CBIR Approach to the Recognition of a Sign Language Alphabet (*)

Dimo Dimov, Alexander Marinov, and Nadezhda Zlateva

***Abstract**: The task of recognizing letters from the sign language alphabet, by means of which the hearing impaired people finger-spell words and proper nouns, is interpreted as a CBIR (Content Based Image Retrieval) problem. An arbitrary input image of given sign (palm gesture) is treated as a sample for search within a database (DB), which contains a large enough set of images (i.e. projections from a sufficient number of view points) for each letter of the sign language alphabet. We assume that the gestures for recognition are static images, which have been appropriately extracted from the input video sequence. In addition, we have at our disposal a CBIR method for image DB access that is simultaneously fast enough and noise-tolerant. The paper describes both the methodology used for building up the DB of image samples and the experimental study for the noise-tolerance of the available CBIR method. The latter is used to acknowledge the applicability of the proposed approach.*

***Key words**: CBIR, sign language recognition, finger-spelling, image databases, image/video analysis.*

1. INTRODUCTION

Sign language (SL) recognition and translation is an active area of research, its main goal being the provision of equal opportunities and integration of the hearing impaired people. The implementation of a computer system that will help the mutual communication between deaf and hearing people is a rather ambitious task, aiming at replacing the third human factor, i.e. the translator. We limit ourselves to the task of translating finger-spelled words to text, in whose basis lies the problem of recognizing the SL alphabet [3].

Defined by their own grammar and rules, sign languages comprise a dynamic set of hand and palm configurations, body movements and positions, and facial expressions [1, 13]. There are different sign languages for almost all known natural languages and/or dialects. The one we will be using and referring to herein is the Bulgarian SL (see Fig.1).

There are two main approaches towards SL recognition:

- *vision-based* approaches which use one or more cameras to capture the gestures and proceed with image analysis stage [9, 10, 11], and
- *glove-based* (instrumental sensor approaches) which, by using additional equipment (the "glove"), aim to bridge over difficulties typical for the vision-based approaches. Though often obtrusive and burdening the user/signer, these approaches provide a more precise tracking of the gesture dynamics [8, 11].

The proposed approach for SL recognition is vision-based, with the usage of (minimum) one video camera for capturing the gestures in dynamics. For the sake of simplicity, we will consider the recognition of a single moving object, i.e. assume that the gesticulating palm/hand is „the biggest spot" in the frames, or at least in the majority of frames incoming from the camera.

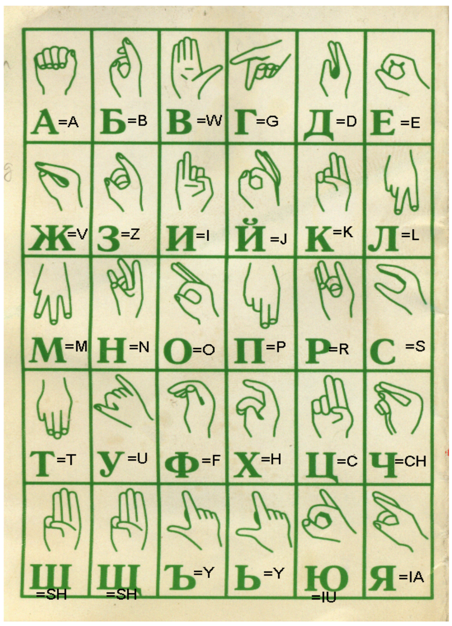


Fig.1. A schema of the Bulgarian SL alphabet

2. BACKGROUND

We will interpret the task of recognizing a gesticulating palm within a frame as a task for the direct comparison of an input example with samples from an image DB (IDB), i.e. as a *CBIR problem*.

(*) This work was supported by following grants of the Institute of Information Technologies (IIT) at Bulgarian Academy of Sciences (BAS): Grant # I-1306/2003 and Grant # BG-GR-17/2005 of the National Science Fund at Bulgarian Ministry of Education & Science t, and Grant #010088/2007 of BAS.

2.1. Essence of the proposed approach

Each gesture is considered as a dynamic 3D object, represented by a series of 2D projections, i.e. static images from the video camera. If an appropriate part of these images, or similar to them, is already stored in a IDB with gesture samples, then we can search into this IDB for the sample that is closest (most similar) to the source/input image. Moreover, we can search for a series of image samples, sorted (in descending order) of their similarity to the input image. Of course, a "dictionary" with a large amount of samples will be needed, whose size we will try to estimate experimentally here.

In the same time, the comparison time with all samples from the dictionary has to be quick enough to assure of operation in real-time. I.e. the realization of the idea is possible, if we have a noise-tolerant and simultaneously fast enough CBIR method for accessing a (large) DB with images. Such CBIR methods are available with the system EFIRS (Effective and Fast Image Retrieval System) developed by IIT-BAS, [5, 6]. Their noise-tolerance covers cases of eventual linear transformations in the input (translation, rotation, scaling), as well as regular noise, and more rough artifact-noise to a certain (preliminarily unknown) level.

In this sense, this research is also directed to the experimental estimation of the integrated noise-tolerance of the available CBIR methods.

2.2. Expected problems and allowed limitations

Two essential problems of the approach can be formulated:

(1) for isolating the object of interest (gesticulating human palm) from the input image (static 2D scene) and/or from the time sequence of similar scenes (video-clip);

(2) for accumulating a representative enough IDB, i.e. a dictionary with image samples of similar gestures.

The first problem is characteristic for the most of the approaches in the area of computer vision, image processing and recognition, therefore, at this stage, we will consider it for apriory resolved.

We will concentrate here on the 2nd problem – gathering of representative samples in the IDB and performing an experiment of evidence for the chosen concept.

Actually, the primary goal of this research was to clarify if the available (developed by us) CBIR approach is appropriate for the recognition of dynamic objects in a video-clip. This goal determined the use case formulation – recognition of the "static" sign language alphabet. Another use case, for example face recognition in a frame, would require much bigger data resource for the experiment, for example ~28 times bigger. That is, for the accumulation of an IDB with the same amount of samples, we would need the unique faces of 28 persons, versus the one signer demonstrating the 28 SL alphabet signs.

Except for being enough in number, the samples from the experimental IDB have to adhere to the general limitation of the available CBIR methods – the images need to be relatively "clean", i.e. to contain the whole object of interest (the palm), in color or gray scale, over an uniform (white) background and if possible be devoid of noise-artifacts from the natural surrounding.

All these limitations, related in fact to the well known problems of segmentation [2, 4, 9, 12, 15], are resolved here in their light form, according to the experiment's specifics. For that purpose, we are using a simplified scene – motionless hand in a gesture over a dark blue curtain, i.e. with opposite color to the signer's skin. Besides, we are hiding the elbow in the curtain folds, so that only the palm remains visible. In this way, we are avoiding the problem of segmenting the part from the elbow to the wrist, which is addressed in [10].

We also managed relatively easy the expected difficulties with the scene's surrounding illumination. Thus, we were thoroughly concentrated on the uniformity of the manual scanning with the camera, row by row, see § 3.1.

At this stage, we are ignoring several possible requirements for the completeness of the experiment:

• The signer's gestures are captured from one instead of many cameras.

• We use only one signer – one palm demonstrating the 28 gestures from the Bulgarian SL alphabet, see Fig. 1.

In addition to that, we have limited ourselves to the recognition of only static sign language symbols, since at this stage the proving experiment only needs a rich enough selection of possible projections. The matter for the retrieval of the appropriate input frames from the dynamics of the input video-clip will be addressed at a later stage.

3. APPROACH

As a possible alternative to the needed experimental IDB, we could use an IDB available on the Internet, [14]. It consists of 2060 images of a palm, gesticulating static signs from the international sign language alphabet. Unfortunately, the presented images are in the "gray-scale" and each of the sign shots is taken frontally. Of course, there are also characteristics that satisfy our requirements: a dark background has been used, which is with approximately the same intensity to the signer's clothing (sleeve) and, therefore, could facilitate the palm segmentation. Each gesture from the SL alphabet is represented by 40÷100 images – in various scales, translation, and rotation in the image plane. Yet, the research goals of this IDB are different: to achieve sign recognition through the use of principal component analysis. It is also lacking the requirements of exposition spatiality. Nevertheless, it was definitely useful in defining our experimental construction.

3.1. Three possible approaches to data gathering

The methodology of gathering data in the dictionary (IDB) is based on the recording of a short video-clip that traces out the object (the gesticulating palm/hand) through uniform scanning (by position and time) in a "wide" enough spatial sector in front of the object. Three possible approaches have been considered:

(1) Static object and moving camera that scans uniformly the needed spatial sector around/in front of the object.

(2) Static camera and moving object that makes uniform motion, in order to expose itself to the camera from all needed points of view.

(3) Static object and static camera.

The third approach, in comparison to the first two, is not directly oriented towards the creation of a video-clip. The latter needs to be computer generated from a few static photos, taken from different positions within the needed spatial sector around/in front of the object. This approach is comfortable from the signer's point of view, but it is unacceptable for our goals. The needed development efforts on generating the film or the final frame sequence (imitating close enough positions of exposure) are unduly expensive.

The second approach is fairly simple from a photographer's point of view. Yet, it requires high precision and a certain level of "acting" skills on the part of the "object-signer", when performing the uniform movements in front of the camera. If the signer is not trained enough for the role in question, there will be the need of additional efforts on "normalizing" the film. In other words, the approach is unacceptable for our goals, even if it is quite popular from the practice of gathering similar data, [7].

The first approach turns out to be the most acceptable one even though it requires an auxiliary construction for recording a video-clip with a conventional camera (Fig.2. illustrates an idea for this construction). Here, the responsibility on providing the needed uniformity of motion within the video-clip falls on the operator – researcher. This is acceptable and natural as we are speaking of a single (possibly a few multiple) session of unvaried scanning procedures.

Thus, in view of the auxiliary construction as of Fig.2, we propose the following method on providing the experimental data for our IDB:

(s1) Fix the "signer-object" in the next necessary pose.

(s2) Scan the needed spatial sector in front of the object: row by row, top-down, moving "zigzag", uniformly row along, and camera always on while scanning. In vertical transitions (from row to row) use an assistant to provide the "synchronous" release of the vertical restriction (a rope) through equally spaced distances *d, vertically on the arc, R=const.* It is desirable that the time for transition (from row to row) does not go beyond 1÷2 seconds.

(s3) Should there be further material (gesture) to capture, go to (s1).

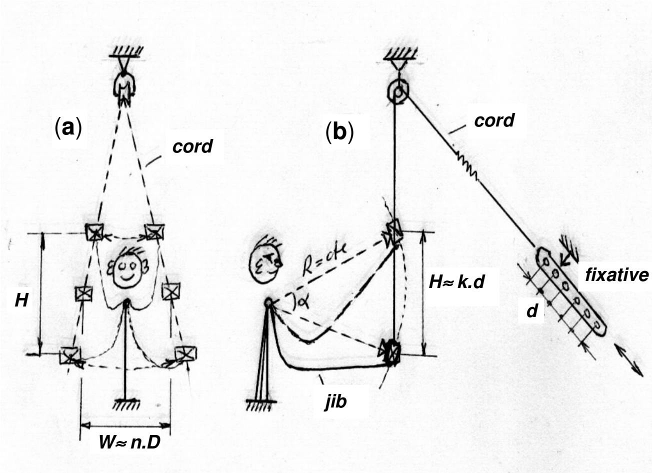


Fig.2. Kinematic schema of the construction for taking "primary" video-clips, using the "static object – moving camera" approach.

3.2. Video capture

Type of the experimental material – video-clips, whose unique frames could serve as samples of the object for recognition. We call these "*primary (video) clips*" and let them meet the following set of requirements:

• there is a separate video-clip for each letter from the SL alphabet (as of Fig.1);

• the video capture is carried out row by row, having each row separated from its neighboring ones by a sequence of empty (black) frames;

• it is advisable that the "manual" scanning with the camera along each row be carried out at a regular speed;

• the average scanning speed along the different rows is tolerated to vary within some narrow bounds.

Primary video-clips processing. Stages:

♦ Separate the motion frames from the entire video-clip.

♦ Derive a representative set of frames from the video sequence by, for example, obtaining those frames that fall within a uniform grid ("square" or "triangular" one) over the spherical sector of scanning. The linear parameter *D* of this grid (Fig.2) will represent the differences (as the number of frames) between the consecutive representative frames, which contain the object samples from the corresponding view points.

We assume that the noise-tolerance of the used CBIR method could also be measured by means of *D*. Thus, in a uniform grid with $D>D_0$, D_0 an admissible lower boundary, the CBIR method would start to err when recognizing by similarity to the samples of the IDB. This D_0 boundary can serve as a measure for the noise-tolerance of the CBIR method in use. It will also determine the number ($\approx D_0^2$) of the needed grid nodes, i.e. the stored representative samples within the IDB.

Fig.3. The construction in action.

We will skip the details on defining the geometrical model of the experiment. We will give some *more-significant* parameters:

♠ radius R of the scanned spatial sector, R=51 cm;

♠ average angle of the spatial sector of scanning – horizontally ≈ 80° and vertically ≈ 115° (in angle degrees);

♠ average speed of scanning, i.e. of the manually moved camera capturing along the scanning rows (horizontal arcs) ≈ 15÷20 cm/s, with camera capture speed of 15÷16 frames per second;

♠ distance d between neighboring rows over the spherical sector, d = 10 cm, Fig.2;

♠ number N of the scanning rows (arcs over the spherical sector), N = 8, Fig.4.

Thus, for an uniform square grid we have the following 2 variants of the possible values of D:

(□a) $D = kd$, (Fig.4a), and

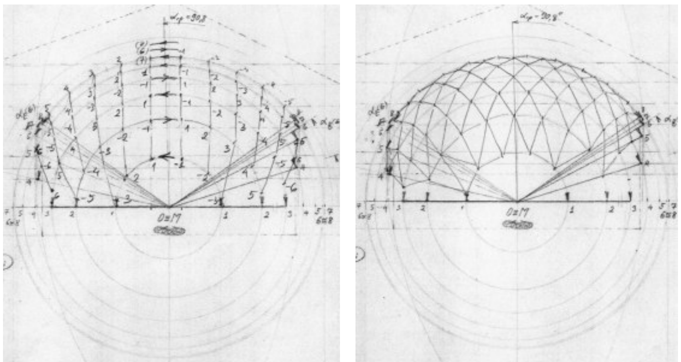
(□b) $D = kd\sqrt{2}$, $k = 1,2,\ldots 7$,

while for a uniform triangular grid, respectively:

(∇a) $D = kd2\sqrt{3}/3$, (Fig.4b), and

(∇b) $D = kd\sqrt{2}/2$, $k = 1,2,\ldots 7$.

In addition, on the basis of these 2 uniform figures – square and triangle – we can construct a large number of rotated versions for the 8 parallel rows (arcs) of scanning, but this is irrelevant for the exposition.



(a) (b)

Fig.4. Spherical sector in front of the camera, scanned row by row. Two main variants for the uniform grid of representative frames.

3.3. Significant frame extraction

The used methodology for estimating the identification numbers of the representative frames relies considerably on the following: during the scanning phase, while proceeding from row to row, the camera lens is manually hidden, i.e. the corresponding frames from the video-clip are almost black. This simplifies the separation of the significant frames (see Fig.5):

- Transform the frames from RGB to Gray scale;

- Create an image of the differences between each two consecutive (Gray) frames from the video sequence;

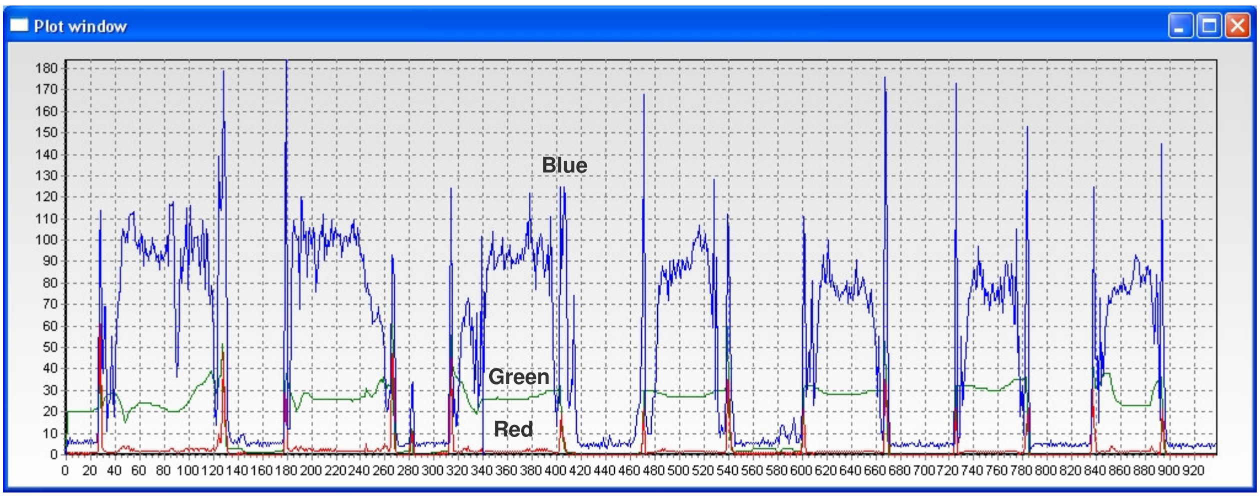


Fig.5. The graphic in Green shows the average intensity of each of the frames from the primary video-clip. Blue displays the maximal, while Red – the average intensity of the differences between each 2 consecutive frames. (R) specifies the zones with useful frames as well as the black zones in between. The useful frames' zones can be made more precise through (G) and (B).

- Calculate the values of: the average intensity of the original frames; the maximal and average intensity of the differences;
- Find a statistical estimation of the corresponding thresholds for the transition from the sequences of insignificant (black) to the sequences of significant (motion) frames.

3.4. Hand region isolation

Some of the more popular approaches in segmenting the hand are: usage of a controlled or known background [12] with a previously acquired background image (i.e. applying background subtraction); using segmentation by motion [9, 2]; or using color segmentation with predefined or generated skin color models [4, 15]. In addition, to alleviate the segmentation and recognition process, multiple camera configurations [9, 10] are used. These contribute with additional information such as image depth, 3D shape and motion and thus are helpful in detecting the hand/finger position and orientation.

The problem of segmenting the hand from each of the motion frames is simplified by the chosen experimental environment – dark blue background, in contrast to the human skin color. Thus, after a RGB to HSV (Hue-Saturation-Value) color scale transformation of the frames, the segmentation can be carried on mainly in the 2D HS-histogram schema (Fig.6), where 2 main color sectors are outlined – blue for the background and beige for the hand. Hence, after a transformation to a (cyclic) histogram by H only, the segmentation task can be reduced to finding the 2 optimally separating thresholds: blue-beige and beige-blue. That is, at this stage we don't need to use information about the motion. Of course, some other peculiarities are considered in parallel, such as the fact that in some frames the hand region contains gleams (too bright pixels), where H is undefined. Here, we also check the illumination (V) of the corresponding pixels, define a threshold, and apply binarization by V.

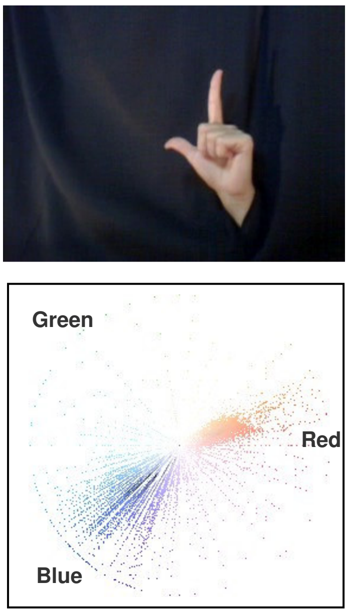


Fig. 6. HS-histogram of a given frame (letter «Ъ»).

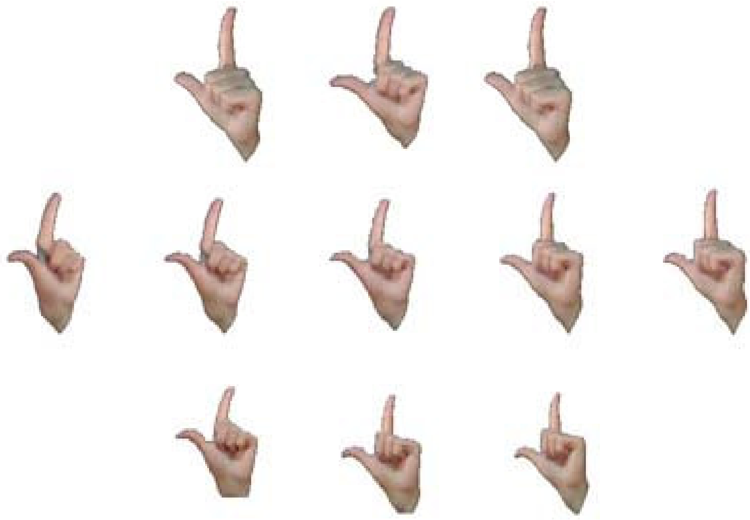


Fig. 7. Several representative frames extracted from the primary video-clip for the letter «Ъ»

In brief, the chosen segmentation algorithm is as follows, for given frame:

(s1) Determine the contour of the blue background, and color the area outside it in black. Thus we exclude those parts from the frame that might accidentally contain parts from the experimental environment (floor, ceiling, walls, etc).
(s2) Compute the HSV space. Apply Otsu binarization on the V histogram (by S < 0,2).
(s3) Test the Hue values of the original pixels corresponding to the black marks after binarization and mark the so determined pixels in white.

(s4) Search for the biggest contour (of maximal area) in the resulted white region that is expected containing the object of interest (palm/hand).
(s5) Mark everything outside the biggest contour (from s4) in black, and the internal pixels – in white.
(s6) Use the resulting image as a mask over the original image, interpreting the black marks as opaque and the white ones – as transparent. In this way, we achieve the palm (hand) segmentation as shown on Fig.7.

4. IDB EXPERIMENTS AND RESULTS

We carry an experimental analysis of the proposed approach through the use of EFIRS [5, 6]. EFIRS is a C/C++ written Windows-XP application operating on an IBM compatible PC. For the objectives of the experiment, we wrote an additional test, which is functionally similar to the SLT (Simple Locate Test) of EFIRS, [6]. The existing IDB structure of EFIRS is used for the generation of the needed IDB of samples for the test. The chosen CBIR access method is PFWT, as described in [6]. The primary video-clips have been acquired through a construction as in Fig. 2.

4.1. Essence of the experiment

For each possible value of the examined parameter D (the basic size of the grid), do:

♣ Generate a separate IDB for EFIRS by loading it with all representative frames for each letter/sign from the SL alphabet (see Fig.1). The representative frames for each letter are chosen according to their position over the square grid within the experimental sector of visibility.

♦ For each square on the grid associated with given primary video-clip, define the closest frame the center of this square. These central frames are obtained from the set of motion frames along each row of the clip, and are uniformly the most distant ones (at a distance of $\approx D/2$) from their corresponding 4 neighbors (the frames at the squares' corners), which have already been stored into the IDB. These central frames are used to provide "the heaviest case" of input precedents for the CBIR search within the IDB.

♥ Carry on a SLT for a CBIR search within an IDB over all square centers, i.e. over "the heaviest cases" of input precedents. Summarize the results for the successful and unsuccessful searches.

4.2. Experiment

At the current stage we have loaded the IDB with the representative frames of only 7 from the signs/letters (Fig. 1), namely: "А", "Б", "В", "Ж", "Ч", "Ц", and "Ъ".

The acquired images for these letters are qualitative enough, as well as enough in number – 344 images altogether, on average 49 per letter – so that we can carry a preliminary experiment with the EFIRS system. The representative frames for each letter have been arranged in a uniform square grid (Fig.4a).

The generated IDB applies to the case of $D=d$, where D is the basic size of the grid, and d – the distance between the scanning rows of the video-clip.

Table 1 contains the results for sign "A". When excluding the boundary row 1, the results can be rated as promising – error rate by letter recognition < 4%.

Thus, at this stage we define the boundary value of D_0, i.e. the noise-tolerance measure of the CBIR approach in the given application (SL recognition), as:

$D_0 = 1d$, *at letter recognition rate* > 96%.

And, the expectations for the future experiment improvement are optimistic:

$D_0 = 2d$ and even $D_0 = 4d$, *at letter recognition rate* > 99%.

Table 1. Letter "A" test results.

Row No.	Frames	Errors (1): Letter err.	Errors (2): row diff.s > ±1	Exact matches	Warns (1) row diff.s = ±1	Warns (2) row diff.s = 0	by w(1, 2) Avrg of position diff.s	by w(1, 2) Avrg of abs. position diff.s
1	73	15	14	3	12	29	6,0	6,8
2	70	5	0	5	16	44	0,7	4,8
3	71	2	1	6	25	37	3,8	5,3
4	45	2	0	8	8	27	0,0	2,4
5	52	0	1	7	13	31	0,7	4,9
6	39	1	0	7	6	25	-3,9	5,6
7	40	0	0	6	6	28	6,4	9,6
8	41	1	1	6	2	31	3,7	6,1
total (2÷8)	**358**	**3 %**	**1 %**	**13 %**	**21 %**	**62 %**	**1,7**	**5,5**

Notes to Table 1:

- "Exact matches" corresponds to the number of samples (for given letter and scan row) loaded into the IDB.
- "Errors (1)" count the errors of type 1 ("erroneously recognized letter"). These are expected to be highest along the boundary of the scanned spatial sector, and lowest – in its central area.
- "Errors (2)" count the errors of type 2 ("greater than ±1 deviation between the both rows, the input and the found"). Their expected behavior is similar to that described for Errors (1).
- "Warns (1)" and "Warns (2)" register the expected situations of "deviation of found rows ≤ ±1". The averaged values "Avrg_of_position_diff(erences)" and "Avrg_of_abs(olute)_position_diff" concern these both recognition situations. It is expected that "Avrg_of_position_diff" would be close to zero (if the model is void of geometric inaccuracies), while "Avrg_of_abs_position_diff" – close to the average distance (in number of frames) between the consecutive representative frames.

4.3. Comments over the results

We can consider as positive the result of $D_0 = d$. It signifies that the size of the needed IDB would be the largest, i.e. ~ 64=8*8 – the number of representative images for each sign letter. The latter is insignificant in terms of the needed resources of (external) memory, since for the proposed CBIR technology, which covers a size of hundreds of thousands of objects, it is enough to only store a key with size of ~0.5KB.

From the viewpoint of efficiency, a search within an IDB that contains 64 samples for each sign letter would require ~ $6=\log_2(64)$ number of accesses to the IDB, while each access would take ~x10ms (according to the efficiency of the memory device (disk)).

Even a result such as $D_0 = 0$ should not be considered hopeless. It signifies that D_0 cannot be found at the chosen accuracy d (i.e. the distance between the rows) of the IDB experiment. The value of D_0 would be a fractional number $D_0=\alpha d$, $\alpha\in(0,1)$.

Actually, by definition the result is $D_0 > 0.0$, i.e. the chosen CBIR technology does not err when searching of an exact match, while the percentage of errors from occasional linear transformations on the input (translation, rotation, and scaling) is < 1.5%. The former is fundamental for the CBIR technology, while the latter has been experimentally proven over a database of a large size ≈58000 of trademark/hallmark images.

Thus, even at an IDB with $D=d$, we can determine more accurately the value of D_0 – at least horizontally, on the basis of the average distance between two consecutive motion frames from a primary video-clip. This distance is defined by the camera capture speed and by the speed of the (manual) scanning by rows.

5. CONCLUSION

The experimental results can be used to measure the applicability of the approach in other similar tasks, for example for the computer recognition of people on the basis of their physiognomy, as a promising alternative to the well-known methods in that area (eigenfaces, neural networks, etc.).

Besides, the proposed approach can be also considered as one more way to CBIR techniques' invariance against accidental projective/perspective transforms of the input images of query.

REFERENCES

[1] American Sign Language. National Institute on Deafness and Other Communication Disorders, http://www.nidcd.nih.gov/health/hearing/asl.asp

[2] Binh N.D., E. Shuichi, and T. Ejima: Real-Time Hand Tracking and Gesture Recognition System, In: Proceedings of GVIP 05 Conference, Dec. 2005, CICC, Cairo, Egypt.

[3] Bowden R., A. Zisserman, T. Kadir, and M. Brady: Vision based interpretation of natural sign languages, In: Exhibition at ICVS03: The 3rd International Conference on Computer Vision Systems, Padova, Graz, Austria, April 2003.

[4] Bradski G.R., and S. Clara: Computer Vision Face Tracking For Use in a Perceptual User Interface, *Intel Technology Journal*, Q2'98, (1998).

[5] Dimov D.T.: Fast Image Retrieval by the Tree of Contours' Content. *Cybernetics and Information Technologies J. of BAS*, 4, 2, (2004), pp.15-29.

[6] Dimov D.: Polar-Fourier-Wavelet Transform for an Effective CBIR, In: Proceedings of the Int. Conf. on Computer Science, 12-15 Oct., 2006, Istanbul, Turkey, Vol.2, (2007), pp.369-374.

[7] Extended Multi-Modal Face Database, http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb

[8] Fang G., W. Gao, X. Chen, C. Wang, and J. Ma: Signer independent continuous sign language recognition based on SRN/HMM, In Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time, Vancouver, Canada, July 2001, pp. 90–95.

[9] Jennings C.: Robust finger tracking with multiple cameras, In: Proceedings Int. Workshop on Recognition, Analysis, and tracking of Faces and Gestures in Real-Time Systems, 1999.

[10] Malassiotis S., N. Aifanti, and M.G. Strintzis: A Gesture Recognition System Using 3D Data, In: Proceedings of IEEE 1st International Symposium on 3D Data Processing Visualization and Transmission, Padova, Italy, June 2002.

[11] Parton B. S.: Sign Language Recognition and Translation: A Multidisciplined Approach From the Field of Artificial Intelligence, *Journal of Deaf Studies and Deaf Education*, 11(1):94-101, 2005, http://jdsde.oxfordjournals.org/cgi/content/full/11/1/94

[12] Segen J., S. Kumar, Shadow gestures: 3D hand pose estimation using a single camera, In: Proceedings of the Computer Vision and Pattern Recognition Conference, CVPR99, v.1, 1999, pp.485-.

[13] Sign Language. Wikipedia the free encyclopedia, http://en.wikipedia.org/wiki/Sign_language

[14] Thomas Moeslund's Gesture Recognition Database, 1996, http://www-prima.inrialpes.fr/FGnet/data/12-MoeslundGesture/database.html

[15] Zhu X., J. Yang, and A. Waibel: Segmenting hands of arbitrary color, In: Proceeding of the 4th International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000, pp.446-453.

ABOUT THE AUTHORS

Dimo Dimov, Assoc. Prof., Ph.D., Institute of Information Technologies, Bulgarian Academy of Sciences, Tel: (+359 2) 8706493, E-mail: dtdim@iinf.bas.bg

Alexander Marinov, Ph.D. Student, Tel: (+359 2) 8771830, E-mail: amarinov@iinf.bas.bg

Nadezhda Zlateva, Ph.D. Student, Tel: (+359 2) 8771830, E-mail: nzlateva@iinf.bas.bg