Improving handwritten signature-based identity prediction through the integration of fuzzy soft-biometric data

Márjory Da Costa-Abreu and Michael Fairhurst
School of Engineering and Digital Arts, University of Kent
Canterbury, Kent CT2 7NT, UK
{*m.c.d.c.abreu, m.c.fairhurst*}*@kent.ac.uk*

***Abstract*—Automated identification of individuals using biometric technologies is finding increasing application in diverse areas, yet designing practical systems can still present significant challenges. Choice of the modality to adopt, the classification/matching techniques best suited to the application, the most effective sensors to use, and so on, are all important considerations, and can help to ameliorate factors which might detract from optimal performance. Less well researched, however, is how to optimise performance by means of exploiting broader-based information often available in a specific task and, in particular, the exploitation of so-called "soft" biometric data is often overlooked. This paper proposes a novel approach to the integration of soft biometric data into an effective processing structure for an identification task by adopting a fuzzy representation of information which is inherently continuous, using subject age as a typical example. Our results show this to be a promising methodology with possible benefits in a number of potentially difficult practical scenarios.**

***Keywords*-Soft-biometrics data representation; Fuzzy age; handwritten signature-based identification**

I. Introduction

The use of biometrics-based systems in the identification of individuals is now increasingly widespread and well established. Many approaches have been explored and therefore many issues have been highlighted for their importance in specific practical implementations, such as: choice of classification and fusion technique [1], choice of modality(ies) [2], possible use of additional information to be considered [3] and so on.

The availability of multiple information sources for biometric data processing can suggest various different strategies by means of which to achieve enhanced performance. The commonly adopted protocol follows some established principles, for example:

- It is generally accepted that the use of multimodal solutions offers a more reliable, flexible and secure approach than where only specific individual modalities are used [1].
- The techniques and devices which perform the data collection should be given special attention as should the characteristics which will be extracted and used in the identification/verification process [4].
- The use of increasingly powerful, but sometimes complex fusion techniques is also a very popular topic of research, mainly directed towards the minimisation of equal error rates (EER) [5].

Beyond these system-level considerations, other approaches can also be adopted. One option is to exploit further information which is available about an individual, but which is not necessarily unique. Such "soft-biometric" information (a typical example is subject age) can be integrated into a biometric solution to aid performance enhancement. While an established option, this approach has been adopted much less frequently than might be expected.

Soft-biometric information, however, can be a very powerful ally of the system designer when developing a biometric-based system. It is normally relatively easy and very cheap to acquire (it generally does not need another sensor for its capture), it is not invasive (the user can give most required soft-biometric information by responding to simple questions readily and often routinely), it is usually inclusive (everybody has this information) and, it is relatively simple to incorporate in the identification process [3].

Any information which is a characteristic of an individual and can be extracted from (or given by) him/her can be used as a soft-biometric as long as it is measurable and/or categorisable and, unlike true biometric information, need not be unique to that individual. Typical examples include age [6], gender [6], handedness [3], height [6], percentage of body fat [7] and so on, and the representation of such data for appropriate processing can be very simple.

Because the potential benefits of incorporating soft-biometrics into the identification process are often overlooked, the work that can be found in the literature has perhaps not explored fully the merits of this source of information. For example, age has been used as a factor in identification tasks, but this is usually represented either as a single numerical value or by considering discrete broader age bands. However, age is a continuous variable, where significant differences attributable to this factor are generally not expected to occur as sharp discontinuities at the boundary between one age value and another. Thus, by using a typical discrete representation, some loss or degradation of information is almost unavoidable. Despite this, some initial

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.221

results based on defining discrete age bands in handwritten signature identification have proved to be encouraging [3].

In this paper, we will investigate the impact of presenting age in a continuous representation (by using a fuzzy age representation), as applied in the handwritten signature modality as an illustrative task. When comparing our results with similar studies found in the literature, we are able to show the potential value of using this alternative fuzzy age representation.

II. Categorisation of soft-biometrics

The choice of which soft-biometrics to use for optimality can be as difficult as the choice of which full biometrics to use, and is very much dependent on the population characteristics and on the modalities which will be used in the system. Furthermore, how to characterise the data provided may also be an issue. The most common ways of categorisation of the soft-biometrics, such as those listed above can be described as follows:

- **Gender:** Subject gender is a very easy item of information to categorise, since *male* and *female* naturally define the two obvious possible categories.
- **Handedness:** The categorisation of "handedness" is also a very easy process. Naturally, *right-handed* and *left-handed* are used as the two possible categories. However, ambidextrous writers may also be considered as a separate category. Arguably, handedness can indeed be considered as a continuous soft-biometric. The ambidextrous user can provide the proportion of use in each hand and this information could be very valuable as an identifier for this small proportion of the population.
- **Age:** This is a potentially more complex area. In a very straightforward way, age can be represented directly as an absolute number (in years). However, it should be noted that such a representation is, in fact, an approximation, because age is a continuous variable. A typical alternative approach divides the population into age bands creating a categorical feature which will indicate to which group each user belongs. The number of groups (typically taken as, for example, *young*, *adult* or *elderly*, < 25, $25 - 60$, > 60) is still very empirical, as there has not been an extensive study of the effects of age in biometrics and, therefore, there is incomplete knowledge of exactly how subject age affects each different modality.
- **Height:** This soft-biometric has been represented as an absolute number that is incorporated either as a selector of user sub-groups or as an extra feature to be added in the feature vector of the biometric modality.
- **Iris colour:** The different possible eye colours may be created as possible categories (eg. *blue*, *green*, *brown*, *black*, etc) to be simply adopted as an additional categorical feature.

Even though the use of soft biometrics to enhance identification performance is not new and a variety of investigations have been reported in the literature (several ways of using this information can been found, such as adding this information as an extra feature, using it as a feature filter, selecting the best classifiers [3], and so on), most of the work reported uses the soft-biometric information in a very simple way and, therefore, its full potential may be missed.

For example, using the absolute value of age, percentage of body fat or height may actually have a negative impact on identification, because these are continuous variables being treated as absolutes. Similarly, iris colour can also, strictly speaking, be considered a continuous feature because it can be very hard to identify the small variations of eye colour which commonly occur.

A single solution is unlikely to encompass the whole spectrum of different soft-biometric categories. Nevertheless, the use of fuzzy logic has the potential of representing more accurately the categories that can be classified as continuous. Section III will show how the *fuzzyfication* of age can improve the exploitation of soft-biometric information.

III. Fuzzyfication of age

The age of an individual is clearly a very important item of information and can be especially relevant in some situations (for example in some legal transactions). In practical terms too, the age of a subject can have an impact on the overall design of a system as well as the planning of the enrolment protocol and update of user information. In this sense, it is important to know which modalities and what information it will be necessary to collect in the first enrolment process as well as the frequency with which re-enrolment will be necessary. Nevertheless, although using broad age groups reflects the biological differences among the biometric samples generated in each different group, it still relies on a "sharp" division, which may be misleading in specific cases.

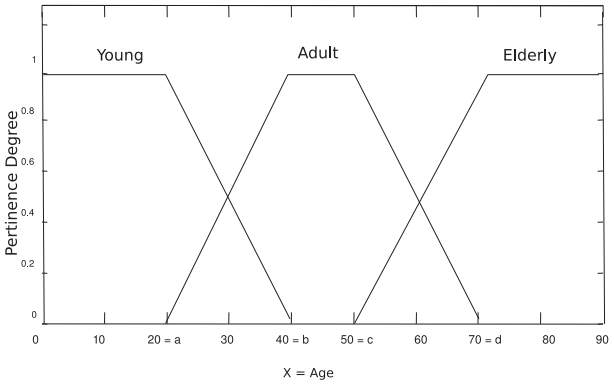

Figure 1. Fuzzyfication of the age

When the soft biometric adopted is age, the two conventional ways in which this information is incorporated in a biometric-based system are either to use an absolute age parameter or to group the users into different, broader, age bands. However, age is not a "sharp" characteristic and the way people age may be different in each age group and, of course, in different modalities. In order to retain this idea of the continuous nature of this information source, we propose the use of fuzzy groupings to represent age. Retaining, for example, the idea of using three basic age groupings, our approach proposes that each user will have a *degree of membership* of each of the three groups, allowing a more flexible representation of the effects of age in the classification process. A Mamdani rule aggregation method is used to represent the age groups [8] using a simple Trapezoidal function (which can be seen in Equation 1) that is defined by a lower limit a, an upper limit d, a lower support limit b, and an upper support limit c, where $a < b < c < d$.

$$\mu_x = \begin{cases} 0 & \text{if} \quad (x < a) or (x > d) \\ \frac{x-a}{b-a} & \text{if} \quad a \leq x \leq b \\ 1 & \text{if} \quad b \leq x \leq c \\ \frac{d-x}{d-c} & \text{if} \quad c \leq x \leq d \end{cases} \tag{1}$$

In this approach, *degrees of pertinence* will be given to the three different age groups with respect to each user. There will be three Trapezoidal functions, one for each "age band", according to the following values:

- *Young*: $a = 0$, $b = 0$, $c = 20$ and $d = 40$
- *Adult*: $a = 20$, $b = 40$, $c = 50$ and $d = 70$
- *Elderly*: $a = 50$, $b = 70$, $c = 100$ and $d = 101$

As an example, if the user is 30 years old, the values assigned to each fuzzy set would be: Young = 0.5, Adult = 0.5 and Elderly = 0.

By using a fuzzy logic approach, the representation of a user's age will be more commensurate with reality and the intrinsic variations in each individual ageing process. Also, the ageing process can be very different according to the biometric modality, while simply choosing fixed age bands will clearly disregard their particularities (eg. The iris modality suffers less as a result of the ageing effects than, say, the handwritten signature).

As the potential importance of soft-biometric information has apparently not been fully exploited to date (perhaps because this kind of data is not considered distinct enough to be used in security-based applications), we use the handwritten signature (often considered a comparatively "weak" modality) as our biometric, in order to show how intelligently integrating soft-biometric data to enhance performance can open up new opportunities in modality choice and system flexibility in practical situations. Moreover, as there has been some work already reported using the handwritten signature supplemented with age information [3], this will provide a useful comparative basis for analysing our results.

IV. Technical specifications

There are relatively few investigations reported which can help to establish a clear view about how effectively and efficiently to incorporate soft-biometrics into biometrics-based processing. The most common approach is simply to add such information as an extra input feature to the traditional feature vector, and this is therefore the technique we have implemented in this work (based on the work presented in [3]).

In order to analyse the real impact of the different representations of the age feature in the handwritten signature identification process, three different representational models were implemented (the two cited in Section II and the one presented in Section III) as follows:

- *Absolute age:* Here, age was represented as a single feature (positive integer) that will simply record the age of the user in years. For example, if the user is 35 years old, the value of the feature *absolute-age* will be 35.
- *Age bands:* Here, age was represented as three binary features (< 25, $25 - 60$ and > 60). The values of each of these features can only be 0 or 1 (and two of these will always be 0 and the third will be 1) where the feature with the value 1 is the age interval to which the user belongs. For example, a 35 year old user would have as age-related features: $< 25 = 0$, $25 - 60 = 1$ and $> 60 = 0$.
- *Fuzzy age:* Finally, in this representation the age will have *pertinence degrees* to all three possible "age bands" (we use three different age functions to be consistent with what is typically found in the literature). These numbers were calculated using the approach presented in Section III.

For comparison, we have adopted the BioSecure database (the same handwritten signature database used in [3]). This database is part of a multimodal database, where each of 79 users provided their information in two sessions. The handwritten signature contains 25 samples for each subject, where 15 are samples of the subject's true signature and 10 are attempts to imitate another user's signature. In this investigation we have used only the 15 genuine signatures of each subject of both sessions (30 samples per user in total). The data were collected using an A4-sized graphics tablet with a density of 500 lines per inch. The 21 representative biometric features were extracted from each signature sample and were chosen to be representative of those known to be commonly adopted in signature processing applications. All the available biometric features are used in the classification process as input to the system. The database provides the user's age which we will adopt in the three ways presented previously.

In order to obtain a broader view of the potential of these new techniques, we have chosen to carry out some experimentation also with an alternative database (containing signature samples collected in a retailing outlet at Hedge End, near Southampton in the UK, hence designated the "Hedge End" database and unfortunately not publicy available). The capture environment of this database was a typical retail outlet, providing a more real-world scenario in which to acquire representative data, in which 359 volunteers, from a cross-section of the general public, took part in the data collection. There are 7428 signature samples in total, where the number of samples from each individual varies between 2 and 79. The data were collected using an A4-sized graphics tablet with a density of 500 lines per inch. Also, 21 features were extracted and the same age groups were used.

In both databases, all the demographic information was verified during the acquisition process and is genuine. The same range of different individual classifiers used in [3] were selected: Fuzzy multi-layer Perceptron (FMLP) [9], Optimised IREP (Incremental reduced error pruning) (JRip) [10], Support vector machines (SVM) [11] and K-Nearest neighbours (KNN) [12] with a ten-fold-cross-validation training scheme [13].

V. Discussion

The principal issue addressed in this paper is the impact of using a more sophisticated representation of soft-biometric information (in our illustrative case, age), in improving the identification of individuals from their signature, and we will analyse the error rates for the individual age groups adopted (< 25, $25 - 60$ and > 60). These results can be seen in Figures 2 and 3 for the BioSecure and Hedge End databases respectively. In a comparison with work previously reported in the literature these results raise some interesting considerations, particularly with respect to the BioSecure data, as follows:

- Clearly, the use of the *Absolute age* representation produces the worst results. This is hardly surprising, and reflects the fact that a single number representation of age is not sufficiently subtle, and is not on an appropriate timescale to capture changes in human characteristics related to the complexities of the ageing process.
- When the individual error rates are obtained by representing age with a broader set of *age bands*, it can be seen that the least affected group is the < 25 group, followed by the > 60 group, while the $25 - 60$ group generates the lowest error rates. This happens simply because the population is unevenly distributed and the errors are proportional to the number of users in each group (34% for the < 25 band, 59% for the $25 - 60$ band and 7% for the > 60 band).
- The results using the *Age bands* representation are very different from the results using the *Absolute age* representation. The error rate is seen to decrease to approximately half, which in itself shows that using a more continuous representation of age can improve the accuracy of the system.
- Another very interesting change is the individual error distribution across the age population is that the highest error rate is generated by the < 25 group followed by the > 60 group and then by the $25 - 60$ group, even though the majority of the population is in the middle group.
- Finally, the error rates obtained using the *Fuzzy age* representation are the lowest of the three cases. Also, it is important to notice that the difference among the individual error rates associated with the different groupings is less than 0.5%, which indicates that the technique works relatively uniformly across the diversity of the population.
- The highest individual error rate with respect to the *Fuzzy age* representation occurs in the < 25 group. It might be suggested that this is because in this age group the signature is relatively unstable and is still forming, and that this is highlighted by the use of a more accurate representation of the subject age.

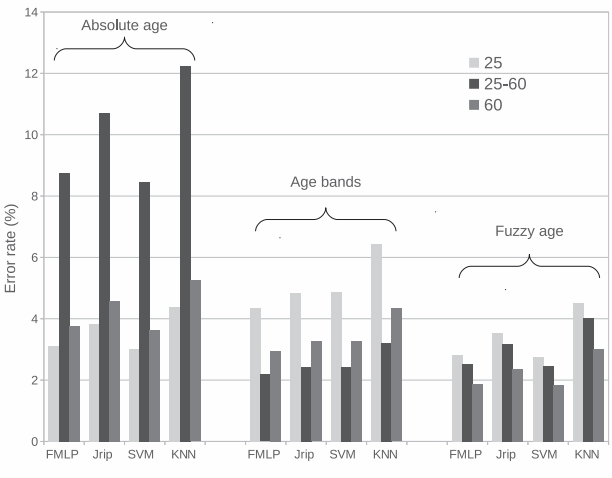

Figure 2. Age relative error percentages for the BioSecure database

These results, however, clearly provide practical hard evidence that by using a more appropriate and meaningful representation of age as a soft biometric, the impact on error-rate performance in a biometric recognition task can be considerable, leading to improvements of around 50%.

As already noted above, the BioSecure database was acquired in a relatively controlled environment, with all the users asked to give the same number of samples. The Hedge End database, on the other hand, was acquired in a fundamentally different and much more "real-world"

operational environment. The data available in this case provide a better representation of the sort of data generated in most real-world scenarios. Also, the nature of the data collection exercise in this case meant that users typically provided a different number of samples, making the training process much more difficult. The age population distribution for this database is 24% (< 25), 43% ($25 - 60$) and 33% (> 60).

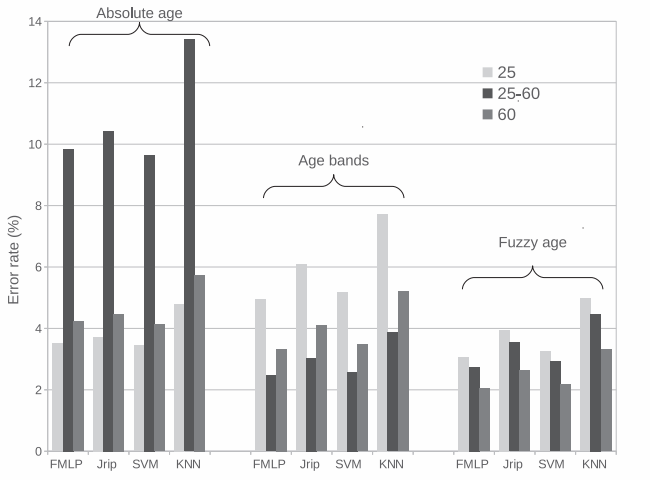

Figure 3. Age relative error percentages for the Hedge End database

When comparing the results of the experiments using the Hedge End database with those using the BioSecure database a very similar error rate behaviour can be seen. Also, the error rates among the three different age representations decrease by half as well as showing exactly the same behaviour for the individual age groups for each different age representation approach. This is very encouraging and again, it is a strong indicator that by using a rather more sophisticated representation of the soft-biometric data, the performance of a biometric-based system can be improved greatly.

For an overall and more detailed accuracy comparison, Table I shows the error rates (and standard deviation) for the individual classifiers for both databases. It is interesting to note the decrease in the standard deviation which can be brought about only by the changing the age representation. It is possible that this happens because the addition of the subject age as an extra feature in itself can have an impact in the user identification process and should be considered, indeed, as representative information.

When comparing the individual performances of the classifiers used, some further conclusions can be drawn, as follows:

- The classifiers performed generally better using the BioSecure database (lower error rates and standard deviations), which would be expected because this database was compiled under more favourable conditions. Nevertheless, this difference is not substantial and decreases when the age representation changes. This is also a very interesting fact and indicates that by only choosing representative features, problems in unbalanced databases can be overcome.
- When comparing the classifiers using the *Fuzzy age* representation and the same database, performing the t-test reveals that only the KNN classifier is statistically inferior to the others with p-values of 2.94E-035 (FMLP vs KNN), 3.01E-016 (Jrip vs KNN), 2.66E-042 (SVM vs KNN) for the BioSecure database and 4.00E-028 (FMLP vs KNN), 1.56E-010 (Jrip vs KNN), 3.95E-028 (SVM vs KNN) for the Hedge End database. In this case, even if we had used a confidence level of 99%, we would have observed similar results. These results show that not only is the classifier choice is important in the classification process, but the way the data is represented (in our case, incorporating age information represented in a more meaningful and appropriate way) made the choice of the classifiers to be adopted much less sensitive).
- Comparing these individual results with other studies of age-based soft-biometric enhancement reported elsewhere in the literature ([3], for example) shows a great improvement achieved in the EER (of around 30%).

Absolute age	BioSecure	Hedge End
FMLP	17.62±3.77	16.58±4.51
Jrip	19.13±5.96	18.62±4.56
SVM	15.13±3.45	17.23±4.58
KNN	21.86±4.36	23.95±4.78
Age bands	BioSecure	Hedge End
FMLP	11.49±3.61	10.75±3.91
Jrip	10.51±3.89	13.21±3.22
SVM	10.57±2.61	10.23±3.54
KNN	13.98±2.83	16.79±4.06
Fuzzy age	BioSecure	Hedge End
FMLP	7.24±1.94	7.86±2.04
Jrip	9.08±1.87	10.14±2.34
SVM	7.03±1.54	8.34±2.06
KNN	11.54±2.07	12.77±2.27

Table I
INDIVIDUAL ERROR RATES AND STANDARD DEVIATION

It is often to be expected that the simplest solutions for biometric-based systems implementation will return poorer levels of performance than can be achieved with much richer feature sets available, or the use of many modalities, or

the use of very sophisticated fusion techniques and so on. However, the results reported in this paper show that making use of a very simple and easily acquired supplementary information source can greatly enhance system performance provided that the information is used appropriately and imaginatively.

Of course, in this work we have so far only considered one soft-biometric category and we have only applied this new concept to enhance handwritten signature-based identification, but the results we have presented are extremely encouraging, and we will extend the scope of our investigation in future work.

VI. Conclusions

This paper has presented a novel approach to the effective representation of soft-biometric information (specifically in our case, age), by representing the available data in a way that can cover all the singularities which can be attached to this very rich, yet perhaps under-used source of identity information.

The improvement in accuracy which is achievable by using a fuzzy representation of a continuous human characteristic, as described in this paper, suggests that this new idea may be a very valuable and more widely applicable benefit of an approach which seeks to exploit the availability of soft-biometrics as a means of enhancing performance. In this case, such an approach provides the opportunity for significant enhancement in a scenario which has important practical implications yet which is often especially limited by the inherent nature of the task domain.

We understand that this is not the only way of using such information, but we believe that by incorporating demographic information, such as fuzzy age (as a specific example of more general soft biometric information), gives the classification process greater reliability, because the ageing process is not sharp and should not be considered as an all-or-nothing factor relative to each age group. Also, our results indicate that choosing the correct representation of the features in general can make the choice of the classifiers easier.

It is important to highlight that the novelty is our new idea of fuzzy representation of a soft biometric such as age (which has not been reported previously), and it is known that using discrete soft biometric categories in such situations presents problems of choosing boundaries. This is important because we demonstrate that our approach can positively impact on performance in the example considered, but may well have an important role to play in other applications with fundamentally different soft-biometric sources.

References

[1] M. Fairhurst and M. Abreu, "Balancing performance factors in multisource biometric processing platforms," *IET Signal Processing*, vol. 3, no. 4, pp. 342–351, July 2009.

[2] S. Dass, Y. Zhu, and A. Jain, "Validating a biometric authentication system: Sample size requirements," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1902–1319, 2006.

[3] M. Abreu and M. Fairhurst, "Enhancing identity prediction using a novel approach to combining hard- and soft-biometric information," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. PP, no. 99, pp. 1–9, 2010.

[4] G. Chetty and M. White, "Multimedia sensor fusion for retrieving identity in biometric access control systems," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 6, pp. 26:1–26:21, November 2010.

[5] M. Fahmy, A. Atyia, and R. Elfouly, "Biometric fusion using enhanced svm classification," in *The International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, ser. IIHMSP 2008, August 2008, pp. 1043–1048.

[6] J. Jain, S. Dass, and N. Karthik, "Soft biometric traits for personal recognition systems," in *The 1st International Conference on Biometric Authentication*, ser. Lecture Notes in Computer Science, 2004, pp. 731–738.

[7] H. Ailisto, E. Vildjiounaite, M. Lindholm, S. Mäkelä, and J. Peltola, "Soft biometrics-combining body weight and fat measurements with fingerprint biometrics," *Pattern Recognition Letter*, vol. 27, no. 5, pp. 325–334, 2006.

[8] K. Damghani, S. Sadi-Nezhad, and M. Aryanezhad, "A modular decision support system for optimum investment selection in presence of uncertainty: Combination of fuzzy mathematical programming and fuzzy rule based system," *Expert Systems with Applications: An International Journal*, vol. 38, pp. 824–834, January 2011.

[9] S. Mitra and S. Pal, "Logical operation based fuzzy mlp for classification and rule generation," *Neural Networks*, vol. 7, no. 2, pp. 353–373, 1994.

[10] J. Furnkranz and G. Widmer, "Incremental reduced error pruning," in *Proceedings the 11st International Conference on Machine Learning*, ser. ICML 1994, New Brunswick, NJ, 1994, pp. 70–77. [Online]. Available: http://www.ai.univie.ac.at/ juffi/publications/ml-94.ps.gz

[11] C. Nello and S. John, "An introduction to support vector machines and other kernel-based learning methods," *Robotics*, vol. 18, no. 6, pp. 687–689, 2000.

[12] A. Arya, "An optimal algorithm for approximate nearest neighbors searching fixed dimensions," *Journal of ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[13] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.