TITLE OF THE INVENTION

METHOD OF LANGUAGE CODING IDENTIFICATION AND DATA FORMAT THEREFOR

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a method of language coding identification and a data format therefor, which identifies language coding of pure text strings, to be applied to microcontrollers and other embedded systems, particularly, relates to a method of recognizing general text scripts of a language to be identified prior to selecting a rendering engine so as to produce an accurate text image.

Description of the Related Art

In order to represent writing characters and scripts on computers, a plurality of encoding schemes has been created so as to associate language characters and scripts with numbers, for instance, ASCII for English and GB18030 for simplified Chinese. However, there exist a large number of different languages in the world. In this regard, there are so many overlaps on encoding such that a same number represents a different character among different languages. Due to the overlaps on encoding, language coding must be identified before an appropriate rendering process is implemented to produce an image of a character or a script.

One example of such a method of language identification is disclosed in the United States Patent No. 5,548,507. According to the U. S. Patent No. 5,548,507, the method is based on a plurality of "Word Frequency Tables (WFTs)". However, the WFTs require lots of system resources.

Another method of language identification is disclosed in the United States

Patent No. 6,157,905. According to the U. S. Patent No. 6,157,905, the language identification is achieved by collecting statistical occurrence of particular data and comparing the statistical occurrence with generalized language characterization. The method is further composed of a special pre-processing step, named "plurality-to-one reductive mapping". The "plurality-to-one reductive mapping" is applied to text strings of a language to be identified prior to generating a language profile and identifying the language.

A further method of language identification is disclosed in the United States Patent No. 6,272,456. According to the U. S. Patent No. 6,272,456, a language is identified by scoring the language based on frequency parameters of matched reference letter sequences (RLS) in N-gram profile. A score of a language is obtained by the longest matched RLS of the language.

Currently, it is necessary for displaying a script of a certain language that appropriate language coding must be selected so as to render a text. In conventional products or systems for language coding, which support multiple languages, the selection of appropriate language coding is based on system language settings. However, only one fixed language coding is normally set at a time, so that a script of other languages can not be displayed correctly. Consequently, displaying scripts of different languages results in much inconvenience, because language setting is manually changed for respective languages.

Accordingly, the method of language identification according to the prior arts causes problems such that extra steps for identifying multiple languages are required and lots of system resources are necessary for identifying language coding.

SUMMARY OF THE INVENTION

Accordingly, in consideration of the above-mentioned problems of the prior arts, an object of the present invention is to provide a method of language coding identification and a data format therefor, which can identify each language automatically and also select appropriate language coding for each language automatically and further minimize system resources.

In order to achieve the above object, the present invention provides, according to an aspect thereof, a method of language coding identification for identifying different language coding systems through phasing an identification process, the method comprising the steps of: checking whether or not a text of a language to be identified is ASCII coded by examining a data value of each byte of a string in the text; applying the double-byte coding identification method to the text for obtaining an identification result in case the step of checking is not successful; applying the Latin-based coding identification method to the text for obtaining an identification result in case the step of applying the double-byte coding identification method is not successful; and applying a single-byte coding identification method to the text for obtaining a final identification result in case the step of applying the Latin-based coding identification method is not successful, wherein the single-byte coding identification method further comprises the steps of: generating a language profile of each supported language coding by analyzing a sample text and extracting occurrence possibility of all two character sequence combinations in the sample text; calculating a confidence value of each language coding by checking the occurrence possibility of all the two character sequence combinations in a text string of the sample text by the respective language profile; and selecting language coding with the highest confidence value as an identification result.

Other object and further features of the present invention will be apparent from the following detailed description when read in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flow chart showing a method of language coding identification according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[Embodiment]

Prior to detailing an embodiment of the present invention, background of the present invention is described first.

A method of language coding identification disclosed in the present invention is based on the international standards for the language characteristics and the language coding rules. A great number of languages are spoken in the world. According to the present invention, the great number of languages is categorized into several language groups. Then whole identification process is phased in several steps, wherein each step contains an identification process for a group of defined languages. Different technologies and algorithms are applied to a different group of languages so as to maximize accuracy of identification. The method just makes use of statistical facts of a language coding instead of linguistic features of the language, so that minimal system resources are required.

Further, the method of language coding identification is only based on a

pure script text, so that scripts of all languages enable to be displayed correctly without changing language setting for each language.

Accordingly, the method of language coding identification disclosed in the present invention enables to be easily applied to any microcontrollers and embedded systems with least system resources.

FIG. 1 is a flow chart showing a statistical stepwise identification (SSI) process that is adopted in a method of language coding identification according to the present invention. As shown in FIG. 1, the SSI process is phased and sequenced in four steps (S1 to S4) such as ASCII (American Standard Code for Information Interchange) coding identification (S1), double-byte coding identification (S2), Latin-based coding identification (S3), and single-byte coding identification (S4).

The ASCII coding identification examines each binary value of all characters in a text of a language to be identified (step S1). In case all binary values in the text are within a range of "0x00" to "0x7F", identification is successful and the text is identified as encoded by the ASCII coding system.

On the contrary, in case a binary value is out of the range of "0x00" to "0x7F", the ASCII coding identification is not successful. The text is conducted to a double-byte coding identification process according to the language coding rules (step S2). The double-byte coding identification mainly supports language coding of East Asian languages such as Chinese, Japanese and Korean, wherein each character of such a language is encoded by a double byte code. Identification of double-byte coding is based on an inherent code range of each coding system.

In other words, for language coding of each language to be identified, the identification is to determine whether or not the two-byte code is a valid character in accordance with the respective language coding rules.

Further, for some languages, the identification is also able to indicate

whether a character is a commonly used character or a less commonly used character. An identification result is concluded on the basis of statistical results of all two-byte characters.

With respect to Chinese, Japanese and Korean characters, actual coding facts are as follows:

1. Simplified Chinese (GB2312)

Simplified Chinese is based on simplified Chinese characters and commonly circulates in People's Republic of China (mainland China) as the standard Chinese, Singapore and also some other countries. The GB2312 contains 7445 characters. Out of a total of 7445 characters, 6763 characters are assigned to simplified Chinese characters. The GB2312 is totally divided into 94 regions, wherein first byte representing each region ranges from "0xA1" to "0xFE". Each region is divided into 94 slots, wherein second byte representing each slot ranges from "0xA1" to "0xFE".

Further, a range from "0xA1A1" to "0xAFFE" is assigned to numerals, ASCII characters, and characters other than Chinese.

Furthermore, a range from "0xB0A1" to "0xD7FE" is assigned to most commonly used Chinese characters, which are defined as level 1 characters and composed of 3755 characters in total.

More, a range from "0xD8A1" to "0xFEFE" is assigned to less commonly used Chinese characters, which are defined as level 2 characters and composed of 3008 characters in total.

2. Traditional Chinese (BIG5)

Traditional Chinese is based on traditional Chinese characters and mostly circulates in Hong Kong and Taiwan (Republic of China). The BIG5 is an encoding system of which first byte ranges from "0xA1" to "0xF9" and second byte ranges from "0x40" to "0x7E" and "0xA1" to "0xFE", and composed of 13868 characters.

In the BIG5, beside characters assigned to ASCII and special characters, 5401

characters are assigned to most commonly used Chinese characters, which are defined as level 1 characters and range from "0xA440" to "0xC67E".

Further, 7652 characters are assigned to less commonly used Chines characters, which are defined as level 2 characters and range from "0xC940" to "0xF9D5".

3. Japanese (Shift-JIS)

The Shift-JIS (SJIS) is a Japanese character encoding system of which first byte ranges from "0x81" to "0x9F" and from "0xE0" to "0xFC" and second byte ranging from "0x40" to "0x7E" and "0x80" to "0xFC".

4. Korean (K-EUC)

The K-EUC is a Korean character encoding system of which first byte ranges from "0x81" to "0xC8" and from "0xCA" to "0xFD" and second byte ranges from "0x41" to "0x5A", "0x61" to "0x7A" and "0x81" to"0xFE", and else second byte ranges from "0xA1" to "0xFE".

Referring back to FIG. 1, in the step S2, in case the double-byte coding identification is not successful, the SSI process is proceeded to a Latin-based coding identification process (step S3). Generally, Latin languages share the same characteristics such that majority of characters in a text are coded by the ASCII coding system. Consequently, following parameters shall be examined for the Latin-based coding identification:

A. Percentage of non-ASCII characters in a text, and

B. Percentage of non-ASCII two character sequence in the text.

In case both percentages of a text of a language to be identified are below a predetermined threshold respectively, the text is identified as a Latin based language coded by the Latin-based coding system.

On the contrary, in case both percentages exceed the predetermined threshold, the text is identified as a non-Latin based language. As a result, the Latin-based coding identification is not successful.

In case the Latin-based coding identification is not successful, the SSI process is proceeded to a single-byte coding identification process (step S4). Target languages of the single-byte coding identification method are all of other alphabetic character based languages, which are not supported by the preceding coding identification methods in the step 1 through the step 3 mentioned above.

In case of a particular language that uses only a small number of characters, it is necessary for such a particular language to apply further processes of counting occurrences of each single character. It is commonly known that combination of characters reveals more language characteristic information.

Accordingly, in the method of language coding identification according to the embodiment of the present invention, a two character sequence is defined as two characters that appear immediately one after another in an input text. In this connection, order of such a sequence is significant in this case. Since not all characters are used equally and frequently in one language, a distribution of two character sequence also turns out to be language or encoding dependent extremely. Consequently, such a characteristic of two character sequence enables to be used in language coding identification. Using a characteristic of two character sequence leads to better confidence in identifying a character encoding system and is very useful for identifying languages in single-byte coding accurately.

According to results of research and study conducted by the inventors of the present invention, non-ASCII two character sequence is possibly 16384 in total. In 16384 sequences, only 1000 to 2000 sequences, that is, about 10 % of 16384 sequences are just used in one language coding.

A method of the single-byte coding identification according to the embodiment of the present invention is to analyze and calculate a statistical

result of occurrences of combination of two character sequence in each supported language, and then to compare the calculated statistical result of occurrences with a statistical result of occurrences in a text of a language to be identified.

The method of the single-byte coding identification is processed as follows:

1. Analyze each sample text of a language to be identified and generate a profile of two character sequence.

This process is a step of preprocessing of the single-byte coding identification. The preprocessing is not included in the identification process. However, a generated profile is a key factor for an identification result. A sample text to be analyzed is previously provided for each language. A number of occurrences of a two-byte non-ASCII character combination in the sample text is calculated, wherein a binary value of the two-byte non-ASCII character is greater than "0x7F". The number of occurrences is stored in a two-dimensional array formed in a predetermined format. To reduce the array size, a result of each two character sequence should be set to 1 bit only.

More specifically, value "0" denotes that the character combination is unlikely to appear in the language. On the other hand, value "1" denotes that the character combination is likely to appear in the language.

Further, bit operations are used to access the array and a data file, so that a result of eight two character sequences enables to be stored in a form of one byte (equal to 8 bits).

Furthermore, each language contains an identification profile in a fixed size of 2048 bytes.

Main algorithm of the step of analyzing each sample text of a language to be identified and generating a profile of two character sequence is as follows:

WHILE (Sample File NOT End)

{

IF ((1st byte>0x7F) && (2nd byte>0x7F))

SET Result Array [1st byte–0x80] [2nd byte–0x80] to 1;

ELSE

SET Result Array [1st byte–0x80] [2nd byte–0x80] to 0;

GOTO next character;

}

The identification profile is defined as an array of [128] [16].

Setting Array [X] [Y] to 1 uses bit operations:

Array [X] [Y] = Array [X] [Y] OR (0x80>>(Y%8))

Setting Array [X] [Y] to 0 uses bit operations:

Array [X] [Y] = Array [X] [Y] AND (NOT(0x80>>(Y%8)))

2. Calculate a confidence value for each language.

This process is a step of mapping the sample text to be identified with profiles of all supported languages, and calculating a confidence value of each language. A number of all occurrences of a combination of two-byte non-ASCII characters in the sample text to be identified are extracted first, wherein a binary value of both characters is more than "0x7F". Then the number of occurrences is compared with the profiles of all supported languages. Finally, each supported language obtains a confidence value.

Algorithm of the step of examining the occurrence for byte X, Y is as follows:

IF (Array [X-0x80] [Y-0x80/8] &0x80>>(Y-0x80%8))

Return (TRUE);

ELSE

Return (FALSE);

3. Decision.

Decision is made on the basis of the confidence value from the final step, that is, language coding having the highest confidence value is defined as an

identification result.

The method of language coding identification according to the present invention provides a simple yet efficient method to identify language coding of a pure text. Pre-defined language groups and individual identification methods of the groups ensure better accuracy of identification of language coding. Extremely low system requirements make the method suitable for less expensive consumer electronic products.

While the invention has been described above with reference to a specific embodiment thereof, it is apparent that many changes, modifications and variations in configuration, materials and the arrangement of equipment and devices can be made without departing form the invention concept disclosed herein.

In addition thereto, it will be apparent to those skilled in the art that various modifications and variations could be made in the method field in the present invention without departing from the scope of the invention.

WHAT IS CLAIMED IS

1. A method of language coding identification for identifying different language coding systems through phasing an identification process, the method comprising the steps of:

checking whether or not a text of a language to be identified is ASCII coded by examining a data value of each byte of a string in the text;

applying the double-byte coding identification method to the text for obtaining an identification result in case the step of checking is not successful;

applying the Latin-based coding identification method to the text for obtaining an identification result in case the step of applying the double-byte coding identification method is not successful; and

applying a single-byte coding identification method to the text for obtaining a final identification result in case the step of applying the Latin-based coding identification method is not successful,

wherein the single-byte coding identification method further comprises the steps of:

generating a language profile of each supported language coding by analyzing a sample text and extracting occurrence possibility of all two character sequence combinations in the sample text;

calculating a confidence value of each language coding by checking the occurrence possibility of all the two character sequence combinations in a text string of the sample text by the respective language profile; and

selecting language coding with the highest confidence value as an identification result.

2. The method of language coding identification as claimed in claim 1, the method comprising one or more steps including the step of applying the

single-byte coding identification method in any sequence order.

3. The method of language coding identification as claimed in claim 1, wherein all the steps are achieved on the basis of a statistical fact of language coding instead of linguistic feature of the language for reducing memory usage and calculation complexity.

4. The method of language coding identification as claimed in claim 1, wherein the double-byte coding identification is achieved by checking all of two-byte characters according to an inherent code rule of each coding system.

5. The method of language coding identification as claimed in claim 1, wherein the Latin-based coding identification is achieved by checking percentage of non-ASCII characters as well as percentage of non-ASCII two character sequence in the text.

6. A data format for the method of language coding identification as claimed in claim 1, wherein the language profile is stored in accordance with a full set of two character sequence combination occurrence possibility bit series.