APPROACH TO THE AUTOMATIC NONVERBAL BEHAVIOR RECOGNITION SYSTEM

Knyazev B.A.[1,2], Latysheva E.Yu. [1,2], Gapanyuk Yu.E. [1,2]
Scientific supervisor: Professor Chernenkiy V.M.
[1]Bauman Moscow State Technical University,
[2]Research and Development Center of Biometric Technology at BMSTU
5, 2-nd Baumanskaya Str. Moscow 105005, Russia
E-mail: bknyazev@bmstu.ru

Introduction

The human's necessity of activity and motions is innate and depends on the age, gender, environment circumstances, biorhythms and many other factors. Activity features reflect the state of physical health, the level of motor, psychological and intellectual development [1]. Human activity and motions are an integral individual characteristic, and if decomposed provides valuable information. Automatic nonverbal behavior (NVB) recognition systems might be applied to solve security problems and medical and psychological diagnosis and treatment tasks as well [2]. The object of this work is to provide an approach to develop the system of automatic nonverbal behavior recognition and semantic annotation, which includes a formalized knowledge model, a model of the NVB features space and machine learning techniques. Let O be the chosen engineering knowledge model; $D = \{d_j\}$ – a 2- and/or 3-dimensional input; $H(O)$ – an entropy of the knowledge model; $S = \{s_i\}, i = 1, N$ – the NVB feature vector in a N-dimensional space.

Thus, the solution of the problem of the development of the automatic NVB recognition and annotation system implies finding the following mappings:

$$\begin{matrix} \varphi|_{N\to min, Q\_seq\to min}\colon D \mapsto S \\ \psi|_{H\to min}\colon S \mapsto O \end{matrix}, \quad (1)$$

where Q_seq – the number of sequential data processing operations.

Knowledge model

Finding ψ requires creating the knowledge model. Ontology models are effective, flexible, allow multilevel representations and Web integration, are rich in samples and API for development [3]. Let $O = (C, P, R, I, A)$ be the ontology, where C – concepts, P – properties, R – relations, I – instances, A – axioms. Then, the entropy of the ontology, which was suggested earlier in [4], is a measure of the uncertainty of information received by querying the block *Reasoner* (fig. 1). The condition $H \to min \equiv II \to max$ (II – quantity of information) of the knowledge model O can be reached iff (if and only if) $R \to max$, that is the number of relations between the resultant instance(-s) and the others must tend to a maximum. This is achieved by extending the intertwined network of C, I and R developed for pose description [2]. The knowledge model is developed using the Web Ontology Language extension OWL DL. Decidability and completeness of this model is achieved due to support of the description logic (DL) included in this language; whereas consistency is justified using the DIG (DL standard) compliant reasoner FaCT++ and Pellet embedded in the ontology designer Protégé.

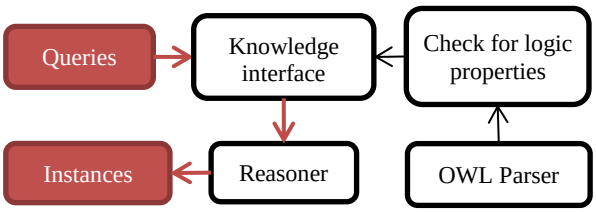

Fig. 1. The knowledge model

Base concepts <owl:Class> and derived instances <rdfs:subClassOf> introduced in the model together with properties <owl:ObjectProperty> and manifold relations between concepts and their instances allow to infer and aggregate knowledge in an automated way and to make a semantic video annotation.

NVB features

Finding φ requires the development of the NVB feature space model based on low-level processing of video frames and the development of a classifier. The techniques of edge and skeleton (of people) detection have been well studied [5] (differential, statistical and contextual detectors, Gabor and Gaussian wavelets, etc.) In addition, our research has shown that using of a Kinect device in particular conditions (illumination, the distance from the object to the device, etc.) yields results as accurate as 85-90%, therefore reasonable would be to focus on the techniques of object description and classification.

Object description and the condition $N \to min$ imply calculation of the vector S, which defines each video frame and NVB uniquely. For this purpose we conduct calculations of the uniform extension of the local binary patterns (uniform LBP) (fig. 2). To make the LBP more robust to image irregularities, such as orientation of the object, illuminations and contrast changes, it is recommended to preprocess images with Gabor filters of different orientations ($\theta \in \{0{:}7\}$) and scale ratios ($\gamma \in \{0{:}4\}$) (fig. 3). Classification then is done using the Euclidean distance. We also suggest using the Hamming distance, the Kullback–Leibler distance and the Fisher's linear discriminant which demonstrated more accurate results [6].

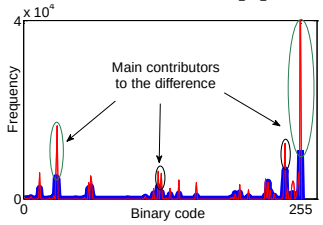

Fig. 2. Comparison of the histograms. Red – a closed eye, blue – an opened eye

The regions of interest (ROI) of video frames are: eyes, eyebrows, lips corners and others. Each region in turn is divided into 2-4 areas W (fig. 3), for which the feature vector (LBP) is computed. The size of the LBP is $P(P-1)+2|_{P=8}=58$, where P – the number of neighboring points to compute one binary code. The best results was achieved using the LBP with following parameters: $P,R=(8,2)$, where R – radius of a circle.

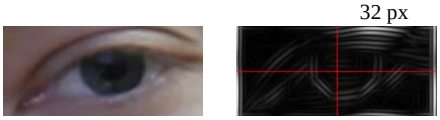

Fig. 3. Image preprocessing (left – original image, right – after the Gabor filter applied)

We also suggest using the histograms of oriented gradients (HOG) and their extensions. They have proved to be more robust to irregularities in shape and color and luminance components of the image, but they demand additional processing since their size is not constant. The size of the HOG depends on the size of the image, cells, and the number of bins; if $W=32\times16$ then its size can reach 3780 elements and more. This compels using reduction techniques, such as principal component (PCA) and linear discriminant (LDA) analyses.

The condition $Q\_seq\rightarrow min$ in the (1) requires using parallel computing. The size of an area for the LBP $W=32\times16$ because the maximum size of thread blocks in a GPU G84, which was used for computations, is 512; the warp size is 32 threads. Thus, each thread block of the GPU is able to calculate the feature vector (LBP) in a whole and independently of other blocks, and without any warp divergence, which is essential for GPU computing. Outcomes of the overall dimension reduction are partly shown in the table 1 (initial vector size equals to an HD frame, that is $N=1920\times1080$), where N – the dimension of a vector; Q_seq – the computational complexity if sequential computing is implemented; Q_par – if parallel is implemented.

Table 1. Feature vector dimension reduction

ROI	N	Q_seq	Q_par
Kinect skeletal joints	20	20	20
Eyes	$2\cdot58\cdot4$	$2\cdot16\cdot32\cdot4$	$2\cdot1\cdot4$
Eyebrows + between	$3\cdot58\cdot2$	$3\cdot16\cdot32\cdot2$	$3\cdot1\cdot2$
Lips corners	$2\cdot58\cdot2$	$2\cdot16\cdot32\cdot2$	$2\cdot1\cdot2$
Σ	**1064**	**9236**	**38**

Experiment and Results

We used two sources of scene data *D*: recorded using a video camera to extract information about facial motions and recorded using a Kinect device to extract information about body movements. Automatic analysis was run for 49 nonverbal features *S* using combinations of the CUDA API and .NET Framework 4.0. Semantic annotations *O* were then built automatically at the speed of $v\leq280$ $ms/frame$. To evaluate the effectiveness of our system false positive and false negative errors were estimated, which were ≤40% and ≤38% respectively for the body features and ≤43%, ≤42% for the facial ones.

Discussion

The mapping ψ also might imply usage of a linear/non-linear classifier, which together with the Fisher's linear discriminant must improve feature vector classification results. Using artificial neural networks (ANN) for pose classification demonstrated results of up to 80% [2]. However, this accuracy is poor for automatic NVB recognition. Furthermore building the ANN showed such disadvantages as laborious training and fitting the neural network weights and low processing speed. For these reasons the SVM (support vector machine) classifier is suggested. Further performance improvements are suggested to be done by parallel computing the feature vector for every ROI and frame independently, if decompressed videos are given.

Conclusion

An approach to solve the problem of the development of the automatic NVB recognition and annotation system is provided, the ontology model is developed, its logical decidability and consistency are justified; the NVB features space model is developed, vector's dimension is reduced, parallelism is presented. Further research will be focused on improving the uniqueness of the NVB feature vector and on improving the classifier in order to enhance poor results.

Acknowledgement

This work was conducted as part of the government contract № 02-10/okr of April 9th, 2010.

References

1. Ilin E. P. Psychomotor human organization: Textbook. - Spb.: Piter, 1st edition, 2003. – 384p.
2. Nekhina A.A., Knyazev B.A., Kashapova L.H., Spiridonov I.N. Applying an ontology approach and Kinect SDK to human posture description// Biomedicine Radioengineering.- 2012.-№12.- P.54-60
3. Bashmakov A.I. Intelligent information technology: Textbook. - M.: MGTU, 2005. - 304p.
4. Calmet J., Daemi A. From entropy to ontology// In AT2AI-4 - Fourth Int. Symposium "From Agent Theory to Agent Implementation" at 17th European Meeting on Cybernetics and Systems Research, Vienna.-2004.
5. Papari G, Petkov N. Review article: Edge and line oriented contour detection: State of the art// Image and Vision Computing. - 2011. - Vol. 29, Iss. 2-3. - P.79-103
6. Petruk V.I., Samorodov A.V., Spiridonov I.N. Applying local binary patterns to face recognition problem// Vestnik MGTU. Seriya Priborostroenie. - 2011.- Spec. edition. Biometric technology. - P.58-63.