METRIC LEARNING BASED AUTOMATIC DETECTION AND SEGMENTATION OF PATTERNED SPECIES

Ankita Shukla, Sarthak Ahuja, Saket Anand

IIIT-Delhi

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. Additionally, 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 a learned metric based clustering. 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. We report qualitative and quantitative results on many images of tigers, leopards and zebras.

***Index Terms*—** Mahalanobis distance, Metric Learning, figure-ground segmentation

1. INTRODUCTION

Figure-ground segmentation techniques play a key role in image processing tasks and in computer vision applications like object recognition [1]. These techniques produce a binary segmentation of an image such that all foreground objects are separated from its background. In this work, we show its application in wildlife photo identification systems [2, 3, 4] for patterned species. These systems help researchers to monitor and study the population of a species over time based on images captured in their natural habitats. These methods uniquely identify an individual of a species by the unique pattern present on their body. For example every tiger is identified by its unique pattern of stripes [3] and similarly leopards by their spots. In [3], a 3D model is manually fitted using key locations on the species' body to capture a pattern that is unaffected by pose or nonlinear deformation caused due to motion. In current identification systems, manual inputs are required to find the animal boundaries precisely for either fitting a 3D model or patch selection for template matching. This makes the process labour intensive for large image dataset. So, in order to utilize these identifiers, a robust and efficient technique to mark the region boundaries containing patterned species is required. The aim of this work is to develop a figure-ground segmentation approach that uses some knowledge about the species to automatically extract the region containing patterned species as accurately as possible.

However, automatic segmentation of region boundaries in camera trapped images is challenging due to background clutter, poor illumination, complex pose and occlusion. Sometimes due to flash of light during image capture visual materials like vegetation and ground plane near the species are illuminated, making the two less distinctive as shown in Fig.1. Also, textural similarity between the camouflaged pattern of the species with vegetation makes segmentation process further intractable.

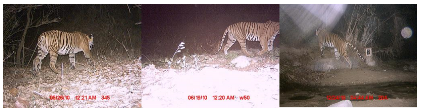

Fig. 1: Example of tiger images showing different illumination conditions and their effects on surrounding.

The rest of the paper is organized as follows: In Section 2, we briefly review the existing approaches for figure-ground segmentation. In Section 3, we discuss preprocessing and feature extraction stage followed by our metric learning formulation and the optimization strategy. In Section 4, we present our segmentation strategy and present evaluation results of our proposed approach on three different patterned species in Section 5. We also present comparison of our metric learning approach with state of the art metric learning algorithm in our image segmentation framework and other segmentation techniques. Finally, we conclude our work in Section 6.

2. LITERATURE REVIEW

Recently, figure-ground segmentation is addressed by interactive segmentation techniques [5, 6, 7, 8, 9] that requires man-

Thanks to WII for providing camera trap images of tigers.

ual input from the user to guide the segmentation process. In [5], a bounding box around the object of interest is drawn to define the region outside the box as background while part inside the box is considered as a combination of foreground and background. Based on the concept of graph cuts an iterative algorithm is developed to get the desired segmentation result. The common theme underlying these approaches is to treat an image as a weighted graph, where each node corresponds to a pixel or image region and an edge between the two captures the likelihood that they belong to the same segment. Since these methods use some label information from the given image, figure-ground segmentation of every image requires manual input defeating our purpose for automatic segmentation. Also, due to feature similarity between the object and background that may arise due to noise or illumination conditions, several iterations of these algorithms are required to achieve the desired segmentation result.

However, methods like [10, 1] require no input for the given image and use SVM models trained on different low level features extracted from training images to classify test image regions as foreground or background. In [1], a supervised foreground/background segmentation approach is developed for object recognition. This method learns both geometric and appearance prior for the task while formulating it as a graph partitioning problem where nodes correspond to superpixels. In [11], figure-ground segmentation is achieved in two steps. In the first step, overlapping windows in test image are assigned a label based on the nearest neighbor in the training images, and in the second step an energy function is minimized over all pixels and their labels to obtain the optimal labeling. Even though these methods achieve desired segmentation results but they do require a lot of training images and many different features like local phase quantization (LPQ) texture feature, GIST features along with spatial information for training purposes. In case of wildlife monitoring, obtaining a lot of training images eventually raises the need to label the images into foreground and background regions.

Recently, [9] overcomes the dependence of supervised learning on large training data by proposing a metric learning based approach that uses some labeled foreground and background superpixels from the given image. The distance metric is learned repeatedly by using state of the art metric learning approach [12] based on the current labeled set. The unlabeled superpixels are labeled as foreground and background using learned metric, while adding the superpixels labeled with high confidence to labeled set to relearn the metric.

In this work, we propose an automatic figure-ground segmentation technique also based on metric learning approach. However, the proposed approach goes a step ahead by uses no label knowledge from the given image to learn the distance metric. The proposed approach uses supervision from training images to learn the distance metric while using far less training data then used in existing supervised leaning paradigms. We also proposes a new metric learning formulation while adopting joint optimization strategy developed in [13] to build our optimization strategy.

3. FEATURE EXTRACTION AND METRIC LEARNING

The proposed image segmentation approach learns a distance function using training samples to guide the segmentation process. In application like visual biometrics, where the object of interest is known apriori, the task of segmentation can be improved by incorporating this information in some form. We learn a distance metric that ensures features corresponding to figure and ground are well separated. In this section, we first discuss the feature representation used in this work followed by our metric learning approach.

3.1. Feature Extraction

All the images used for testing and training are organized as superpixels by using SLIC [14] segmentation algorithm. We use texture features to distinguish the object of interest i.e a patterned species from its background. The images are operated with a filter bank [15] of 48 filters at different scales and orientations. Thus, every pixel in the image has a 48 dimensional response vector. The response vectors corresponding to figure and ground superpixels from the test images are clustered using K-means algorithm into 20 clusters each. Finally, each superpixel is represent by a 40 dimensional quantized vector by mapping each of its pixel response to its nearest cluster center. The generated feature vectors are then used to create similarity and dissimilarity pair constraints required for our metric learning approach.

3.2. Background and Notations

We follow the following notations in our work. The features extracted from figure and ground superpixels are represented by a set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, with $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$. Figure features are denoted with class label $\ell_i = 1$ and background features with label $\ell_i = 2$. We create pairwise similarity and dissimilarity constraints where set of similar pairs $\mathcal{C}_s = \{(i,j) | \ell_i = \ell_j,\ i,j \in \{1, \ldots, m\}\}$ contains pairs of points which have the same class label. Analogously, the set of dissimilar pairs is defined $\mathcal{C}_d = \{(i,j) : \ell_i \neq \ell_j,\ i,j \in \{1, \ldots, m\}\}$ and combined set of similar and dissimilar point pairs is denoted by $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_d$. The space of PSD matrices is denoted by $\mathcal{S}_+^n$, $\mathbb{R}_+^n = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{x} \geq \mathbf{0}\}$ represents the positive orthant. The space of orthonormal matrices is the Stiefel Manifold $\mathcal{S}_{n,p} = \{\mathbf{U} \in \mathbb{R}^{n \times p} | \mathbf{U}^\top \mathbf{U} = \mathbf{I}_p\}$, where $\mathbf{I}_p$ is $p \times p$ identity matrix.

In this work, we learn a Mahalanobis distance function that parametrizes euclidean distance between two features $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$, by a positive semidefinite matrix and is given by

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

Here, $\mathbf{M} \succeq 0$ is a positive semidefinite (PSD) matrix. Learning the Mahalanobis distance matrix $\mathbf{M}$ is equivalent to learning the scaling and rotation of the feature space that captures the underlying structure of data. Here, we represent matrix $M = \mathbf{U}\mathbf{W}\mathbf{U}^\top$, with $\mathbf{W} = \text{Diag}(\mathbf{w})$, where $\mathbf{U} \in \mathcal{S}_{n,p}$ is the orthonormal matrix of eigenvectors and $\mathbf{w} \in \mathbb{R}^p_+$ is the vector of nonnegative eigenvalues.

3.3. Metric Learning Algorithm

The objective of metric learning algorithm is to transform the input space that ensures that the distance between similar point pairs is reduced while dissimilar point pairs are moved far apart. Our metric learning formulation uses a hinge loss function to account for violating constraints and a l2 norm regularizer to ensure smooth changes in learned distance metric. The proposed metric learning formulation is given by

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

Here, the term $[x]_+ = \max(0, x)$ captures the hinge loss, the vectors $\mathbf{z}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ represent the difference vectors for the constraint pairs $\{i, j\} \in \mathcal{C}$, the target distance b_{ij} takes value s or d based on whether $(i, j) \in \mathcal{C}_s$ or $(i, j) \in \mathcal{C}_d$. $\alpha > 0$ is a regularization parameter, $y_{ij} = 1$ if $(i, j) \in \mathcal{C}_s$ and $y_{ij} = -1$ for $(i, j) \in \mathcal{C}_d$.

The proposed formulation in (2) is non convex due to the domain of the orthogonal matrix $\mathbf{U}$. Similar to [13], we follow a joint optimization strategy that solves for eigenvalues by optimization on the positive orthant and the corresponding eigenvectors matrix by optimization on the Stiefel manifold alternatively.

Therefore, keeping $\mathbf{U}$ fixed, subproblem for optimizing $\mathbf{w}$ is given by

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

Here, 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 by decoupling the two in easier sub problems. We therefore adapt a ADMM based approach to solve for $\mathbf{w}$. Now, with the updated $\mathbf{w}$, the subproblem for $\mathbf{U}$ is given by

$$\mathbf{U}^{t+1} = \underset{\mathbf{U} \in \mathcal{S}_{n,p}}{\text{argmin}} \sum_{i,j=1}^{m} \left[y_{ij} (\mathbf{z}_{ij}^\top \mathbf{U} \text{Diag}(\mathbf{w}) \mathbf{U}^\top \mathbf{z}_{ij} - b_{ij}) \right]_+ \qquad (4)$$

We employ a constraint preserving update scheme proposed by Wen and Yin *et al.* [16] and follow [13] for solving $\mathbf{U}$ as an instance of optimization on the Stiefel Manifold. The learned distance metric is then given by $M = \mathbf{U}\text{Diag}(\mathbf{w})\mathbf{U}^\top$.

4. SEGMENTATION STRATEGY

In this section, we lay down the steps involved in figure-ground segmentation of a test image once the Mahalanobis distance matrix $\mathbf{M}$ is learned using training images. The segmentation strategy is summarized in Fig. 4.

4.1. Mean Shift Segmentation

Initially, the query image is preprocessed and features are extracted for each superpixel region as discussed in Section 3.1. The texture feature corresponding to each superpixel is concatenated with mean spatial co-ordinates of the superpixel. These features are provided as an input to Mean Shift algorithm where euclidean distance is replaced by Mahalanobis distance for texture feature similarity. As a result, the test image is segmented into regions where superpixels with similar texture and spatial features are grouped together.

4.2. Distance Map Generation

In order to detect cluster that correspond to patterned species, we generate a distance map by computing Mahalanobis distance of each cluster region with patterned species' feature extracted from one of the training images. The test image cluster with the least distance is marked as region of interest while others are marked as background generating a binary mask. The distance map for a given test image is shown in Fig. 2c, where decrease in brightness denotes increase in distance.

4.3. Morphological Operations

In most of the cases, textural similarity of pattern species with some of the regions in the lower part of the image comprising of vegetation and other clutter is observed. So, the cluster that corresponds to region of interest also considers smaller regions from background as well. For all the test images, we perform a set of morphological operations. First, a dilation operation is applied to eliminate small gaps in the region of interest. These regions arise at the boundaries where some pixels from figure are grouped with background pixels to form a background superpixel. A connected component operation is then performed on the dilated image, where the largest component corresponds to the patterned species. However, other connected components occur due to background superpixels that have texture similarity with pattern species are labeled as background.

5. EXPERIMENTAL EVALUATION

We evaluate the performance of our proposed approach on three different patterned species: tiger, leopard and zebra. We also compare performance of our approach with other segmentation techniques : Graph cut [17], GrabCut [5], Random Walker [18] and Lossy Compression. The proposed metric learning algorithm is also compared with state of the art metric learning algorithm ITML [12] in segmentation framework.

The ground truths for all the images is created by interactive segmentation tool [1]. The training data for metric learning on tiger images is extracted from two labeled images to account for variations in background clutter and illumination and only one image each is used in case of others. The similarity and dissimilarity pair constraints are created by randomly selected 20 figure and 20 ground superpixels from training data in each of the three cases.

Tiger Images: We use the database collected by Wildlife Institute of India (WII) by camera traps located in their habitats. All the images are rescaled to 256×342. Segmentation strategy is evaluated on 30 tiger images and pixel wise precision/recall and segmentation accuracy are reported.

Leopard Images: The leopard images used for training and testing are collected from the Internet. The images are operated with Gabor fitter bank and feature vectors are generated using the same clustering based approach as discussed in Section 3.1. Since, we could not find a large database of images, we only show quantitative results in Fig 3.

Zebra [19]: Similar to tigers, zebras also have characteristic pattern of stripes. The texture features based on LM filter bank are extracted as discussed in Section 3.1. The proposed approach is evaluated on 30 zebra images and pixel wise precision/recall and segmentation accuracy are reported.

5.1. Comparison Results and Discussion

GrabCut [20] and Random Walker [18]. We use interactive tools for GrabCut [2] and Random Walker [3]. GrabCut is initialized with a bounding box by marking a rectangle close to the pattern species all the region from forehead to tail. For Random Walker, we provide 20 seeds for ground and figure region each. The results are given in Table **??**. However, these methods do not complement our aim of automatic figure-ground separation since user input is required to process every image.

Graph Cut [21]. Due to visual similarity of ground region with patterned species, we created three nodes so that ground and patterned species are separated. So, graph cut fails to perform in camera trapped images effected by illumination and vegetation patterns similar to patterned species.

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

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

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

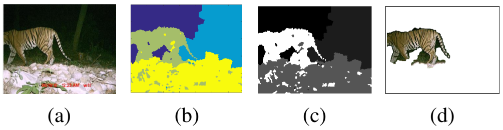

Fig. 2: Steps involved in Segmentation. (a) Test Image (b) Mean Shift Segmentation (different colors denote different clusters). (c) Distance Map (d) Segmentation Result after morphological operations on cluster with minimum distance

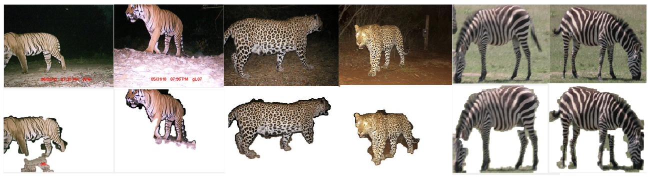

Fig. 3: Segmentation Results.

6. CONCLUSION

In this work, we proposed a novel figure-ground segmentation approach to aid the identification process of patterned species. The proposed approach learned a Mahalanobis distance metric using features extracted from training images to guide the segmentation process. The metric learning algorithm used fe The distance metric learning app The segmentation approach uses only two labeled images for training purpose as oppose to other supervised approaches. We adapted Mahalanobis distance metric to distinguish figure and ground features.

7. REFERENCES

[1] Amir Rosenfeld and Daphna Weinshall, "Extracting foreground masks towards object recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1371–1378.

[2] Sanaul Hoque, MA Azhar, and Farzin Deravi, "Zoometrics-biometric identification of wildlife using natural body marks," *International Journal of Bio-Science and Bio-Technology*, vol. 3, no. 3, pp. 45–53, 2011.

[3] Lex Hiby, Phil Lovell, Narendra Patil, N Samba Kumar, Arjun M Gopalaswamy, and K Ullas Karanth, "A tiger cannot change its stripes: using a three-dimensional model to match

Method	Precision	Recall	Segmentation Accuracy(%)
GrabCut [20]	89	90	93
RW [18]	36	89	69
Graph Cut[21]	32	81	72
ITML [12]			
Ours	81	93	93.64

Table 1: Results on the tiger images dataset.

Method	Precision	Recall	Segmentation Accuracy(%)
GrabCut [20]	89	90	93
RW [18]	36	89	69
Graph Cut[21]	32	81	72
ITML [12]			
Ours	81	93	93.64

Table 2: Results on the Zebra images dataset.

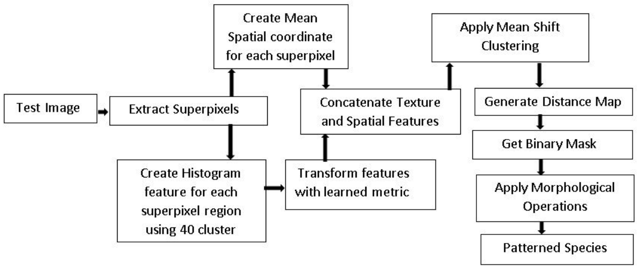
Test Image
Extract Superpixels
Create Mean Spatial coordinate for each superpixel
Create Histogram feature for each superpixel region using 40 cluster
Transform features with learned metric
Concatenate Texture and Spatial Features
Apply Mean Shift Clustering
Generate Distance Map
Get Binary Mask
Apply Morphological Operations
Patterned Species

Fig. 4: Figure-ground segmentation of a test image.

images of living tigers and tiger skins," *Biology Letters*, pp. rsbl–2009, 2009.

[4] Artem Zhelezniakov, Tuomas Eerola, Meeri Koivuniemi, Miina Auttila, Riikka Levänen, Marja Niemi, Mervi Kunnasranta, and Heikki Kälviäinen, "Segmentation of saimaa ringed seals for identification purposes," in *Advances in Visual Computing*, pp. 227–236. Springer, 2015.

[5] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[6] Andrew Blake, Carsten Rother, Matthew Brown, Patrick Perez, and Philip Torr, "Interactive image segmentation using an adaptive gmmrf model," in *Computer Vision-ECCV 2004*, pp. 428–441. Springer, 2004.

[7] Victor Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp, "Image segmentation with a bounding box prior," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 277–284.

[8] Anat Levin and Yair Weiss, "Learning to combine bottom-up and top-down segmentation," in *Computer Vision–ECCV 2006*, pp. 581–594. Springer, 2006.

[9] Wenbin Li, Yinghuan Shi, Wanqi Yang, Hao Wang, and Yang Gao, "Interactive image segmentation via cascaded metric learning," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2900–2904.

[10] Luca Bertelli, Tianli Yu, Diem Vu, and Burak Gokturk, "Kernelized structural svm learning for supervised object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2153–2160.

[11] Daniel Kuettel and Vittorio Ferrari, "Figure-ground segmentation by transferring window masks," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 558–565.

[12] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 209–216.

[13] Ankita Shukla and Saket Anand, "Distance metric learning by optimization on the stiefel manifold," in *DIFF-CV Workshop, co-located with BMVC*, 2015, pp. 7.1–7.10.

[14] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[15] Thomas Leung and Jitendra Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International journal of computer vision*, vol. 43, no. 1, pp. 29–44, 2001.

[16] Zaiwen Wen and Wotao Yin, "A feasible method for optimization with orthogonality constraints," *Mathematical Programming*, vol. 142, no. 1-2, pp. 397–434, 2013.

[17] Yuri Y Boykov and Marie-Pierre Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, 2001, vol. 1, pp. 105–112.

[18] Leo Grady, "Random walks for image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 11, pp. 1768–1783, 2006.

[19] Mayank Lahiri, Chayant Tantipathananandh, Rosemary Warungu, Daniel I Rubenstein, and Tanya Y Berger-Wolf, "Biometric animal databases from field photographs: Identification of individual zebra in the wild," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011, p. 6.

[20] Peng Wang, "grabcut–interactive foreground extraction," .

[21] Yuri Boykov and Gareth Funka-Lea, "Graph cuts and efficient nd image segmentation," *International journal of computer vision*, vol. 70, no. 2, pp. 109–131, 2006.