Fly Wing Biometrics

Using Genetic & Evolutionary Feature Extraction

Michael Payne*, Jonathan Turner*, Joseph Shelton*, Joshua Adams*, Joi Carter*,
Henry Williams*, Caresse Hansen*, Ian Dworkin†, Gerry Dozier*

*Department of Computer Science, North Carolina A&T State University
1601 East Market St. Greensboro, NC, 27411
†Department of Zoology, Michigan State University, East Lansing, Michigan 48824

{mepayne1| jturner | jashelt1 | jcadams2 | jncarte1 | hcwillia | cdhansen}@aggies.ncat.edu,
gvdozier@ncat.edu, idworkin@msu.edu

***Abstract*—Genetic and Evolutionary Feature Extraction (GEFE), introduced by Shelton et al. [1], [2], [3], use genetic and evolutionary computation to evolve Local Binary Pattern (LBP) based feature extractors for facial recognition. In this paper, we use GEFE in an effort to classify male and female *Drosophila melanogaster* by the texture of their wings. To our knowledge, gender classification of the drosophila melanogaster via its wing has not been performed. This research has the potential to simplify the work of geneticists who work with the drosophila melanogaster. Our results show that GEFE outperforms both LBP and Eigenwing methods in terms of accuracy as well as computational complexity.**

Keywords - Biometrics, Drosophila, Feature Extraction, Genetic Algorithms (GA), Genetic and Evolutionary Feature Extraction (GEFE), Genetic and Evolutionary Computations (GECs), Local Binary Pattern (LBP)

I. INTRODUCTION

Drosophila melanogaster (the common fruit fly) has been a genetic model organism for over 100 years. It has an almost complete global distribution as a human commensal, and has a rapid (10-14 day) life cycle. Because of this, Drosophila remains an important tool for bio-medical research given the ease of which it can be used to screen for new mutations. The shape of the Drosophila wing has recently become an important model trait for genetics and evolution [6], [7], [8].

The field of biometrics is devoted to identifying individuals based on physical, chemical, and/or behavioral characteristics [5]. Although biometric recognition techniques are traditionally applied to human subjects, in this paper we apply a well-known technique to classify fly wings. This research explores the classification of male and female Drosophila by the texture of their wings. In this work, we use a genetic and evolutionary computation (GEC) known as Genetic & Evolutionary Feature Extraction with Machine Learning ($GEFE_{ML}$). $GEFE_{ML}$ introduced by Shelton et al. in [9], evolves Local Binary Pattern-based feature extractors for facial recognition. The fly wings of male and females are similar enough so that not even an entomologist can tell them apart. However, $GEFE_{ML}$ will be shown to have a promising effectiveness of classifying genders.

The remainder of this paper is as follows. Section II provides a background of the Local Binary Pattern feature extraction, Eigenwing, and $GEFE_{ML}$ techniques. In Section III, we present our experiments and in Section IV, we provide our results. In Section V, we provide our conclusions and future work.

II. FEATURE EXTRACTION/SELECTION METHODS

A. Local Binary Patterns (LBP)

The LBP feature extraction method is a technique proposed by Ojala et al. [10]. This technique can be used to classify textures in images. LBP works by evenly distributing uniform, non-overlapping patches across the surface of an image. The LBP method is typically used as follows. If a pixel (center pixel) is surrounded another pixel (neighbor pixel) on all sides (3x3 arrangement of pixels), LBP can be performed by taking the difference of the center pixel and each neighboring pixel. If the difference between the center pixel and neighboring pixel is less than 0, then a value of 0 is used to represent the neighbor pixel. Otherwise, a value of 1 is used to represent the neighbor pixel. These neighboring pixel representations are then concatenated together to form a binary string. The resulting string is seen as a form of pattern or texture. Figure 1 shows an example of LBP is calculated for one pixel within a patch.

250	43	187
121	125	159
15	101	202

125	-82	62
-4		34
-110	-24	77

1	0	1
0		1
0	0	1

Figure 1: Center Pixel Resulting in 10111000

If 8 neighboring pixels were being used for the LBP method, then there would be 256 possible binary patterns. This

number is produced from by 2^n, where n is the number of neighbors. To reduce the number of binary patterns, they can be divided into uniform and non-uniform patterns. A uniform pattern as a bit string whose bits, when compared in sequential order and circularly shift values no more than two times. A non-uniform pattern is defined as a bit string that has more than two bit shifts or changes when comparing circularly, and sequentially [10]. For example, in Figure 1 starting from the top left corner going clockwise, the bit string 10111000 has a total of four changes and is deemed a non-uniform pattern. The changes are from the first to second bit; second to third; fifth to sixth; and eight back to the first. An example of a uniform bit pattern would be 10111111 with two changes from the first to second bit and the second to third bit.

With uniform and non-uniform patterns, the number of patterns can be reduced to 59 binary patterns, which are represented by a specific bin in the histogram. There are now 58 unique bins, one for each uniform binary pattern and one bin for the non-uniform patterns. Each histogram has 59 bins for each patch on the image.

When LBP is applied to the pixels of a patch, a histogram is created that represents the frequency of each unique texture pattern for that particular patch. The histograms of every patch on an image are then concatenated together to form a unique set of features, called a feature vector (or feature template), which represents that image. If a FE has 24 patches and each patch has 59 bins, the resulting FV would have 1416 features [10].

During the process of recognition, a probe template, p, is compared to a gallery set of feature vectors $H = \{h_0, h_1, \ldots, h_{k-1}\}$ using the (Manhattan) City Block distance metric. This distance is a numerical representation of the distinction between two biometric instances and can be calculated using (1):

$$d = \sum_{i=0}^{n} |p_i - h_{k,i}| \quad (1)$$

where d is the distance between two subjects, p is the probe feature template, $H=\{h_0, h_1, \ldots, h_{k-1}\}$ is the gallery feature set, n is total number of features, i is the index of the feature, and k is the k^{th} individual in the gallery. The subject, h_k, is considered a match to p when the distance between the two vectors is the smallest compared to all other subjects in H.

B. Eigenwing

Eigenwing is similar to the Eigenface approach introduced by Turk and Pentland in [11] and is a dimensionality reduction technique that is based on Principal Component Analysis (PCA) for fly wing recognition. PCA derives eigenvectors from the covariance matrix of the probability distribution of the high-dimensional vector space of images [11], [12]. These eigenvectors are then ordered according to how much of the variation is present in the data they contain. Features that are present in the original image to a higher degree have a greater share of the corresponding Eigenwing in the summation of the Eigenwings. On the other hand, if the particular feature is to a lower degree or not visible at all in the original image, then the corresponding Eigenwing should contribute a smaller part all the way done to zero to the sum of Eigenwings [11], [12], [13]. The higher the eigen value (or weight), the more discriminant the vector is in relation to a given set. The reconstructed original image is equal to a sum of all Eigenwings, with each weight saying to what degree a specific Eigenwing is present in the original image. For this experiment, the top 10% of eigen vectors were used. This was decided by looking at the eigenvalues. The eigenvectors corresponding to small eigenvalues contain less information about detailed differences in comparison to a higher level of discrimination by the eigenvectors with larger eigenvalues.

C. $GEFE_{ML}$

Genetic and Evolutionary Feature Extraction (GEFE) is a technique that evolves LBP-based feature extractors (FEs). Unlike a standard LBP method (SLBPM), GEFE evolves FEs that consists of a set of patches in a variety of locations on an image. A feature extractor, fe_i, is represented as 6-tuple with 5 sets and 1 single value represented as $<X_i,Y_i,W_i,H_i,M_i,f_i>$. Each of the patches in a particular FE, fe_i, are designed using the values in the 6-tuple. The X_i and Y_i sets hold the <X,Y> points of the center of each patch in fe_i, while the sets W_i and H_i contains the width and heights of the patches. The set M_i denotes a masking value for each patch in fe_i. Though there can be multiple patches defined by the 6-tuple, a patch's specific masking value determines whether the features extracted by that patch are included in the resulting feature vector (FV).

Once a FE has been configured, it is then evaluated and assigned a fitness. The fitness, f_i, is determined by applying fe_i to a dataset of images. The result of this process is a set of FVs, one for each image. The FVs are then separated equally by class into a probe set (FVs to be matched), and a gallery set (FVs that each probe can match to). The FVs in the probe set are compared to the FVs in the gallery set using the Manhattan distance measure. When comparing a single probe FV to the gallery FVs, the smallest Manhattan distance is considered to be a match. If a probe FV is incorrectly matched with a gallery FV, then fe_i is said to cause an error. The number of errors is accumulated along with the percentage of the image used. The resulting fitness, f_i, is ten times the number of errors (ε), added to the percent of surface area (ζ), shown below in (2).

$$f_i = 10\varepsilon + \zeta \quad (2)$$

Unlike the SLBPM, GEFE is able to partition images using overlapping, uniform patches that cover a relatively small

portion of the image. This reduces the amount of surface area needed for the calculations.

Cross validation [14], [15], [16] in Genetic and Evolutionary Feature Extraction – Machine Learning ($GEFE_{ML}$) is done by initially generating a population of random FEs. Every candidate FE, fe_i, is then evaluated on the training set and additionally evaluated on a validation set. The validation set consists of mutually exclusive subjects that are not related to the training set. The results of the FEs on the validation set do not affect the training of FEs. The FE with the best results on the validation set is stored as FE*. FE* is only updated when a new candidate FE performs better on the validation set than the currently stored FE*. As a result of cross validation, FE* should generalize better on unseen subjects as opposed to the best performing FE on the training set. Figure 2 provides a flowchart for cross validation in $GEFE_{ML}$ [9].

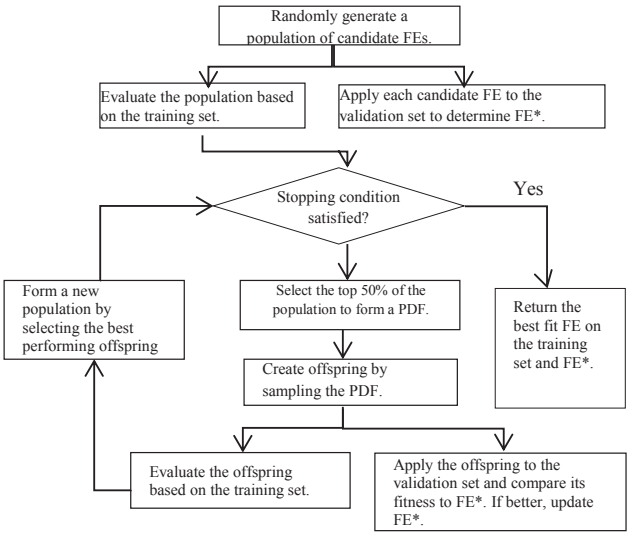


Figure 2: Cross Validation in $GEFE_{ML}$ [9]

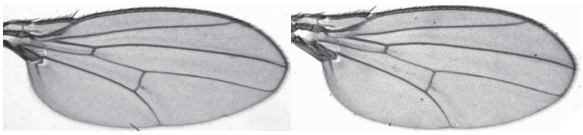
$GEFE_{ML}$ is an instance of an Estimation of Distribution Algorithm (EDA). This is based on results presented in [3] where the EDA instance of GEFE had a statistically better performance than a Steady State Genetic Algorithm (SSGA) instance of GEFE [3], [18]. Figure 3 provides a pseudo code version of an EDA [19]. In an EDA, a population of candidate (FEs) is created and evaluated. Next, a probability of density function (PDF) is created using the top 50% best performing FEs on the training set. The PDF is then sampled to create a number of offspring equal to the population. However, a user specified number elites (best individuals of the previous population) are allowed to survive. The new offspring are then evaluated. Finally, a new population is created from the offspring and elites of the previous population. This process continues until the algorithm converges on a solution or a user defined number of function evaluations have passed.

```
compute EDA{
t = 0;
initialize pop(t)
evaluate pop(t)
While (Not done) {
        S(t) = selected subpopulation the
        best individuals;
        Build a probability Density Function, PDF(t),
        of S(t);
        Sample PDF(t) to generate O(t);
        Replace P(t)- Elites with O(t);
        t = t+1;
  }
}
```

Figure 3: Pseudo-code for the GEFE EDA [3]

III. EXPERIMENT

The image dataset, supplied by Ian Dworkin (Dworkin Lab) at Michigan State University, will be referred to as the Dworkin Lab Fly Wing (DLFW) dataset. The images were collected during a large-scale experiment which examined how mutations of two biological signaling pathways influenced wing shape [20], [21]. The DLFW dataset consists of 2,453 images of fly wings. Each image is either an instance of a male or female fly wing. Since this experiment only looks at determining the gender of a wing from the normal genotype, all wings with expressed mutations were removed. A visual inspection of each of the remaining wings removed images with partial wings, multiple wings, and large occlusions. The resulting dataset was a subset of the DLFW that contained 656 wings with a single image per subject (DLFW-656). In the DLFW-656 dataset, 328 of the subjects were female and the remaining 328 subjects were male. Each image in the database is normalized to a grey scale image with a resolution of 614 by 266 pixels and a horizontal orientation.



Female Fly Wing Male Fly Wing

Figure 4: Male and Female Fly Wings

The 656 subjects were divided into three subsets: DLFW-300, consisting 150 subjects in the probe and 150 in the gallery; DLFW-200, consisting of 100 subjects in the probe and 100 in the gallery; and DLFW-156, consisting 78 subjects in the probe and 78 in the gallery. DLFW-300 was used as a training set, DLFW-200 was used as a validation set, and DLFW-156 was used as a test set to confirm that the FEs generalize well to unseen subjects. All datasets were split evenly between and female and male subjects.

For this experiment, we compared the Standard Local Binary Pattern (SLBPM) and Eigenwing methods with the $GEFE_{ML}$ technique. The SLBPM used a set of 24 patches (four

rows by six columns) that were uniform, non-overlapping, and covered the entire image. $GEFE_{ML}$ used a set of 24 patches that were uniform but had the capability to overlap patches on top of each other and be turned off.

IV. Results

For our results, 30 EDA instances of $GEFE_{ML}$ were executed. All instances were run for 1000 function evaluations evolving a population of 20 FEs keeping a single elite. The PDF was created from the top 50% of the best performing candidate solutions according to fitness. The EDA instances of $GEFE_{ML}$ implemented were part of the eXploration Toolset for the Optimization of Launch and Space Systems (XTOOLSS) application [3], [22]. Table I shows the performance of SLBPM and Eigenwing compared to the average performance of $GEFE_{ML}$.

TABLE I
EXPERIMENTAL RESULTS FOR
SLBPM, EIGENWING, AND $GEFE_{ML}$ METHODS

Method	Accuracy	Best Accuracy	Avg Surface Area
SLBPM	56.41%	56.41%	100.00%
Eigenwing	43.58%	43.58%	100.00%
$GEFE_{ML}$	73.16%	80.77%	68.04%

The feature extractors created by $GEFE_{ML}$, with an average accuracy of 73.16% outperformed their SLBPM and Eigenwing counterparts with accuracies of 58.33% and 43.58% respectively. An ANOVA test was performed between the three techniques and the results showed a statistical difference with a 95% confidence.

In terms of feature extraction, it is interesting to see that all of the patches covered a portion of the wing where curvatures were present. This may suggest that there is more discriminatory information in this region that can be used to differentiate between male and female Drosophila. Figure 5 shows an approximate positioning of patches for one of the best feature extractors created using $GEFE_{ML}$.

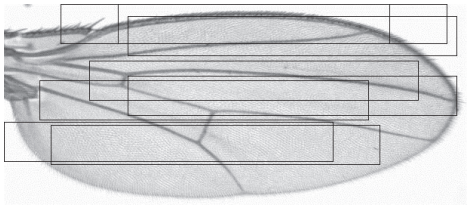


Figure 5: Location of patches generated by $GEFE_{ML}$

V. Conclusions & Future Work

In this paper, Genetic & Evolutionary Feature Extraction–Machine Learning was used to classify *Drosophila melanogaster* as either female or male by the texture of their wings. Though our method outperforms other methods of classification while reducing the number of features needed for classification, we believe that more improvements can be made. The ability to classify gender of the drosophila melanogaster just using their wings can not only simplify research done with them, but it may be possible to discern other things about a fly using just its wing. Our future work will be devoted towards developing methods to increase the classification rate of evolved feature extractors.

Acknowledgment

This research was funded by the office of the Director of National Intelligence (ODNI), Center for Academic Excellence (CAE) for the multi-university, Center for Advanced Studies in Identity Sciences (CASIS) and by the National Science Foundation (NSF), Science & Technology Center: Bio/Computational Evolution in Action Consortium (BEACON). The authors would like to thank the ODNI and the NSF for their support of this research.

References

[1] Shelton, J., Dozier, G., Bryant, K., Adams, J., Popplewell, K., Abegaz, T., & Ricanek, K. (2011, March). Genetic based LBP feature extraction and selection for facial recognition. In *Proceedings of the 49th Annual Southeast Regional Conference* (pp. 197-200). ACM.

[2] Shelton, J., Dozier, G., Bryant, K., Smalls, L., Adams, J., Popplewell, K., & Ricanek, K. (2011, April). Comparison of Genetic-based Feature Extraction Methods for Facial Recognition. In *Midwest Artificial Intelligence and Cognitive Science Conference* (p. 216).

[3] Shelton, J., Dozier, G., Bryant, K., Smalls, L., Adams, J., Popplewell, K., Abegaz, T., Woodard, D., and Ricanek, K. (2011). " Genetic and Evolutionary Feature Extraction via X-TOOLSS" in The 8th annual International Conference on Genetic and Evolutionary Methods (GEM).

[4] Liu, J., Tang, Y. Y., & Cao, Y. C. (1997). An evolutionary autonomous agents approach to image feature extraction. *Evolutionary Computation, IEEE Transactions on*, *1*(2), 141-158.

[5] Ross, A., Nandakumar, K., & Jain, A. (2008). Introduction to multibiometrics.*Handbook of Biometrics*, 271-292.

[6] Dworkin, I., Anderson, J. A., Idaghdour, Y., Parker, E. K., Stone, E. A., & Gibson, G. (2011). The Effects of Weak Genetic Perturbations on the Transcriptome of the Wing Imaginal Disc and Its Association With Wing Shape in Drosophila melanogaster. *Genetics*, *187*(4), 1171-1184.

[7] Dworkin, I., Palsson, A., & Gibson, G. (2005). Replication of an Egfr-wing shape association in a wild-caught cohort of Drosophila melanogaster.*Genetics*, *169*(4), 2115-2125.

[8] Pitchers, W., Pool, J. E., & Dworkin, I. (2012). ALTITUDINAL CLINAL VARIATION IN WING SIZE AND SHAPE IN AFRICAN DROSOPHILA MELANOGASTER: ONE CLINE OR MANY? Evolution. doi:doi:10.1111/j.1558-5646.2012.01774.x

[9] Shelton, J., Bryant, K., Abrams, S., Small, L., Adams, J., Leflore, D., ... & Dozier, G. (2012). Genetic & Evolutionary Biometric Security: Disposable Feature Extractors for Mitigating Biometric Replay Attacks. *Procedia Computer Science*, *8*, 351-360.

[10] Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(7), 971-987.

[11] Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*(12), 2037-2041.

[12] Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, *3*(1), 71-86.

[13] Turk, M. A., & Pentland, A. P. (1991, June). Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on* (pp. 586-591). IEEE.

[14] Lata, Y. V., Tungathurthi, C. K. B., Rao, H. R. M., Govardhan, A., & Reddy, L. P. (2009). Facial recognition using eigenfaces by PCA. *International Journal of Recent Trends in Engineering*, *1*(1), 587-590.

[15] Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill Companies, Inc.

[16] Haykin, S. (1999). *Neural Networks: a Comprehensive Foundation*, 2nd Edition. Prentice-Hall, Inc. New Jersey.

[17] Jang, J.-S. R., Sun, C.-T. Sun, and Mizutani, Z. (1997). Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Prentice-Hall, Inc. New Jersey

[18] L. Davis, Handbook of Genetic Algorithms . New York: VanNostrand Reinhold, 1991.

[19] Larranaga, P., & Lozano, J. A. (2002). Estimation of distribution algorithms: A new tool for evolutionary computation (Vol. 2). Springer.

[20] Dworkin, I., & Gibson, G. (2006). Epidermal growth factor receptor and transforming growth factor-beta signaling contributes to variation for wing shape in Drosophila melanogaster. *Genetics*, *173*(3), 1417–1431. doi:10.1534/genetics.105.053868

[21] Debat, V., Debelle, A., & Dworkin, I. (2009). Plasticity, canalization, and developmental stability of the Drosophila wing: joint effects of mutations and developmental temperature. *Evolution*, *63*(11), 2864–2876. doi:10.1111/j.1558-5646.2009.00774.x

[22] M. L. Tinker, G. Dozier, and A. Garrett, "*The exploratory toolset for the optimization of launch and space systems (x-toolss)*," http://nxt.ncat.edu/, 2010