METRIC LEARNING BASED AUTOMATIC SEGMENTATION OF PATTERNED SPECIES

Ankita Shukla, Saket Anand

IIIT-Delhi
New Delhi, India

ABSTRACT

Many species in the wild exhibit a visual pattern that can be used to uniquely identify an individual. This observation has recently led to *visual animal biometrics* become a rapidly growing application area of computer vision. Customized software tools for animal biometrics already employ vision based techniques to recognize individuals in images taken in uncontrolled environments. However, most existing tools require the user to localize the animals for accurate identification. In this work, we propose a figure/ground segmentation method that automatically extracts out the animal in an image. Our method relies on a semi-supervised metric learning algorithm that uses a small amount of training data without compromising generalization performance. We design a simple pipeline comprising of superpixel segmentation, texture based feature extraction followed by mean shift clustering using the learned metric. We show that our approach can yield competitive results for figure/ground segmentation of patterned animals in images taken in the wild, often under extreme illumination conditions.

***Index Terms*—** Metric learning, figure/ground segmentation, patterned species

1. INTRODUCTION

Visual animal biometrics [1, 2, 3] is an upcoming and important application area of computer vision. Many species exhibit a visual pattern, which can be used to identify individuals. For instance tigers can be identified uniquely by their stripes [2] and leopards can be identified by their rosettes. Biologists prefer such non-invasive methods for identification as these lead to minimal stress for the animals, while reducing cost and risks of field trips, animal capture and mounting/dismounting of tracking devices.

For endangered species like tigers, identification of individuals from camera trap images are crucial for monitoring and tracking population, localization of an individual and forensics to control poaching. Customized software tools exist that perform identification of individuals [2], however, they require intensive interaction to achieve reasonable accuracy. For each query image, manual effort is needed to mark several keypoints around the animal's silhouette, which is tedious considering several thousand query images are generated during a typical population census drive.

Fig. 1: Example camera trap images of tigers depicting illumination variations.

Segmenting out the animal's silhouette can be done by figure-ground segmentation, for which several techniques exist, both in unsupervised and the (semi-)supervised category. These techniques have been studied well as they play a key role in vision based tasks like object recognition [4]. The goal is to produce a binary segmentation separating the foreground object from its background. To cope with variability in images, many of the semi-supervised techniques take an interactive approach while others rely on classifiers trained on extensive training data.

Fig. 1 shows typical images from our camera trap tiger dataset, where the use of external lighting sources often result in similar color and texture properties for the tiger and ground regions. In case of camera trap images, unsupervised techniques typically fail due to background clutter, complex illumination effects, nonrigid variations in animal pose and occlusions. Supervised approaches use training data to learn classifiers that are robust to these variations. However, the performance is crucially dependent on the quantity of training data. Since large amounts of training data is expensive to collect, fully supervised methods are not suitable for this application.

In this work, we take a semi-supervised approach that relies on limited training data (1-2 images). We first generate a superpixel representation of the image, which allows efficient processing at query time. As our domain comprises images of patterned animals, we creating a feature space that captures the texture properties. The training data is used to learn a discriminative Mahalanobis distance based metric offline. At query time the learned metric is used with mean shift clustering [5] to perform an automatic figure-ground segmentation. Using a combination of mean shift and metric learning increases the robustness of our figure-ground segmentation pro-

Thanks to WII for providing camera trap images of tigers.

cess against clutter and illumination variations. Our method does not require any user input for the query image and relies only on a few training images for metric learning.

The paper is organized as follows. In Section 2, we briefly review the existing approaches for figure-ground segmentation. In Section 3, we discuss our preprocessing and metric learning strategy. We present our overall figure-ground segmentation technique in Section 4, followed by experimental results and comparisons for three different patterned species in Section 5. Finally, we conclude our work in Section 6.

2. LITERATURE REVIEW

Many figure-ground segmentation techniques work with an interactive framework [6, 7, 8, 9, 10, 11], which requires manual input from the user to guide the segmentation process. In [6], a bounding box around the object of interest is drawn to define the region outside the box as background while the part within the box is considered as a combination of foreground and background. An iterative algorithm based on graph cuts is used to obtain the final segmentation result. Since these methods use manual input in each query image, this approach becomes tedious when the query set is large.

Recently, Li et al. [10] proposed an interactive figure-ground segmentation approach by employing metric learning repeatedly in the feature space. The user provides scribbles indicating foreground and background pixels, which are then used to identify superpixels to be used as training data for the metric learning algorithm. Using the learned metric, previously unlabeled superpixels are added to the training set and the metric learning process is repeated. While empirical results in [10] are promising, the stopping criteria for iterative metric learning is not discussed. Such an approach may not converge, especially in case of camera trap images, which are often affected by high noise or extreme illumination as shown in Fig. 1.

Several techniques like [12, 4] avoid manual input by training a classifier like SVM on different low-level features extracted from training images to classify test image as figure/ground. In [4], a supervised approach is proposed for object recognition, which learns both geometric and appearance based priors to partition a superpixel graph using graph-cut based energy minimization. The authors take a two-step approach in [13], where first a set of overlapping windows in the query image are assigned a label based on their nearest neighbors in the training set. The second step minimizes an energy function using a graph-cut based approach to obtain the optimal segmentation. These methods perform well in terms of segmentation results, however, they rely heavily on large training sets for which several features are combined like local phase quantization (LPQ) texture feature and GIST features along with spatial information.

Unlike the techniques discussed above that train classifiers using large amounts of data, we use very less training data to learn a distance metric. Our mean shift clustering based approach leverages the learned metric for automatic figure/ground segmentation. While any metric learning algorithm can be introduced in our pipeline, we compare our segmentation results using two popular metric learning approaches [15, 16]. We also develop a modified version of a recent metric learning technique proposed in [17] and show its benefits for our application.

3. FEATURE EXTRACTION AND METRIC LEARNING

For segmenting camera trap images of patterned species, we rely mainly on the texture features. In this section, we describe our feature space representation followed by the metric learning approach.

3.1. Feature Extraction

All the images used for testing and training are first oversegmented using SLIC superpixels [18]. The texture features are extracted using a filter bank [19] of 48 filters at different scales and orientations. Thus, every pixel in the image has a 48-dimensional response vector. We use the labeled training images and map pixels from each class to this feature space and perform k-means clustering with $k = 20$. The cluster centers from *each class* are then concatenated for vector quantization of the feature space.

Each superpixel in the oversegmented image is represented by a histogram of texture features. The texture features corresponding to each constituent pixel is assigned to a histogram bin based on its closest cluster center. This texture histogram is normalized to have unit ℓ_1-norm and is our final feature vector.

3.2. Metric Learning Background

We represent figure and ground feature vectors by a set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, with $\mathbf{x}_i \in \mathbb{R}^n, \;\; i = 1, \ldots, m$. A Mahalanobis distance metric between two feature vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$, is given by

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} \qquad (1)$$

where, $\mathbf{M} \succeq 0$ is a symmetric positive semidefinite (PSD) matrix. The PSD constraint ensures that $d_{\mathbf{M}}(\cdot, \cdot)$ is a valid distance metric.

Mahalanobis metric learning algorithms search for a $\mathbf{M} \succeq 0$ that captures a linear transformation of the feature space such that similar points have small distances, while dissimilar points have large distances between them. This problem is formulated as an optimization problem with a semidefiniteness constraint on $\mathbf{M}$.

Large Margin Nearest Neighbor (LMNN) [16] takes a margin maximization approach and learns a low-rank $\mathbf{M}$ that is tuned for k-NN classification, and hence does not perform well with mean shift based segmentation. Information Theoretic Metric Learning (ITML) [15], on the other hand, optimizes a $\log\det$ based objective, which implicitly maintains the positive definiteness of $\mathbf{M}$. Since the $\log\det$ function does not permit a low-rank $\mathbf{M}$, the learned metric

turns out to be sensitive to noise and does not generalize well. Our empirical comparisons in Section 5 support this intuition.

3.3. Metric Learning Algorithm

In order to overcome the difficulties with LMNN and ITML, we use a modified version of a recent approach, Stiefel Manifold based Metric Learning (SMML) proposed in [17]. SMML parametrizes the rank-p matrix $\mathbf{M} = \mathbf{U}\mathbf{W}\mathbf{U}^\top$ using its diagonal eigenvalue matrix $\mathbf{W}_{p\times p}$ and orthonormal eigenvector matrix $\mathbf{U}_{n\times p}$. This parametrization permits learning a low-rank $\mathbf{M}$, by driving some of its eigenvalues to zero. Intuitively, the eigenvectors corresponding to nondiscriminative directions in the feature space are assigned zero eigenvalues, thus promoting discriminative metric learning and feature selection.

We denote the set of similar and dissimilar feature point pairs as $\mathcal{C}_s$ and $\mathcal{C}_d$ respectively. The similarity set $\mathcal{C}_s$ contains feature point pairs, both coming from the same class (foreground or background). The dissimilarity set $\mathcal{C}_d$ contains point pairs that come from different classes (one from foreground and one from background). The combined set $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_d$ is used to generate constraints for the metric learning problem.

We modify the metric learning formulation of [17] and formulate an unconstrained optimization problem. The two-term objective contains a hinge loss function to control distance constraint violations and a ℓ_2-norm regularizer to ensure smoothness. The formulation is given by

$$\min_{\mathbf{U}\in\mathcal{S}_{n,p},\mathbf{w}\in\mathbb{R}^p_+} \quad \sum_{i,j=1}^{m}\Big[y_{ij}(\mathbf{z}_{ij}^\top\mathbf{U}\mathrm{Diag}(\mathbf{w})\mathbf{U}^\top\mathbf{z}_{ij} - b_{ij})\Big]_+ + \alpha\left\|\mathbf{w}-\mathbf{w}_0\right\|_2^2 \tag{2}$$

Here, the term $[x]_+ = \max(0, x)$ is the hinge loss term that captures the degree of violation of constraints, and $\mathbf{w} = \mathrm{Diag}(\mathbf{W})$, with the initial value $\mathbf{w} = \mathbf{1}$ corresponding to $\mathbf{M} = \mathbf{I}$. The vectors $\mathbf{z}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ represent the difference vectors for the constraint pairs $(i, j) \in \mathcal{C}$ and b_{ij} are the corresponding target distances. The constants y_{ij} are indicator variables identifying similarity or dissimilarity pairs by respectively taking 1 or -1 and $\alpha > 0$ is a regularization parameter. The domains for $\mathbf{w}$ and $\mathbf{U}$ are $\mathbb{R}^p_+$, the nonnegative orthant of $\mathbb{R}^p$ and $\mathcal{S}_{n,p}$ as the space of all $n \times p$ matrices with orthonormal columns (a.k.a. Stiefel Manifold).

Due to space limitations, we are unable to describe the detailed algorithm for optimizing (2). We use an alternating strategy similar to [20]. Therefore, keeping $\mathbf{U}$ fixed, we solve the subproblem for $\mathbf{w}$

$$\min_{\mathbf{w}\in\mathbb{R}^p_+} \quad \sum_{i,j=1}^{m}\left[y_{ij}(\mathbf{a}_{ij}^\top\mathrm{Diag}(\mathbf{w})\mathbf{a}_{ij} - b_{ij})\right]_+ + \alpha\left\|\mathbf{w}-\mathbf{w}_0\right\|_2^2 \tag{3}$$

where we replace $\mathbf{U}^\top\mathbf{z}_{ij} = \mathbf{a}_{ij}$ for notational convenience. The objective function in (3), is sum of two terms and can be solved efficiently by using an ADMM based approach as in [20]. Next, we solve the following subproblem for $\mathbf{U}$

$$\min_{\mathbf{U}\in\mathcal{S}_{n,p}} \sum_{i,j=1}^{m}\Big[y_{ij}(\mathbf{z}_{ij}^\top\mathbf{U}\mathrm{Diag}(\mathbf{w})\mathbf{U}^\top\mathbf{z}_{ij} - b_{ij})\Big]_+ \tag{4}$$

For solving (4), we follow the approach of [17] to solve $\mathbf{U}$ as an instance of optimization on the Stiefel Manifold. The two steps (3) and (4) are repeated until convergence. $\widehat{\mathbf{M}} = \mathbf{U}\mathrm{Diag}(\mathbf{w})\mathbf{U}^\top$ is the learned metric.

(a) (b) (c) (d)

Fig. 2: Steps involved in Segmentation. (a) Test Image (b) Mean Shift Segmentation (different colors denote different clusters). (c) Distance Map (d) Segmentation Result after morphological operations on cluster with minimum distance

4. SEGMENTATION STRATEGY

In this section, we describe the steps involved in processing a query image. The oversegmentation and feature extraction is done as explained in Section 3.1.

4.1. Mean Shift Segmentation

The feature vectors representing the superpixels are concatenated with the centroid of the superpixel in image space. Mean shift clustering is used in this concatenated feature space, with the learned metric $\widehat{\mathbf{M}}$ used to compute distance between texture feature vectors, while the spatial distance is computed as the ℓ_2 distance. For all experiments, we use a variable bandwidth mean shift with the pointwise bandwidth as the 3-nearest neighbor distance. The step forms clusters of superpixels exhibiting texture and spatial similarity. An example of a test image after mean shift clustering is shown in Fig 2b.

4.2. Distance Map Generation

To identify the cluster that corresponds to patterned species, we generate a distance map by computing the average Mahalanobis distance of each cluster with the foreground feature vector extracted from the training images. An example of an inverted distance map for a test image is shown in Fig. 2c, where darker pixels have larger distances. The cluster with the least average distance is marked as the region containing animal while others are marked as background to generate a binary mask.

4.3. Morphological Operations

The cluster marked as animal is often effected by illuminated vegetation and clutter of some background superpixels. We perform two morphological operations on the binary image to obtain the final segmentation mask [21]. The binary image is first operated with a dilation operation then a connected component operation is performed on the dilated image. The largest connected component corresponds to the animal, while other connected components occur due to illuminated background superpixels and hence are discarded.

Method	Tiger (30 images)			Zebra (30 images)		
	Precision (%)	Recall (%)	Segmentation Accuracy(%)	Precision (%)	Recall (%)	Segmentation Accuracy(%)
GrabCut [6]	69.11	90.04	93.51	**98.10**	92.74	**96.43**
RW [22]	36.58	89.27	69.57	97.59	65.23	60.39
Graph Cut[23]	32.48	81.87	72.17	94	59.26	57.03
ITML [15]	30.71	64.57	65.01	64.92	92.77	66.76
LMNN [16]	51.75	67.93	72.19	50.03	70.93	73.14
Euclidean	26.37	33.98	71.51	75.86	80.95	78.61
ml-FigSeg	**78.04**	**93.49**	**93.61**	90.33	**93.72**	94.59

Table 1: Average Precision/Recall and Segmentation Accuracy on Tiger and Zebra Dataset (best in bold)

5. EXPERIMENTAL EVALUATION

We evaluate the performance of our approach on three patterned species: tiger, leopard and zebra and compare with other segmentation techniques : Graph cut [24], GrabCut [6] and Random Walker [22]. We compare the effectiveness of our learned distance metric with Euclidean distance as a baseline and metrics learned by two popular approaches ITML [15] and LMNN [16].

Datasets: We evaluate our metric learning based figure-ground segmentation approach (**ml-FigSeg**) on 30 tiger and 30 zebra [25] images. The leopard images for training and testing are collected from the web. The ground truths for all the images are created using interactive segmentation tool [1].

Feature Representation: Since, both zebras and tigers have characteristic pattern of stripes. We use texture features based on LM filter bank to represent our feature vectors as discussed in Section 3.1. In case of leopards, we follow the same feature extraction and representation while images are operated with a Gabor filter bank [2] instead.

Training Data and Metric learning: The distance metric learning for leopard and zebra is formulated as a two class problem: figure and ground. We use only one labeled image to extract 20 feature vectors from each class and generate similarity and dissimilarity pair constraints for metric learning. For tiger images, these feature vectors are selected from two labeled images and metric learning is formulated as a three class problem : figure, upper background and lower background. We divide background in two subcategories to handle the similarity between tiger and illuminated lower background.

Evaluation metric: We report average pixel-wise precision/recall and segmentation accuracy for tiger and zebra images in Table 1. Since a large database of leopard images was not publicly available, we only report qualitative results in Fig 3.

GrabCut [6] and Random Walker [22]: We use interactive tools for GrabCut [3] and Random Walker [4]. GrabCut is initialized by marking a rectangle around the animal. For RW, we provide 40 seeds for ground and figure regions each. These methods however, do not complement our aim of automatic figure-ground separation since user input is required to process every image.

ITML [15] and LMNN [16]: We follow the same experimental setup to evaluate the performance of these approaches in our application. The results indicate that our approach outperforms the two. We inspected the eigenvalues for the three learned metrics and observed that our approach led to a very sparse vector $\mathbf{w}$, whereas the other two methods had significant number of nonzero eigenvalues. This possibly limits the generalization of the learned metric to new noisy images.

Fig. 3: Qualitative segmentation results.

6. CONCLUSION

In this work, we proposed a novel figure-ground segmentation approach to aid visual identification of pattern species by automatically separating the animal from its background. We used a small training set to learn a discriminative metric to guide the segmentation process. First, the learned metric is used in mean shift clustering to partition the image into clusters, where similar superpixels cluster together. In the second step, the animal is detected by localizing the cluster that is most similar (least distance) to the animal features used for training. The proposed approach performed effectively in images with extreme illumination conditions and achieved better performance than popularly used metric learning algorithms. We also showed that generic figure ground segmentation techniques are not effective and fail due to similarity between the animal's pattern and the background.

[1] Available at http://kspace.cdvp.dcu.ie/public/interactive-segmentation

[2] Available at http://in.mathworks.com/matlabcentral/fileexchange/44630-gabor-feature-extraction

[3] Available at http://grabcut.weebly.com/code.html

[4] Available at http://fastrw.cs.sfu.ca/

7. REFERENCES

[1] S. Hoque, MA Azhar, and F. Deravi, "Zoometrics-biometric identification of wildlife using natural body marks," *International Journal of Bio-Science and Bio-Technology*, 2011.

[2] L. Hiby, P. Lovell, N. Patil, N. S. Kumar, A. M. Gopalaswamy, and K. U. Karanth, "A tiger cannot change its stripes: using a three-dimensional model to match images of living tigers and tiger skins," 2009.

[3] A. Zhelezniakov, T. Eerola, M. Koivuniemi, M. Auttila, R. Levänen, M. Niemi, M. Kunnasranta, and H. Kälviäinen, "Segmentation of saimaa ringed seals for identification purposes," in *Advances in Visual Computing*, pp. 227–236. Springer, 2015.

[4] A. Rosenfeld and D. Weinshall, "Extracting foreground masks towards object recognition," in *Proc. ICCV*, 2011, pp. 1371–1378.

[5] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE PAMI*, vol. 24, no. 5, pp. 603–619, 2002.

[6] C Rother, V. Kolmogorov, and A Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, 2004.

[7] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive gmmrf model," in *Proc. ECCV*, pp. 428–441. 2004.

[8] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *Proc. 12th IEEE ICCV*, 2009, pp. 277–284.

[9] A. Levin and Y. Weiss, "Learning to combine bottom-up and top-down segmentation," in *Proc. ECCV*, pp. 581–594. Springer, 2006.

[10] W. Li, Y. Shi, W. Yang, H. Wang, and Y. Gao, "Interactive image segmentation via cascaded metric learning," in *Proc. IEEE ICIP*, 2015, pp. 2900–2904.

[11] Jia Xu, Maxwell Collins, and Vikas Singh, "Incorporating user interaction and topological constraints within contour completion via discrete calculus," in *CVPR*, 2013, pp. 1886–1893.

[12] L. Bertelli, T. Yu, D. Vu, and B. Gokturk, "Kernelized structural svm learning for supervised object segmentation," in *Proc. CVPR*, 2011, pp. 2153–2160.

[13] D. Kuettel and V. Ferrari, "Figure-ground segmentation by transferring window masks," in *Proc. CVPR*, 2012, pp. 558–565.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE PAMI*, vol. 38, no. 1, pp. 142–158, 2016.

[15] Jason V Davis, B. Kulis, P. Jain, S. Sra, and I. S Dhillon, "Information-theoretic metric learning," in *Proc. 24th ACM ICML*, 2007.

[16] K. Q Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, pp. 207–244, 2009.

[17] A. Shukla and S. Anand, "Distance metric learning by optimization on the stiefel manifold," in *DIFF-CV Workshop, co-located with BMVC*, 2015.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE PAMI*, 2012.

[19] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *IJCV*, vol. 43, no. 1, pp. 29–44, 2001.

[20] W. Liu, C. Mu, R. Ji, S. Ma, J. R. Smith, and S.-F. Chang, "Low-rank similarity metric learning in high dimensions," in *AAAI*, Austin, Texas, USA, 2015.

[21] R. M. Haralick, SR Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *SPIE MILESTONE SERIES MS*, vol. 127, pp. 71–89, 1996.

[22] L. Grady, "Random walks for image segmentation," *IEEE PAMI*, vol. 28, no. 11, pp. 1768–1783, 2006.

[23] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *IJCV*, 2006.

[24] Y. Y Boykov and Marie-Pierre Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proc. 8th ICCV*, 2001, vol. 1, pp. 105–112.

[25] M. Lahiri, C. Tantipathananandh, R. Warungu, D. I. Rubenstein, and T. Y. Berger-Wolf, "Biometric animal databases from field photographs: Identification of individual zebra in the wild," in *1st ACM ICMR*, 2011, p. 6.