Residential Load Pattern Analysis for Smart Grid Applications based on Audio Feature EEUPC

Yunzhi Wang¹, ², Xiangdong Wang¹, Yueliang Qian¹, Haiyong Luo¹, Fujiang Ge³, Yuhang Yang³, Yingju Xia³

¹ Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, China
² 2 Graduate University of Chinese Academy of Sciences, China
³ 3 Fujitsu Research & Development Center Co., LTD.

Abstract

The smart grid is an important application field of the Internet of things. This paper presents a method of user electricity consumption pattern analysis for smart grid applications based on the audio feature EEUPC. A novel similarity function based on EEUPC is adapted to better support clustering analysis of residential load patterns. The EEUPC similarity exploits features of peaks and valleys on curves instead of directly comparing values as the Euclidean distance does, and can obtain better performance for clustering analysis. Moreover, the approach proposed in this paper not only performs load pattern clustering, but also extracts a typical pattern for each cluster and gives suggestions towards wiser power consumption for each typical pattern. Experimental results demonstrate that the EEUPC similarity is more consistent with human judgment than the Euclidean distance and higher clustering performance can be achieved for residential electric load data.

Keywords

Internet of Things, Smart Grid, Electric load pattern analysis, EEUPC, EEUPC similarity

1 Introduction

The smart grid is an intelligent electrical power management system inherited from the conception of the Internet of Things. It is based on the physical electricity network, and benefits humanity by advanced technologies highly integrated, such as sensors, automation control and decision support. In the field of smart grid, electric load analysis has attracted considerable attention of researchers in recent years. According to the result of electric load analysis, electricity suppliers are able to improve power energy supply and distribution. What's more, electric load analysis is closely linked with consumers, helping them understand their own needs and make an arrangement of power energy consuming more wisely.

Current work on electric load analysis mainly includes two aspects. On one hand, many researchers analyze the impact of various factors on the electric load in order to facilitate load forecasting. Numerous methods have been proposed, such as Kalman filtering analysis, regression analysis, exponential smoothing forecasting, expert systems, fuzzy prediction, gray model, optimal combination forecasting, artificial neural networks, rough sets algorithm, fuzzy clustering, particle swarm optimization, and genetic algorithm. Based on these algorithms, researchers intend to figure out the relationships between the electric load and factors such as weather, economic growth and so on, and using the factors with high relevance, higher accuracy can be achieved in load forecasting.

On the other hand, there is also much research effort on user electric load pattern analysis, of which current research mainly focuses on clustering and classification of load patterns (daily or monthly load curves in practice). The purpose of load pattern clustering analysis is to group users' load patterns into several typical classes and thus help electricity suppliers get better knowledge of their customers and
customize their supply strategies. For example, many researchers perform clustering analysis on load patterns of industrial electricity customers such as companies and factories [1]. They compare clustering results with economic type codes of customers, indicating that electric power load patterns can be effectively distinguished by pattern modes and the results are approximately consistent with industry types. The limitation of this sort of research lies in three aspects: First, most current research work focuses on clustering of load data of industrial customers other than ordinary residential customers, and conclusions of this type of research cannot suit residential customers as consumption habits of industrial customers and residential customers are of considerable difference. However, clustering analysis of residential customers is of great significance, since the domestic load occupies a large part of total electricity consumption, and is usually not as stable as industrial consumption. Secondly, current methods only yield the result of clustering analysis and support of decision-making is not provided. To make decisions, users have to analyze the clusters of patterns manually to extract useful information. Thirdly, most current methods which deal with industrial load data use the Euclidean distance as the distance measurement of load patterns. However, for residential load data which are more unstable, similar patterns with consistent peaks and valleys may yield low similarity due to the difference in value, thus making results of clustering not satisfactory. Therefore, a different distance metric is needed, which measures the similarity in terms of the shape of the load curve (e. g. peaks and valleys on the curve) instead of simply comparing the values.

In this paper, an approach for residential electric load pattern analysis is proposed. The method focuses on analysis of residential electric load patterns and proposes a novel similarity function based on the audio feature EEUPC (which is named EEUPC similarity). The EEUPC distance exploits features of peaks and valleys on curves instead of directly comparing values as Euclidean distance does, and can obtain better performance for clustering analysis. Moreover, the approach proposed in this paper not only performs load pattern clustering, but also extracts a typical pattern for each cluster and gives suggestions of wiser consumptions with lower cost for each typical pattern.

The rest of this paper is organized as follows. In Section 2, related work on electric load analysis is presented. In Section 3, the electric load pattern clustering method based on the audio feature EEUPC is described in detail. And Section 4 presents the method for typical load pattern analysis after clustering. Experimental results are given in Section 5. And finally, conclusions are drawn in Section 6.

2 Related Work

As mentioned above, there are mainly two kinds of research work on electric load analysis: relevance analysis of climatic and economic factors for load forecasting, and user load pattern analysis.

In recent years, there is a large amount of academic research on the analysis of factors associated with load forecasting [2-6]. Hor et al. [3] analyzed the impact of weather variables on monthly electricity demand in England and Wales using a multiple regression model. Weather variables considered includes degree days, enthalpy latent days, and relative humidity. Mori and Kobyashi [4] used the fuzzy inference system to forecast electricity load [4]. They proposed a method for constructing an optimal structure of the simplified fuzzy inference that minimizes model errors and the number of the membership functions to grasp nonlinear behavior of power system short-term loads. Apart from these, ANN (artificial neural networks) is also widely used in multi-variable electricity load analysis and forecasting. In [5], an algorithm using cascaded ANN together with historical load and weather data is proposed to forecast half-hourly power system load for the next 24 hours. The ANNs were trained and tested on the electric power system of Kuwait. Some other researchers analyzed electric load from the aspect of economy and society. For instance, Su and Song [6] tried to make a comparative analysis of the weights of influencing factors of electrical energy consumption by weighted least squares (WLS) and quantile regression (QR).

For user load pattern analysis, current research work mainly focuses on clustering and classification of consumers' load data [1, 7]. The most common scheme is as presented in [7], in which load data clustering analysis is adopted to generate user load profiles for industrial consumers. Under the
scheme, Euclidean distance is adopted as the distance measurement for clustering and various clustering algorithms may be used for clustering analysis. In [7], three algorithms, namely, K-means clustering, K-centers clustering and hierarchical clustering were compared to each other and experimental results showed that hierarchical clustering algorithm performed better than other algorithms. Also, for each cluster, a typical load pattern was extracted and compared with traditional economic industrial clusters. Other clustering algorithms such as fuzzy analogy were also used for clustering [8]. There is also research work on analysis of electricity consumption patterns for making decisions in distribution management and price strategies. These researches mainly focus on analyzing peaks and valleys in the electricity consumption pattern to decide better price strategy [9, 10]. Those analyses are performed on consumption data of a city or county, and do not give information in terms of groups of users.

3 Residential Electric Load Pattern Clustering based on the Audio Feature EEUPC

The aim of clustering analysis is to classify data into categories where data in the same category are similar and data in different categories have a greater difference. Therefore, clustering analysis is on the basis of similarity measurement of data as it indicates differences between data while clustering.

There are many similarity and distance measurements to be taken into account for clustering analysis, such as the Euclidean distance, Minkowski distance and Mahalanobis distance, etc. As mentioned above, most current methods for electric load data clustering use the Euclidean distance which measures the difference between load values. However, for residential load data which are more unstable than industrial data, Euclidean distance cannot depict the similarity well as similar patterns with consistent peaks and valleys may yield low similarity due to the difference in value. In this paper, we propose a new similarity measurement function (the EEUPC similarity) based on EEUPC (Energy Envelop Unit Position and Confidence), which is an audio feature for efficient audio clip similarity measurement first used in audio retrieval [11, 12]. Unlike the Euclidean distance, the EEUPC similarity focuses on the shapes, especially peaks and valleys of curves and does not calculate similarity strictly according to the values on the curve. Since the electric load curve is different from audio data in terms of factors such as value range, degree of instability, etc., the calculation of EEUPC similarity used in this paper is a modified version of that for audio data [12]. The definition and calculation of the similarity is detailed in the rest of this section.

3.1 The EEUPC Similarity

In the definition of EEUPC, the curve of energy is referred to as the energy envelope [12]. In terms of shape of the curve, it can be observed that energy envelopes could be divided into units each contains one major peak and two low endpoints. Fig. 1 shows an example of a residential consumer's power energy curve on all weekdays in a month. It can be seen that the curve can be divided into three units, in each of which exists a major peak of load value. Notice that in the first unit there are actually two peaks. However, the first peak is relatively short in time and not high enough to distinguish from the other peak, which implies that there may be noise or instability. Therefore, the two peaks are considered as one major peak to avoid influences of noise. In fact, this is the most important feature of EEUPC according to other unit segmentation methods.
To segment units of energy envelope, values of a detection function are calculated first. The detection function is defined as follows and is used to detect the maximum energy difference among $J$ neighbor points after each point of the whole curve.

$$d_i = \max_{j=1,...,J} (E_{i+j} - E_i)$$

(1)

Where $E_i$ denotes the electric load value of the $i^{th}$ point on the curve, and the value of $J$ is to be decided by experiments. Notice that $E_{i+j} - E_j$ is used instead of $E_{i+j} / E_j$ used in audio retrieval [12] due to the difference in the value range between load data and audio data.

According to the detection function, energy envelope units can be segmented. In order to improve accuracy, segmentation confidence is adopted instead of binary thresholding. The confidence is calculated as

$$P(i) = \begin{cases} 1, & d_i \geq T_2 \\ \frac{d_i - T_1}{T_2 - T_1}, & T_1 < d_i < T_2 \\ 0, & d_i \leq T_1 \end{cases}$$

(2)

where $d_i$ denotes the detection function value on the $i^{th}$ point on the curve, and $T_1$ and $T_2$ are pre-determined thresholds. After the segmentation confidence calculation, points with non-zero confidence are recorded as segmentation positions, and segments between these segment positions are recorded as segmentation units. The segmentation positions and confidence values are used together as the EEUPC (Energy Envelope Unit Positions and Confidence) representation of the electric load curve, which is denoted as $U = (u_1, p_1), (u_2, p_2), ..., (u_n, p_n)$, where $U$ denotes a load curve, $n$ is the number of energy envelope units, and $u_i$ and $p_i$ denote the position and confidence value of the $i^{th}$ unit, respectively. The procedure of energy envelope unit segmentation is illustrated in Fig. 2. In the graph illustrating the segmentation of energy envelope units, the position and height of each vertical line shows the position and the confidence value of each energy envelope unit.
Fig. 2 Procedure of energy envelope unit segmentation.

Using the EEUPC representation of load curves, the EEUPC similarity can be calculated as follows. Suppose that there are two load curves, which can be represented by EEUPC as $U = \{ (u_1, p_1), (u_2, p_2), \ldots, (u_m, p_m) \}$, and $V = \{ (v_1, q_1), (v_2, q_2), \ldots, (v_n, q_n) \}$, where $u_i$, $v_j$ and $p_i$, $q_j$ ($i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$) denote positions and confidences of unit segmentation, respectively. For each segmentation position $u_i$ in $U$, if there exists $v_j$ in $V$ satisfying that $|u_i - v_j| < T$, where $T$ is a pre-determined threshold, then $u_i$ is said to be detected, and the detection confidence $p_i = \min \{ p_i, q_j \}$. Then, similarity based on EEUPC of the two curves $U$ and $V$ is calculated as

$$S(U, V) = \frac{2R(U, V)P(U, V)}{R(U, V) + P(U, V)}$$ (3)

$$R(U, V) = \sum_i p_i \frac{1}{\sum_{j=1}^n p_j}, \quad P(U, V) = \sum_i p_i \frac{1}{\sum_{j=1}^n q_j}$$ (4)

It can be seen that $R(U, V)$ and $P(U, V)$ are similar to the widely used metrics of recall and precision, and $S(U, V)$ can be seen as the F1 value of $R$ and $P$. Therefore, the EEUPC similarity actually calculates the consistence of unit segmentations between the two curves. Since the similarity depends on both the position and confidence (which essentially implies the magnitude of the peak), it considers both the position and height of the peak in a relatively approximate way instead of directly comparing values.

### 3.2 Residential Electric Load Pattern Clustering

Based on the EEUPC similarity, clustering of residential electric load patterns is performed. The aim of load pattern clustering is to cluster load patterns of different consumers into several classes to better understand the consumers' behavior and support decision making. The load pattern of each residential consumer is represented by the load curve within a certain period. In our work, clustering is performed on daily load curves to explore the user load pattern within one day, where each point on the curve stands for the load of the hour, making totally 24 points on the curve.
Before clustering, some pre-processing is needed for the load data [7]. First, for each consumer, a daily load curve should be obtained for clustering. Obviously, using the load curve of a certain day may incorporate random error into the results. Therefore, the ordinary method adopted is to average daily load curves within a period (e.g. a month or a year) to generate an average daily load curve. Furthermore, since the load on weekdays and weekends may differ considerably from each other, load curve clustering is performed separately for weekday curves and weekend curves and averaging is performed separately accordingly.

In order to emphasize the trend of a curve and to weaken the influence of absolute values, daily load data should be normalized before clustering. Suppose that load value of time \( i \) is denoted as \( E_i \), the normalized value is calculated as

\[
E'_i = \frac{E_i - E_{\min}}{E_{\max} - E_{\min}}
\]  

(5)

where \( E_{\max} \) and \( E_{\min} \) are the maximum and minimum load values on the curve.

For electric load pattern clustering, there are many commonly used clustering methods such as model-based methods, intensity-based methods and so on. Among them most widely used are K-means algorithm, K-center algorithm and hierarchical clustering. Which method to use depends on features of data. In our work, after tests on different methods, we choose hierarchical clustering as our final clustering method. In this algorithm, each load curve forms a cluster initially, and then in every step of the clustering procedure, the nearest two curves (which means that the two curves have the smallest distance or largest similarity) are found and emerged into one cluster. The procedure ends when the total number of clusters reduced to the cluster number pre-determined.

In the clustering process, distances between clusters can be calculated in different ways. Generally speaking, the most commonly used one is single-linkage, that is, when two clusters contain more than one curve, the distance between each point in cluster 1 and each point in cluster 2 are calculated and the minimum of all these distances is chosen as the distance between cluster 1 and cluster 2. Meanwhile, there are also other methods such as choosing the maximum of all distances (referred to as complete-linkage), and choosing the average distance (referred to as average-linkage). In our work, with comparison of all these methods, we finally used averaged-linkage as our distance calculating method. Suppose that there are two clusters \( P \) and \( Q \). The average distance \( D_{PQ} \) is calculated as

\[
D_{PQ}^2 = \frac{1}{n_P n_Q} \sum_{X_i \in P} \sum_{X_j \in Q} d_{ij}^2
\]  

(6)

where \( n_P \) is the number of all curves in cluster \( P \), and \( n_Q \) is the number of all data points in cluster \( Q \). \( X_i \) indicates the \( i^{th} \) curve in cluster \( P \), \( X_j \) indicates the \( j^{th} curve belongs to cluster \( Q \), and \( d_{ij} \) is the distance between \( X_i \) and \( X_j \).

4 Typical Load Pattern Analysis

After load pattern clustering, consumer load patterns are clustered into several classes, and further analysis is needed to explore the characteristics of each class. Some researchers end their work by presenting the classes to system users and leave the analysis and decision making to humans. Other research work extracts a typical pattern for each cluster, but does not perform automatic analysis on the typical pattern and thus cannot support the browse and retrieval of peaks and valleys of power consumption for each cluster of consumers.

In this paper, in addition to load pattern clustering, a method of typical load pattern analysis is
proposed to extract useful information from each cluster and support browse, retrieval of those information. The main idea is to extract a typical pattern of each cluster, and extract information about peaks and valleys of power consumption. Extracted information can benefit decision making of power supplies and can also help consumers better understand their electric load patterns and better arrange their daily electricity consumptions.

4.1 Typical Load Pattern Extraction

Typical patterns can represent characteristics of the electric load pattern of a class of users. By analyzing typical patterns instead of performing separate analyses on each customer, influence of abnormal load patterns can be avoided and more robust conclusions can be achieved.

To extract the typical load patterns from each cluster, the method of averaging is used as in [7]. For a cluster $P$, the typical load pattern $p_t$ is calculated as

$$E_{tj} = \frac{1}{n} \sum_{p_i \in P} E_{ij}, \quad j = 1, 2, \ldots, N$$

(7)

where $N$ denotes the dimension of the load pattern (for example, in our work, $N=24$ since daily load pattern with data of 24 hours are used.), $n$ is the number of load patterns in cluster $P$, $p_i$ is the $j^{th}$ load pattern in cluster $P$, and $E_{ij}$ and $E_{tj}$ are the $j^{th}$ load values of $p_i$ and the typical pattern $p_t$, respectively.

Despite the fact that a typical pattern is not a real electric load curve, it reveals the characteristic of the electricity consuming habit of users in a cluster. Fig. 3 shows an example of typical pattern extraction where all 190 daily load patterns in a cluster and the typical pattern (the bold curve in the figure) extracted by averaging are plotted. It can be seen that the typical pattern shows basic trends of most of the 190 patterns.
4.2 Analysis of the Typical Load Pattern

As mentioned above, typical patterns can be analyzed to indicate the features of clusters. In this paper, a method for analysis of the typical load pattern is proposed, which can provide two kinds of information for the typical load pattern of a cluster: First, the proportion that the consumers in the cluster take in all consumers is calculated to indicate the significance of the typical load pattern. Secondly, the typical load pattern curve is segmented into energy envelope units and the peaks of units are extracted and shown to system users for further decision making.

Fig. 4 shows an example of typical load pattern analysis, where Fig. 4(a) is a typical load pattern curve and Fig. 4(b) shows the segmentation result of the typical pattern. It can be seen in Fig. 4(b) that a peak of power consumption in the pattern may be extracted which appears at the hour between 17 o’clock and 23 o’clock. Therefore, the information extracted from the analysis will be presented to the system user in the following form.

“This type of consumers takes a proportion of 38.0% in all consumers.
There is one electric load peak in the typical load pattern.
The load peak appears in 17:00 -23:00.
The maximal load appears at 21:00.
It might be encouraged that the customers consume electric power from 0:00 to 17:00. ”

With the aid of automatic analysis and suggestion, power suppliers are able to better understand their customers and make decision in power management and price strategies. Moreover, with the suggestions and impact of corresponding price strategies, electricity consumers are able to arrange power consuming plans wisely.
5 Experimental Results Analysis

To evaluate the method proposed in this paper, experiments are conducted using real electric load data recorded by smart meters. The data set includes daily load data of 500 household consumers of a community in Beijing within the year 2009. Each daily load datum is represented by a load curve with 24 points, which denotes the load in each hour within the day. As detailed in 3.2, for each consumer, all load curves in weekdays are averaged to generate the load pattern of the consumer, resulting in 500 load patterns. Clustering analysis are performed on the 500 load patterns. For most clustering algorithms, including the hierarchical clustering algorithm used in our work, the cluster number should be designated. In our work, after a number of experiments, the results showed that a cluster number of 7 to 10 seemed appropriate. In this paper, experimental results of cluster number 10 will be presented since results of other parameters are quite similar.

5.1 Euclidean Distance vs. EEUPC Similarity

To figure out the difference between Euclidean distance and the EEUPC similarity, both the Euclidean distance and the EEUPC similarity are calculated for each pair of the 500 consumers. We observed, compared and analyzed the top 100 pairs of each distance (similarity) measurement and came to the conclusion that the EEUPC similarity yields results more reasonable and consistent with human judgments. As mentioned above, this is because the EEUPC similarity emphasizes the shape of curves, especially the difference between peaks and valleys, and relatively weakens the impact of absolute values of the power loads.
As an example, Fig. 5 shows a pair of daily load curves. For the pair of curves, the EEUPC similarity value $S_{\text{EEUPC}} = 0.3262$. To be compared with the Euclidean distance, we also calculate a $\text{EEUPC distance}$ measurement, i. e., $\alpha_{\text{EEUPC}} = 1 - S_{\text{EEUPC}} = 0.6738$. At the same time, the Euclidean distance between the two curves $\alpha_{\text{Euclidean}} = 0.2907$. It can be seen from the results that the Euclidean distance is quite small and the EEUPC distance is quite large. The ranks of the two distance values among all distance values are consistent with these results, namely, the Euclidean distance ranks within the 100 smallest Euclidean distance values, while the EEUPC distance is larger than thousands of EEUPC distances.

The above example, along with many other similar cases we observed in observation and analysis, demonstrated the fact that the EEUPC similarity (distance) is more consistent with human comprehension in terms of similarity for load curves. As for the above example, from the view of humans, the two curves apparently, have different peaks within different periods: The peak of curve 1 occurs in the evening while the peak of curve 2 occurs in the middle of the day. In fact, the EEUPC similarity is calculated in a similar way to this human judgment by considering both the position and the confidence of peaks. On the contrary, the calculation of Euclidean distance uses directly the absolute value of each point, and yields a low distance value since the values on the two curves are close except for several special points, which are the peak values and should have been the focus when considering the differences between the two curves.

![Curve 1](image1)

![Curve 2](image2)

Fig.5 Two load curves for which the Euclidean distance is 0.2907 and the EEUPC distance is 0.6738.

**5.2 Experimental Results of Load Pattern Clustering**

As mentioned above, all the 500 residential daily load patterns are clustered into 10 classes using the EEUPC similarity and hierarchical clustering algorithm as detailed in Section 3. For comparison, we
also clustered all 500 data into 10 classes using the Euclidean distance with the same clustering method. The results are shown in Fig. 6 and Fig. 7. In each graph thin lines show all load curves in the cluster and the bold line is the typical load pattern obtained by averaging all the curves.

From these experimental results it can be seen that compared to the Euclidean distance, the EEUPC-similarity-based method can divide data more evenly. On the contrary, when the Euclidean distance is used, there are some clusters that contain few curves, making these clusters 'abnormal' ones, and therefore lack of representativeness. The reason for this result is that the Euclidean distance use directly the values and the distance between some special patterns and other patterns will be quite large, whereas EEUPC mainly considers peaks and valleys and even special patterns may share several same peaks and valleys with other patterns. To reduce the impact of "bad data", in our experiment, we tried to delete the curves one or two of which formed a single cluster and performed clustering for the rest of the data. However, the similar results occurred again with some clusters contain few curves, which indicated that the result was related to the using of the Euclidean distance instead of some bad data.

Moreover, it can be seen from Fig. 6 that the patterns in each cluster have similar shapes, especially in terms of peaks and valleys, and the typical load pattern extracted can represent the patterns well. On the other hand, patterns in different clusters differ much in shape. For example, in Fig. 6, the patterns in the first cluster have only 1 peak, and the patterns in the second cluster have 3 peaks. As for the Euclidean-distance-based method, clustering results reveal zigzagging curves with more than 5 peaks a day, indicating that the Euclidean distance is too sensitive to instable data, as is mentioned in earlier part of this paper.
Fig. 6 Results of clustering based on EEUPC similarity
Fig. 7 Result of clustering based on Euclidean distance

5.3 Experimental Results of Typical Pattern Analysis

As detailed in 4.2, in our experiment, typical pattern analysis was performed for each typical pattern extracted from each cluster. And descriptive information was also given by the system, addressing characteristics of the class of consumers and corresponding suggestions for electricity consumption. Fig. 8 shows two clusters (cluster 2 and cluster 5 in Fig. 6) as examples and the descriptions output by the system are as follows.
Fig. 8. Two clusters for typical load pattern analysis.

Description for Fig. 8(a):

"This type of customers takes a proportion of 15.2% in all consumers. There are three electric load peaks in the typical load pattern. And they appear in 6:00-9:00, 12:00-13:00, and 19:00-23:00. The maximal load appears at 22:00, and the second highest peak appears at 8:00. It might be encouraged that the customers consume electric power from 0:00 to 5:00."

Description for Fig. 8(b):

"This type of customers takes a proportion of 7.6% in all consumers. There are two electric load peaks in the typical load pattern. And they appear in 11:00-13:00, and 18:00-23:00. The maximal load appears at 21:00, and the second highest peak appears at 12:00. It might be encouraged that the customers consume electric power from 0:00 to 8:00."

6 Conclusions

In this paper, an approach for residential electric load pattern analysis is presented. A novel similarity function based on the audio feature EEUPC (which is named EEUPC similarity) is proposed to better support clustering analysis. The EEUPC distance exploits features of peaks and valleys on curves instead of directly comparing values as the Euclidean distance does, and can obtain better performance for clustering analysis. Moreover, the approach proposed in this paper not only performs load pattern clustering, but also extracts the typical pattern for each cluster and gives suggestion of wiser consumptions with lower cost for each typical pattern. Experimental results demonstrate that the EEUPC similarity is more consistent with human judgment than Euclidean distance and better clustering performance can be achieved.
References


