Assessment of Predictive Clustering Trees on 2D-Image-Based Ear Recognition

Žiga Emeršič[1], Peter Peer[1], Ivica Dimitrovski[2]

[1]*Faculty of Computer and Information Science, University of Ljubljana*
Večna pot 113, SI-1000 Ljubljana, Slovenia
[2]*Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje*
Ruger Boskovik 16, MK-1000 Skopje, Macedonia
E-mails: {ziga.emersic, peter.peer}@fri.uni-lj, ivica.dimitrovski@finki.ukim.mk

Abstract

In the last decade person recognition based on various biometric metrics have steadily been gaining on popularity. The same holds for machine learning approaches and various image classification and retrieval techniques. However, many techniques rely on distinguishing between significantly dissimilar images, which is often not the case in person recognition. Person recognition based on images relies on detecting minor differences and not global appearance of an image. To test if retrieval approaches based on bag-of-words fail in the task of biometric recognition we evaluated the following procedure. Ear images were used to extract Scale Invariant Feature Transform feature vectors. These vectors were then fed into forest of Predictive Clustering Trees, k-means and approximate k-means; and then compared to baseline system where only distances between plain descriptors are compared. While these methods have been proven to perform well in image with significantly different content, the results show that these methods do not perform well under the task of ear recognition.

1 Introduction

Biometric features that uniquely define persons can be obtained from multiple sources: sound, images, videos, smell and others. Arguably the most widely available and useful are 2D images of biometric modalities: images of face, ear, iris etc. Features can be acquired and represented from images in different ways. If we use and compare images as a whole we consider data that is not informative or even impacts distinctiveness negatively: images contain noise (varied illumination conditions, distortions during image acquisition etc.) and unwanted objects (hair, earrings etc.) or the target object is not entirely visible. Therefore we want to capture only relevant data, that distinguishes target objects as well as possible. Furthermore we want to store the data in a way that is most concise and easy to compare and then compare the acquired data as well as possible.

In this paper we compare baseline system of comparing plain description vectors to the following group: bag-of-visual-words representation using k-means clustering (and approximate k-means), Predictive Clustering Trees (PCT) [1–3] and combination of both. In both cases source vectors are obtained using Scale Invariant Feature Transform (SIFT) [4] descriptors. The goal was to evaluate how k-means and PCTs performs with the task of ear recognition where details are important. In [3] the authors report superior performance in both computational aspect as well as overall accuracy. However, the nature of the data they used is different to images intended for biometric recognition.

To the best of our knowledge ear recognition has never been applied to PCTs, whereas k-means was only used in combination with some holistic approaches such as in [5]. Multi-clustering search strategy was used on ears [6] with promising results. However, Cascaded Pose Regression (CPR) ear normalization technique was applied on ear datasets that are less challenging than data we used.

2 Methodology

Procedure consists of six major steps, visualized in Fig. 1: acquiring images, processing images, extracting features, three different types of image retrieval, additional post-processing of acquired feature vectors and finally performance evaluation. The procedure is as follows:

- Dataset acquisition: images were acquired from the internet.
- Image preprocessing: all images are transformed to gray-scale and resized to uniform size of 100×100 pixels.
- Features acquisition: feature vectors are calculated on grids with a step of 10 pixels. For feature description Scale Invariant Feature Transform (SIFT) [4] is used.
- Image retrieval: this step is skipped in the baseline experiments where descriptors are compared directly. However in other four cases the following methods are used: PCT, k-means, k-means with PCT, approximate k-means. After this the following two steps are taken: (1) Histogram calculation—each item in histogram represents the frequency of each word for each image and (2) Weights calculation—weights are calculated using td—idf scoring described in Section 2.
- Vectors normalization: all vectors are L^2-normalized.

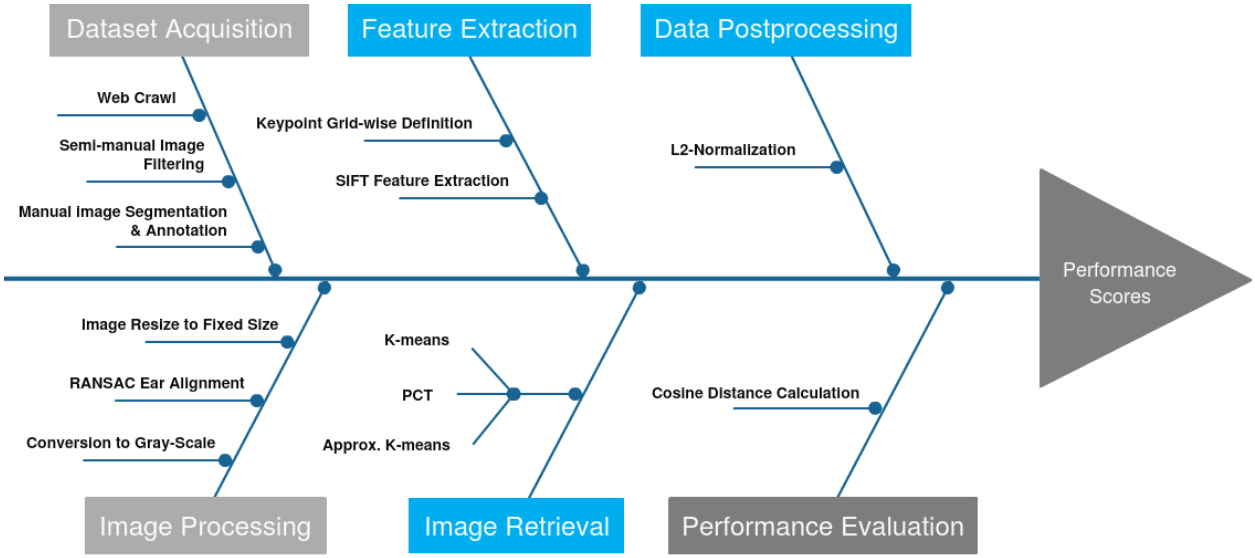

Figure 1: Diagram showing procedure.

- Distance measurement & performance evaluation: here Euclidean and cosine distances are used to compare vectors. The comparison serves for the final performance evaluation.

In order to achieve good recognition performance only selected areas need to be acknowledged and described. The field of selecting such areas (or keypoint detection) is beyond the scope of this paper, and the reader is referred to [7, 8]. In our experiments, described in Section 3, we used predefined grid of points instead of using keypoint detector (e.g. the detection part of SIFT). Another aspect is the selected areas' description. This can be done using one of many existing feature descriptors [9]. In this work we compare and asses performance on SIFT, more specifically Dense SIFT.

Scale Invariant Feature Transform—SIFT is widely used in object detection, texture classification, photo stitching (in combination with RANSAC), object classification etc., even though the algorithm is more than a decade old and patented. It still remains in the group of state-of-the-art feature detectors and descriptors. It was proposed by Lowe [4] and proved to be fairly robust to scaling and occlusions, specifically in ears as well [10–12].. The algorithm provides both detector and descriptor; however, here we focus on the descriptor part of the algorithm only. In our experiments we use dense grid as keypoints for feature extraction instead of using SIFT detector.

In our experiments we use Random Forests (RF) of Predictive Clustering Trees (PCT) [3, 13] which produce codebook on which td-idf weighting is then applied. Visual codebook is built by clustering all descriptor vectors from all images. Each cluster then represents a visual word, while all words present the visual dictionary. After that each image descriptor is paired with the visual word from the codebook. This results in images being described by a n-dimensional histogram, where each dimension corresponds to one visual word, and the value to number of descriptors that match that visual word. However, the prerequisite here is that there is a sufficient similarity/repeatability between the descriptors: if no or very small amount of descriptors repeat, this means, that final histograms will have non-distinctive shapes. This is the core problem of using PCT for a task where small details in images are important. After that step the images are ranked using tf-idf scores.

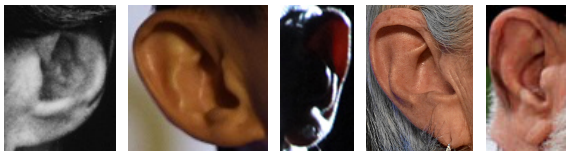

Figure 2: Five sample images from the AWE dataset representing large variability in images.

Term frequency (tf) is defined as a number of occurrences of a term in a document, divided by number of all terms in a document. Inverse document frequency (idf) is defined as $log(\frac{D}{d})$, where D is number of all documents, and d a number of documents that contained the searched term.

We have also used k-means [14] and approximate k-means clustering algorithms [15]

3 Experiments

We used AWE ear dataset which is freely available from `awe.fri.uni-lj.si` and was first presented in [16]. It contains 1000 cropped ear images of 100 persons, with 10 images per person. The images contain high variation of poses, illumination conditions, occlusions, image quality and image sizes—the database is challenging. Sample images from the dataset are shown in Fig. 2.

DSIFT parameters were experimentally set. Grid point size 10 pixels, patch size 16×16 pixels and bin size 8. During the experiments we noticed that few, if any, SIFT descriptors completly match between image instance of

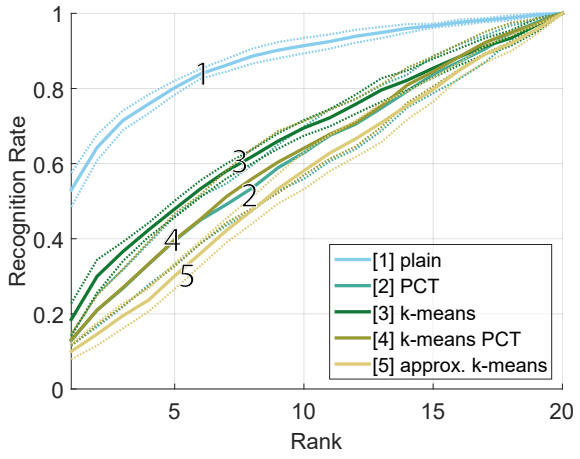

Figure 3: Cumulative Match Curve (CMC) showing that comparing plain SIFT feature vector outperforms other methods.

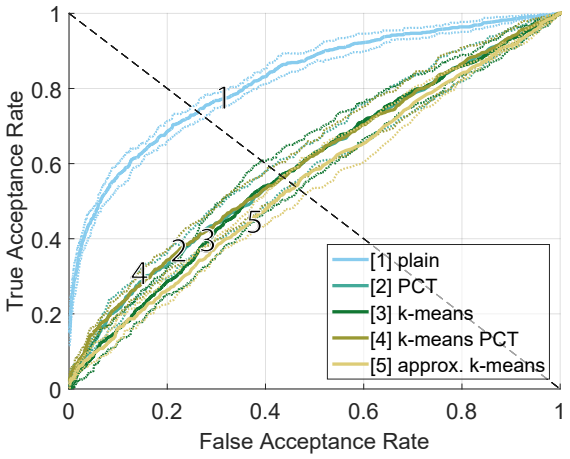

Figure 4: ROC curve showing that comparing plain SIFT feature vector significantly outperforms other methods. Intersections with dashed diagonal line represent equal error rate.

Table 1: Results of identification experiments showing plain SIFT descriptors outperforming other approaches in Rank-1 performance.

Method	**Rank-1 [%]**
[1] plain	53.00 ± 4.51
[2] PCT	12.90 ± 1.28
[3] k-means	18.50 ± 4.30
[4] k-means PCT	13.10 ± 1.59
[5] aprx. k-means	10.00 ± 2.10

Table 2: Results of verification experiments, showing that in both measures Equal Error Rate (EER) and Area Under the ROC Curve (AUC) methods fail.

Method	**EER [%]**	**AUC [%]**
[1] plain	26.30 ± 3.64	82.22 ± 1.94
[2] PCT	43.18 ± 3.33	59.62 ± 3.61
[3] k-means	43.25 ± 5.96	58.24 ± 4.33
[4] k-means PCT	43.73 ± 3.78	59.83 ± 3.61
[5] aprx. k-means	46.07 ± 3.05	55.41 ± 3.07

the same class. This means that using SIFT detector and then matching features results in only few matching descriptors at best—rendering this approach close to useless. After the descriptors were acquired PCT, k-means and approximate k-means procedures followed. In all three cases CLUS implementation was used. CLUS can be freely downloaded for educational purposes from `dtai.cs.kuleuven.be/clus/`. For random forest of PCTs we used tree ensembles of 4 PCTs, with maximum tree depth of 14. Random forest of PCTs was used to improve robustness and to improve discriminative, as suggested by [3]. For k-means 512 clusters were used. This number was set by an expert in accordance with the nature of the data used. This descriptors are obtained in a regular bag-of-visual-words setup. All the DSIFT descriptors from all the images were clustered and then histograms were calculated based on these clusters for each image.

4 Results & Discussion

Here we report identification and verification experiments. Identification experiments are here reported with Cumulative Match-score Curves (CMCs) [17]. All results are given with standard deviation over 5 folds of 5-fold cross validation, where $1/5$ of data is used as a test set. Verification experiments we report with Receiver Operating Characteristics (ROC) curves [17]. For reporting performance of verification experiments we use the following measures: Equal Error Rate (EER), Verification Rate @ 1% False Acceptance Rate and Area Under the Curve (AUC). Results of identification experiments are shown in Fig. 3 and in Table 1. Results show that all methods fail compared to plain descriptors comparison. Random classifier would achieve 10% rank-1 recognition rates. This means that approximate k-means performs randomly and therefore removes all information from the data—fails to capture any underlying logic in data. PCT with 12.9% rank-1 recognition rate perform slightly better, but is outperformed with k-means—18.5% and k-means PCT 13.1%. Plain SIFT descriptors outperform all of the listed methods significantly with 53.0%. However, when standard deviation is taken into the account, all methods (with the exception of plain SIFTs) perform comparably.

Verification results are shown in Fig. 4 and in Table 2. Conclusions here are the same as for the identification experiments. All methods fail compared to plain SIFT description vectors which achieve 26.3% EER, 28.4% V@1% and 82.2% AUC. Again, there is significant margin between plain SIFTs and other approaches: with EER values around 43% with standard deviation taken into the account, 0 to 6% V@1% and AUC below 60%.

In Fig. 3 and Fig. 4 dashed lines represent standard deviation over 5 folds. For more in-depth analysis of performance curves and measures the reader is referred to [17, 18].

K-means clustering gives best results when data sets are well separated from each other. Due to our difficult dataset this was not the case, which could explain to a large extend the bad performance.

5 Conclusion

The experiments have shown that none of the approaches outperform plain image descriptors comparisons and thus confirmed the initial hypothesis. We have shown that the bag-of-visual-words approach is not suited for this kind of problems in which we need to distinguish between very specific details, which get lost in the process and overall image appearance is considered. Furthermore, in the clustering phase the discriminative information is additionally removed and all the positional location of SIFT descriptors is lost. However, it would be feasible to use these approaches in biometric modality detection. Localizing and segmenting ears is much closer to tasks where PCTs and k-means have been proven to work well.

Acknowledgement

Special thanks to Dragi Kocev and Sašo Džeroski for the initial idea, all the knowledge shared, and all the help. Without them this paper would not exist.

References

[1] H. Blockeel and L. De Raedt, "Top-down induction of first-order logical decision trees," *Artificial intelligence*, vol. 101, no. 1, pp. 285–297, 1998.

[2] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Tree ensembles for predicting structured outputs," *Pattern Recognition*, vol. 46, no. 3, pp. 817–833, 2013.

[3] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Fast and scalable image retrieval using predictive clustering trees," in *International Conference on Discovery Science*. Springer, 2013, pp. 33–48.

[4] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] C. Jayachandra and B. S. Patel, "An ear recognition approach using edge detection," *Technology*, vol. 1, no. 1, pp. 38–45, 2013.

[6] A. Pflug, C. Busch, and A. Ross, "2D ear classification based on unsupervised clustering," in *Proceedings of the International Joint Conference on Biometrics*. IEEE, 2014, pp. 1–8.

[7] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, no. 10, pp. 1771–1787, 2008.

[8] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and trends® in computer graphics and vision*, vol. 3, no. 3, pp. 177–280, 2008.

[9] A. Pflug and C. Busch, "Ear biometrics: a survey of detection, feature extraction and recognition methods," *Biometrics*, vol. 1, no. 2, pp. 114–129, 2012.

[10] J. D. Bustard and M. S. Nixon, "Toward unconstrained ear recognition from two-dimensional images," *Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 40, no. 3, pp. 486–494, 2010.

[11] B. Arbab-Zavar and M. S. Nixon, "Robust log-gabor filter for ear biometrics," in *Proceedings of the International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[12] K. Dewi and T. Yahagi, "Ear photo recognition using scale invariant keypoints." in *Proceedings of the Computational Intelligence*, 2006, pp. 253–258.

[13] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Ensembles of multi-objective decision trees," in *European Conference on Machine Learning*. Springer, 2007, pp. 624–631.

[14] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[16] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: More that a survey," *Accepted to Neurocomputing*, 2016.

[17] A. Jain, A. Ross, and K. Nandakumar, *Introduction to biometrics*. Springer Science & Business Media, 2011.

[18] B. DeCann and A. Ross, "Relating ROC and CMC Curves via the Biometric Menagerie," in *Proceedings of the International Conference on Biometrics: Theory, Applications and Systems*. IEEE, 2013, pp. 1–8.