

The upper row of Fig. 21 shows the connected components detected by fixed thresholding (size-normalized). A closer view of the plate characters is shown in the middle row; their sizes vary from 10×31 to 13×35 squared pixels. The bottom row shows one of the observations, its corresponding model, and a comparison of their respective canonical forms at 100×100 resolution. The Hausdorff distance in the

canonical frame in this case is 4 pixels.

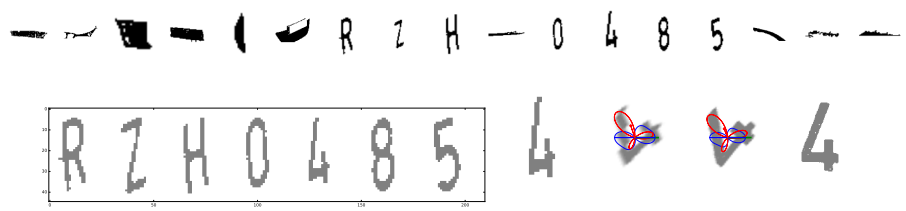

Figure 21: Top: detected regions; bottom-left: true targets; bottom-right: comparison of observed and model images in the original and canonical frames.

The following table shows the labels predicted by the considered classification methods with several rejection thresholds:

```
      ground truth:   ? ? ? ? ? ? R Z H ? O 4 8 5 ? ? ?
        I-SK d<0.3:   ? ? W ? ? ? R Z H I 0 4 8 5 ? ? ?
       C-XOR d<700:   ? ? ? ? ? ? R Z H I O 4 8 5 ? ? ?
  C-Hausdorff d<4:    ? ? I ? ? ? ? Z H ? O 4 8 5 ? ? ?
  C-Hausdorff d<8:    I ? I I A ? R Z H ? O 4 8 5 ? 1 ?
I-Fourier d <0.12:    1 P D 1 B ? R 8 H 1 I 4 I 2 B 4 1
```

The rotation-invariant Fourier descriptor of whitened shapes obtains very poor results while I-SK, symmetric difference (C-XOR) and C-Hausdorff distance in the canonical frame correctly identify all targets. The character `R` is not detected by Hausdorff distance with a low threshold due to the noisy protuberance caused by the simple gray level thresholding method employed (Fig. 22). With a lower rejection rate all true characters are correctly classified but we also get several false positives. In a realistic application most false positives can be easily removed by exploiting pose consistency and plate grammar.

The results are satisfactory for such low resolution images, as nothing was assumed about location, size, orientation, or slant of the observed shapes.

5.7. Out-of-the-box preprocessing

Finally we have evaluated the canonicalization method as a generic preprocessing step for statistical classification techniques. The goal is to get affine invariance for free,

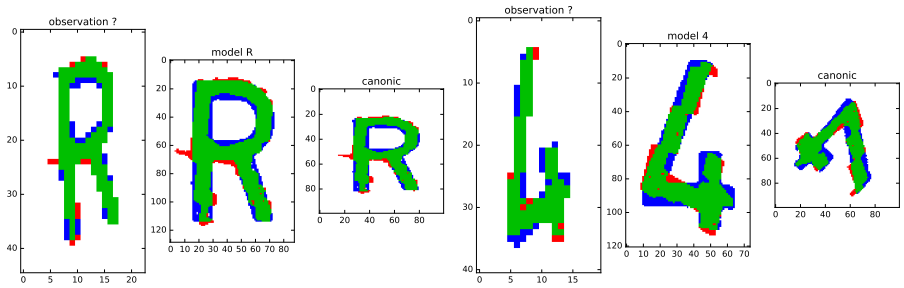

Figure 22: Examples of shape alignment in the plate recognition experiment.

without the need of augmenting the training images with a large number of representative (real or synthetic) rotated, scaled, and tilted views. Specifically, we have studied the results of our method as a direct replacement of the isotropic size-normalization and centering preprocessing step in the MNIST handwritten digit benchmark [48]. We have canonicalized the whole dataset and, for maximum simplicity, we use a very small ϵ_s to get just a single canonical version per digit (Fig. 23). This is risky, as certain digits have some orientation ambiguity, but we let the classifier deal with this variability.

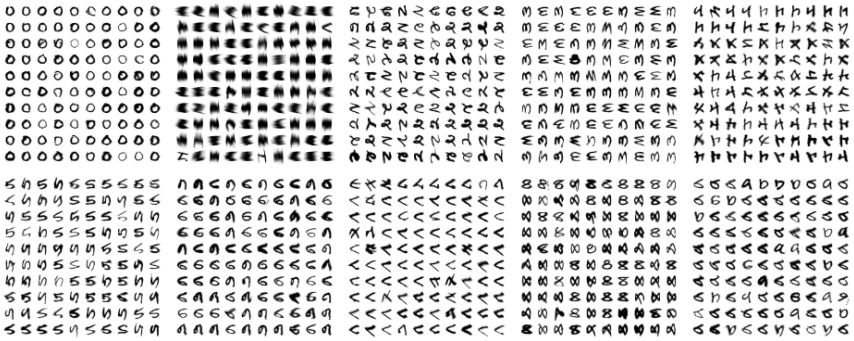

Figure 23: Samples of the canonicalized MNIST dataset

In a first experiment we have trained a full-covariance Gaussian classifier working on the 40 principal components of the global population. This machine has a high accuracy and speed relative to simplicity. For the original MNIST dataset we get 96.25%

accuracy almost instantaneously. For evaluation of the alternative canonical preprocessing step we merge classes 6 and 9 as they are very similar to rotated versions of each other, and the alignment transformation can be used later to distinguish them. In these conditions we get 93.4% accuracy.

This can be improved using more expressive classifiers like Support Vector Machines or Artificial Neural Networks, at the cost of longer training times. For example, a deep convolutional network with two convolution layers of 32 and 64 filters and a full connected layer with 1024 elements can be easily trained (using tensorflow or any similar system running on a modern GPU) to get 97.0% accuracy on the ten classes (98.2% if we merge 6 and 9), and the following confusion matrix:

	0	1	2	3	4	5	6	7	8	9
0	976	0	0	0	1	1	1	0	0	1
1	0	1125	1	0	2	0	0	7	0	0
2	3	1	1016	0	1	0	0	8	3	0
3	3	0	1	996	2	2	0	2	3	1
4	0	1	1	3	968	2	1	1	2	3
5	1	0	0	3	7	868	3	1	2	7
6	2	0	0	1	3	4	902	16	1	29
7	2	2	13	2	2	3	0	1004	0	0
8	1	0	2	5	4	0	2	0	958	2
9	4	1	0	0	7	14	80	1	12	890

The same network architecture working with the standard preprocessing step obtains 99.2% accuracy. This small performance degradation is acceptable to provide affine invariance, which is conveniently achieved by plugging a generic preprocessing module. Fig. 24 shows a demonstration of digit classification in a real image.

For comparison with a pure data-driven approach, we trained a Spatial Transformer Network [31] to directly classifiy weak perspective views (30° tilt and 40° FOV) of the handwritten characters. We augmented the MNIST dataset with 96 synthetic versions of each sample, corresponding to 8 rotations $\times$ 12 tilt orientations. The $60\,000$ original training samples were extended to $5\,820\,000$ instances, occupying 17GB of single precision float numbers. We randomly selected $10\,000$ for testing from the $970\,000$ examples in the extended test set. The network architecture consists of an initial affine transformer followed by two convolutional layers (with 32 filters of size 3×3, stride 2×2, no maxpool) and two fully connected layers (with $1\,024$ and 10 elements). The optimization method was Adam with learning rate 0.001 and 0.2 dropout rate. The

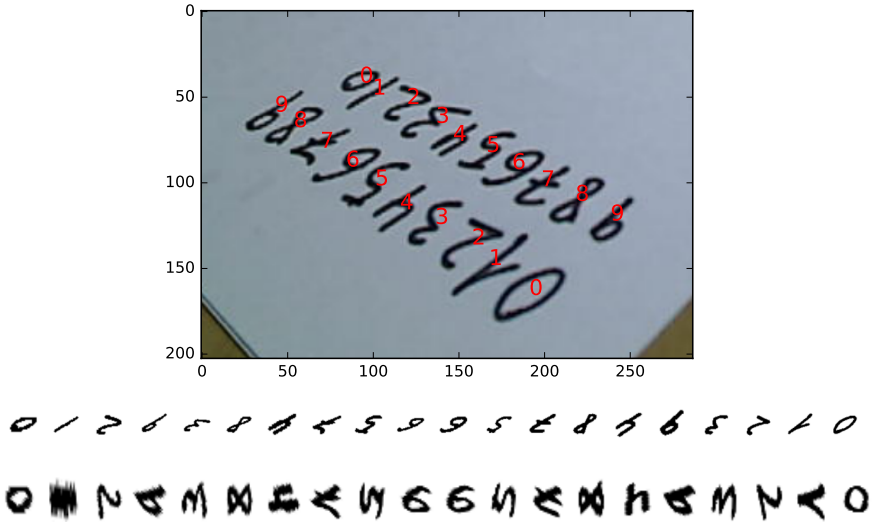

Figure 24: Classification results for a set of handwritten digits taken in a perspective view. The bottom rows show the original and canonicalized shapes.

accuracy of the best result obtained after 36 epochs ($\sim$ 4 hours of GPU time using NVIDIA GeForce GTX 1080) was 0.967. Such long learning time and large storage requirements clearly indicate that a purely empirical approach to normalization is not competitive with the proposed closed-form canonicalization method.

5.8. Discussion

The goal of these experiments is not to break any record for character recognition but to study the stability, efficiency, and practical utility of the proposed canonicalization method. It is also not possible to extract definitive conclusions from particular case studies, as the relative advantages of the different methods depend on the peculiarities of each problem. In any case, our experiments provide strong evidence that, for natural images of rigid shapes, the proposed canonicalization method comfortably tolerates perspective views of the target images occupying the equivalent to 30° FOV with tilt angles up to 45°. The breaking point can be roughly defined by the thumb rule TILT $\times$ FOV > 1500 squared degree. For other kind of nonlinear deformations, such as those

caused by handwriting, the method is useful to automatically achieve affine invariance in statistical classifiers.

In general no discrimination loss is introduced by the method in addition to that produced by the possible increase of intrinsic class overlapping produced by the expansion of the domain to the whole affine group.

5.9. Reproducible research

The code developed for this work is available online as an open source package, together with several illustrative *jupyter* notebooks[5].

6. Conclusions

In this paper we have presented an efficient and stable affine planar canonicalization method based on the properties of the third and fourth-order whitened moments. Previous methods used iterative optimization, obtained multiple canonical solutions, and did not fully exploit the rich geometric information that can be inferred from the moments. Our method is based on the analysis of the joint curve of marginal skewness and kurtosis for each direction (the so-called SK *signature*), which can be efficiently obtained in closed form from the raw moments. This signature is very sensitive to anisotropy, and can be used to define a single canonical version of the figure. The main contributions of this work are the following:

1. A closed-form algorithm for the relative extremes of the marginal moments $K(\theta)$ and $S(\theta)$, and for the average $\overline{\mathcal{S}}$. The orientation ambiguity of the multiple extrema of K is resolved by the direction of $\overline{\mathcal{S}}$.
2. A selection method of multiple canonical versions for weakly orientable shapes based on simple morphological features of the SK signature. Using safe selection parameters we obtain a significant speedup in registration time versus considering all four minima of K. The SK descriptor can also be used to reject wrong candidate matchings quickly.

[5] `http://dis.um.es/~alberto/canonic.html`

3. A comprehensive set of experiments to evaluate the robustness of the method to non-affine distortion caused by natural imaging conditions. The canonical versions are stable and can be used for shape classification under perspective views occupying the equivalent to 30^o FOV with tilt angles up to 45^o.
4. A study of the registration error quality in different reference frames. We have found that alignment in the canonical frame, the most efficient, is in practice as good as the theoretically optimal error measured in the physical sensor frame.

We have also demonstrated that canonicalization is a general and practical preprocessing step to achieve affine invariance in statistical pattern recognition. This is in accordance with the design principle that machine learning should be dedicated to capture variability that cannot be explained by a theoretical model.

Future work includes the extension of this approach to canonicalization of 3D regions, improved robustness to perspective distortion, automatic segmentation of noisy regions with ill-defined boundaries, automatic parameter tuning, and extension to textured and colored images.

Acknowledgements

This work was supported by the Spanish MINECO, as well as European Commission FEDER funds, under grant TIN2015-66972-C5-3-R. The authors are grateful for the constructive and valuable comments from the reviewers.

References

[1] M. K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Info. Theory IT-8 (1962) 179–187.

[2] C. T. Zahn, R. Z. Roskies, Fourier descriptors for plane closed curves, IEEE Transactions on computers 100 (3) (1972) 269–281.

[3] M. A. Rodrigues (Ed.), Invariants for pattern recognition and classification, World Scientific, 2000.

[4] J. Flusser, T. Suk, Pattern recognition by affine moment invariants, Pattern Recognition 26 (1) (1993) 167–174.

[5] J. Flusser, T. Suk, B. Zitov, Moments and Moment Invariants in Pattern Recognition, Wiley, 2009.

[6] D. Zhang, G. Lu, Review of shape representation and description techniques, Pattern Recognition 37 (1) (2004) 1–19.

[7] I. Dryden, K. Mardia, Statistical shape analysis, John Wiley & Sons, Ltd., 1998.

[8] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE TPAMI 24 (4) (2002) 509–522.

[9] S. Baker, I. Matthews, Lucas-kanade 20 years on: A unifying framework, International Journal of Computer Vision 56 (3) (2004) 221–255.

[10] R. C. Veltkamp, Shape matching: similarity measures and algorithms, in: Shape Modeling and Applications, SMI 2001 International Conference on., IEEE, 2001, pp. 188–197.

[11] X. Huang, N. Paragios, D. N. Metaxas, Shape registration in implicit spaces using information theory and free form deformations, IEEE TPAMI 28 (8) (2006) 1303–1318.

[12] J. S. Marques, A. J. Abrantes, Shape alignment optimal initial point and pose estimation, Pattern Recognition Letters 18 (1) (1997) 49 – 53.

[13] N. Paragios, M. Rousson, V. Ramesh, Matching distance functions: a shape-to-area variational approach for global-to-local registration, in: European Conference on Computer Vision, Springer, 2002, pp. 775–789.

[14] D. G. Kendall, A survey of the statistical theory of shape, Statistical Science 4 (2) (1989) pp. 87–99.

[15] D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge, Comparing images using the Hausdorff distance, IEEE TPAMI 15 (9) (1993) 850–863.

[16] A. Amor-Martinez, A. Ruiz, F. Moreno-Noguer, A. Sanfeliu, On-board real-time pose estimation for uavs using deformable visual contour registration, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2014, pp. 2595–2601.

[17] J. Nemeth, C. Domokos, Z. Kato, Recovering planar homographies between 2d shapes, in: Proc ICCV, 2009.

[18] A. Ruiz, P. E. López de Teruel, L. Fernández, Robust homography estimation from planar contours based on convexity, in: ECCV'06, 2006, pp. 107–120.

[19] P. K. Jain, Homography estimation from planar contours, in: 3D Data Processing, Visualization, and Transmission, Third International Symposium on, 2006, pp. 877–884.

[20] M. Zhai, S. Fu, Z. Jing, Homography estimation from planar contours in image sequence, Optical Engineering 49 (3) (2010) 037202–037202.

[21] S. Kuthirrummal, C. V. Jawahar, P. J. Narayanan, Fourier domain representation of planar curves for recognition in multiple views, Pattern Recognition 37 (2004) 739–754.

[22] T. F. Cootes, C. J. Taylor, et al., Statistical models of appearance for computer vision (2004).

[23] A. Blake, M. Isard, Active Contours, Springer, 1998.

[24] C. Small, The Statistical Theory of Shape, Springer-Verlag, Inc., 1996.

[25] D. Bryner, A. Srivastava, E. Klassen, Affine-invariant, elastic shape analysis of planar contours, in: Proceedings of the CVPR, 2012.

[26] E. G. Learned-Miller, Data driven image models through continuous joint alignment, IEEE TPAMI 28 (2) (2006) 236–250.

[27] G. B. Huang, V. Jain, E. Learned-Miller, Unsupervised joint alignment of complex images, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[28] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[29] J. L. Long, N. Zhang, T. Darrell, Do convnets learn correspondence?, in: Advances in Neural Information Processing Systems, 2014, pp. 1601–1609.

[30] G. Huang, M. Mattar, H. Lee, E. G. Learned-Miller, Learning to align from scratch, in: Advances in Neural Information Processing Systems, 2012, pp. 764–772.

[31] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.

[32] B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, Human-level concept learning through probabilistic program induction, Science 350 (6266) (2015) 1332–1338.

[33] T. Collins, A. Bartoli, Infinitesimal plane-based pose estimation, International Journal of Computer Vision 109 (3) (2014) 252–286.

[34] E. Begelfor, M. Werman, Affine invariance revisited, in: CVPR'06, CVPR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 2087–2094.

[35] J. Sprinzak, M. Werman, Affine point matching, Pattern Recognition Letters 15 (4) (1994) 337–339.

[36] Š. Obdržálek, J. Matas, Object recognition using local affine frames on maximally stable extremal regions, in: Toward Category-Level Object Recognition, Springer, 2006, pp. 83–104.

[37] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. V. Gool, A comparison of affine region detectors, International Journal of Computer Vision 65 (1) (2005) 43–72.

[38] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural networks 13 (4) (2000) 411–430.

[39] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE transactions on Neural Networks 10 (3) (1999) 626–634.

[40] V. Zarzoso, P. Comon, Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size, IEEE Transactions on Neural Networks 21 (2) (2010) 248–261.

[41] Y. Mei, D. Androutsos, Robust affine invariant shape image retrieval using the ICA Zernike moment shape descriptor, in: ICIP09, 2009, pp. 1065–1068.

[42] Y. Mei, D. Androutsos, Affine invariant shape descriptors: The ICA-Fourier descriptor and the PCA-Fourier descriptor, in: ICPR08, 2008, pp. 1–4.

[43] X. Huang, B. Wang, L. Zhang, A new scheme for extraction of affine invariant descriptor and affine motion estimation based on independent component analysis, Pattern Recognition Letters 26 (9) (2005) 1244 – 1255.

[44] L. Zhang, X. Huang, Applications for affine invariant descriptor and affine parameter estimation based on two-source ICA, Journal of Mathematical Modelling and Algorithms 5 (2006) 505–523.

[45] J. Heikkilä, Pattern matching with affine moment descriptors, Pattern Recognition 37 (9) (2004) 1825–1834.

[46] Y. Mei, Robust affine invariant shape descriptors, PhD dissertation, Ryerson University, Library and Archives Canada, 2010.

[47] A. Rosenfeld, J. L. Pfaltz, Distance functions on digital pictures, Pattern Recognition 1 (1) (1968) 33–61.

[48] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

Appendix I: Whitened moments

The raw moments of a raster image can be directly computed by contraction with a precomputed auxiliary multidimensional array containing values $x^p y^q$. In principle, we can obtain μ_{pq} by the same method from an intermediate whitened image. However, this method has several drawbacks: the required auxiliary warping is costly and introduces some accuracy loss, and, in noisy scenes we must be careful not to distort noise distribution. It is therefore better to obtain μ_{pq} in closed form from m_{pq}. Note that some standard software packages like OpenCV usually compute normalized image moments only up to order three and do not enforce the whitening condition $\mu_{11} = 0$.

We assume that the image has been normalized to $m_{00} = 1$. The whitening transformation can be expressed as

$$\begin{aligned} x' &= ax + by + e \\ y' &= cy + f \end{aligned} \tag{21}$$

where

$$\sigma_{xx} = m_{20} - m_{10}^2, \quad \sigma_{yy} = m_{02} - m_{01}^2, \quad \sigma_{xy} = m_{11} - m_{10}m_{01},$$
$$a = \sqrt{\frac{\sigma_{yy}}{\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2}}, \quad b = -a\frac{\sigma_{xy}}{\sigma_{yy}}, \quad c = \frac{1}{\sqrt{\sigma_{yy}}}, \tag{22}$$
$$e = -am_{10} - bm_{01}, \quad f = -cm_{01}.$$

Therefore

$$\mu_{pq} = E_S\{(ax + by + e)^p (cy + d)^q\}. \tag{23}$$

This involves the product of a trinomial and a binomial expansion:

$$\begin{aligned} \mu_{pq} &= \sum_{(i+j+k=p)} \sum_{(l+n=q)} \binom{p}{i,j,k} \binom{q}{l,n} E_S\{(ax)^i (by)^j (e)^k (cy)^l (d)^n\} \\ &= \sum_{(i+j+k=p)} \sum_{(l+n=q)} \binom{p}{i,j,k} \binom{q}{l,n} a^i b^j e^k c^l d^n m_{i,j+l}. \end{aligned} \tag{24}$$

The number of terms $(p+2)(p+1)(q+1)/2$ grows fast[6] but is not large for our purposes (e.g. μ_{40} has 15 terms, μ_{31} has 20 terms, and μ_{22} has 18 terms), and of course negligible in comparison with the cost of image warping.

(In practice it is better to express the trinomial as two binomial steps, sharing five auxiliary (e.g. centered) moments, so that the total number of terms for the third and fourth order can be reduced from 132 to 105.)

Appendix II: Contour representation

The raw moments can be efficiently obtained from the boundary of a planar figure using Green's Theorem:

$$m_{pq} = \iint_{\mathcal{S}} x^p y^q dxdy = \frac{1}{2} \oint_Z \left[\frac{x^p y^{q+1}}{q+1} dx - \frac{x^{p+1} y^q}{p+1} dy \right]. \tag{25}$$

Contours extracted from digital images are represented by sequences of points $Z = \{z_0, z_1, \ldots z_{n-1}\}$, where $z_k = (x_k, y_k)$. For notational convenience we add a closing

[6] $(p+1)(p+2)/2$ is the number of elements of the base of a Pascal pyramid of depth p, and $q+1$ is the number of elements of the Pascal triangle of depth q.

vertex $z_n = z_0$. In a piecewise linear approximation, the segment from z_k to z_{k+1} is parameterized as $z(t) = z_k + (z_{k+1} - z_k)t$, where $t \in (0, 1)$. Then the contour integral for a raw moment m_{pq} can be decomposed as

$$m_{pq} = \frac{1}{2} \sum_{k=0}^{n-1} \int_{z_k}^{z_{k+1}} \frac{x(t)^p y(t)^{q+1}}{q+1} \frac{dx}{dt} dt - \frac{x(t)^{p+1} y(t)^q}{p+1} \frac{dy}{dt} dt = \sum_{k=0}^{n-1} C_k, \qquad (26)$$

where each segment contribution C_k can be expressed as

$$C_k = \frac{u_k P_{p,q+1}(x_k, u_k, y_k, v_k)}{2(q+1)} - \frac{v_k P_{p+1,q}(x_k, u_k, y_k, v_k)}{2(p+1)} \qquad (27)$$

in terms of

$$\begin{aligned} P_{n,m}(x, u, y, v) &\equiv \int_0^1 (x + ut)^n (y + vt)^m dt = \\ &= \sum_{j=0}^{n} \sum_{k=0}^{m} \binom{n}{j} \binom{m}{k} \frac{x^j u^{n-j} y^k v^{m-k}}{n + m - j - k + 1}, \end{aligned} \qquad (28)$$

where $u_k = x_{k+1} - x_k$ and $v_k = y_{k+1} - y_k$.

Moment m_{pq} contains $(p+1)(q+2)+(p+2)(q+1)$ terms per node. For example, m_{40} contains 16 terms, m_{31} contains 22 terms, and m_{22} contains 24 terms.

Appendix III: Modified Hausdorff distance

As shown in Fig. 11, the alignment error measured in the common canonical frame suffers anisotropic deformations which may degrade classification accuracy. For some similarity functions it is possible to keep information about the original metrics so that the 'canonical' error can be converted back to physical sensor units without the need of any additional warping. For example, for shape similarity based on symmetric difference we just need to correct the weight of the area elements using the constant Jacobian of the affine canonicalization transformations.

Hausdorff distance is more complex, as the distance transform of every model must be evaluated in (or warped to) the input frame. This is costly for on-line classification but a fast acceptable approximation can be achieved as follows. For every shape R (model or target) we precompute the canonicalization transformation C_R, the distance

transform D_R, the canonical shape $C_R R$, and the warped distance $C_R D_R$ (Fig. 25). (For binary regions a single warp is needed, as the original shape can be recovered from the distance transform by thresholding.)

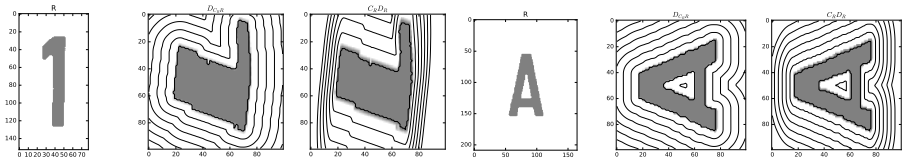

Figure 25: Comparison of distance in the canonical frame $D_{C_R R}$, and the canonicalized original distance $C_R D_R$.

Let $T = C_A^{-1} C_M$ the alignment transformation of model M to target A. By abuse of notation, symbols A and M denote both the regions and the corresponding indicator functions, and warping is expressed by juxtaposition. The operator $\odot$ denotes the Hadamard product (element by element multiplication). With these conventions the Hausdorff distance between TM and A can be written as

$$\begin{aligned} d_H(A, TM) &= \max(\max_{x \in A} \min_{y \in TM} d(x, y), \max_{x \in TM} \min_{y \in A} d(x, y)) \\ &= \max[A \odot D_{TM} | TM \odot D_A]. \end{aligned} \tag{29}$$

The block $TM \odot D_A$ can be evaluated in the canonical frame for every target-model pair from precomputed information. D_A is computed once and $C_A D_A$ is valid to be checked against all the model set.

$$\begin{aligned} \max(TM \odot D_A) &= \max(C_A TM \odot C_A D_A) = \\ &= \max(C_M M \odot C_A D_A). \end{aligned} \tag{30}$$

Unfortunately, $A \odot D_{TM}$ cannot be expressed as element-wise products of precomputed items. A costly specific warp and distance transform is needed for each target A and model M. However, an efficient 'symmetric' variant of Hausdorff distance can be explored in which the block $A \odot D_{TM}$ is replaced by the analogous term $C_A A \odot C_M D_M$:

$$d_{H'}(A, B) = \max[C_A A \odot C_M D_M | C_M M \odot C_A D_A]. \quad (31)$$

Both parts have physical meaning: the right part of the distance is evaluated in the original frame of the target image and the left part is evaluated in the frame of the model. The distance transforms must be normalized to figure size (e.g., using λ_1) for a common measurement unit independent of image resolution.

For classification purposes our experiments do not show any conclusive difference with respect to the true distance. And, perhaps surprisingly, the Hausdorff distance in the canonical frame is no worse than the distance measured in the physical sensor frame.