INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES
BULGARIAN ACADEMY OF SCIENCES

1869

# From 2D CBIR to Multiple View Based 3D Recognition

a presentation by D. Dimov
(Assoc. Prof., PhD, IICT, BAS)

at Computer Vision and Multimedia Lab
University of Pavia
June 04, 2013



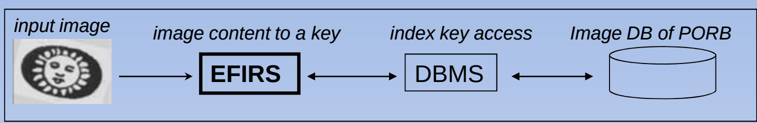
## Contents

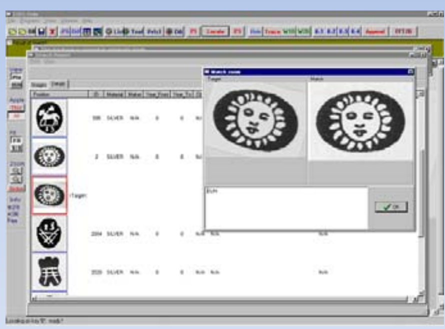
# Introduction

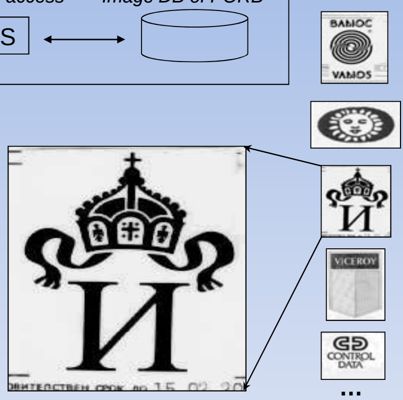
- If you have a CBIR method to access an image DB (IDB), and the method is both fast enough and sufficiently noise-tolerant, then you can apply the simple “Multiple View 3-dimensional Recognition” (MV3R) approach presented hereinafter.
- MV3R treats an arbitrary input image of the object of interest as a search sample within an IDB that is supposed to contain a large enough sub-set of views of the same object, i.e. appearances from a sufficient number of viewpoints. Each such sub-set is assigned to the appropriate person’s data.
- The PFWT access method of EFIRS can play the role of basic CBIR, see also FANTIR from the previous presentation (29.05.2013).
- MV3R has been successfully experimented for two cases of recognition, namely for human faces, and for hand signs (of a given sign-language alphabet).
- Can we extend the application area of this type of 3D recognition technique ?
- At least 3 surveys [1, 2, 3] can be referenced for the CBIR early years [3], and CBIR current state-of-the art as well [2,3]. View based approaches like MV3R are considered very promising ,[1].



# … (only) to remember how EFIRS operates



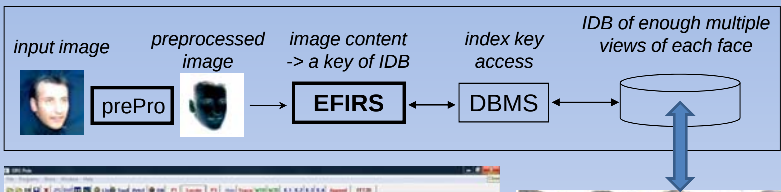


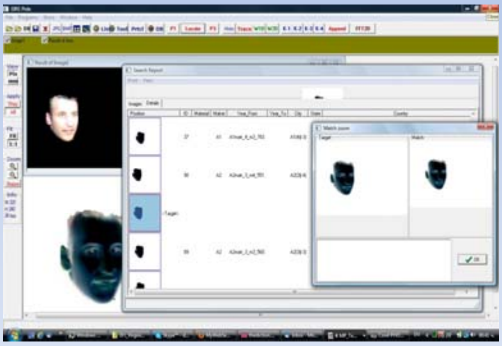

*A result of EFIRS in operation*

*A few images from the DB of PORB (the Patent Office of Republic of Bulgaria)*

FANTIR & EFIRS should be considered in common: FANTIR – as theoretic foundation of our CBIR technology experimented through the EFIRS software.

## Let us see how FANTIR/EFIRS perform face recognition



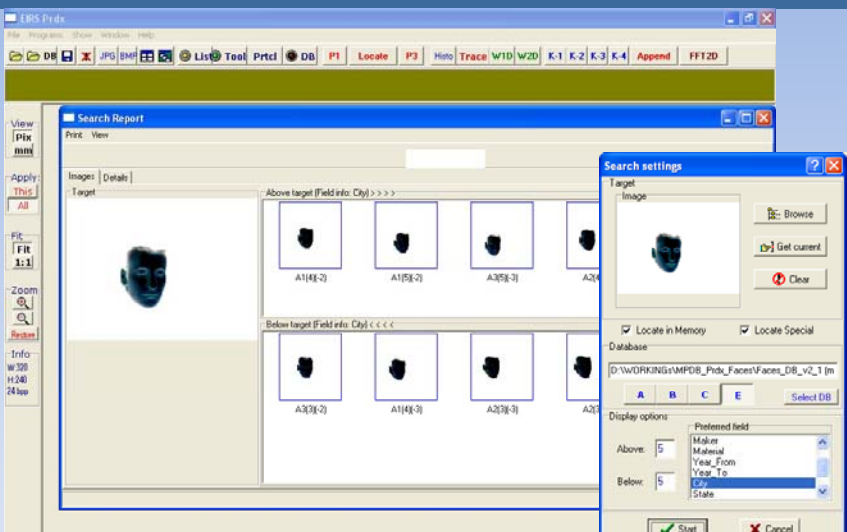



The EFIRS result: a series of view-candidates found, most of them associated with the right person.

A few images from the IDB of face views (about 8*8 = 64 views per face) and 3*5 are shown here for visibility.



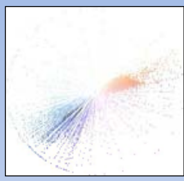
## Each face view is searched like a hallmark image…




One more screenshot of EFIRS in operation – the brief version of the result (in the middle), and the search engine ready to start (on the right).

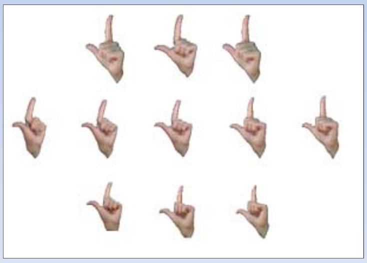
# EFIRS: two applications experimented

For recognition of *hand signs* of a given sign-language alphabet, [d4]



For *face* recognition [d1, d2]

The object segmentation is considered an outside problem. Here, a multi-thresholding of the Hue-histogram of each frame was used [d5].



The principle of using FANTIR/EFIRS in both cases is one and the same.



# One and the same construction is used for 3D data gathering

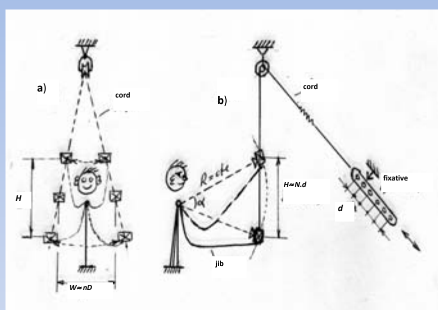
The view samples of given 3D object that will represent it in the IDB should be regularly positioned in the stereo segment of possible appearance of the object towards the user camera.

Taking into account the available mechanical construction (see the schema on the right), we chose among 3 approaches for the necessary data gathering:

1) ***Static object and moving camera***: sufficiently comfortable for the actors participating in the data gathering stage.

(2) *Static camera and moving object*: difficulties for the actors, their movements is hardly to be evaluated precisely. This scenario match the real cases of recognition, but not the data gathering.

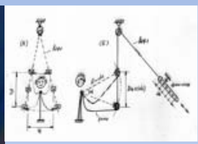
(3) *Static object and static camera* (very limited approach).




A kinematic schema of the construction for taking the “primary” video-clips, using the “static object & moving camera” approach.

Show a video-clip, e.g.: DSC00069_Nadya(1).3GP, and results of its processing as well.

## 3D data gathering: “static object & moving camera”



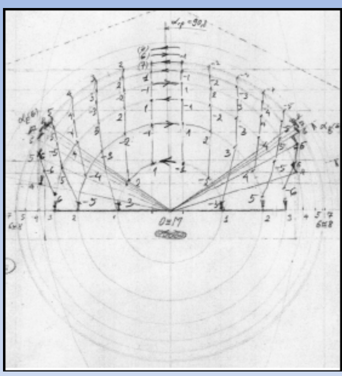
The construction in action:
- Nadia plays an actor, her face will be scanned;
- Alek, as an operator, will scan row by row, hiding the camera lens at each row start/end;
- while, one more people (me) will assist dropping the cord to next down row. Thus, the camera moves in a “zig-zag” top-down making a primary video clip.

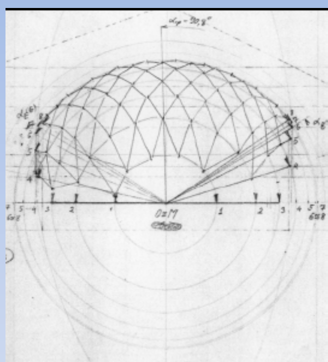

A table of a Bulgarian Sign Language alphabet

This “exotic” approach (and construction) can be easily substituted by a 3D scanner, Time-of-Flight camera, Kinect camera, etc.

6/4/2013 From 2D CBIR to Multiple View Based 3D Recognition D. Dimov 9 http://www.iict.bas.bg

## The spherical sector of viewpoint directions





- Spherical sector in front of the camera that is scanned row by row. Two main variants for the uniform grid of representative frames: (a) a square grid, and (b) a triangular one.
- The sector should be regularly covered by viewpoint directions of possible appearance of the human face towards the camera position at learning time (i.e. 3D data gathering). ( And segmented faces are represented internally by their color inversion ☺).
- The angular distance between each 2 neighboring views that we have to choose for the IDB of current application, depends on the noise tolerance that EFIRS assures.

6/4/2013 From 2D CBIR to Multiple View Based 3D Recognition D. Dimov 10 http://www.iict.bas.bg

## Two essential problems have been solved

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

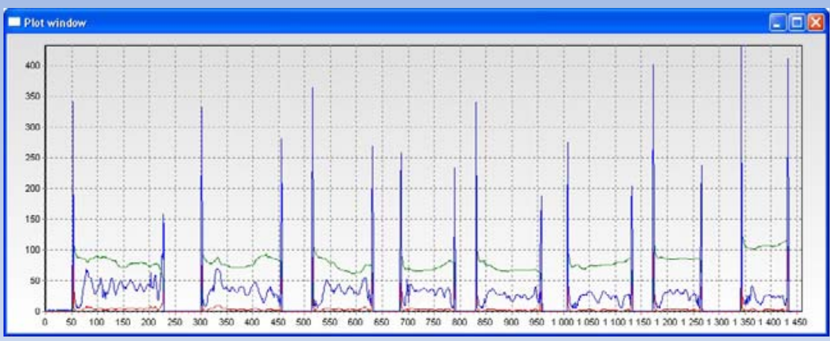
(2) For accumulating a sufficiently representative IDB, i.e. a dictionary of object views:
- image samples of similar gestures (for SL application);
- enough face views for each person of interest (for face recognition).

So far, in both cases we concentrated on the second problem – the IDB gathering as well as experiments of evidence for the chosen concept.

For the present, to lighten the 1st problem (the object segmentation) we use a dark blue curtain that dominates as a background, which color is opposite to the human skin.



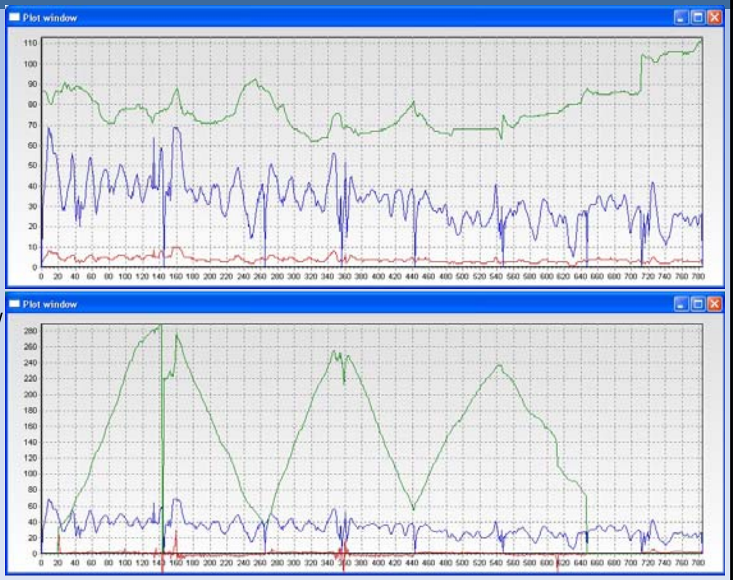
## Some details of video-clips processing




The primary frames for a given face (numbered on the horizontal axis). The Green graphic shows the average image intensity per frame, the Blue one shows the maximal intensity of the differences between two consecutive frames, while the Red one – the average intensity. (R) specifies the zones of significant frames as well as the black zones in between. The significant frame zones can be evaluated more precisely through (G) and (B). (The camera speed is 15 fps).

## … some details of video-clips processing

The significant frame zones only, the black frames zones (indicated herein only by the max-top-down row changes) are already removed.




The significant frames with evaluated the row scan position progress (in Green) in dependence on the evaluated manual scan speed (in Red) along the row scan direction. The Blue is (again) the maximal intensity per couple of subtracted frames.



## Experimental results

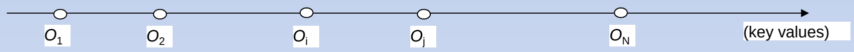
| Faces IDs | Rows scanned | Significant frames | Represent. frames | % Errors (1) ‖ (2) | % Errors (1): Letter errs. | % Errors (2): row diff.s > ±1 | % Exact matches | % Warns (1) row diff.s = ±1 | % Warns (2) row diff.s = 0 |
|---|---|---|---|---|---|---|---|---|---|
| A1 | 6 | 538 | 63 | 3.2 | 1.7 | 1.5 | 11.3 | 21.0 | 64.5 |
| A2 | 6 | 572 | 60 | 2.6 | 1.2 | 1.4 | 10.5 | 20.3 | 66.6 |
| A3 | 5 | 488 | 53 | 11.1 | 9.0 | 2.0 | 10.7 | 15.2 | 63.1 |
| Ao | 5 | 312 | 53 | 9.9 | 9.9 | 0.0 | 16.3 | 11.5 | 62.2 |
| Au | 5 | 328 | 54 | 8.8 | 8.8 | 0.0 | 15.9 | 10.1 | 65.2 |
| DD | 5 | 260 | 54 | 16.2 | 14.2 | 1.9 | 19.6 | 0.4 | 63.8 |
| GA | 4 | 220 | 41 | 0.0 | 0.0 | 0.0 | 18.6 | 5.9 | 75.5 |
| GG | 3 | 214 | 33 | 4.2 | 4.2 | 0.0 | 15.0 | 10.3 | 70.6 |
| HT | 6 | 454 | 61 | 14.3 | 13.2 | 1.1 | 12.8 | 14.5 | 58.4 |
| IH | 5 | 282 | 48 | 0.7 | 0.7 | 0.0 | 16.7 | 7.1 | 75.5 |
| LB | 6 | 392 | 60 | 0.8 | 0.0 | 0.8 | 14.8 | 18.9 | 65.6 |
| LI | 6 | 388 | 62 | 3.1 | 1.5 | 1.5 | 16.0 | 11.6 | 69.3 |
| LK | 3 | 215 | 42 | 17.7 | 16.7 | 0.9 | 16.7 | 11.2 | 54.4 |
| MP | 6 | 391 | 61 | 3.6 | 3.6 | 0.0 | 15.3 | 6.4 | 74.7 |
| MV | 7 | 463 | 71 | 4.8 | 4.8 | 0.0 | 14.9 | 13.2 | 67.2 |
| N1 | 6 | 342 | 62 | 4.4 | 2.3 | 2.0 | 16.4 | 10.5 | 68.7 |
| N2 | 8 | 481 | 76 | 3.5 | 3.1 | 0.4 | 14.8 | 9.6 | 72.1 |
| N3 | 7 | 478 | 74 | 6.1 | 3.1 | 2.9 | 13.6 | 22.6 | 57.7 |
| PK | 6 | 399 | (63) | 9.0 | 5.5 | 3.5 | 14.5 | 14.8 | 61.7 |
| SK | 5 | 312 | 53 | 12.5 | 11.9 | 0.6 | 16.0 | 5.8 | 65.7 |
| V1 | 5 | 304 | 53 | 6.6 | 4.3 | 2.3 | 17.1 | 15.8 | 60.5 |
| V2 | 5 | 344 | 54 | 7.8 | 5.8 | 2.0 | 14.5 | 16.0 | 61.6 |
| | | | | | | | | | |
| avrg % | - | - | - | **6.6** | 5.3 | 1.2 | 14.6 | 13.4 | 65.5 |

Recognition rate achieved ~ 93.4% over an IDB for face views of 18 persons. As future improvement, an extra preprocessing is envisaged (for face normalization, including removal of noisy attributes like hair, moustaches, glasses, etc.).

## EFIRS: linear order of IDB keys (image signatures)

For each image of interest, EFIRS generates an appropriate key, i.e. a visual signature, see [2], but 1-dimentional (1D) one. Just this key is applied for fast search into an IDB, via the available index access method of the used DBMS (Data Base Management System).
The access time is ~ $\log_2(N)$, $N$ the number of images written into the IDB.

Thus, the all Objects (images) in IDB can be considered ordered linearly by their keys, each one derived via the chosen CBIR technology. We recommend the PFWT (Polar-Fourier-Wavelet Transform) of EFIRS [d3]:




PFWT assures very high recognition ability among the IDB images:

-Thus, very close images can be interpreted as almost equal ones, differing only because of accidental noise added on them.

- I.e. each image that is enough far from its both neighbors (left & right ones) can be interpreted as a (narrow) class of recognition.

6/4/2013 From 2D CBIR to Multiple View Based 3D Recognition D. Dimov http://www.iict.bas.bg 15

## ... linear order of IDB keys (image signatures)

By analogy, a larger group of almost isolated images, i.e. lying on a line-cut of the keys line, can be interpreted as a class of similar images. Unfortunately, this trivial 1D-distance, i.e. the absolute difference between two keys, does not cover the majority of cases for recognition.
Really, the *noise tolerance of ERIFS* is performed by a combinatorial way over the similar order of errors like the chosen order for the coordinates of the feature space.

And, the more important here is that…

6/4/2013 From 2D CBIR to Multiple View Based 3D Recognition D. Dimov http://www.iict.bas.bg 16

## More important characteristics of the proposed MV3R method

Vision-based approach for Face or Hand Posture Recognition, which uses one video camera for capturing the object.

For the sake of simplicity, we consider the recognition of a *single* object in the image scene, i.e. we assume that the face or the hand sign to be recognized is 'the biggest spot' in the video frames, or at least in the majority of frames incoming from the camera.

A given object in front of the video camera is considered as a dynamic 3D object, represented by a series of 2D appearance projections, i.e. video frames (or static images) from the camera.

If an appropriate part of these images, or similar to them, is already stored in an IDB, then we can search (very quickly) into this IDB for the image that is the nearest (most similar) to the observed input frame image.

Moreover, we can also locate a series of images, sorted in descending order of their similarity to the input image.

The IDB is built as an *object view dictionary* that uses the CBIR access methods of FANTIR/EFIRS.



## More important characteristics…

The necessary multiple views of a given object are arranged in a 2D scheme only for aims of explanation. Really, they are kept in the IDB in a way that the used DBMS defines, and are recognized by their keys generated by EFIRS.

The image views representing the 3D objects of interest are necessary to be written in the IDB only for visualization (or retrieval) aims. For the seek of recognition it is sufficiently to have in the IDB only the object-view-keys generated by EFFIRS.

Of course, in view of possible cooperation with the RSVP (Rapid Serial Visual Presentation) of CVML, the image views should be available. In this case they (and/or respective FANTIR/EFIRS keys) can be organized into an appropriate visual structure for each object (or class of objects) of interest.

## Conclusion: Open problems to be solved

Weight/Probability approaches for combination of the result – the increase of the total recognition rate over the series of similar images retrieved for a given query.

More effective approaches for 3D data gathering of the aimed IDB:

- 3D scanner (it will be available soon in the frames of AComIn project)
- Time-of-flight camera (CVML has good experience in the topic)
- Kinect camera (it is not very expensive and an appropriate device)

… while the simple video-camera can (or is profitably to) remain only as basic input device for real-time recognition.

Other…



## Conclusion: Other possible applications of MV3R

- Pass control systems: for the personnel of a small or larger office;
- Pass control systems: for wanted people that are prohibited to enter given large meeting (e.g. some soccer/football fens, etc.);
- Parking security systems: at the parking entrance, where cars appear generally by their “face” (or their “back”);
- For car type/model recognition in real time on the streets, highways, etc.
- In cultural heritage preservation of archeology – for MV3R in large DB of 3D objects, for conservation and/or reconstruction of broken archeological object by available debris, etc.;
- In biometrics – to understand the way of how human vision helps brain to learn new objects and/or to recognize already known ones, using *Eye tracking* approaches to structure the respective elements (characteristic spots) of given 3D object in an IDB;
- To combine with the RSVP (Rapid Serial Visual Presentation) method of CVML.

What is the most promising application of a common interest (of CVML and IICT), e.g., considering [e1, e2, e3, e4, e5], …

## Some References

[1] Datta, R., D. Joshi, j. Li, and J.Z. Wang, Image Retrieval: Ideas Influences, and Trends of the New Age, ACM Computing Surveys, Vol. 40, No. 2 Article 5, April , 2008. pp. 5:1-5:60.
[2] Tangelder, J.W.H., and R. C. Veltkamp, A Survey of Content Based 3D Shape Retrieval Methods, Multimedia Tools Appl. (2008) 39: 441-471, DOI 10.1007/s11042-007-0181-0.
[3] Smeulders, A. W. M., M. Worring, S. Santini, A. Gupta, and R. Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Trans. on PAMI*, **22**, 12, (2000),1349-1380.
…
[e1] Marco Porta, Stefania Ricotti, Calet Jimenez Perez, Emotional E-Learning through Eye Tracking,…

[e2] Virginio Cantoni, Calet Jimenez Perez, Marco Porta, Stefania Ricotti, Exploiting Eye Tracking in Advanced E-Learning Systems, …

[e3] Piercarlo Dondi, Luca Lombardi, and Marco Porta, Human-Computer Interaction through Time-of-Flight and RGB Cameras, …

[e4] Simone Corsato, Mauro Mosconi, Marco Porta, An Eye Tracking Approach to Image Search Activities Using RSVP Display Techniques,…

[e5] P. Dondi, L. Lombardi, L.Cinque, RDVideo: A New Lossless Video Codec on GPU, …
…
[d1] Dimov D., N. Zlateva , and A. Marinov: CBIR over Multiple Projections of 3D Objects, In: Fierrez, J., Ortega-Garcia, J.; et al (Eds.), Proceedings of the Joint COST 2101 & 2102 BioID_MultiComm'09, Sep. 16-18, 2009 Madrid, Spain, Springer Verlag, LNCS 5707, pp.146-153.



## References

[d2] Dimov D.T., N. Zlateva, and A. Marinov: CBIR Approach to Face Recognition, In Proc. of A&I'08 Conf., 01-04.10.2008, Sofia, Workshop on Multisensor signal, image and data processing, 02.10.2008, pp. IV.21-26.
[d3] Dimov D., Rapid and Reliable Content Based Image Retrieval. Proceedings of NATO ASI, Multisensor Data and Information Processing for Rapid and Robust Situation and Threat Assessment, 16-27 May 2005, Albena-Bulgaria, IOS Press, pp. 384-395 (2007)
[d4] Dimov D., A. Marinov and N. Zlateva: CBIR Approach to the Recognition of a Sign Language Alphabet, In: Proceedings of CompSysTech'2007, June 14-15, Ruse, 2007, pp.V.2.1-9, http://www.compsystech.org/

[d5] Dimov D., and L. Laskov: Cyclic Histogram Thresholding and Multithresholding, In: Proceedings of CompSysTech'09, June 18-19, 2009, Ruse, Bulgaria, 2009 - ACM International Conference Proceeding Series; Vol. 433, pp.II.5.1-8, http://portal.acm.org

IICT

# Acknowledgements:

This research is expected to be further developed in the frames of the AComIn project of IICT-BAS (http://iict.bas.bg/acomin/), and in a collaboration with the CVML of University of Pavia.

**THANK YOU !**

6/4/2013 From 2D CBIR to Multiple View Based 3D Recognition D. Dimov http://www.iict.bas.bg 23