Acta Technica *61*, No. 4A/2016, 11–24

Research on static image feature extraction based on hierarchical structure and sparse representation[1]

Juan Xiong[2]

Abstract. A new technique to improve feature extraction task is proposed. The main benefit of using scaling process is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Moreover, the adoption of this developed technique achieves best accuracy with low computational complexity for object recognition. Note that, through experiment, the method has been proven to be accurate at 63.75 %. We concluded that the recognition of image is hard and complicated not only because of the handwriting ambiguity but also due to the similarity between characters and their positions in a word. This paper has presented a simple scheme for binary template selection in a context of feature extraction based on entropy. It is experimentally shown that the proposed algorithm can be employed to select an effective template. In addition, this method uses low computational processes, which provide hierarchical sparse method (HSM) to increase recognition rate within fewer computations and short time.

Key words. Static image, feature extraction, hierarchical structure, sparse representation.

1. Introduction

The first stage in image recognition is how to make a computer know the content of image, regarding that computer can only process mathematical computations. Scientists proposed to compute the likelihood between digital images, but the remaining problem is the likelihood of what [1–2]. Composed of hundreds of pixels, an image set incurs a huge amount of calculations that computers can hardly afford. Depending on the scientific fact mentioned previously, in computing with images, it is more suitable to work with both the notions of digital image and analog image. The image function is a mathematical model that is frequently used in analysis where it is profitable to consider the object (i.e. image) as a function of two variables. Consequently, for analyzing images, all of functional analysis is then available. The

[1]The author acknowledges the National Natural Science Foundation of China (Grant 51578109 and Grant 51121005).

[2]Huanghuai University, Henan, 463000, China

digital image is just a 2D rectangular matrix of discrete values. In order to allowing the image to be stored in a 2D computer memory structure, both image space and intensity range are quantified into a discrete set of values [3–5]. Figure 1 shows the instance of mathematical vision of image.

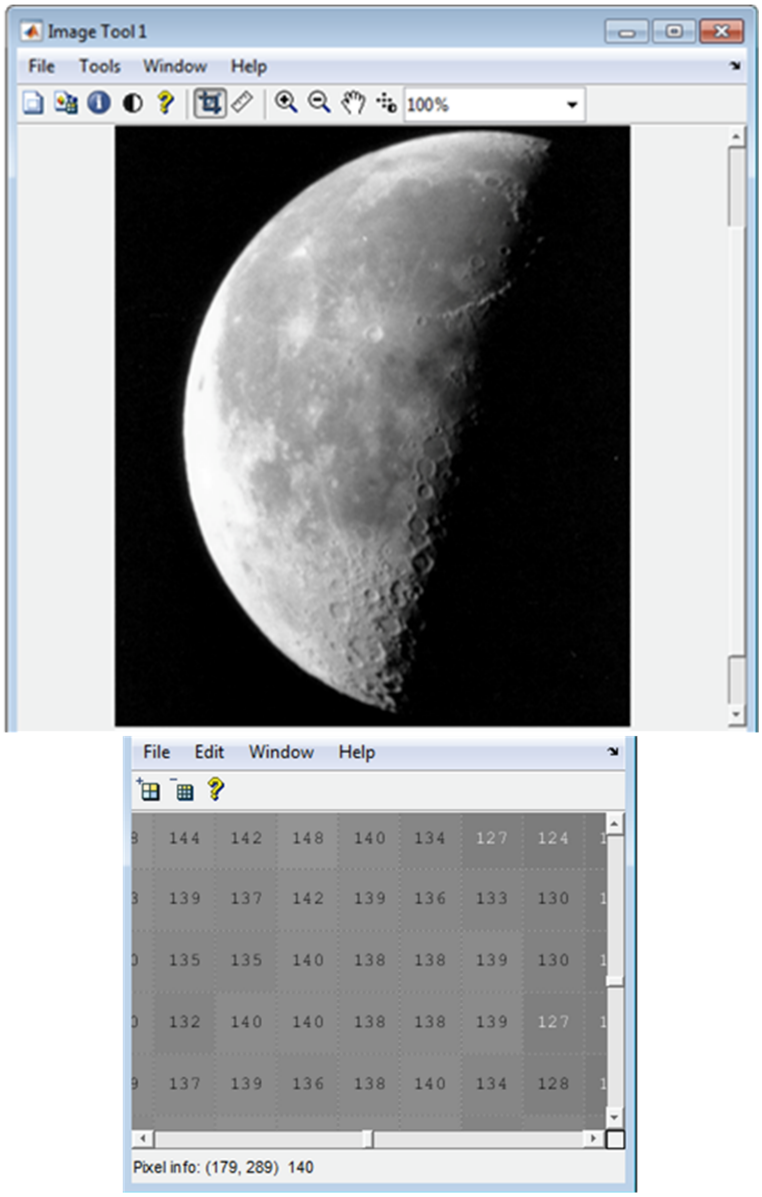

Fig. 1. Instance of mathematical vision of image

2. Feature extraction

The significance of feature extraction is mainly due to three reasons [6]:

Data Reduction: Feature extraction can be viewed as a powerful data reduction tool via reducing the space of measurement and storage requirements. For instance, when a machine learning program is given too many variables to operate, a large number of features do not always lead to better decision and the running time space storage also increases. Therefore, it is necessary to select a much smaller number of features as they are more important and more relevant.

Automatic Investigation and Data Mining: In many classical applications, discriminatory and informative features are often selected as a priori experts in the field of feature extraction, i.e., investigators pick out which are the most important variables to build a model. More and more often in modern data-mining applications, however, there is a growing demand for fully automated "black-box" type of prediction models that are capable of identifying the important features on their own. The need for such automated systems are arises for two reasons. On one hand, there are economic needs to process large amounts of data in a short period of time with a little manual supervision. On the other hand, sometimes the problem and the data are so novel moreover, there are simple no field experts who can understand the data well enough and be able to pick out the important variables prior to the analysis. Under such circumstances, automatic exploratory data analysis becomes the key. Instead of relying on pre-conceived ideas, there is a need (as well as interest) to let the data speak for it.

Data Visualization: The last but not the least application of feature extraction that shares the flavor of exploratory data analysis is data visualization. Here, this concept can be best understood by considering examples of its applications. The human eye has an amazing capability in recognizing systematic patterns in the data. At the same time, human eyes are usually unable to make good sense for data if it is more than three dimensions. To maximize the use of the highly developed human faculty in visual identification, we often wish to identify two or three of the most informative features in the data so that we can plot the data in a reduced space.

3. Hierarchical sparse method and algorithm

The basic assumption underlying hierarchical learning algorithms is that each input can be decomposed to a hierarchy of parts of increasing size with increasing semantic complexity. In fact, the hierarchy is useful for reducing the sample complexity of the problem. Given the representation of the smallest parts, the hierarchical architecture recursively builds at each layer a representation of the next larger parts by using a combination of sparse coding and pooling operations. Intuitively, the sparse coding step induces discrimination while the pooling step induces invariance in the architecture. Thus, the alternating applications of the sparse coding and pooling operations yield a complex representation of the input data with nontrivial discrimination and invariance properties. As natural images, can be sparsely represented by a set of localized, oriented filters, therefore, by imposing the norm

regularization on representation coefficients, sparse coding can be solved efficiently.

Recent progress in computer vision has demonstrated that sparse coding is an effective tool for representing visual data at different levels, e.g. image classification and image delousing. In simple word, a code is sparse if most of its components are zero. The idea is based on a simple concept that high dimensional signals can be represented as a linear combination of very small number of basis function taken from dictionary. Commonly, intensity record is an 8-bit (1-byte) number which permits values of 0 to 255. It is worth mentioning that 256 different levels is generally enough to satisfy the consumer and also mostly represents the precision available from the sensor, as well as bytes suitable for computers. The following definitions are intended to clarify important concepts and also to establish notation used through this research. An image is generally 3D, but mostly represented in 2D on the computer. Analog images are 2D images $F(x, y)$ which have infinite precision in spatial parameters x and y and infinite precision in intensity at each spatial point (x, y). Digital images are 2D images [row; col] represented by a discrete 2D matrix of intensity samples, each of which being represented by using a limited precision. It can be stored in physical memory (like hard drives) and is easier to process. Raster images are represented as a 2D array of pixels. A pixel is the smallest visual element of a picture. The resolution is defined as the total number of pixels in a picture. Aspect Ratio refers to the ratio of width to height of a picture. Binary images are the digital images that comprised of two possible colors for each pixel (i.e. white (1) and black (0)). Gray scale images are comprised of only shades of gray (i.e. no color) in between white (255) and black (0). Color images are the digital images that are formed by a combination of different colors for each pixel. The depth of an image denotes the number of shades of color in between 1 and 0 in a picture. A coordinate system must be used to address individual pixels of an image (as shown in Fig.i̇); to operate on it in a computer program, refer to it in a mathematical formula, or to address its device-relative coordinates. The mathematical model of an image as a function of two real spatial parameters is enormously useful in both describing images and defining operations on them. A picture function is a mathematical representation $f(x, y)$ of a picture as a function of two spatial variables x and y, where symbols x and y are real values defining points of the picture and $f(x, y)$ is usually a real value describing the intensity of the picture at point x, y.

Formally, if x is a column signal and D is the dictionary (whose columns are the atom signals), the sparse representation of x is obtained by carrying out the following optimization

$$\min \|s\|_0 \, s \bullet t \bullet x = Ds \,, \tag{1}$$

where s is the sparse representation of x and $\|\|_0$ is the pseudo norm which counts the non-zero entries. The nonlinear mapping approach has been defined as the following nonlinear mapping function

$$N^v = f(x^v, T_{\mathrm{u}}) \tag{2}$$

where T_{u} is the image patch of size u. The single matrix T_u is defined as

$$T_u = [t_1, t_2, ..., t_m] = \begin{bmatrix} t_{11} & t_{21} & ... & t_{m1} \\ t_{12} & t_{22} & ... & t_{m2} \\ ... & ... & ... & ... \\ t_{1n} & t_{2n} & ... & t_{mn} \end{bmatrix},$$

where t_i is the ith candidate. Figure 2 illustrates the sparse coding operation. Through this research, in particular, the goal of sparse coding is to represent a training image signal x approximately as a weighted linear combination of small numbers of dictionary (e.g. basis vectors). Generally, in the class of hierarchical architectures that we have considered in suggested technique, the inputs to the sparse coding operation will in general have different lengths from layer to layer. To cope with this problem, we have defined the sparse coding S on a sufficiently large space which contains all the possible inputs which works with the restrictions of S on the appropriate domains of the inputs.

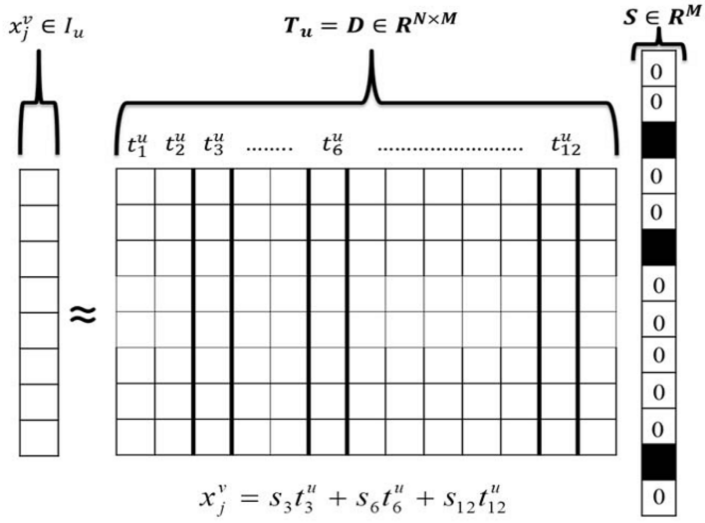

Fig. 2. Scheme of sparse coding operation

A spatial pooling stage is a very important step in many of the computer vision architecture. Since it combines the responses of feature detectors obtained at adjacent locations into some statistic that summarizes the joint distribution of the features over some region of interest. It is worth mentioning that, the pooling operation is typically an average, a max, a sum, or more rarely some other commutative (i.e., independent of the order of the contributing features) combination rules. Meanwhile, the pooling operation can be described as a function that summarizes the content of a sequence of values with a single value, similar to aggregation functions used in voting schemes and database systems. Following sparse coding, the inputs to the

pooling operation generally have different lengths at different layers but the actions of pooling operation on input values is not related to the layers of sparse coding. We now turn to describe the mathematical framework formalizes the hierarchical structure of the architecture, that each input is composed of parts of increasing size. Figure 3 shows the domains of nested patch.

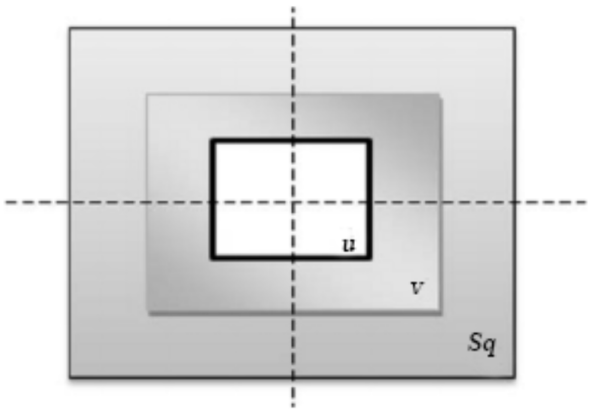

Fig. 3. Domains of Nested patch

4. Experiment result and data analysis

Feature extraction is a long-standing research topic in computer vision. It has become the main focus and objective of most researches in the areas of both computer vision and machine learning because the good feature extraction is a central to achieve high performance in any computer vision task. Nevertheless, there is still a need to develop efficient feature extraction algorithm that can represent an informative properties of an object. This paper is concerned with problems of developed hierarchical feature extraction method to perform a successful recognition target. The latter refers to approach of computer Science interested in giving the computer human learning capability. In other words, how to build an efficient predictive model using a computer? In this part, overview feature extraction is considered. Definition of Feature Extraction: Indeed, feature can be defined as a scale on which human can easily recognize a collection of objects. On the other hand, feature extraction can be defined as the problem of finding the most relevant and informative set of features to improve the data representation for classification and regression task. Actually, we are able to extract informative features in our everyday lives. For example, we can easily identify person's sex from a distance, without examining full characteristics of the person. This is because a certain signature for the two genders was known, e.g., body shape, hair style, or perhaps a combination of the two. In other words, we can say that, it is not necessary for us to process all the charac-

teristics of items to be capable to recognize them. In this sense, the goal of feature extraction method is finding feature which is informative and relevant in order to give the computer its ability to understand and simulate the operation of the human vision system. To produce such plots, feature extraction is the crucial analytical step. Feature Extraction and Feature Selection Feature extraction is one of the key steps in both computer vision and machine learning. It becomes the focus of much research, because a good feature extraction is a central to achieve high performance in any computer vision task. Actually, feature extraction includes simplifying the amount of resources required to distinguish a large set of data accurately. Practically, feature extraction concept can be decomposed into two consecutive phases: feature construction and feature selection. On one hand, in feature construction the step obtaining all features that appears reasonable but it causes increase in the dimensionality of the data and thereby immerses the relevant information into a sea of possibly irrelevant, noisy or redundant features. Here, we can point out some of generic feature construction approaches including: basic linear transforms of the input variables (PCA/SVD, LDA); clustering; singular value decomposition (SVD); applying simple functions to subsets variable like products to create monomials; more sophisticated linear transforms like spectral transforms, wavelet convolutions or transforms of kernels.

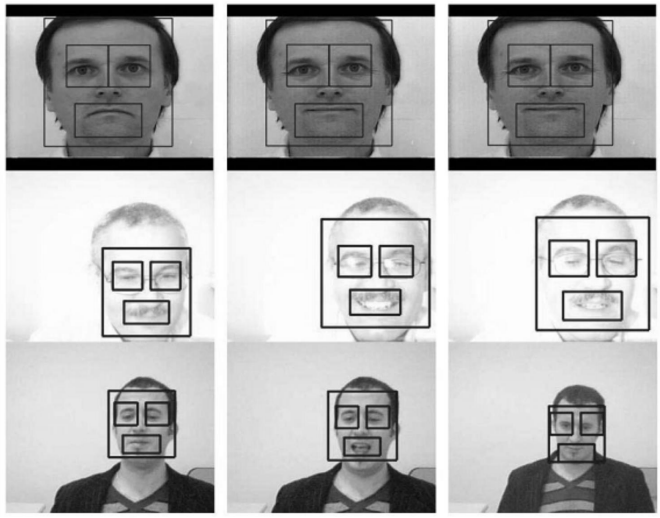

Fig. 4. Example of feature extraction for face recognition problem

Nested Feature Subset Selection Methods A number of learning machines extract features as part of the learning process. Practically, there are two types of nested methods: (1) backward elimination styles and (2) forward selection styles. These contain neural networks whose internal nodes are feature extractors. Figure 4 shows the example of feature extraction for face recognition problem. Definition of Feature Hierarchies: is a technique used features composed of image patches during a

learning step. Indeed, such tactic was often based on natural modeling, motivated by the structure of the primate visual cortex. Mainly due two reasons this algorithm is successful: First, they detect common object components that characterize the different objects within the class and secondly, the components are combined in a way that allows differences can be learned from training data. To make this notion clearer this example can be taken: the part itself (such as an eye in face detection) is decomposed into own optimal components (e.g. eye corner, eye pupil, eyelid, etc.), and the allowed variations in the configuration of the sub-parts are learned from the training data (an example is given in Figure 5.

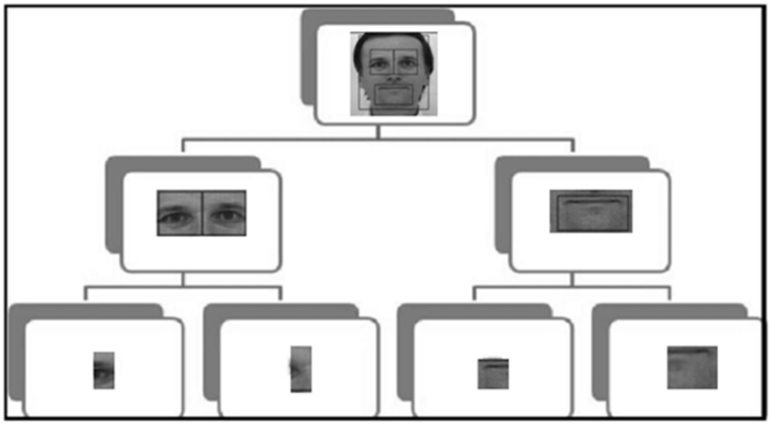

Fig. 5. Examples of the hierarchies used in the proposed algorithm

The hierarchical architecture has a key semantic component; a dictionary of templates that is usually learned from data. The templates will play the role of the sparsely through the proposed model, where this template is used as set of the dictionary in the sparse coding operation. The advantage to this idea of representation is the template set which is adapted to the data. The template also links the architecture to the underlying distribution of the input data. One way to create template is to sample from the probability distribution on the function space. Finally, the paper determines template sets as $\mathrm{T}u \subset \mathrm{Im}(u)$ and $\mathrm{T}v \subset \mathrm{Im}(v)$, that are considered to be finite, discrete, and endowed with the uniform probability measure (see Fig. 6). Actually, the success for the sparse representation features depends heavily on a good choice of dictionary.

Experimental result show that for some classes of signals, learned dictionaries can be benefited from template sets, which ultimately lead to a similar/better recognition performance in comparison with other classical methods. For more detail, please refer to [7–10]. In this paper, we applied proposed method on the two domains images and speech.

For the domain of images, we evaluated hierarchical sparse on the well-known MNIST digit recognition benchmark, and COIL-30 dataset. In the other hand for the speech, isolated words speech recognition is selected. We tested our recognition algorithm using training data and testing data from two distinct vocabularies. After

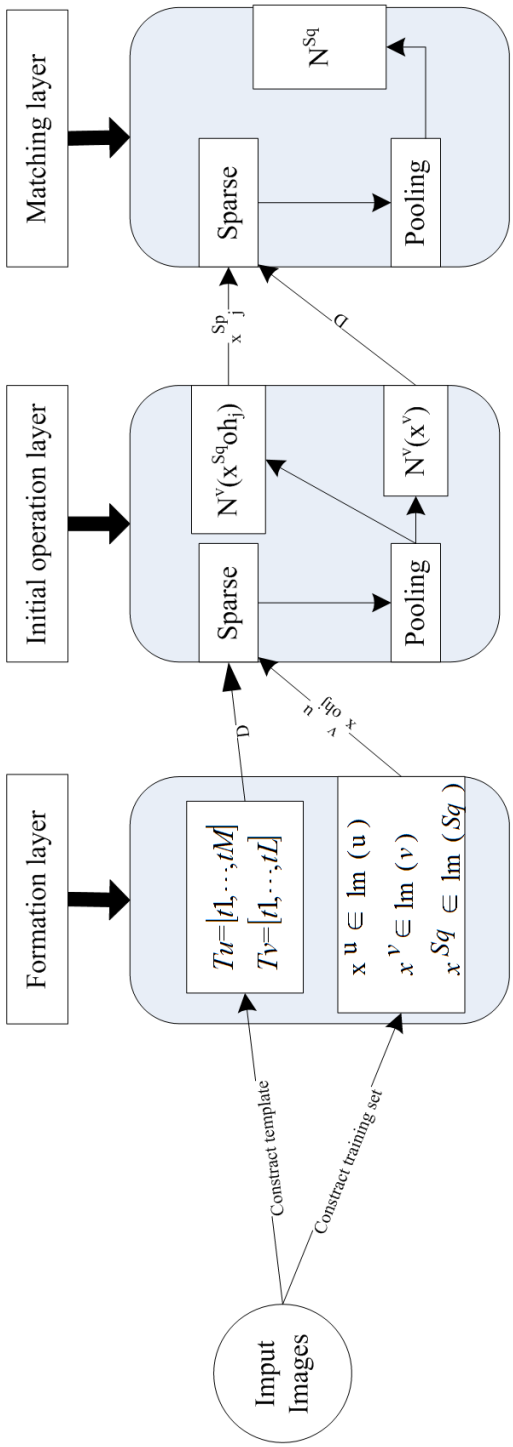

Fig. 6. Layered proposed hierarchical structure of sparse representation method

that we used the representation as a feature extraction step for a classification algorithm such as Support Vector Machines (SVM) and 1-Nearest Neighbors (1NN). We first verified developed method with object recognition experiments using the M-NIST hand-written digit recognition benchmark, where there are 70.000 data examples, and each is of 28×28 gray scale images. In the experiments, we used some images randomly selected from the MNIST data set. We considered eight classes of images: 2s through 9s (see Fig. 7).

Fig. 7. Eighteen instances from the set of training examples for Coil-30

The digits in this dataset include a small amount of natural translation, found in a corpus containing the handwriting of human subjects. The labeled image sets that we have used contain 5 examples per class, while the out-of-sample test sets contain 30 examples per class. The Tu and Tv are template sets constructed by randomly extracting 500 image patches (of size u and/or v) from images, which are not used in the train or test sets (in experiments we set $D = Tu$). For the digit dataset, templates of size 10×10 pixels are large enough to include semi-circles and distinct stroke intersections, while larger templates, closer to 20×20, are seen to include nearly full digits were more discriminatory structure is present. For the experiments we set the first layer template $u = 11 \times 11$ pixels and the second layer template $v = 19 \times 19$ pixels. After the features are learned we can obtain classification accuracy by applying a $k-NN$ with $k = 1$ and SVM note that classifier are averaged over 50 random test sets, holding the training and template sets fixed. Experiment is performed under the same Smale's environment, comparison of results shows that the

developed model consistently outperforms the Smale's methods. Here we adapted our model to the case of one-dimensional of length n. We built a template in this setting by considering patches that are segmentation of original signal(i.e. the word is segmented into sub-word units as shown in Fig. 8, and the transformations are taken to be all possible translations. In the experiment, the used dataset consists of seven different names of fruit consists of "apple", "banana", "kiwi", "lime", "orange", "peach", and "pineapple". The algorithm is tested for the percentage of accuracy. We tested the ten utterances of each seven words, while the training was done five utterances of each seven words (i.e. the first 5 utterances in the corpus are kept as training set. The left utterances are used for testing.) We tested proposed method for the speech signals as an input instead of the images; features extracted from the speech signal are passed to each word as shown in Fig. 9.

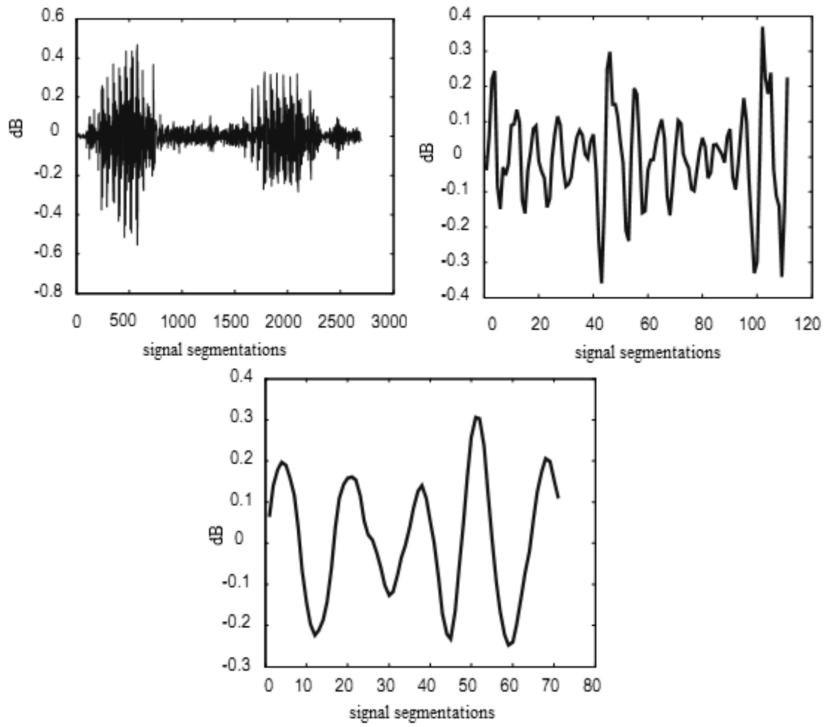

Fig. 8. One example from the set of template (Isolated Word Speech): up left–full word (Apple) of size 2694, up right–subword (Apple) of size 110, bottom–subword (Apple) of size 70

5. Conclusion

Feature extraction algorithms considered as a pillar key task to make vision modeling systems fully operative. It has been effectively utilized to minimize the

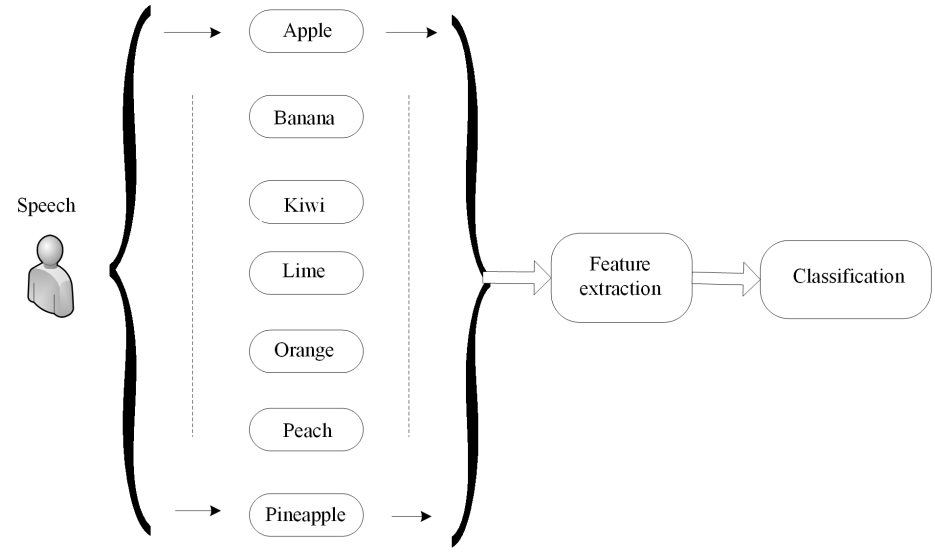

Fig. 9. Isolated Word Speech recognition process

computation difficulty and to perform ideal classification through extraction of the significant pattern information. An essential component of a successful classification system is the selection of an effective object features. Throughout this paper, we seek to tease out basic principles that underlie the recent in hierarchical feature extraction method. The paper introduces a new algorithm for template selection based on the entropy concept. Algorithm suggests picking of the template of more information and discards the templates of less information. The proposed method provides HSM with better discriminatory ability. Experimental results show that the introduced method achieves good performance in template selection with fewer computation processes and shorter time.

References

[1] F. J. Pulido, L. Mandow, J. L. Perez-de-la-Cruz: *Dimensionality reduction in multiobjective shortest path search.* Computers & Operations Research *64* (2015), 60 to 70.

[2] A. A. Agafonov, V. V. Myasnikov: *Method for the reliable shortest path search in time-dependent stochastic networks and its application to GIS-based traffic control.* Computer Optics *40* (2016), No. 2, 275–283 (in Russian).

[3] F. J. Pulido, L. Mandow, J. L. Perez-De-La-Cruz: *Multiobjective shortest path problems with lexicographic goal-based preferences.* European Journal of Operational Research *239* (2014), No. 1, 89–101.

[4] M. Maire, X. Y. Stella, P. Perona: *Reconstructive sparse code transfer for contour detection and semantic labeling.* Lecture Notes in Computer Science, *90069* (2014), 273–287.

[5] S. Zhang, X. Xu, L. Lu, Y. Chen: *Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems.* Proc. IEEE Globecom 2014—Wireless Networking Symposium, 8–12 Dec. 2014, Austin, TX, USA, 4782–4787.

[6] D. Ni, H. Ma: *Hyperspectral image classification via sparse code histogram.* IEEE Geoscience and Remote Sensing Letters *12* (2015), No. 9, 1–5.

[7] K. Badni: *Windows and mirrors – interaction design, digital art, and the myth of transparency.* The Design Journal (An International Journal for All Aspects of Design) *7* (2004), No. 1, 57–58.

[8] X. W. Liu: *Environmental art design based on digital technology.* Applied Mechanics & Materials *543–547* (2014), 4145–4148.

[9] S. Dhakal, A. Bayesteh, S. Hranilovic, A. Mobasher, T. Sexton: *Sparse codes for MIMO channel and detector alternatives for sparse code.* Patent PCT/US2011/055361, `http://www.google.st/patents/US20130182791`.

Received November 16, 2016