Mid-level Image Representation for Fruit Crop Pest Identification

Matheus Macedo*, and Fabio A. Faria*
*GIBIS Lab., Institute of Science and Technology, Federal University of Sao Paulo, Sao Jose dos Campos, Brazil,
Email: matheus.macedo.leonardo@gmail.com and ffaria@unifesp.br

***Abstract*—Fruit flies are of huge biological and economic importance for the farming of different countries in the World, especially for Brazil. Brazil is the third largest fruit producer in the world with 44 million tons in 2016. The direct and indirect losses caused by fruit flies can exceed USD 2 billion, putting these pests as one of the biggest problems of the world agriculture. In Brazil, it is estimated that the economic losses directly related to production, the cost of pest control and in the loss of export markets, are between USD 120 and 200 million/year. We propose to apply mid-level image representations based on local descriptors for fruit fly identification tasks of three species of the genus *Anastrepha*. In our experiments, several local image descriptors based on keypoints and machine learning techniques have been compared in the target task. Furthermore, the proposed approaches have achieved excellent effectiveness results when compared with a state-of-art technique.**

I. Introduction

The fruit flies belong to the Tephritidae family, which comprises approximately 5,000 species. They are distributed all over the world and several species are important agricultural pests. The damages are caused by the larvae that feed inside the fruit, making them unfit for consumption and commercialization. In addition to direct damage to fruits, some species of fruit flies are of quarantine importance, that is, they hamper the international market for fresh fruits. The country where the quarantine pest does not occur imposes customs barriers for the importation of commodities from the country, in which the pest is present.

Among the fruit flies economically important in the Americas are the species of the *Anastrepha*. This genus is the most diverse in America tropics and subtropics with approximately 300 known species, of which 120 are recorded in Brazil [1]. However, few species are economically important in Brazil namely the South American fruit fly *Anastrepha fraterculus* (Wiedemann), the West India fruit fly *Anastrepha obliqua* (Macquart), and the guava fruit fly *Anastrepha striata* Schiner. These three species are considered pests of quarantine significance by many regulatory agencies.

Identification of species is a crucial step for the development studies on biology such as distribution, damage, quarantine, and control. The identification of *Anastrepha* species are based on wing pattern, and mostly on the aculeus (the piercing part of the female ovipositor). However, the species boundaries of some fruit fly complexes are difficult to be delimited. *Anastrepha fraterculus* is the most emblematic case of a cryptic species complex in the Americas, because it is a major pest only in some areas of its occurrence, which ranges from Mexico to northern Argentina [2]. Thus, misidentifications can be of serious problem for the implementation of quarantine restrictions, integrated pest management, and other control programs [3].

This work aims to propose the use of mid-level representations based on local image descriptors for fruit fly identification. Furthermore, it compares the effectiveness results of different machine learning techniques using those representations to support the development of a real-time system for fruit fly identification of the genus *Anastrepha*. This system can be a good solution for a quick and precise identification, reducing the time and costs in performing and assisting the few experts in their tasks. Finally, the proposed approaches can be incorporated into other systems already existing in the literature.

II. Local Feature Extraction

Local feature extraction usually includes two distinct steps [4]: feature detection and feature description. Feature detection consists in finding a set of interest points, or salient regions in the image that are invariant to a range of image transformations. Feature description consists in obtaining robust local descriptors from the detected features. In the following we briefly introduce the detectors/descriptors evaluated in this work.

A. Scale-Invariant Feature Transform (SIFT)

Proposed by Lowe [5], this is the most well-known and widely used local descriptor for visual recognition tasks. SIFT is both a feature detector (based upon Differences-of-Gaussians, or DoG), and a feature descriptor. As a descriptor, it computes a histogram of gradient (HoG) locations and orientations. The resulting descriptor is 128-dimensional feature vector, which is invariant to scale, rotation, affine transformations, and partially invariant to illumination changes.

B. Speeded Up Robust Features (SURF)

It was proposed by Bay et al. [6] as an accelerated version of SIFT. SURF is also both a detector (based upon the determinant of the Hessian matrix, also known as Fast-Hessian feature detector) and descriptor. As a descriptor, it describes a distribution of Haar-wavelet responses within the interest

point neighborhood. The SURF descriptor is based on similar properties of localized information and gradient distribution as SIFT, with a complexity stripped down even further. Only 64 dimensions are used, reducing the time for feature computation and matching, and increasing simultaneously the robustness.

C. Binary Robust Independent Elementary Features (BRIEF)

Presented in 2010 by Calonder et al. [7], BRIEF was the first binary descriptor published. It consists mainly in generating binary strings from simple pixel intensity value comparisons over an image patch smoothed using a Gaussian kernel. The patches are usually obtained with the Fast-Hessian detector, but it is not limited only to the use of this feature detector. We employed the STAR detector, derived from CenSurE (Center Surround Extremas) detector [8] and FAST (Features from Accelerated Segment Test) detector [9]. The bit-length of the BRIEF descriptor are 128, 256 (default), or 512 and due to their correspondence in bytes they can also be referred as BRIEF-16, BRIEF-32 and BRIEF-644, respectively.

D. Oriented FAST and Rotated BRIEF (ORB)

As the name itself suggest, ORB [10] combines and extends on the concepts of FAST and BRIEF, reducing sensitivity to noise and having rotational invariance. The ORB detector is essentially a multi-scale FAST with orientation, while the ORB descriptor uses a learning process to determine the spatial arrangement of binary tests, decorrelating BRIEF features under rotational invariance. This makes the nearest neighbor search during matching less error-prone. The learning algorithm search for a set of 256 uncorrelated tests, which produce a 256 bit string, the ORB descriptor size.

E. Binary Robust Invariant Scalable Keypoints (BRISK)

Proposed by Leutenegger et al. [11], BRISK is a fast descriptor which uses symmetric sampling pattern (composed out of concentric rings) for intensity tests. The BRISK detector is based on the AGAST (Adaptive and Generic Accelerated Segment Test) detector [12], which is an extension of a faster performance version of the FAST detector. To describe the features, pairs of pixels around the interest point are separated into two subsets: short-distance and long-distance pairs. BRISK uses the long-distance pairs to estimate the patch orientation and the short-distance pairs to construct the descriptor itself through pixel intensity comparisons. BRISK descriptor is composed of a bit-string of length 512, i.e., a 64-dimensional feature vector.

F. Fast Retina Keypoint (FREAK)

Inspired by the human visual system, FREAK [13] uses a retinal sampled pattern for intensity tests. Similar to BRISK, FREAK applies the same AGAST feature detector. The FREAK descriptor is constructed by evaluating 43 weighted Gaussians at locations around the interest point, leading to 903 possible pairs. A learning algorithm similar to ORB is applied to find the 512 most relevant pairs and build the FREAK bit string.

III. Mid-level Feature Extraction through BossaNova Approach

Mid-level feature extraction aims at transforming local descriptors into a global and richer image representation of intermediate complexity [14]. The standard pipeline to get mid-level features can be broken into two steps: coding and pooling. The *coding* step quantifies the local descriptors according to a visual dictionary of k visual words, which is usually built by clustering a set of local descriptors (e.g., k-means clustering algorithm). The *pooling* step aggregates the codes obtained into a single feature vector.

In the Bag of Visual Words (BoVW) [15], [16], the most popular mid-level image representation, the coding step associates the local descriptors to the closest element in the visual dictionary (called hard-assignment coding), and the pooling takes the average of those codes (called average pooling). Since the pooling operation compacts all the information contained in the individually encoded local descriptors into a single feature vector, that step is critical for BoVW-based representations. In general, the objective of pooling is to summarize the information contained in the individually encoded descriptors into a single feature vector, preserving important information while discarding irrelevant detail [17].

Over the years, BoVW representation has been extended for both steps of coding and pooling [18], [19]. Avila et al. [19] introduced the BossaNova mid-level image representation. To the best of our knowledge, this is the first time that it is applied to fruit fly identification. In our experiments, we kept the default BossaNova parameter values the same as in [19]. Figure 1 shows the main pipeline using BossaNova on insect identification task.

IV. Experimental Settings, Results and Discussion

A. Dataset

The dataset used in this work is composed of 301 images and divided into three different categories: *A. fraterculus* (100), *A. obliqua* (101), and *A. sororcula* (100).

It consists of pictures of specimens reared from samples of fruit trees in experimental and commercial orchards in the state of São Paulo, Brazil, stored in the Department of Entomology and Acarology ESALQ, Piracicaba, SP, Brazil and in the Biological Institute, Campinas, SP, Brazil. It is important to recall that a 5-fold cross-validation protocol has been adopted in our experiments. Figure 2 shows examples of the three species used in this work.

B. Effectiveness Analysis

In this section, we have performed a comparative study among nine learning techniques: Multiple Layer Perceptron (MLP), Naïve Bayes (NB), Decision Tree (DT), Naïve Bayes Tree (NBT), k-Nearest Neighbor (kNN) with $k = \{1, 3, 5\}$, Simple Logistic (SL), and Support Vector Machine (SVM) using polynomial kernel. The implementation of the machine learning techniques are available in the WEKA[1] data min-

[1]http://www.cs.waikato.ac.nz/~ml/weka (As of July, 2017).

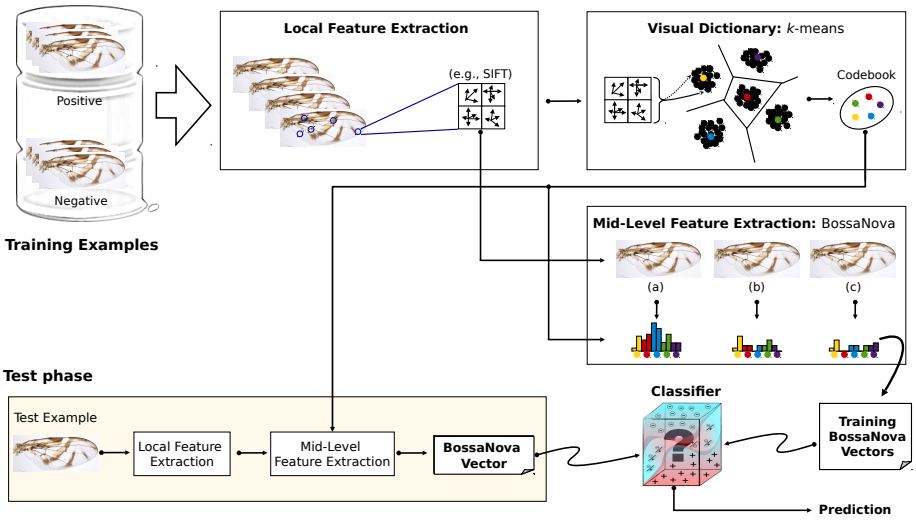

Fig. 1. The main pipeline of BossaNova. **Local Feature Extraction**: robust local descriptors (e.g., SIFT, SURF, BRIEF, ORB) are obtained from the detected features. **Mid-Level Feature Extraction**: BossaNova descriptors creates the feature vectors for the images using a visual dictionary (k-means with Euclidean distance is run over a sample local features, the final centroids are used as visual words). **Decision Model Training**: During the training-phase, the BossaNova vectors of annotated images are employed to train a decision model using a machine learning method. **Decision Model Prediction**: The trained model employs the BossaNova feature vectors of an image to predict on the positive or negative classes.

A. fraterculus *A. obliqua* *A. sororcula*

Fig. 2. Example of wings of each specie studied. Extracted from [20].

ing library. All machine learning techniques were used with default parameters which means we did not optimize them whatsoever.

Table I shows effectiveness results for all local descriptors and learning techniques. Furthermore, the BRIEF, BRISK, FREAK, ORB, F-SIFT and F-SURF local descriptors have been performed with FAST feature detector. SIFT and SURF descriptor used the original implementation, described in the Section II.

In the first experiment, we can observe that BRIEF and F-SIFT descriptors have achieved four of the best effectiveness results among nine released learning techniques (in blue). FREAK descriptor has achieved one best result using simple logistic (SL). Furthermore, we can observe that BRIEF descriptor achieved the best average accuracy (84.7%) with lower confidence interval (4.4).

In the second experiment, it possible to note that multi layer perceptron (MLP) technique has achieved seven of the best effectiveness results among eight local descriptors (in gray cell) released in this work. SVM technique has achieved one better effectiveness result with 90.4% of mean accuracy using ORB descriptor. In addition, MLP technique using F-SIFT descriptor was the best tuple (descriptor+learning technique) performed in this work with 94.7% of mean accuracy (in blue text and gray cell). Finally, we can verify that MLP and SVM techniques were the best learning techniques with average accuracy of 88.9% and 87.7%, respectively.

C. The Best Approaches

We also compared the best learning techniques for each local descriptor (rows in the Table I), BRIEF+MLP, BRISK+MLP, FREAK+MLP, ORB+SVM, SIFT+MLP, SURF+MLP, F-SIFT+MLP, and F-SURF+MLP. Furthermore, the baseline technique LCH+SVM proposed in [20] has been added in this experiment.

Figure 3 shows the effectiveness results among the best tuples (descriptor+learning technique) and the best baseline existing in the literature. Although F-SIFT+MLP (in blue) has achieved the best mean accuracy, when we compute the confidence interval with significance level of 0.05, it is possible to observe that there is no statistically significant difference among our seven approaches and the baseline from the literature LCH+SVM (in red). However, it is very important to note that the baseline achieved excellent effectiveness results by extracting color features from enhancemented images (e.g., segmentation and dilation operations) [20]. Our approaches have been applied on the original images from the dataset. Therefore, our approaches might be used in real-time systems for insect identification tasks with no the use of any image enhancement operation.

V. Conclusion

In this work, we proposed the use of a mid-level image representation approach for insect identification of three species of the genus *Anastrepha* using different local descriptors based

TABLE I

EFFECTIVENESS RESULTS (IN %) AMONG EIGHT LOCAL DESCRIPTORS AND NINE MACHINE LEARNING TECHNIQUES FOR A 5-FOLD CROSS-VALIDATION PROTOCOL. IN BLUE ARE THE BEST IMAGE DESCRIPTORS FOR EACH MACHINE LEARNING TECHNIQUE. IN GRAY CELL ARE THE BEST MACHINE LEARNING TECHNIQUES FOR EACH IMAGE DESCRIPTOR.

Descriptor	Machine Learning Techniques									Average	CI
	MLP	NB	DT	NBT	kNN1	kNN3	kNN5	SL	SVM		
BRIEF [7]	92.0	**73.5**	**79.8**	78.8	86.4	**90.0**	**88.1**	82.7	90.7	**84.7**	**4.4**
BRISK [11]	87.4	51.5	74.4	69.1	78.4	74.4	71.4	79.7	87.0	74.8	7.5
FREAK [13]	88.4	55.8	67.8	70.4	76.7	75.1	73.7	**84.0**	85.0	75.2	7.0
ORB [10]	90.0	61.8	75.1	74.4	86.7	84.4	82.4	82.1	90.4	80.8	6.3
SIFT [5]	84.7	53.8	62.5	61.8	67.2	68.4	68.5	75.1	84.4	69.6	7.1
SURF [6]	89.4	63.5	62.5	70.1	77.4	78.7	79.7	66.8	87.4	75.0	6.9
F-SIFT	**94.7**	62.1	76.7	**82.1**	**87.4**	85.1	84.4	82.4	**93.7**	83.2	6.7
F-SURF	84.7	49.2	65.8	66.1	76.1	74.1	72.1	80.4	83.4	72.4	7.7
Average	**88.9**	58.9	70.6	71.6	79.5	78.8	77.5	79.2	**87.7**		
CI	**2.4**	5.5	4.7	4.57	4.8	5.0	4.9	4.0	**2.5**		

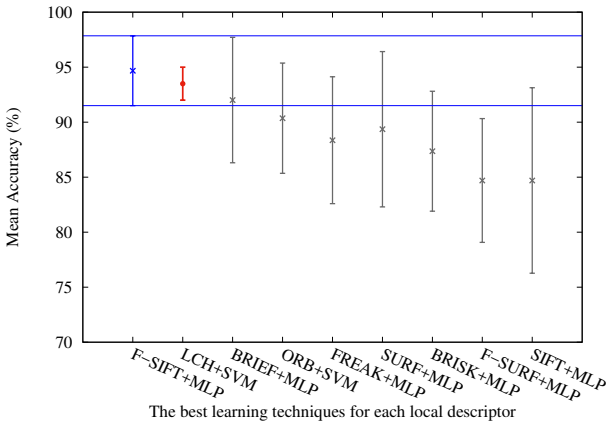

Fig. 3. Effectiveness results for each local image descriptor with 95% confidence interval (CI), i.e, a significance level of 0.05. In blue is MLP using BRIEF descriptor that has achieved the best mean accuracy.

on keypoints. Also, we performed two robust analysis to support the development of a real-time system for fruit fly identification. In the first, an effectiveness analysis among eight local descriptors and nine learning techniques was performed to verify the behavior of the tuples (descriptor+learning technique) in the fruit fly identification task. In this experiment, we observed that BRIEF and F-SIFT achieved the best results of mean accuracy among all of released local descriptors. Moreover, MLP and SVM techniques achieved to be the best learning techniques with higher average accuracy values and lower confidence interval. In the second experiment, we compared the best learning techniques for each local descriptor against the best state-of-the-art baseline from the literature. In this experiment, we verified that even though there are statistical differences among our approach based on mid-level image representation and baseline, our approach might be directly applied on the original images with no require any enhancement operation. Therefore, we conclude that this work will support the development of a real-time system for fruit fly specie identification of the genus *Anastrepha*.

ACKNOWLEDGMENT

This work is partially financed by CNPq Universal Project (408919/2016-7).

REFERENCES

[1] Zucchi, R. A., "Fruit flies in Brazil: *Anastrepha* species and their host plants and parasitoids," http://www.lea.esalq.usp.br/anastrepha/, 2008.

[2] M. K. Schutze, M. Virgilio, A. Norrbom, and A. R. Clarke, "Tephritid integrative taxonomy: Where we are now, with a focus on the resolution of three tropical fruit fly species complexes," *Annual Review of Entomology*, vol. 62, no. 1, pp. 147–164, 2017.

[3] B. A. McPheron, "Population genetics and cryptic species," *Area-wide Control of Fruit Flies and Other Insect Pests*, pp. 483–490, 2000.

[4] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE TPAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *CVIU*, vol. 110, no. 3, pp. 346–359, 2008.

[7] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *ECCV*, 2010, pp. 778–792.

[8] M. Agrawal, K. Konolige, and M. R. Blas, "CenSurE: Center surround extremas for realtime feature detection and matching," in *ECCV*, 2008, pp. 102–115.

[9] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *ECCV*, 2006, pp. 430–443.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *ICCV*, 2011, pp. 2564–2571.

[11] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *ICCV*, 2011, pp. 2548–2555.

[12] E. Mair, G. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *ECCV*, 2010, pp. 183–196.

[13] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *CVPR*, 2012, pp. 510–517.

[14] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *CVPR*, 2010, pp. 2559–2566.

[15] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477.

[16] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in *ECCV*, 2004, pp. 1–22.

[17] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *ICML*, 2010, pp. 111–118.

[18] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "BOSSA: Extended BoW formalism for image classification," in *ICIP*, 2011, pp. 2909–2912.

[19] ——, "Pooling in image representation: the visual codeword point of view," *CVIU*, vol. 117, no. 5, pp. 453–465, 2013.

[20] F. Faria, P. Perre, R. Zucchi, L. Jorge, T. Lewinsohn, A. Rocha, and R. da S. Torres, "Automatic identification of fruit flies (diptera: Tephritidae)," *JVCIR*, vol. 25, no. 7, pp. 1516–1527, 2014.