International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184 Volume 4, Number 5(2015), pp.244-258

http://www.meacse.org/ijcar

Identical Twins Face Recognition System Based on Time Series Motif Discovery

Mahmoud A. Elgamal, Esam A. Khan and Sameer M. Shaarawy
The Custodian of the Two Holy Mosques Institute for Hajj and Omra Research, Umm Al -Qura University, Saudi Arabia

Abstract

In the last few decades the birth rate of twins has been increased, thus the need for an accurate biometric system to precisely determine the identity of a person who has an identical twin became of great interest especially in criminal cases. Recent researches has showed the performance of automatic face recognition technology fails drastically in case of identical twin siblings compared with other unrelated persons. In this paper, we propose a new technique for identifying identical twins using their time series which is unique feature for image. The characteristic time series dimensionality reduced by using Piecewise Linear Approximation(PLA) method. A recognition and matching experiment conducted on an identical twins of high-similarity correlation coefficient score 0.96 and the methodology was able to distinguish between them even with fewer number of features; 8-features.

Keywords: Time Series Motif, Face Detection, Dimensionality Reduction, Piecewise Linear Approximation.

1 Introduction

The increase of multiple births in the last few decades associated with the increase in the use of fertility therapies and the older age of childbearing[17]. The twin birth rate increased at an average rate of 3% between years 1990 and 2004[17]. This increase has created a big demand for biometric identification systems, that can accurately determine twin's identity. Recognition of facial images of identical twin siblings poses a considerable challenge for any face recognition algorithm because of the strong similarity between the face images. Some researches has showed that the performance of automated face recognition systems fails drastically in case of twin images compared with unrelated persons [14]. The degradation is shown to be far more drastic for face than other biometrics like, iris and fingerprints[17].

Recent studies, have proved that facial recognition is also advantageous in distinguishing identical twins[11, 15, 22, 28].

The concept of time series motifs was first proposed in 2002 by J. Lin[12], at the same time, clear descriptions and definitions of the related concepts about time series motifs was given in details; e.g., k-motifs [4], the trivial match and son. subsequently, more researchers begin to focus on the study of time series motifs mining. In recent years many sophisticated papers on the topic were published in top journals and conferences, e.g., Knowledge Discovery and Data Mining (KDD) journal[2], The Very Large Data Bases(VLDB) journal[5], IEEE International Conference on Data Mining(ICDM) [13], ...etc. Furthermore, research results have also been applied in medicine, environmental studies[3], biology[1], telemedicine[7] as well as weather prediction[12] and other fields.

Time series motifs first appeared in the biomedical sequence analysis and were used to describe structural characteristics of biological sequences. Its significance lies in that frequently occurring patterns are often able to reflect some important features of the original sequences, such as

the special structures of biological sequences, important words in the voice sequences and special behaviors of robot activities.

In this paper, we utilize the time series motif to develop an efficient method for recognition and matching of identical twin images. First the facial part is extracted, using an appropriate compression method to reduce its dimensionality. To judge the efficiency of the method a simulation test conducted using a very similar twin images and the result was satisfactory. The paper is organized as follows, section(2) introduce the proposed system, section(3) Viola-Jones Face detection, section(4) Time series of an image, section(5) Dimensionality reduction, section(6), Motif discovery algorithm, section(7) Experimental Results, and finally conclusion in section (8).

2 Proposed System

To compare identical twins using facial identity, we just need the face of the twins for matching. Our data base consists of half profile images, therefore we apply facial extraction using Viola-Jones technique. The resulting facial image transformed to time series using image histogram, the resulting time series needs to be compressed with the aid of the appropriate method. There are many methods for compression or feature extraction like, Discrete Fourier Transformation(DFT), Discrete Wavelet Transformation(DWT), Piecewise Aggregate Approximation(PAA) and Piecewise Linear Approximation(PLA). PLA is our choice as it the most used feature extraction technique in time series data mining [10]. The detection system figure (1) composed of two parts, in the first part(left) the twin image faces extracted from the frame, converted to time series. Applying PLA-method to extract features from those time series and finally store them in a data base system for later use. In the second part(right) the twin query processed in similar fashion, in order to search in the data base system.

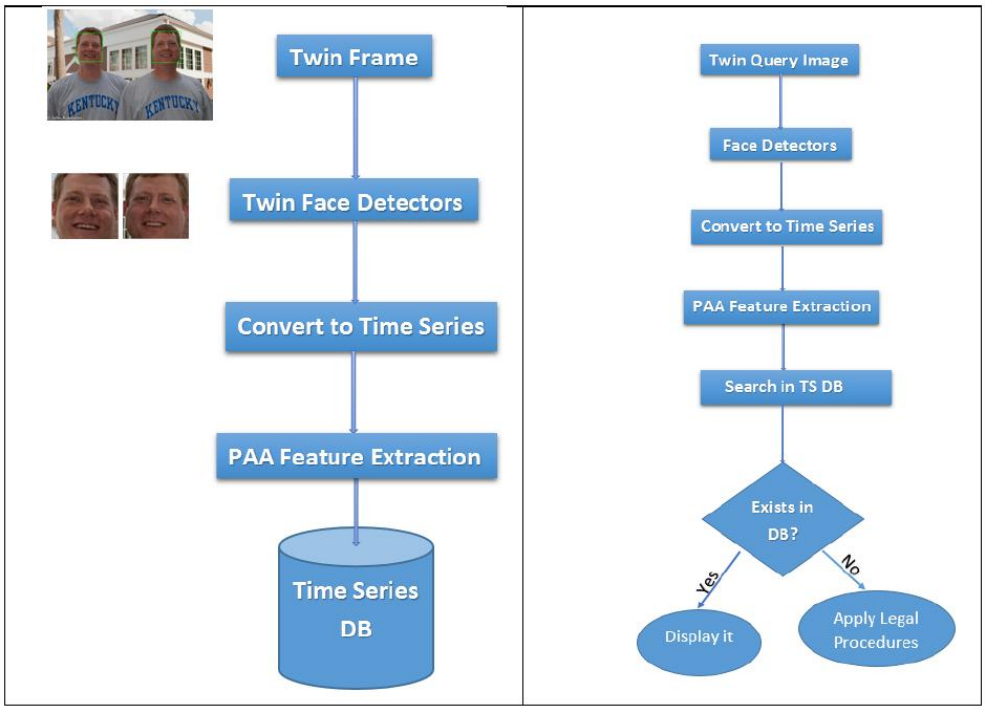

Figure 1: (left) extract twin faces, convert to time series, apply PAA, and store in DB. (right) query twin's image process and search in DB.

3 Viola-Jones Face detection

The Viola-Jones object detection framework is the first object detection framework to provide competitive object detection rates in real-time proposed in 2001 by Paul Viola and Michael Jones [25]. Although it can be trained to detect a variety of object classes, it was motivated primarily

by the problem of face detection. The technique relies on the use of simple Haar-like features that are evaluated quickly through the use of a new image representation[16]. Based on the concept of an Integral Image it generates a large set of features and uses the boosting algorithm AdaBoost to reduce the over-complete set and the introduction of a degenerative tree of the boosted classifiers provides for robust and fast interferences. In the technique only simple rectangular (Haar-like) features are used, reminiscent to Haar basis functions. These features are equivalent to intensity difference readings and are quite easy to compute. Figure(2) illustrates the four different types of features used in the framework.

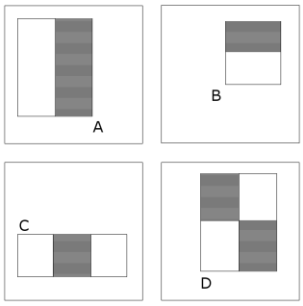

Figure 2: Feature types used by Viola and Jones.

Applying Viola- Jones to sample twins image, we get the reult shown in figure(3).

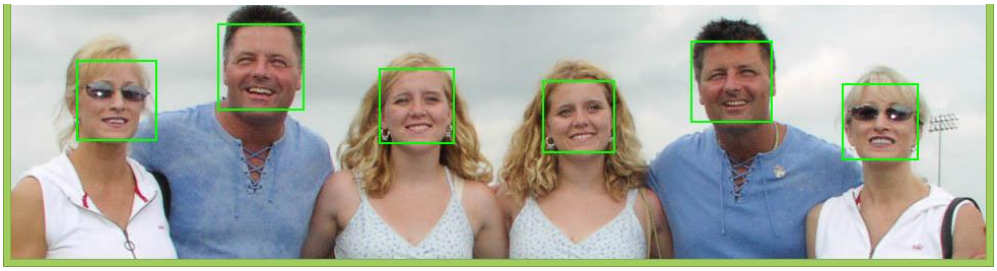

Figure 3: Face detection using Viola-Jones.

4 Time series of an image

A time series is a sequence $X = (x_1, x_2, \cdots, x_n)$, where n is the number of observations. Tracking the behavior of a specific phenomenon/data in time can produce important information. Time series can be very long, sometimes containing a billions of observations [8].

An image can be converted to time series from image color histograms as in figure 4. The color content of face images utilized as an efficient method to generate face image time series based on their color content.

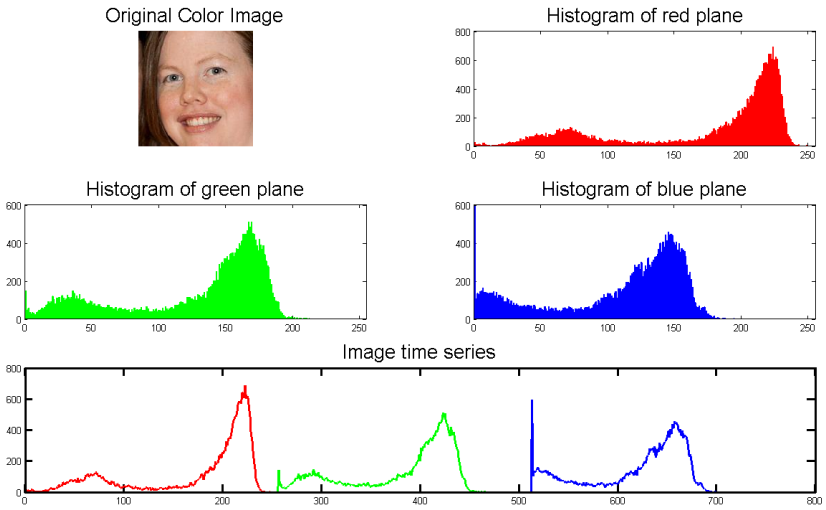

Figure 4: Conversion of an image from the RGB (Red, Green, Blue) image color histograms, to a time series.

In 2002, Lin Jessica first proposed the concept of time series motifs mining [12]. A time series motif is a set of subsequences (i.e. segments) of a time series, which are very similar to each other [12] in their shapes. Figure5 illustrates an example of a motif. The red and blue time series shown overlapped on one another are the motifs. The motifs are so similar that it is implausible that they happened at random and therefore, deserve a further exploration.

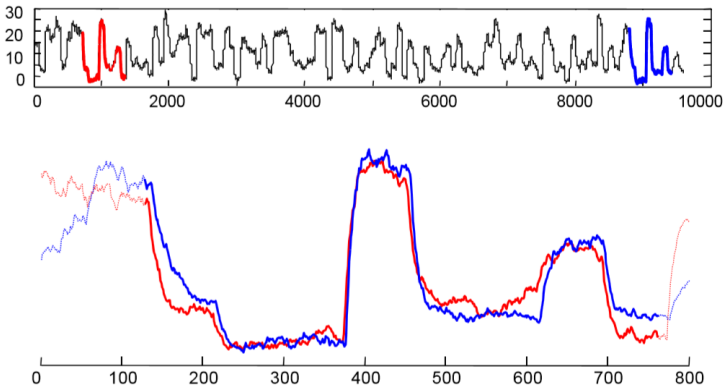

Figure 5: (top) A motif of length 640 beginning at locations 589 and 8,895. (bottom) by overlaying the two motifs we can see how remarkably similar they are to each other.

Motif discovery is a core subroutine in many research projects on activity discovery [18][19], with applications in elder care [20], surveillance and sports training. In addition, there has been a recent explosion of interest in motifs from the graphics and animation communities, where motifs are used for finding transition sequences to allow just a few motion capture sequences to be stitched together in an endless cycle [2].

5 Dimensionality reduction

The key aspect to achieve efficiency, when mining time series data, is to work with a data representation that is lighter than the raw data. This can be done by reducing the dimensionality of data, still maintaining its main properties. An important feature to be considered, when choosing a representation, is the lower bounding property. Given two raw representations of the time series T and S, by this property, after establishing a true distance measure d_{true} for the raw data (such as the Euclidean distance), the distance $d_{feature}$ between two time series, in the reduced space, $R(T)$ and

$R(S)$, have to be always less or equal than d_{true}:

$$d_{feature}(R(T), R(S)) \leq d_{true}(T, S) \quad (1)$$

If a dimensionality reduction techniques ensures that the reduced representation of a time series satisfies such a property, we can assume that the similarity matching in the reduced space maintains its meaning. Moreover, we can take advantage of indexing structure such as GEMINI [6] to perform speed-up search even avoiding false negative results. In the following subsections, we will review the main dimensionality reduction techniques that preserve the lower bounding property.

5.1 Feature Extraction using Piecewise Linear Approximation

Piecewise Linear Approximation is perhaps the most used feature extraction technique in time series data mining[10], the idea is to approximate the input sequence using a desired number of straight lines or segments see figure(6). As the number of segments is much smaller than n, a high level of compression can be achieved.

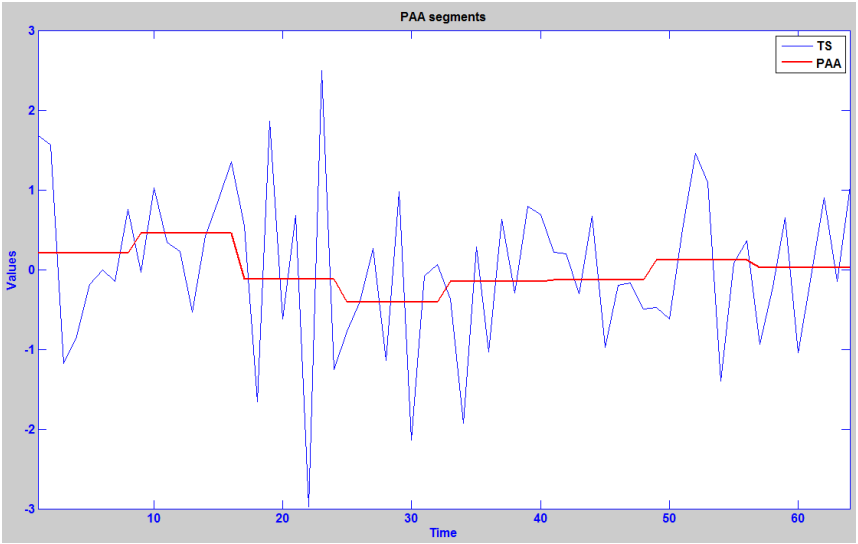

Figure 6: Time series(of length 64) is reduced to 8-dimensions using PAA representation.

Piecewise Linear Approximation techniques classified into three categories, (1) sliding window: a window is grown until a specified error threshold is reached this method is used to process data on-line. (2) top-down: starts initially by approximating the entire time series with one segment. It then recursively partitions the segment until all segments fall within a specified error threshold, this method is used to process data off-line. (3) bottom-up: starts initially with the finest grain approximation possible, it then iteratively merges segments until some error threshold is met and also used to process data off-line.

For all techniques, approximating lines are calculated using linear interpolation, the bottom-Up approach is generally considered the best overall[10] and it has a time complexity of $O(n \times n/(N-1))$, where $N-1$ is the number of segments. Pseudo code for the generic bottom-up segmentation algorithm is shown in listing(1).

Algorithm 1 :Seg_TS = Bottom-Up (T, max_error)

```
for i ← 1 : 2 : length(T) do
    Seg_TS = concat(Seg_TS, create_segment(T[i : i + 1]))
end for

for i ← 1 : length(Seg_TS) − 1 do
    merge_cost(i) = calculate_error([merge(Seg_TS(i), Seg_TS(i + 1))])
end for

while min(merge_cost) < max_error do
    index = min(merge_cost)
    Seg_TS(index) = merge(Seg_TS(index), Seg_TS(index+1)))
    delete(Seg_TS(index+1))
    merge_cost(index) = calculate_error( merge(Seg_TS(index),
                                          Seg_TS(index+1)) )
    merge_cost(index-1) =calculate_error( merge(Seg_TS(index-1),
                                          Seg_TS(index)) )
end while
```

Listing(1) Pseudo code for the generic bottom-up segmentation algorithm.

Having introduced the new representations of time series data, we can now define distance measures on them. By far the most common distance measure for time series is the Euclidean distance[9]. Given two time series Q and C of the same length n, equation(4) defines their Euclidean distance, and Figure(3-upper) illustrates a visual intuition of the measure.

$$D(Q,C) = \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2} \tag{4}$$

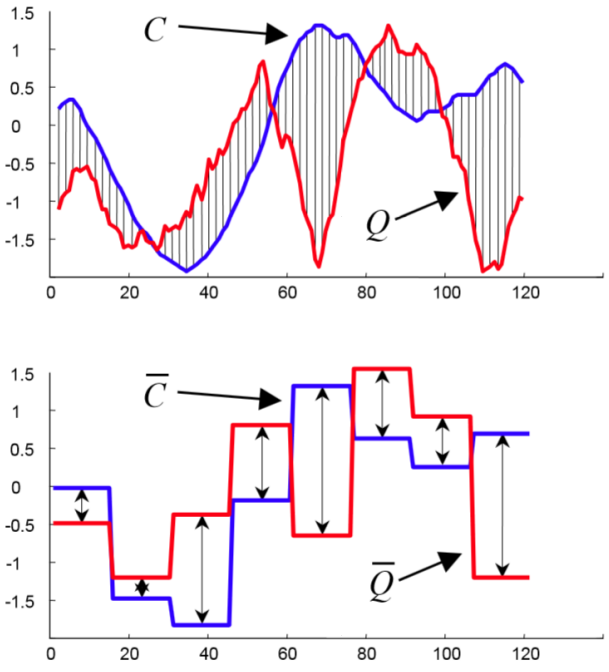

Figure 7: (upper)The Euclidean distance between two time series can be visualized as the square root of the sum of the squared differences of each pair of corresponding points. (lower) The distance measure defined for the PAA approximation can be seen as the square root of the sum of the squared differences between each pair of corresponding PAA coefficients, multiplied by the square root of the compression rate.

If we transform the original subsequences into PAA representations, Q and C , using equation(3), we can then obtain a lower bounding approximation of the Euclidean distance between the original subsequences by

$$DR(\overline{Q},\overline{C}) = \sqrt{\frac{N}{n}}\sqrt{\sum_{i=1}^{N}(\overline{q}_i - \overline{c}_i)^2} \tag{5}$$

This measure is illustrated in Figure(7-lower).

6 Motif discovery algorithm

In this section, we will describe the Motif discovery algorithm in detail which is used to discover motif based on generated image time series. In this algorithm [4], we extend the triangular inequality pruning method to preprocess the time series dataset and utilize an optimized matrix structure to improve the efficiency of this algorithm. First of all, we randomly select a time series as the reference time series marked by T_1 from time series dataset. Then we calculate the Euclidean distances from other time series to T_1. After that, according to the Euclidean distances, we make the linear arrangement of these time series as shown in figure 7.

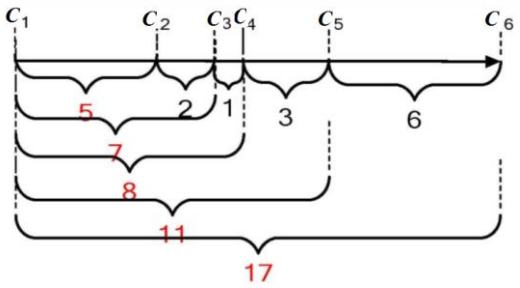

Figure 8: Linear arrangement of time series according to Euclidean distances.

Figure 8 presents the differences of Euclidean distances between each pair of consecutive time series. In our algorithm, we use the notation $lower\_bound(C_i, C_j)\,(1 < i < j)$ to denote the difference between $D(C_1, C_j)$ and $D(C_1, C_i)$. For example, based on the triangle inequality principle, the difference between both sides of the triangle must be smaller than the third side. So $lower\_bound(C_2, C_3)$ must be smaller than the actual Euclidean distance between C_2 and C_3. If the $lower\_bound(C_2, C_3)$ is greater than the range R, we don't need to calculate the actual Euclidean distance. Certainly, $C_2, C_4, C_5, and C_6$ are not in the same motif. Because the lower bounds from C_2 to other time series are greater than R. This is a very important point which needs to be emphasized. We extensively leverage the triangle inequality pruning method to make the preprocessing on the distances and realize the pruning quickly. Secondly, on the basis of the preprocessing on the distances between time series, we can construct the time series matrix. According to the symmetry of the Euclidean distance, the matrix $C[][]$ is a symmetric matrix. When most of element values in the matrix are 0, $C[][]$ is a sparse matrix. So we use a compressed storage structure: Triple sequence table. At last, based on the previous operations, we implement our Motif

Discovery Algorithm. In our Motif Discovery Algorithm, the triple sequence table Euc_dist stores the values of row, col and distance, and the $C\_count[]$ stores the number of time series the distances of which to C_i are less than R. Then we look for the element with the largest value in array $C\_count[]$ and add the time series.

Algorithm 2 describes the process of discovering motif. In Algorithm 8, motif_center represents the center of motif. First, we look for the position where the value is the maximum in C_count (lines $2-6$). Then we use the corresponding subsequence as the center of motif (lines $7-8$), and find all the subsequences whose distances to this center subsequence are smaller than R (lines $10-19$). Finally, we regard all subsequences as motif.

Algorithm 2 :Motif Discovery Algorithm

```
1: Initialize max ← 1
2: for i ← 2, m do
3:     if C_count[i] > C_count[max] then
4:         max ← i                                   ▷ get motif center
5:     end if
6: end for
7: motif_center ← C_max
8: add motif_center to motif
9: ▷ find the time series in Euc_dist, the distance of each time series to C_max is
   less than R
10: for i ← 1, lengthofEuc_dist do
11:     if Euc_dist[i].row == max then
12:         k ← Euc_dist[i].col
13:         add C_k to motif
14:     end if
15:     if Euc_dist[i].col == max then
16:         k ← Euc_dist[i].row
17:         add C_k to motif
18:     end if
19: end for
20: return motif
```

Listing(2) Pseudo code for the Motif Discovery Algorithm.

7 Experimental Results

Our experiments were conducted on identical twin image data collected from [24, 25, 26, 27], the database contains 50 pairs of identical twins (100 subjects), each subject contains one image.

The algorithm was coded using MatLab version 8.1.0.604 (R2013a), run on a PC computer with a 3.2 GHZ Intel I3 processor and 2 GB of RAM. The program interface is shown in Figure(9).

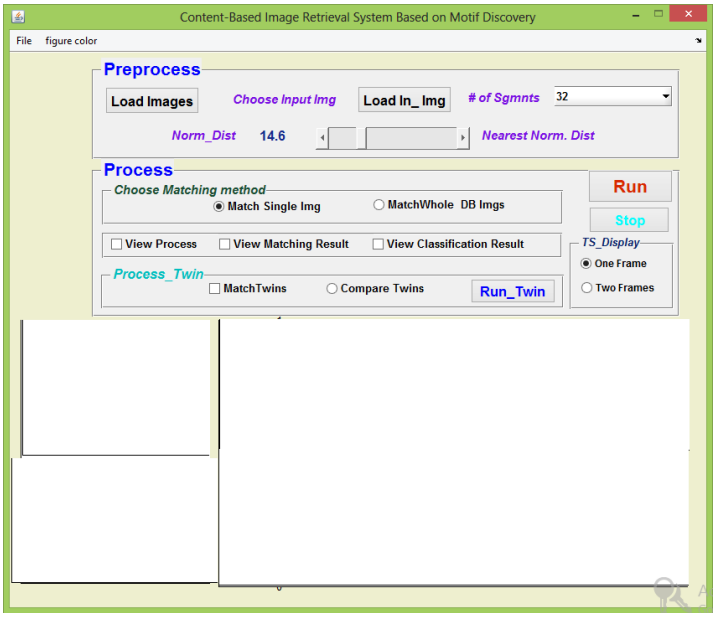

Figure 9: Friendly user interface of the program GUI.

7.1 Twin Similarity

Similarity can be roughly described as the measure of how much two or more objects are alike. Similarity can also be seen as the numerical distance between multiple data objects that are typically represented as value between the range of 0 (not similar at all) and 1 (completely similar). The measure of similarity must fall within the range of 0 and 1 and symmetry.

There are many similarity metrics, like; Euclidean distance, Pearson's correlation coefficient, Jaccard similarity coefficient, Tanimoto coefficient, and Cosine similarity[23]. Here we will adopt Pearson's correlation coefficient, correlation is the measure of the linear relationship between the attributes of two objects. Pearson's correlation coefficient is one such measure between two objects, A and B, such that:

$$r = \frac{cov(A,B)}{\sigma_A \sigma_B} \tag{6}$$

where σ_A and σ_B represents the standard deviation of the data set A and B respectively. Figure(10) shows twin images with correlation coefficient $r = 0.95943$ for the original time series and $r = 0.99576$ for the compressed version of length 8.

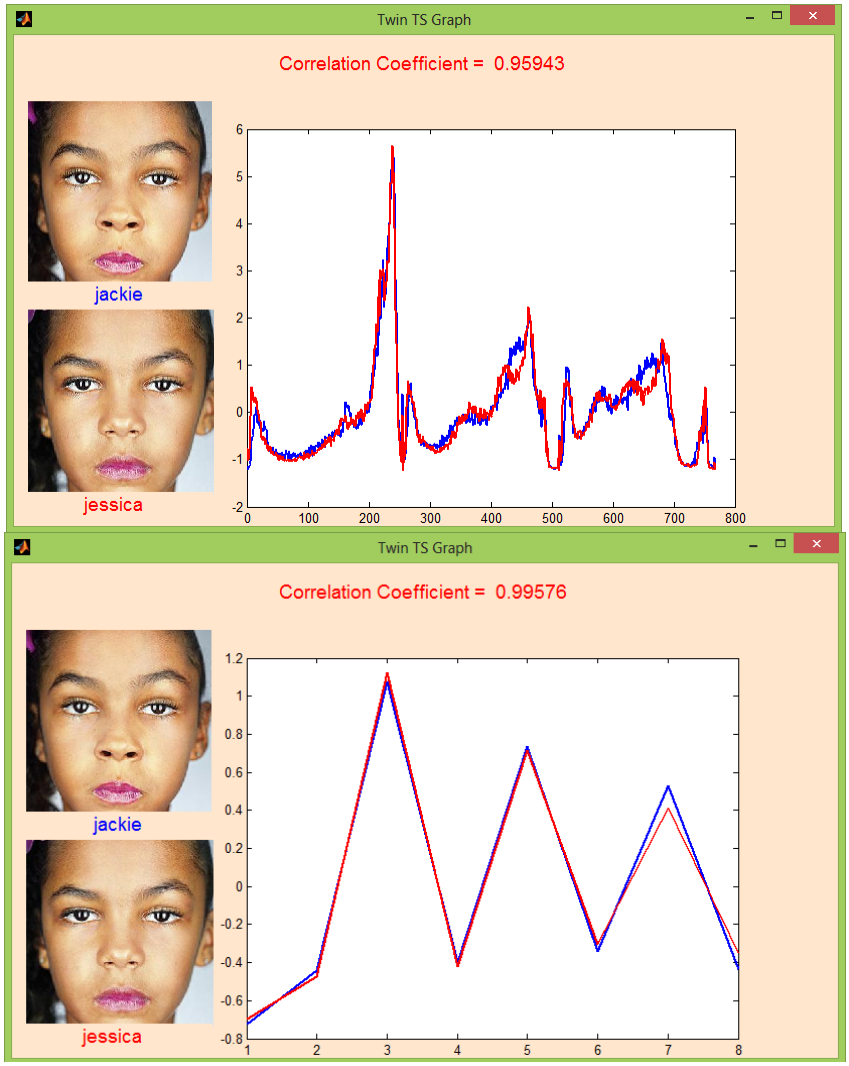

Figure 10: Correlation coefficient using the whole time series and using the extracted one of length 8.

7.2 Twin Matching

In this matching experiment, we take as input the above identical images as they have high similarity score; $:0.96$ and run the program to get the result shown in figure (11).

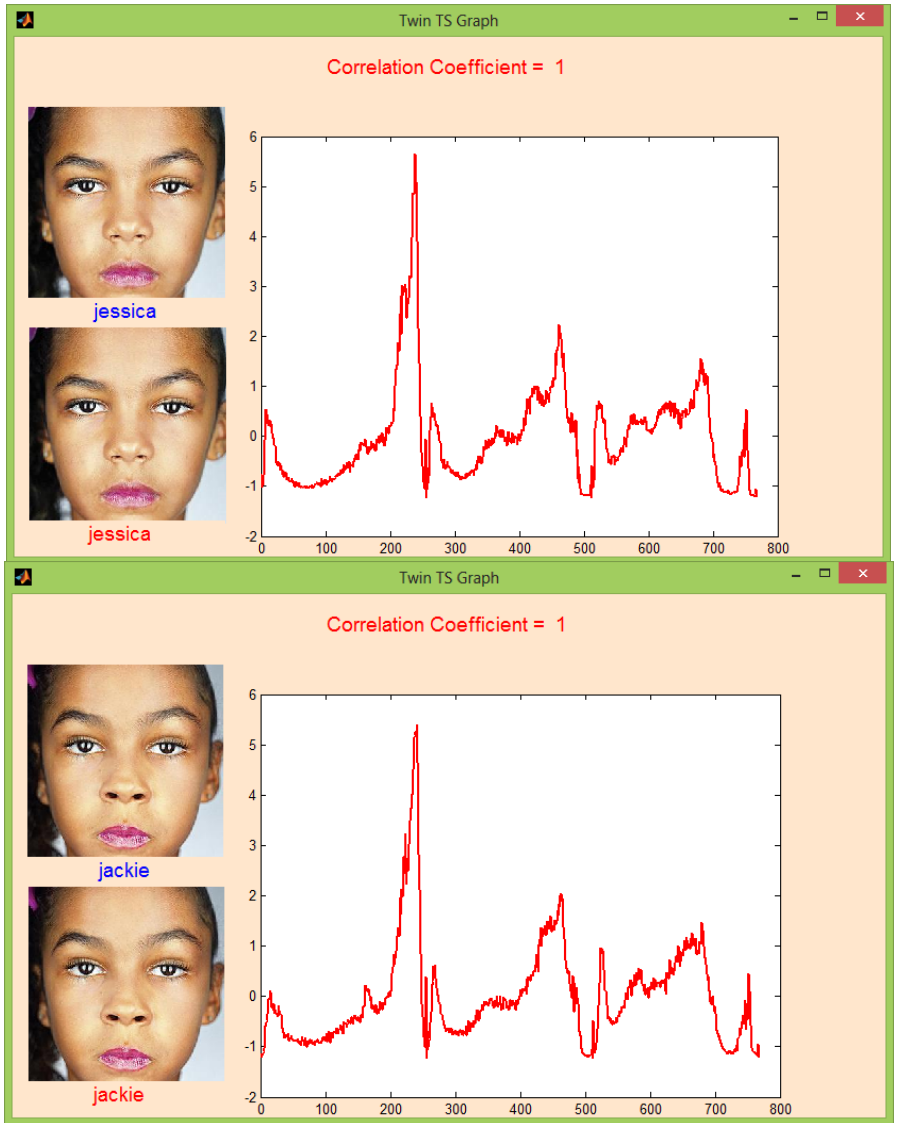

Figure 11: (upper)matched image of identical twin;Jessica, (lower)matched image of identical twin;Jackie.

7.3 Twin Matching using small number of features

Here we employed PAA-feature extraction technique (feature vectors of length 8) to test the program for the previous matching experiment and got the result in figure(12)

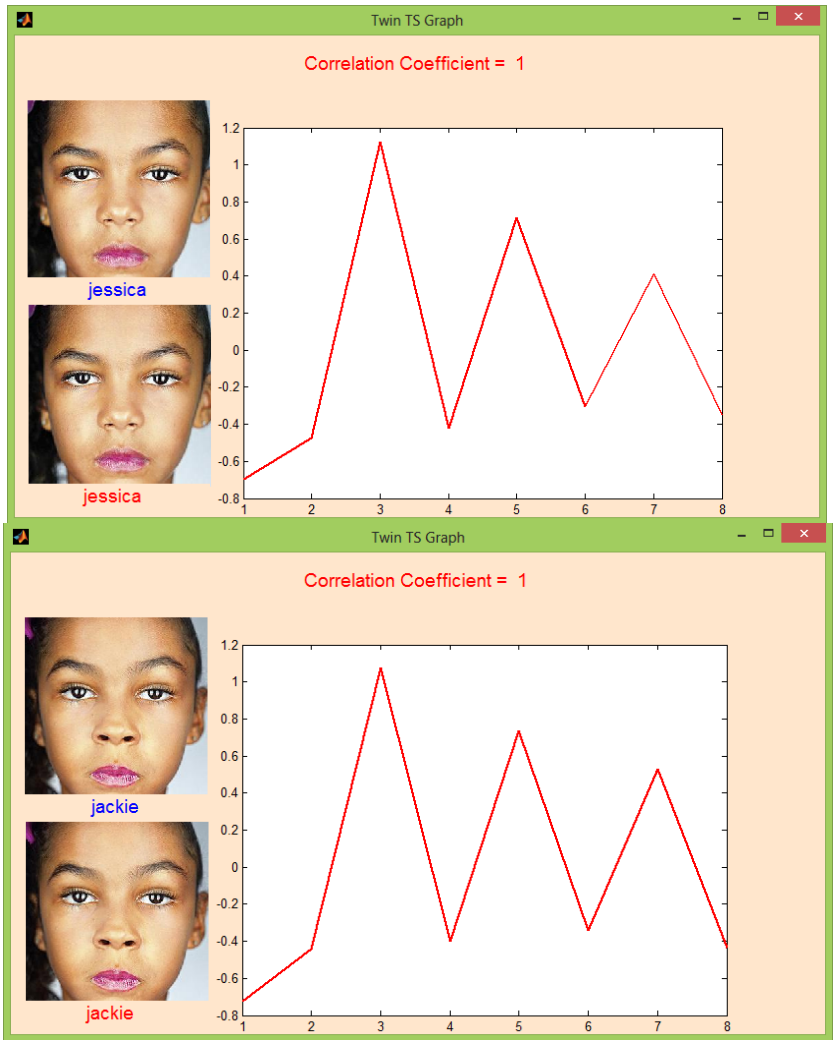

Figure 12: matched image of identical twin using feature vectors of length 8.

7.4 Performance Evaluation

We evaluate the performance of the proposed method in terms of precision, recall, and accuracy see[21]. Image retrieval system has the goal to retrieve relevant images while not retrieving irrelevant ones. The measures of performance used in image retrieval borrowd from the field of document information retrieval and are based on two primary figures of merit: precision and recall.

- Precision(P) is the number of relevant documents retrieved by the system divided by the total number of documents retrieved(i.e., true positives plus false alarms).

$$P = \frac{TP}{TP + FP} \tag{7}$$

- Recall(R) is the number of relevant documents retrieved by the system divided by the total number of relevant documents in the data base(which should have been retrieved).

$$R = \frac{TP}{TP + FN} \tag{8}$$

Precision can be interpreted as a measure of exactness, whereas recall provides a measure of

completeness.

- Accuracy(A) is the probability that the retrieval is correctly performed

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

where,

TP (True Positive) - correctly classified positive,

TN (True Negative) - correctly classified negative,

FP (False Positive) - incorrectly classified negative, and

FN (False Negative) - incorrectly classified positive.

TP	*TN*	*FP*	*FN*	*P(%)*	*R(%)*	*A(%)*
100	20	0	0	100	100	100

Figure 13: Performance of the used techniques.

8 Conclusion

In this paper, the problem of identifying the identical twin image has been addressed using the most recent methodology and techniques. The proposed system composed of two stages. In the first stage, the whole images are preprocessed to have only the facial parts, which are converted into time series and their dimensionality are reduced using PAA-technique and the results are stored in a database system. A simulation using the proposed method showed a good result for identical twins having correlation coefficient of $r = 0.95943$ which is not easy to discriminate using other biometric technology. It may be necessary to test the method on more image data.

References

[1] I. Androulakis,.I. WU,.I. Vitolo, and C. Roth, "Selecting maximally informative genes to enable temporal expression profiling analysis," Proc. of Foundations of Systems Biology in Engineering, 2005.

[2] Beaudoin, P., van de Panne, M., Poulin, P., and Coros, S., "Motion-Motif Graphs, Symposium on Computer Animation," Eurographics Association, Dublin, Ireland, 2008.

[3] B. Celly, and V. Zordan, "Animated people textures, " Proc. of 17th International Conference on Computer Animation and Social Agents (CASA), 2004.

[4] L.H. Chi, H.H. Chi, Y.C. Feng, S.L. Wang, and Z.S. Cao, "Comprehensive and efficient discovery of time series motifs," Journal of Zhejiang University - Science C, China, 2011, pp. 1000-1009.

[5] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic discovery of time series motifs," Proc. of the 9th International Conference on Knowledge Discovery and Data mining (KDD'03), 2003, pp. 493-498.

[6] B.K. Yi and C. Faloutsos. "Fast time sequence indexing for arbitrary lp norms.", In VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases, pages 385-394, San Francisco, CA, USA, 2000.

[7] T. Guyet, C. Garbay and M. Dojat, "Knowledge construction from time series data using a collaborative exploration system," Journal of Biomedical Informatics, 2007, pp. 40(6): 672-687.

[8] Hegland, M. Clarke, W. & kahn, M. "Mining the MACHO dataset, Computer Physics Comunications, Vol142(1-3), December 15. pp.22-28."

[9] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. "Dimensionality reduction for fast similarity search in large time series databases.", Knowledge and Information Systems, 3(3):263-286, 2001.

[10] E.Keogh, S. Chu, D. Hart, and M. J. Pazzani. "An on-line algorithm for segmenting time series.", In ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining, pages 289-296, Washington, DC, USA, 2001.

[11] Le N., Khoa L., Seshadri K., and Savvides M., A facial aging approach to identification of identical twins, Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference.

[12] J . Lin, E. Keogh, S. Lonardi, and P. Patel, "Finding motifs in time series, the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining," Edmonton, Alberta, Canada, 2002 , pp. 53-68.

[13] A. McGovern, D. Rosendahl, A. Kruger, M. Beaton, R. Brown, and K. Droegemeier, "Understanding the formation of tornadoes through data mining," 5th Conference on Artificial Intelligence and its Applications to Environmental Sciences at the American Meteorological Society, 2007.

[14] Martin, J. A., Kung, H.-C., Mathews, T. J., Hoyert, D. L., Strobino, D. M., Guyer, B., and Sutton, S. R., "Annual summary of vital statistics: 2006," Pediatrics, 788-801 (2008).

[15] Nejati H. et all,Wonder ears: Identification of identical twins from ear images, Pattern Recognition (ICPR), 2012.

[16] Nida Aslam, Irfanullah, K. K. Loo, and Roohullah " Limitation and Challenges- Image/Video Search & Retrieval", JDCTA 2009.

[17] Z. Sun, A . A. Paulino,J . Feng,Z . Chai, T. Tan,a nd A. K. Jain,"A study of multibiometric traits of identical twins," in Proc. SPIE, Biometric Technology for Human Identification VII, April 2010.

[18] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell, "Unsupervised activity discovery and characterization from event- streams. In In Proc. of the 21st Conference on Uncertainty in Artificial Intelligence (UAI05, 2005). "

[19] D. Minnen, T. Starner, M. Essa, and C. Isbell, "Discovering characteristic actions from on-body sensor data. In Wearable Computers, 2006 10th IEEE International Symposium on, pages 11-18, 2006. "

[20] A. Vahdatpour, N. Amini, and M. Sarrafzadeh, "Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In Proceedings of the 21st international jont conference on Artificial intelligence, IJCAI'09, pages 1261-1266, 2009. "

[21] Olson, David L.; and Delen, Dursun (2008), "Advanced Data Mining Techniques", Springer, 1st edition (February 1, 2008), page 138, ISBN 3-540-76916-1.

[22] Paone J.R., Flynn P.J., Philips P.J., Bowyer K.W., Double Trouble: Differentiating Identical Twins by Face Recognition, Information Forensics and Security, IEEE Transactions 2014.

[23] Data Mining Portfolio:Similarity Techniques,"http://mines.humanoriented.com/

classes/2010/fall/ csci568/portfolio_exports/bfindley/similarity.html"

[24] "http://ngm.nationalgeographic.com/2012/01/twins/schoeller-photography"

[25] "http://www.businessinsider.com/how-smoking-ages-the-face-of-identical-twins-2013-11"

[26] "http://www.dailymail.co.uk/news/article-2333645/Can-tell-difference-Photographer-captures-striking-similarities-identical-twins.html"

[27] "http://www.fastcodesign.com/1672590/do-these-identical-twins-look-the-same-to-you#1"

[28] Zhang H., Tang C., Li X.; Kong, A.W.K, "A study of similarity between genetically identical body vein patterns", Computational Intelligence in Biometrics and Identity Management (CIBIM), 2014 IEEE Symposium.