Sill image object categorization using 3D models

Raluca Diana Petre, Titus Zaharia

Sill Image Object Categorization Using 3D Models

Subtitle as needed *(paper subtitle)*

Raluca-Diana Petre
ARTEMIS Department
Institut TELECOM; TELECOM SudParis
Evry, France
Raluca-Diana.Petre@it-sudparis.eu

Titus Zaharia
ARTEMIS Department
Institut TELECOM; TELECOM SudParis
Evry, France
Titus.Zaharia@it-sudparis.eu

***Abstract*— This paper proposes a novel recognition scheme algorithm for semantic labeling of 2D object present in still images. The principle consists of matching unknown 2D objects with categorized 3D models in order to associate the semantics of the 3D object to the image. We tested our new recognition framework by using the MPEG-7 and Princeton 3D model databases in order to label unknown images randomly selected from the web. Experiments show that such a system can achieve recognition rate up to 70.4%.**

Keywords-indexing and retrieval; object classification; 2D and 3D shape descriptors; 2D/3D indexing.

I. INTRODUCTION

Nowadays, the amount of multimedia content available for the general public is permanently increasing. Disposing of powerful search and retrieval methods becomes a key issue for efficient indexing and intelligent access to AV material. This paper addresses the problem of automatic 2D object recognition, which is of crucial importance, since identifying automatically the semantics of the elements present in an image allows a machine to easily retrieve the required content.

The great majority of the existing approaches make use on machine learning (ML) techniques. However, the methods based on ML require large and already label training databases.

As today numerous 3D graphical model repositories are available, we have developed a method that exploits the information included in categorized 3D model databases and thus avoiding the ML. Therefore, we use 2D/3D indexing and matching algorithms in order to find the 3D model which is the most similar to a 2D object (Figure 1.). Thus, the semantics of the 3D model can be transferred to the unknown 2D object allowing its identification.

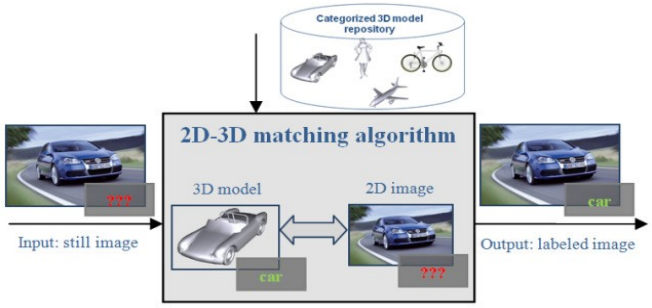

Figure 1. 2D object categorisation using 3D labeld models.

This paper is structured as follows. In section II we are going to briefly present the related work. The 2D/3D indexing principle and the proposed approaches are described in section III. The experimental protocol as well as the results are presented and analyzed in section IV. Finally, section V concludes the paper and presents our future work.

II. RELATED WORK

In the field of object recognition, a large part of the approaches is based on machine learning [1], [2]. Such algorithms allow to automatically learn to recognize complex structures. They include two main steps: the learning and the recognition. In the learning stage the system needs a large variety of example data from which it can extracts information about classes. Further, in the recognition stage, the system exploits the information driven in the learning step in order to identify new data. As two similar objects may have very different appearances, due to their color, texture, 3D pose etc., the learning database has to present objects covering all of these possibilities.

However, a more stable feature is the shape of the objects. Thus, using synthetic 3D models and exploiting only the shape information, the issues related to the appearance can be avoided.

The idea of using classified 3D models for real object recognition purposes was recently exploited by Toshev *et al.* in [3] and by Liebelt *et al.* in [4]. However, in [3] the algorithm aim at recognizing 2D objects automatically segmented not from images but from videos. Thus, the unknown object to be identified is represented by a set containing several images. Each 3D model used in the recognition stage is described by a set of $N=20$ views with the help of shape context descriptor [5]. These projections are selected by k-mean clustering of 500 evenly distributed views around the model.

The algorithm proposed by Liebelt *et al.* makes use of textured 3D models (in contrast to the approach presented in our present work and the one proposed in [3] which exploits only the shape information of the models). The *a priori* information is extracted from views of the 3D models and organized for each class as a visual codebook of $K=2000$ clusters of appearance features.

In our present work we propose a new image recognition framework that is evaluated on larger databases (up to 23 query

categories including the 2 respectively 3 classes tested in [3] and [4]) as described in section IV.

III. THE 2D/3D INGEXING

The principle of 2D/3D indexing consists in presenting the 3D model as a set of 2D views. These views are binary images and correspond to 3D-to-2D projections from several viewing angles (Figure 2.). As the projections obtained by using opposite directions represent one the mirror reflection of the other, all the viewing angles lie in half of the virtual space ($z \geq 0$).

Further, each view is characterized with the help of a set of 2D shape descriptors. In order to allow matching between 3D models and 2D objects, the same shape descriptor is used for the query objects, which are first manually segmented from still image.

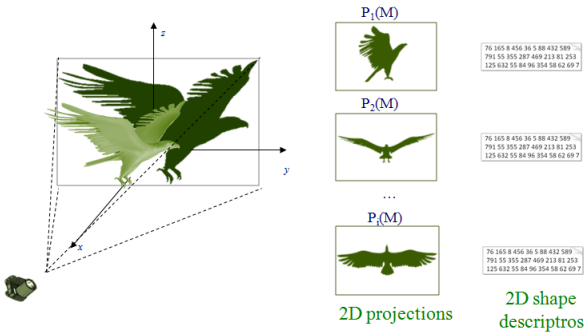

Figure 2. The principle of 2D/3D indexing

A. The 3D-to-2D projection

Before generating the set of projections, each 3D object is normalized in size and 3D pose. Firstly, the 3D model is resized to fit the unit sphere. Then, the model is turned in order to align the three axes of inertia (computed using the Principal Component Analysis – PCA [6]) with the coordinate system.

Further, the 3D model is rendered using N viewing angles by positioning the camera in N different places around the object and orienting it toward the coordinate system origin (which coincides with the center of the object) (Figure 3.). For each viewing direction n_i (i=1…N) results a 2D binary projection P_i(M) of the model M.

There are several strategies that can be used in order to acquire the set of projections $\{P_i(M)\}$. Each strategy is characterized by a number N of projections and by a set $\{n_i\}$ of viewing directions (meaning that for a given number of projection N there is an infinity of sets $\{n_i\}$ of viewing directions).

A first approach, suggested by the MPEG-7 Multiview descriptor, is based on the idea that the most representative views are those corresponding to the projections on the three principal planes. Optionally, if we consider the eight octants described by the principal planes, then we can use the bisectors of these octants as viewing directions (Figure 4.). From now on, we will refer to these techniques as PCA3, respectively PCA7.

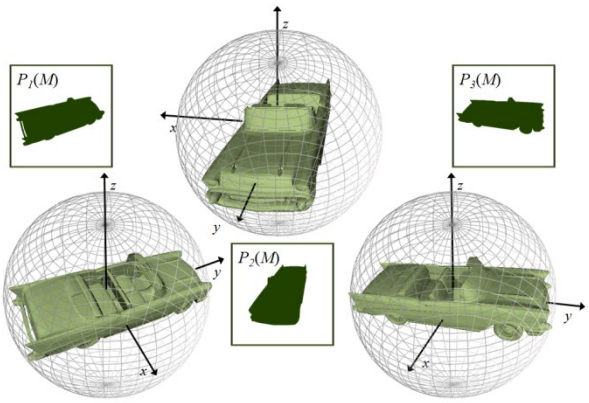

Figure 3. A 3D model M rendered from three diferent viewing angles and the corresponding binary projections $P_i(M)$.

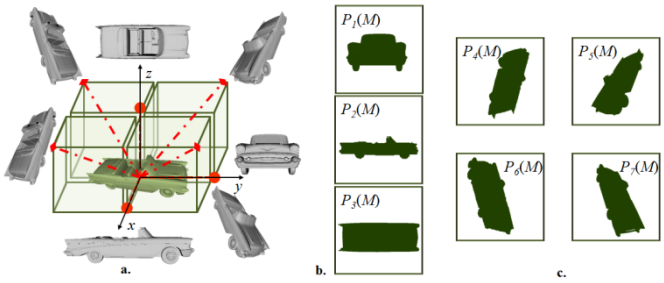

Figure 4. PCA-bsed strategy of projection; a. the seven viewing directions and the corresponding views of the 3D model. b; the views on the principal planes (PCA3); c. the four secondary views.

A second approach places the camera on the vertexes of a dodecahedron surrounding the 3D model [7]. In order to obtain additional views, the edges of the dodecahedron are successively divided, resulting into 3, 9 and 33 vertexes (and implicitly the same number of views) (Figure 5.). These strategies are called *OCTA3*, *OCTA9* and *OCTA33*.

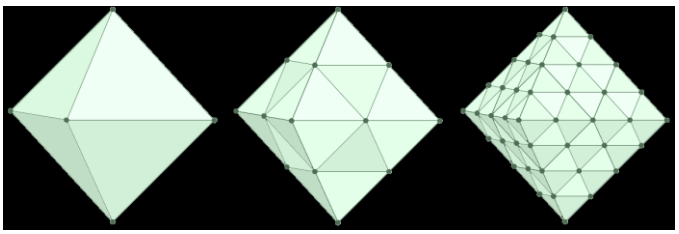

Figure 5. Successively subdivision of the octahedron faces

Finally, a uniform repartition of the viewing angles around the model can be obtained by using the vertexes of a regular dodecahedron (Figure 6. , as suggested for the Light Field Descriptor (LFD) [8]. Two sub-cases are possible. For the first one we have placed the cameras uniformly around the canonical representation of the object. This strategy will be referred as LFDPCA. Finally, we have used the same repartition of the camera given by the dodecahedron, but we have applied a random rotation of the 3D model (strategy referred by *LFD*). This choice is justified by the fact that the objects in real images are represented in a quasi-random pose.

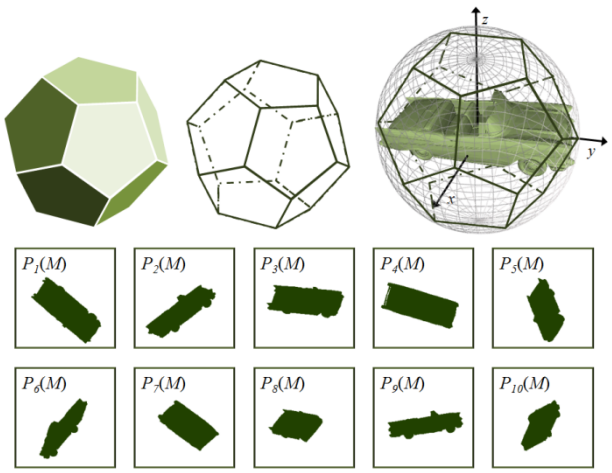

Figure 6. The evenly distribution of the viewing angles obtained by placing the camera on the verticies of a dodecahedron.

B. 2D Shape Descriptors

Once we have obtained a set of projections $\{P_i(M)\}$, we have to assign a 2D shape descriptor for each of these views. In the case of binary images, there are two features that can be exploited: the exterior contour and the corresponding region of support. In our work, we have considered two contour-based and two region based descriptors, briefly recalled here below.

Let us start with the Contour Shape (CS) descriptor proposed by the MPEG-7 standard [9], [10]. In order to obtain the CS descriptor, the first stage is to extract the curve representing the exterior contour of the 2D object. Further, the contour is successively convolved with a Gaussian kernel in order to obtain its representation in the Contour Scale Space (CSS) [1]. The curvature peaks are determined using a multi-scale analysis. The values of the curvature and the corresponding position in curvilinear abscise will compose the CS descriptor. The similarity measure used to compare two CSS representation is based on a matching procedure which takes into account the cost of fitted and unfitted curvatures peaks [12].

A second descriptor, also adopted in the MPEG-7 Description Scheme, is the Region Shape (RS) descriptor. Here, the image is decomposed using a basis of 2D Angular Radial Transform (ART) functions [13]. Thus, the 2D object is represented as a weighted sum of 34 ART basis functions. The RS descriptor is composed of the 34 corresponding weights. When comparing two ART representations, their distance is given by the L_1 distance between the absolute values of the ART coefficients.

A second region-based descriptor is based on the Hough Transform (HT) [14], [15]. In order to obtain the HT, each point p representing the 2D object in the xOy Cartesian space is associated with a family of lines $l(s, \theta)$, where θ represents the angles between the Ox axe and the line l and s is the distance from the coordinate system origin to the line. Thus, the region function of the 2D object can be represented in the (s, θ) cumulative space. A simple L_1 distance is used as similarity measure between two HT representations.

Finally, we also propose a second contour-based descriptor, so-called Angle Histogram (AH). The contour of the 2D object is first extracted and sub-sampled in N successive 2D points. For each point i is computed the angle α_i defined by the points i-n, i, and i+n. Having the ensemble of angles α_i, i=1…N for a given step n, we can compute the angular histogram. When using low values of the step n, the three considered points are close one from the other and the histogram describes the local variations of the contour. When using larger values of the step n, the angular histogram encodes the global behavior of the contour. The 2D AH descriptor is composed by concatenating several histograms which are described by different values of the step n. In the present work we have used 18-bins for each angular histogram and 5 different histograms to create the AH descriptors. The distance between two objects encoded by the AH descriptor is given by the L_1 distance between the corresponding coefficients.

In order to evaluate the above presented descriptors (*i.e.*, CS, RS, HT and AH) and projection strategies (PCA3, PCA7, OCTA9, OCTA33, LFD and LFDPCA), we have established an evaluation protocol described in the following section.

IV. Experimental Results

A. Evaluation protocol

The experiments we have carried out aim at finding the performance of a 2D shape recognition approach based on 2D/3D indexing.

In order to transfer the semantics from a classified 3D model to an unknown 2D object represented by an image (Figure 1.), we need a tool allowing the matching between 2D and 3D content. By using the 2D/3D indexing methods presented in section III, a 2D object O can be compared to a 3D model M. Their distance $d(O, M)$ is given by the minimum distance between the object and each projection $P_i(M)$:

$$d(O,M) = \min_i d\big(O, P_i(M)\big). \quad (1)$$

In order to identify the class of a 2D object, it is compared against all the 3D models from the considered repository. Further, the models M_i are sorted by decreasing order of similarity and only the first N_M of them are retained. Each model retained votes for a category and thus we can identify the k most represented classes (C_1…C_k) among the first models. Finally, one or several classes are associated to the unknown image. If one of these classes coincides with the category to which belongs the image, then we can state that the recognition has succeeded.

The evaluation measure that we have used is the recognition rate (RR), defined as the percentage of cases when the recognition has succeeded. Depending on the number k of categories that are taken into account, different $RR(k)$ values are obtained. In our experiments we have considered $RR(1)$, $RR(2)$ and $RR(3)$.

Two different 3D model repositories were used in order to evaluate the performance of the recognition system. The first one is the MPEG-7 3D Model database [16] and consists of 362 models semantically divided in 23 categories. The second

database is the Princeton Shape Benchmark (PSB) [17] which includes 1814 models classified in 161 categories.

The experiments have been carried out on a 2D objects database consisting in 115 images randomly chosen from the web (5 images for each MPEG-7 category). The interest objects were manually segmented from each image before the extraction of the shape descriptor. When considering the PSB, only 65 of these 2D objects have been tested.

Tables 1 and 2 present the recognition rates obtained when using the MPEG-7 respectively the PSB databases.

TABLE I. RECOGNITION RATE USING THE MPEG-7 DATABASE

CS	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	33,9	34,8	37,4	33,9	33,9	37,4	37,4
RR(2)	41,7	53,9	52,2	50,4	41,7	51,3	51,3
RR(3)	53,9	61,7	**59,1**	60,0	53,9	56,5	**60,0**

RS	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	24,3	22,6	28,7	27,0	24,3	26,1	30,4
RR(2)	36,5	37,4	40,9	37,4	36,5	42,6	46,1
RR(3)	40,9	45,2	**46,1**	45,2	40,9	50,4	**54,8**

AH	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	30,4	35,7	44,3	42,6	30,4	32,2	38,3
RR(2)	47,8	55,7	60,9	56,5	47,8	48,7	60,0
RR(3)	56,5	61,7	**67,0**	62,6	56,5	60,0	**70,4**

HT	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	18,3	20,9	27,0	24,3	18,3	28,7	34,8
RR(2)	27,0	29,6	35,7	30,4	27,0	36,5	41,7
RR(3)	37,4	37,4	**46,1**	35,7	37,4	43,5	**49,6**

We observe that in the most cases LFD and OCTA33 projection strategies yield the maximal performances in terms of recognition rates. The difference of the scores obtained with the LFD and with the LFDPCA strategies may be explained by the fact that a part of the 3D models present symmetries (as the cars, the airplanes, the humanoids…). The views generated by the LFDPCA strategies are obtained using couples of symmetrical positions of the camera and thus results pairs of mirror-reflected images. Therefore, the redundancy appeared in the LFDPCA reduce the number of useful views from 10 to 5.

For the MPEG-7 database we reach 60% recognition rate for the CS descriptor and 70.4% for AH while only 49.6% when using HT and 54.8% when RS was employed. When using the PSB, the same tendency can be observed: the descriptors providing best recognition rates are CS and AH with scores of 64.6% for CS and 60% for AH.

The fact that we accept several possible classes as response means that the system is able to reduce the number of candidate categories from 161 (in the case of PSB database) to k=1,2,3. In a second stage of our work we intend to implement an algorithm able to select among these k proposed classes.

Finally, in order to be useful, a recognition system has to dispose of appropriate interface allowing user interaction. The platform that we propose is illustrated in figures 7 and 8.

The user has the possibility to select one or several 2D/3D indexing methods (among those presented in section III) and to choose a query image. For each indexing method, the system returns the first three categories that are the most probable for that query and also sort all the 3D models from the considered database by decreasing similarity with the given example.

TABLE II. TABLEAU 1: RECOGNITION RATEUSING THE PSB

CS	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	32,3	41,5	40,0	41,5	32,3	41,5	44,6
RR(2)	43,1	53,8	53,8	50,8	43,1	49,2	58,5
RR(3)	49,2	58,5	**58,5**	55,4	49,2	56,9	**64,6**

RS	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	26,2	20,0	23,1	24,6	26,2	29,2	32,3
RR(2)	30,8	27,7	32,3	41,5	30,8	43,1	40,0
RR(3)	38,5	35,4	38,5	41,5	38,5	46,2	**46,2**

AH	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	27,7	40,0	40,0	36,9	27,7	35,4	44,6
RR(2)	40,0	50,8	49,2	53,8	40,0	50,8	52,3
RR(3)	49,2	55,4	52,3	58,5	49,2	**60,0**	53,8

HT	PCA3	PCA7	LFD	LFDPCA	OCTA3	OCTA9	OCTA33
RR(1)	10,8	12,3	21,5	18,5	10,8	26,2	26,2
RR(2)	12,3	15,4	32,3	23,1	12,3	32,3	33,8
RR(3)	15,4	20,0	36,9	24,6	15,4	35,4	**40,0**

Figures 7 and 8 show two examples of queries representing a helicopter and respectively a formula 1 car when using CS and AH descriptors and OCT33 and LFD projection strategies.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel approach for 2D object categorization. Experimental results show that the information contained in the 3D models can be exploited in order to semantically label 2D objects segmented from images. Several descriptors have been tested and for both 3D model databases the two contour-based descriptors (CS and AH) provided best recognition rates. Among the projection strategies presented in section III.A, LFD and OCTA33 are those giving in most of the cases the best result.

In order to increase the recognition rate, for our future work we intend to exploit the complementarities between the descriptors by combining them. Also, we intend to develop a second algorithm which will select a unique response among those proposed by the current system.

Figure 7. 2D/3D retrieval and classification with the proposed system. Upper images represent the query object segmented from the image and the original image. Lower, we can observe the views of the 3D models retrieved in first positions. For each 2D/3D indexing method the three most probable categories are illustrated by the three icons (in this example a helicopter, a tree and a motorcycle for the CS descriptor and a helicopter, an airplane and a tank for the AH descriptor).

Figure 8. 2D/3D retrieval and classification with the proposed system. Query representing a formula 1 car.

ACKNOWLEDGMENT

This work has been performed within the framework of the UBIMEDIA Research Lab, between Institut TELECOM and Alcatel-Lucent Bell-Labs.

REFERENCES

[1] Mitchell, T. M., 1997. Machine Learning. New York: McGraw-Hill.

[2] Xue, M., Zhu, C., A Study and Application on Machine Learning of Artificial Intelligence, International Joint Conference on Artificial Intelligence, pp. 272, July 2009.

[3] A. Toshev, A. Makadia, and K. Daniilidis: Shape-based Object Recognition in Videos Using 3D Synthetic Object Models, IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009.

[4] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D Feature Maps. In IEEE CVPR, 2008.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. IEEE TPAMI, 24(4):509–522, 2002.

[6] R.A. Schwengerdt, Remote Sensing: Models and Methods for Image Processing, 2nd. Ed., Academic Press, 1997.

[7] Petre, R., Zaharia, T., Preteux, F., "An overview of view-based 2D/3D indexing methods", Proceedings of Mathematics of Data/Image Coding, Compression, and Encryption with Applications XII, volume 7799, August 2010.

[8] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen and Ming Ouhyoung, On visual similarity based 3D model retrieval, Computer Graphics Forum, vol. 22, no. 3, pp. 223-232, 2003.

[9] M. Bober, "MPEG-7 Visual Shape Descriptors", IEEE Transaction on Circuits and Systems for Video Technology, Volume 11, Issue 6, pp. 716-719, August 2002 .

[10] B.S. Manjunath, Phillipe Salembier, Thomas Sikora, "Introduction to MPEG-7: Multimedia Content Description Interface", John Wiley & Sons, Inc., New York, NY, 2002.

[11] F. Mokhtarian, A.K. Mackworth, "A Theory of Multiscale, Curvature-Based Shape Representation for Planar Curves", IEEE Transaction on Pattern Analysis and Machine Intelligence, pp. 789-805, August 1992.

[12] ISO/IEC 15938-3: 2002, MPEG-7-Visual, Information Technology – Multimedia content description interface – Part 3: Visual, 2002.

[13] W.-Y. Kim, Y.-S. Kim, "A New Region-Based Shape Descriptor", ISO/IEC MPEG99/M5472, Maui, Hawaii, December 1999.

[14] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. Commun. ACM, 15(1):11–15, 1972.

[15] Hart, P.E.: How the Hough transform was invented" IEEE Signal Processing Magazine, November 2009.

[16] T. Zaharia, F. Prêteux, 3D versus 2D/3D Shape Descriptors: A Comparative study, In SPIE Conf. on Image Processing: Algorithms and Systems, Vol. 2004 , Toulouse, France, January 2004.

[17] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser, "The Princeton Shape Benchmark", Shape Modeling International, Genova, Italy, June 2004.