2D/3D SEMANTIC CATEGORIZATION OF VISUAL OBJECTS

Raluca Diana Petre [1,2], *Titus Zaharia* [1]

[1] ARTEMIS Departement; Institut TELECOM, TELECOM SudParis; Evry, France
UMR CNRS 8145 – MAP5
[2] Alcatel-Lucent Bell Labs France
{Raluca-Diana.Petre, Titus.Zaharia}@it-sudparis.eu

ABSTRACT

In the context of content-based indexing applications, the automatic classification and interpretation of visual content is a key issue that needs to be solved. This paper proposes a novel approach for semantic video object interpretation. The principle consists of exploiting the *a priori* information contained in categorized 3D model data sets, in order to transfer the semantic labels from such models to unknown video objects. Each 3D model is represented as a set of 2D views, described with the help of shape descriptors. A matching technique is used in order to perform an association between categorized 3D models and 2D video objects. The experimental evaluation shows the interest of our approach, which yields recognition rates of up to 92.5%.

Index Terms— 2D/3D indexing, object classification, video indexing, 3D model, shape descriptors

1. INTRODUCTION

Over the last decades, digital technologies have known a spectacular evolution, which made them more and more accessible for the general public. Thus, the amount of multimedia content (still images, videos, 2D/3D graphics) is increasing exponentially. Within this context, the access to the material of interest for a user became a real challenge. Disposing of powerful search and retrieval methods becomes mandatory for efficient indexing and intelligent access to audio-video material. In this context, content-based indexing methods propose an interesting alternative to classic, textual annotations. Their main advantage is that they overcome the linguistic barriers by focusing on the content information. Also, automatic indexing avoids the tedious and highly subjective process of manual annotation.

The objective of automatic object categorization is to determine, without human interaction, the semantic meaning of an object present in an image or video. Most popular approaches are based on machine learning (ML) techniques [1], [2] in order to accomplish the object classification purpose.

Difficulties arise, however, when a large number of categories is involved. In order to guarantee the discrimination capacity, a large amount of features has to be exploited. However, such a solution has a strong impact on the associated computational complexity, which may become intractable [3]. As it is difficult to extract features with a large generalization capacity, this issue has to be overcome with the use of a large variety of examples in the training phase. In addition, a given object may present very different appearances due to the pose variation. Thus, the amount of training examples is further increased because different instances of the same object are needed.

The aim of our work is to avoid the ML techniques by exploiting the information contained in categorized 3D models. In this paper we proposed a new recognition framework designed to automatically assign semantic labels to video objects.

The rest of this paper is organized as follows. Related work is briefly presented in the second section. In Section 3, we present the 2D/3D indexing principle and the adopted methods. The model-based video object recognition framework is presented in Section 4, while the experimental results are detailed in Section 5. Finally, Section 6 concludes the paper and opens perspectives of future research.

2. RELATED WORK

Research on automatic object classification is mainly based on ML techniques **[**4**]**. ML approaches can be divided into two main families: supervised and unsupervised techniques. In the first case, the system aims at finding the function which better discriminates between several sets of labeled data. This function is further employed to classify new cases. For some examples of supervised ML approaches, the reader is invited to refer [5], [6], [7]. Supervised approaches may be very accurate [8], but their main limitation is related to the over-fitting problem [9]. In addition, sufficiently large training sets with already classified objects are requested.

The second family of ML methods allows training from partially or completely unlabelled data. Some commonly used unsupervised machine learning methods are K-means, mixture methods, K-Nearest Neighbor… Several unsupervised ML methods are proposed in [10] and [11]. In terms of performances, the unsupervised methods are less accurate than the supervised machine learning methods.

Even if most object recognition approaches rely on ML techniques, the idea of using categorized 3D models for recognition purposes has been also investigated more recently.

In [12], authors use textured 3D models in the recognition process. For each class, a visual codebook of K=2000 clusters is constructed by extracting appearance features from the views of the 3D models. The obtained codebooks are used in the still image recognition process.

In [13], authors use non-textured 3D models in order to categorize 2D objects segmented from videos and represented by a set of frames. Each 3D model is also represented by a set of 20 views, determined by k-means clustering of 500 evenly distributed projections. Finally, a matching procedure is used in order to determine the relation between the video objects and the 3D model's projections and to estimate the pose of the visual object in each one of the selected frames.

In this paper we propose a different recognition framework which aims to associate semantic labels to video objects and which extends our previous work on still image object recognition, introduced in [14]. Similarly to the work presented in [12] and [13], we exploit in the recognition process the information contained in 3D models. In contrast with the approach introduced in [12], we use only non-textured models and rely the recognition process exclusively on shape features, because the texture of real objects may present important intra-class variations. Compared to the work presented in [12] and [13], our approach makes it possible to deal with a larger number of categories, which are in the same time more complex in terms of shape.

Different 2D/3D shape-based indexing methods are considered, as described in the following section.

3. SHAPE-BASED 2D/3D INDEXING

Let us first recall the general principle of 2D/3D indexing methods.

3.1. The principle of 2D/3D indexation

The basic principle of 2D/3D indexing approach is to represent a 3D model, denoted by M, as a set of 2D projections $\{Pr_i(M)\}$, obtained from different angles of view. The main advantage of using 2D/3D indexing techniques is that they allow comparing a 3D model with other 3D models but also with 2D objects extracted from still images or videos, based on the following hypothesis: if two models are similar, then they should present similar views.

The set of views that represent the 3D model is obtained by considering a set of viewing angles (*i.e.* positions of the camera in the 3D space). In order to obtain a unique set of views, whatever the object's size, position and orientation, each model M is first centered in the origin of the Cartesian system and resized to fit the unit sphere. Then, a Principal Component Analysis (PCA) [15] is performed in order to compute the axes of inertia of the 3D model. The rotation invariance is achieved by aligning the 3D model's axes of inertia with the Cartesian system.

The model is then projected and rendered in 2D from N_{Pr} different viewing angles, thus resulting the set of 2D projections $\{Pr_i(M)\}$, where i=1..N_{Pr}. Each projection is a binary image, which is also called silhouette or view. Two silhouettes of a 3D model obtained from opposite directions represent one the mirror reflection of the other. Thus, in order to reduce the redundancy, the viewing angles should cover only half of the bounding space.

Finally, each projection is described with the help of a 2D shape descriptor. The set of all descriptors is associated to the 3D model.

In order to fully implement a 2D/3D indexing approach, several aspects have to be specified.

First, the number N_{Pr} of viewing angles used to obtain the projections has to be carefully chosen, since a large number of silhouettes provides a more complete description while increasing the computational and storage costs. Also, there are several strategies for the repartition of the viewing angles around the 3D model. Two main hypotheses are used for the repartition. The first one assumes that the most important views are those corresponding to the projection on the principal planes. The second hypothesis is to consider all the views equals by uniformly distributing the cameras around the 3D model.

In the next section we present different strategies of projection and description retained in our work.

3.2. The proposed 2D/3D indexing methods

3.2.1. The viewing angle selection

Several viewing angle selection strategies may be considered.

Let us start with the MPEG-7 approach [16], which relies on PCA and assumes that the most significant views of a 3D model are those corresponding to the first three principle planes. In the following, this strategy will be referred to as PCA3. For a more complete description, the four bisectors of the eight octants defined by the principal planes are also considered, resulting in a total of seven views (PCA7 strategy).

The second strategy uses the vertices of a regular dodecahedron in order to obtain a uniform distribution of the viewing angles. This distribution strategy was first introduced in [17] where the authors present the Light Field Descriptor (LFD). Thus, this strategy will be denotes byLFD. Here, we can have two subcases. In the first one the 3D model is aligned w.r.t. the coordinate system (and implicitly to the dodecahedron used for camera distribution). This case will be referred to as LFDPCA. In the second case, the 3D model has an arbitrary position in the virtual space (strategy called simply LFD).

Finally, the third strategy uses as angles of view the vertices of a regular octahedron, whose faces are recursively subdivided [18]. According to the subdivision level, 3, 9 and respectively 33 views are obtained. These strategies are referred to as OCTA3, OCTA9 and OCTA33. As the octahedron is aligned w.r.t the coordinate system, the first three views corresponds to the projections on the first principal planes, so OCTA3 and PCA3 strategies are equivalent.

By using the above-presented viewing angle distribution, a set of N_{Pr} projections is obtained. Further, each projection is described by using a 2D shape descriptor. Let us now briefly recall the 2D shape descriptors retained in our work.

3.2.1. The 2D shape description

The choice of the 2D shape description methods relies on our previous evaluation [14] that proved that the contour-based descriptors outperform those exploiting the support region of the object. Thus, we chose to retain only the Contour Shape and the Angle Histogram descriptors.

The Contour Shape (CS) descriptor [19], proposed by the MPEG-7 standard [20], [21], [22] uses the contour scale space (CSS) representation. The contour of the shape is first filtered using different Gaussian kernels, resulting in a set of several smoothed contours. For each of them, the curvilinear positions of the inflexion points are computed. The curvature peaks are determined and for each peak the corresponding curvature value and curvilinear position are retained as CS descriptor. The associated similarity measure used to compare two images represented by their CS descriptors is based on a matching procedure that takes into account the cost of fitted and unfitted curvature peaks [20].

The second retained descriptor is the so-called Angle Histogram (AH) introduced in [14]. The contour of the 2D object is first extracted and then sampled in a fixed number of points $\{S_i\}$. Further, the angular distribution of sets of three samples ($S_{i-\alpha}$, S_i, $S_{i+\alpha}$) is computed and represented as a histogram. Depending on the distance α between the considered samples, the angular histogram encodes the local or the global behavior of the shape. Thus, five different angular histograms (representing local and global features) are computed and concatenated in order to obtain the AH descriptor. The associated similarity measure is the L_1 distance computed between the two AH representations.

The 2D/3D indexing methods are further integrated in the recognition framework described in the next section.

4. VIDEO OBJECT RECOGNITION FRAMEWORK

The goal of the video object recognition framework is to associate semantic labels to the objects present in a video.

Figure 1 represents an overview of the video object recognition framework. A 3D model categorized database is supposed to be available. Each model is described using a 2D/3D indexing technique.

In order to reduce the computation complexity, for an input video V. a set of N_F frames $\{F_i(V)\}$ (presenting the object of interest in different poses) is selected. Such a selection may be done by considering a frame-clustering algorithm such as those introduced in [23], [24].

Further, the object of interest is segmented from each one of the N_F retained frame, resulting a set $\{P_i(VO)\}$ of poses that represent the video object VO. For object extraction, we have considered an image segmentation algorithm similar to the one presented in [25].

Each pose $P_i(VO)$ of the video object VO is described using the same 2D shape descriptor that was used for the 3D model's projections $Pr_i(M)$ and further compared with each projection of the 3D model by computing the distance $d(P_i(VO), Pr_j(M))$ associated with the considered descriptor. The distance between a pose P_i and a 3D model M is given by the minimum distance between that pose and all the projections of the model, as described in equation (1).

$$d(\mathrm{P}_i(VO), M) = \min_j d(\mathrm{P}_i(VO), \mathrm{Pr}_j(M)); \ \mathrm{i} = 1:\mathrm{N}_F, \mathrm{j} = 1:\mathrm{N}_{\mathrm{Pr}}. \quad (1)$$

Where N_{pr} is the number of projections and N_F is the number of poses (appearances) of the video object VO.

The similarity between a video object VO (represented by the set of poses $\{P_i(VO)\}$) and a 3D model is defined as the sum of the distances between each pose and the model:

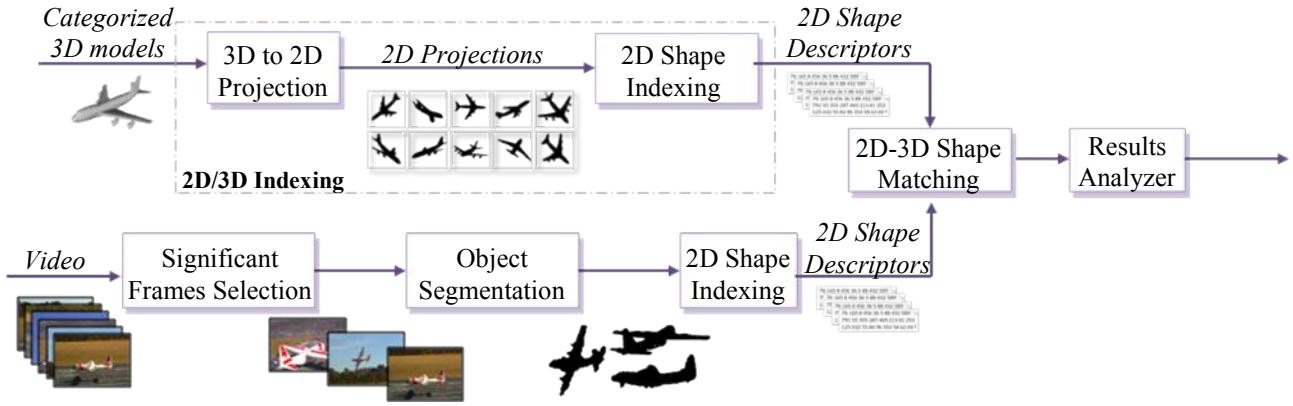

Figure 1: Video object recognition framework

$$d(VO, M) = \sum_{i=1}^{N_F} d(\mathrm{P}_i(VO), M). \quad (2)$$

Finally, the results analyzer module uses these distances in order to establish which are the most probable categories that fit the input video object. First, the N most similar 3D models from the database are retrieved. Then, for the N retained models we count the number of occurrences for each class. Finally, the N_C most represented classes are proposed to the user.

The performances of the proposed recognition framework are analyzed in the next section.

5. EXPERIMENTAL EVALUATION

The experiments have been carried out on a database of 40 videos selected from Internet and including the following 8 object categories: airplanes, cars, chess pieces, helicopters, humanoids, motorcycles, pistols and tanks. Each category is present in 5 different videos with different appearances, and for each video N_F=3 frames were selected (Figure 2).

Figure 2: Sample of video frames and extracted objects

For the categorized 3D content, we have considered the MPEG-7 3D model dataset [16], which consists of 362 models divided into 23 semantic categories (humanoids, airplanes, helicopters, cars, race cars, trees (with and without leafs), rifles, pistols, missiles, letters…).

The performance measure adopted is the recognition rate, denoted by $RR(N_C)$, and defined as the percentage of cases where the correct category is proposed within the top N_C most represented categories. In our experiments we have considered Nc=1, 2, 3.

First, we have evaluated the performances when performing the recognition process from a single image. In this case, we have considered each pose of the video object independently from the others.

The algorithm has been run on an Intel Xeon machine with 2.8GHz and 12GB RAM, under a Windows 7 platform. When the AH descriptor was employed, the still object recognition process took between 40ms and 260ms for PCA3, respectively OCTA33 viewing angle selection. In the case of CS descriptor, the time was between 150ms with PCA3 and 510ms with the OCTA33 projection strategy. The recognition process includes the extraction of the still object descriptor, the distance computation and analysis of the proposed output categories.

TABLE I. shows the recognition rates obtained for the 120 still image objects when employing the CS (TABLE I.a) and the AH (TABLE I.b) descriptors.

As in our previous work [14], here again we observe that in most cases LFD and OCTA33 viewing angle selection strategies provide maximal performances, with $RR(3)$ scores up to 76.67% for the CS descriptor and 75.0% for the AH descriptor.

TABLE I. STILL IMAGE OBJECTS RECOGNITION RATE

a.

CS	PCA3	PCA7	LFD	LFDPCA	OCTA9	OCTA33
RR(1)	38,33	50,83	55,83	45,83	50,00	53,33
RR(2)	55,00	65,00	68,33	64,17	60,83	64,17
RR(3)	65,83	73,33	75,00	71,67	68,33	**76,67**

b.

AH	PCA3	PCA7	LFD	LFDPCA	OCTA9	OCTA33
RR(1)	29,17	43,33	43,33	45,83	40,00	38,33
RR(2)	40,83	60,83	66,67	61,67	54,17	62,50
RR(3)	50,83	70,83	**75,00**	72,50	66,67	68,33

In a second time, we have applied the recognition process with three different poses per video object. The obtained recognition rates are presented in TABLE II. We can observe that increasing the number of views for each query from one to three leads to a gain of up to 16%. Thus, a score $RR(3)$ of 92.5% is obtained for the CS descriptor, with the OCTA33 projection strategy. When using the AH descriptor, the best recognition rate is of 90.00% with the PCA7 projection strategy.

TABLE II. VIDEO OBJECTS RECOGNITION RATE

a.

CS	PCA3	PCA7	LFD	LFDPCA	OCTA9	OCTA33
RR(1)	47,50	67,50	70,00	65,00	70,00	72,50
RR(2)	65,00	75,00	80,00	77,50	80,00	82,50
RR(3)	77,50	80,00	85,00	82,50	85,00	**92,50**

b.

AH	PCA3	PCA7	LFD	LFDPCA	OCTA9	OCTA33
RR(1)	32,50	55,00	45,00	52,50	50,00	55,00
RR(2)	45,00	77,50	70,00	60,00	70,00	80,00
RR(3)	57,50	**90,00**	85,00	75,00	75,00	82,50

The recognition rates detailed per object category are presented in TABLE III. Here, we have considered only the two best performing projection strategies (*i.e.* LFD and OCTA33) with the CS descriptor. The classes airplane, car and humanoid present 100% recognition rates even when considering only the first retrieved category (*i.e.* $RR(1)$ score).

The pistol class achieves recognition rates of only 40% for the LFD projection strategy and 60% for the OCTA33 projection strategy.

The proposed approach uses very compact descriptors, with low complexity similarity measures. Thus, it allows us to reduce the searching space from 23 to *Nc*=1, 2, 3 categories in a very simple yet effective manner. Thereby, the proposed approach can be exploited as a searching space reduction phase for more complex recognition algorithms.

TABLE III. VIDEO OBJECTS RECOGNITION RATE

CS	LFD			OCTA33		
	RR(1)	*RR*(2)	*RR*(3)	*RR*(1)	*RR*(2)	*RR*(3)
airplane	100	100	100	100	100	100
car	100	100	100	100	100	100
chess	80	80	80	80	80	80
helicopter	60	80	100	80	100	100
humanoid	100	100	100	80	100	100
motorcycle	60	60	60	20	40	100
pistol	0	40	40	60	60	60
tank	60	80	100	60	80	100

6. CONCLUSIONS AND FUTURE WORK

In this paper we have address the issue of video object categorization. We proposed a new recognition algorithm, which exploits the *a priori* information contained in categorized 3D models. Two contour-based representations have been used (CS and AH descriptors), for describing the shape information.

The experiment proved that high recognition rates (up to 92.5%) can be achieved by selecting from the video only three instances of the object.

In our future work we intend to use additional information, such as internal edges and/or interest points in order to obtain a more discriminant description.

7. ACKNOLEDGEMENT

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

8. REFERENCES

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

[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] Li, Ling, Data complexity in machine learning and novel classification algorithms. Dissertation (Ph.D.), California Institute of Technology, 2006.

[4] D. Lu, Q.Weng, A survey of image classification methods and techniques for improving classification performance, International jurnal of Remote Sensing, Volume 28, No.5, pp. 823-870, 2007.

[5] A. Bosch, A. Zisserman, and X. Mu~noz. Image classication using random forests and ferns. In International Conference on Computer Vision, Rio de Janeiro, Brazil, Oct. 2007.

[6] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In Neural Information Processing Systems Conference, Vancouver, BC, Canada, Dec. 2006.

[7] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In European Conference on Computer Vision, volume 3951 of Lecture Notes in Computer Science, po. 1-15, Graz, Austria, May 2006.

[8] Deselaers, T., Heigold, G., Ney, H., Object classification by fusing SVMs and Gaussian mixtures, Vol. 43, Issue 7, pp. 2476-2484, July 2010.

[9] Pados, G.A., Papantoni-Kazakos, P., A note on the estimation of the generalization error and prevention of overfitting [machine learning], IEEE Conference on Neural Networks, volume 1, pp 321, July 1994.

[10] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In Proc. ECCV, pages 18–32, 2000.

[11] R. Fergus, P. Perona, A. Zisserman. Object class recognition by unsupervised scale-invariant learning, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2 (2003), pp. 264-271, June 2003.

[12] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D Feature Maps. In IEEE CVPR, pp. 1-8, 2008.

[13] 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, Volume 60, No. 2, pp. 91-110, Miami, FL, 2009.

[14] Petre, R. D., Zaharia, T., "3D models-based semantic labeling of 2D objects", International Conference on Digital Image Computing: Techniques and Applications, pp. 152-157, Noosa, QLD, Australia, December 2011.

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

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

[17] 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.

[18] 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, pp. 779904, August 2010.

[19] F. Mokhtarian, A.K. Mackworth, A Theory of Multiscale, Curvature-Based Shape Representation for Planar Curves, IEEE Transaction on Pattern Analysis and Machine Intelligence, Volume 14, No. 8, pp. 789-805, August 1992.

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

[21] 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 .

[22] B.S. Manjunath, Phillipe Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, John Wiley & Sons, Inc., Volume 1, New York, NY, 2002.

[23] R.G. Tapu, T. Zaharia, High Level Video Temporal Segmentation, Proceeding of SVC'11 Proceedings of the 7th international conference on Advances in visual computing, Volume 1, pp. 224-235, September 2011

[24] Rasheed, Z.; Shah, M.; Detection and Representation of Scenes in Videos, IEEE transactions on Multimedia, Issue 6, pp. 1097-1105, December 2005.

[25] Protiere, A.; Sapiro, G.; Interactive Image Segmentation via Adaptive Weighted Distances, IEEE Transactions on Image Processing, Issue 4, pp. 1046-1057,April 2007.