Part-Aware Descriptor Classifier for Trustworthy Species Detection

Supplementary Material

0.1. Shape Descriptors (ϕ_s)

For each anatomical mask (e.g., head, thorax, abdomen), a set of shape descriptors is computed to capture morphological and geometric properties of the segmented region. Let $\mathcal{R}$ denote a connected region corresponding to a binary mask M for a single part (we ignore writing $\mathcal{R}_i$ for i^{th} part for better readability). We compute the $(p,q)^{th}$ raw moments $m_{pq} = \sum_x \sum_y x^p y^q M(x,y)$ where $M(x,y) = 1$ if pixel (x,y) belongs to the region $\mathcal{R}$, and 0 otherwise. We compute the central moments $\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q M(x,y)$ where $(\bar{x},\bar{y})$ is the centroid. are invariant to the position of the object in the image. The second-order central moments include variance along the x-axis(μ_{20}), variance along the y-axis (μ_{02}), and covariance between x and y (μ_{11}). These three moments form the covariance matrix of the region:

$$\Sigma = \begin{bmatrix} \mu_{20} & \mu_{11} \\ \mu_{11} & \mu_{02} \end{bmatrix} \tag{1}$$

The eigenvalues λ_1 and λ_2 of Σ correspond to the squared lengths of the semi-major and semi-minor axes of the best-fitting ellipse:

$$\lambda_{1,2} = \frac{\mu_{20}+\mu_{02}}{2} \pm \frac{1}{2}\sqrt{4\mu_{11}^2 + (\mu_{20}-\mu_{02})^2} \tag{2}$$

From the second-order moments of the fitted ellipse, the lengths of the major and minor axes are computed as:

$$L_{\text{major}} = 2\sqrt{2}\sqrt{\frac{\mu_{20}+\mu_{02}+\sqrt{4\mu_{11}^2+(\mu_{20}-\mu_{02})^2}}{n}} \tag{3}$$

$$L_{\text{minor}} = 2\sqrt{2}\sqrt{\frac{\mu_{20}+\mu_{02}-\sqrt{4\mu_{11}^2+(\mu_{20}-\mu_{02})^2}}{n}} \tag{4}$$

where n is the number of pixels in $\mathcal{R}$. To achieve scale invariance, central moments are normalized as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}, \quad \text{where} \quad \gamma = 1 + \frac{p+q}{2} \tag{5}$$

These *normalized central moments* are invariant to translation and scaling, forming the basis of more complex invariant descriptors. These moments serve as the foundation for several derived shape descriptors, such as:

- **Area** (A): Total number of pixels within the region, $A = \sum_{x,y\in\mathcal{R}} M(x,y)$.
- **Perimeter** (P): Total boundary length of the region, $P = \sum_{x,y\in\partial\mathcal{R}} M(x,y)$ where $\partial\mathcal{R}$ denotes the contour of the region $\mathcal{R}$.
- **Aspect Ratio** (AR): Ratio between the lengths of the major and minor axes, AR $= \frac{L_{\text{major}}}{L_{\text{minor}}}$.
- **Extent (E):** Ratio between the region area and its bounding box area, $E = \frac{A}{A_{\text{bbox}}}$, where A_{bbox}corresponds to the area of the minimal bounding rectangle of the object.
- **Solidity** (S): Ratio of region area to its convex hull area, $S = \frac{A}{A_{\text{convex}}}$ where A_{convex} corresponds to the area of the convex hull of the object.
- **Eccentricity** (e): Elongation based on the second central moments $e = \sqrt{1-\frac{\lambda_2}{\lambda_1}}$ where $\lambda_1 \geq \lambda_2$ are eigenvalues of the covariance matrix Σ.
- **Orientation** (θ): The dominant orientation of the shape, given by the angle between the x-axis and the major axis of the fitted ellipse:

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right).$$

- **Circularity** (C): How close the shape is to a circle, $C = \frac{4\pi A}{P^2}$.
- **Compactness** (C_m): Inverse of the normalized second moment of the shape: $C_m = \frac{A}{\pi r_{\max}^2}$.
- **Elongation** (E): Ratio of major to minor axis lengths of the fitted ellipse: $E = \frac{L_{\text{major}}}{L_{\text{minor}}}$.
- **Hu Moments** (Φ_i): Using the normalized central moments, Hu [2] proposed seven higher-order combinations $\{\Phi_1, \ldots, \Phi_7\}$ that are invariant to translation, scale, and rotation. Hu moments capture higher-order asymmetries and structural irregularities of bee body parts, providing complementary global shape cues beyond simple geometric descriptors.
- **Zernike Moments** [7]: Zernike moments provide an orthogonal basis for representing the shape of a region within the unit disk. For each binary mask M_i for the i^{th} anatomical part, Zernike polynomials up to order $n_{\max} = 12$ are computed, yielding 25 complex moments A_{nm}. The magnitudes $|A_{nm}|$ form a 25-dimensional descriptor vector $z_i \in \mathbb{R}^{25}$, which is invariant to translation, rotation, and scaling, and effectively captures smooth global shape variations.
- **SIFT** [3]: The scale-invariant feature transform (SIFT) generates keypoints and 128-dimensional descriptors that capture local gradient patterns and distinctive keypoints in the body part region. The raw set of SIFT descriptors is used directly without quantization into a visual vocabulary.
- **Fourier Descriptors (FD)** [9]: To represent contour-based shape characteristics, the boundary of each body part is sampled and expressed as a complex sequence

$c(k) = x(k) + j\,y(k)$. The one-dimensional discrete Fourier transform (DFT) is applied to obtain frequency-domain coefficients. The first 20 non-DC coefficients are retained after normalization by the magnitude of the first harmonic, ensuring invariance to scale and rotation:

$$f_i^{\text{FD}} = \left[\frac{|c(2)|}{|c(1)|}, \frac{|c(3)|}{|c(1)|}, \ldots, \frac{|c(21)|}{|c(1)|}\right]^T \in \mathbb{R}^{20}.$$

Fourier descriptors effectively encode boundary smoothness and periodic deformations of the segmented anatomical structures.

0.2. Appearance and Quality Descriptors

We compute a set of appearance and no-reference image quality descriptors that capture perceptual distortions and visual fidelity of each segmented body part crop:

- **Brightness (B)**: The mean luminance of the grayscale body part crop ($I_{\text{gray}} \odot M_i$), computed as:

$$B = \frac{1}{|\mathcal{R}_i|} \sum_{(x,y)\in\mathcal{R}_i} I_{\text{gray}}(x, y),$$

  where $\mathcal{R}_i$ denotes the set of pixels within the i-th region.
- **Contrast (C)**: The local contrast is measured as the standard deviation of luminance values within the region:

$$C = \sqrt{\frac{1}{|\mathcal{R}_i|} \sum_{(x,y)\in\mathcal{R}_i} (I_{\text{gray}}(x, y) - B)^2}.$$

- **Sharpness (S)**: Sharpness is estimated using the variance of the Laplacian of the grayscale image:

$$S = \text{Var}(\nabla^2 I_{\text{gray}}),$$

  where ∇^2 denotes the Laplacian operator.
- **Colorfulness** [1]: Let R, G, B be the matrix of the red, green, and blue channels of an image I, respectively. The colorfulness score is computed as,

$$F = \sqrt{SD(\alpha_1)^2 + SD(\alpha_2)^2} + 0.3\sqrt{\bar{\alpha_1}^2 + \bar{\alpha_2}^2}$$

  where $\alpha_1 = R - G$, $\alpha_2 = \frac{1}{2}(R + G) - B$ and SD denotes the standard deviation.
- **Entropy** [6]: Texture complexity is measured using Shannon entropy, $H = -\sum_k p_k \log p_k$ where p_k is the probability of grayscale intensity level k within $\mathcal{R}_i$.
- **BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator)** [4]: It is computed directly on the masked crop without resizing. It models natural scene statistics using Mean Subtracted Contrast Normalized (MSCN) coefficients and pairwise products to quantify deviations from natural image regularities. The resulting scalar score reflects perceptual quality degradation such as blur, noise, and compression artifacts, where lower values indicate higher visual quality.
- **NIQE (Natural Image Quality Evaluator)** [5]: NIQE measures image naturalness by comparing statistical features of the crop against a multivariate Gaussian model derived from high-quality natural images. It produces a scalar score $n_i \in \mathbb{R}$, with lower values indicating more natural, undistorted appearances.
- **PIQE (Perception-based Image Quality Evaluator)** [8]: PIQE estimates perceptual degradation by analyzing block-wise distortions and computing a perceptual distortion mask. The final scalar score $p_i \in \mathbb{R}$ reflects visual quality degradation, where smaller values correspond to higher perceptual quality.

0.3. Inter-part Descriptors

We also measure inter-part ratios to express relative body proportions across k body parts. For each pair of parts (i, j), the inter-part ratio set:

$$\mathbf{r}_c = \phi_c(M) = \{r_{ij}, \hat{r}_i \mid i, j \in \{1, \ldots, k\},\, i \neq j\} \quad (6)$$

where $r_{ij} = \frac{A_i}{A_j}$ is the ratio of their respective areas and $\hat{r}_i = \frac{A_i}{\sum_{m=1}^{k} A_m}$ is the ratio with respect to the total body area. For n parts (including full-body), we have $(n-1)$ part-to-total and $\binom{n-1}{2}$ part-to-part ratios. Hence, the total number of inter-part descriptors is, $n_c = (n-1) + \binom{n-1}{2}$. These ratios provide a comprehensive representation of relative proportions and scaling relationships across all body parts, extending the morphometric feature vector that captures size, shape, elongation, convexity, and inter-part proportionality.

0.4. Total number of descriptors.

Let d_i be the total number of shape and visual features for the i^{th} part. For n parts (including full-body), the total number of descriptors for the full mask M is,

$$d_M = nd_i + \binom{n-1}{2} + n - 1$$

For each part mask M_i, we have 10 basic shape features ("area", "perimeter", "aspect_ratio", "extent", "solidity", "eccentricity", "orientation", "circularity", "elongation", "compactness"), 130 SIFT features (number of keypoints, size of the keypoints, and 128 SIFT descriptors), 33 ORB features (number of keypoints and 32 ORB descriptors), 7 Hu moments, 25 Zernike moments (for order 8), 20 fourier descriptors, and 8 visual features (brightness, contrast, sharpness, colorfulness, entropym BRISQUE, NIQE, and PIQE). Therefore, the total number of descriptors for each part i is, $10 + 130 + 33 + 7 + 25 + 30 + 8 = 233$.

For the Beemachine dataset, we have three parts and one full-body mask. Therefore, the total number of descriptors is $233 \times 4 + \binom{4-1}{2} + 4 - 1 = 938$. For the CUB dataset, we have 11 parts and one full-body mask. Therefore, the total

number of descriptors is $233 \times 12 + \binom{12-1}{2} + 12 - 1 = 2862$. For the Fish-Vista dataset, we have 9 parts and one full-body mask. Therefore, the total number of descriptors is $233 \times 10 + \binom{10-1}{2} + 10 - 1 = 2375$.

References

[1] David Hasler and Sabine E. Süsstrunk. Measuring colourfulness in natural images. In *Proceedings of the IS&T/SPIE Electronic Imaging: Human Vision and Electronic Imaging VIII*, pages 87–95. SPIE, 2003. 2

[2] M.K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962. 1

[3] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 1

[4] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012. 2

[5] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. In *IEEE Signal Processing Letters*, pages 209–212. IEEE, 2012. 2

[6] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. 2

[7] Michael R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 1980. 1

[8] Nagesh Venkatanath, R. Krishna Babu, and K. Srinivasa Reddy. Blind image quality evaluation using perception-based image quality evaluator (piqe). In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3579–3583. IEEE, 2015. 2

[9] Charles T. Zahn and Ralph Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, 1972. 1