Video Object Recognition and Modeling by SIFT Matching Optimization

Alessandro Bruno, Luca Greco and Marco La Cascia
Dipartimento di Ingegneria Chimica, Gestionale, Informatica, Meccanica,
Università degli studi di Palermo, Palermo, Italy

Keywords: Object Modeling, Video Query, Object Recognition.

Abstract: In this paper we present a novel technique for object modeling and object recognition in video. Given a set of videos containing 360 degrees views of objects we compute a model for each object, then we analyze short videos to determine if the object depicted in the video is one of the modeled objects. The object model is built from a video spanning a 360 degree view of the object taken against a uniform background. In order to create the object model, the proposed techniques selects a few representative frames from each video and local features of such frames. The object recognition is performed selecting a few frames from the query video, extracting local features from each frame and looking for matches in all the representative frames constituting the models of all the objects. If the number of matches exceed a fixed threshold the corresponding object is considered the recognized objects .To evaluate our approach we acquired a dataset of 25 videos representing 25 different objects and used these videos to build the objects model. Then we took 25 test videos containing only one of the known objects and 5 videos containing only unknown objects. Experiments showed that, despite a significant compression in the model, recognition results are satisfactory.

1 INTRODUCTION

The ever-increasing popularity of mobile devices such as smartphones and digital cameras, enables new classes of dedicated applications of image analysis such as mobile visual search, image cropping, object detection, object recognition, data representation (object modeling), etc.... Object modeling and object recognition are two of the most important issues in the field of computer vision.

Object modeling aims to give a compact and complete representation of an object. Object models can be used for many computer vision applications such as object recognition and object indexing in large database.

Object recognition is the core problem of learning visual object categories and visual object instance. Researchers of computer vision considered two types of recognition: the specific object case and the generic category case. In the specific case the goal is to identify instances of a particular object. In the generic category case the goal is to recognize different instances of objects as belonging to the same conceptual class. In this paper we focused our attention on the first case (the specific instance of a particular object). More in details we developed a new technique for video object recognition and modeling (data representation).

Matching and learning visual objects is a challenge on a number of fronts. The instances of the same object can appear very differently depending on variables such as illumination conditions, object pose, camera viewpoint, partial occlusions, backgroud clutter.

Object recognition is accomplished by finding a correspondence between certain features of the image and comparable features of the object model. The two most important issues that a method must address are what constitutes a feature, and how is the correspondence found between image features and model features. Some methods use global features, which summarize information about the entire visible portion of an object, other methods use local features invariant to affine transforms such as local keypoints descriptors (Lowe, 2004).

We focus our work on methods that use local features, such as local keypoints descriptors such as SIFT.

Bruno A., Greco L. and La Cascia M..
Video Object Recognition and Modeling by SIFT Matching Optimization.
DOI: 10.5220/0004828006620670
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 662-670
ISBN: 978-989-758-018-5

The contributes of our paper are: a new technique for object modeling; a new method for video object recognition based on objects matching; a new video dataset that consists of 360 degree video collection of thirty objects (CVIPLab, 2013).
We suppose to analyze the case in which a person take a video of an object with a videocamera and then wants to know information about the object. The scenario of our system is to upload the video to a system able to recognize the video object taken by the videocamera.

We developed a new model for video objects by giving a very compact and complete description of the object. We also developed a new video object recognition based on object matching that achieves very good results in terms of accuracy.

The rest of this papers is organized as follows: in section 2 we describe the related work of the state of the arts in object modeling and object recognition; in section 3 a detailed description of the video object models dataset is given; in section 4 we describe the proposed method for object recognition; in section 5 we show the experimental results; the section 6 ends the paper with some conclusions and future works.

2 RELATED WORKS

In this section we show the most popular method for object modeling and object recognition with particular attention to video oriented methods.

2.1 Object Modeling

The most important factors in object retrieval are the data representation (modeling) and the search (matching) strategy. In (Li, 1999) the authors use multiresolution modeling because it preserves necessary details when they are appropriate at various scales. Features such as color, texture, shape are used to build object models, more particularly GHT (the Generalized Hough Transform) is adopted over the others shape representations because it is robust againts noise and occlusion. Moreover it can be applied hierarchically to describe the object at multiple resolution.

In recogniton kernel (Li,1996) based method, the features of an object are extracted at levels that are the most appropriate to yield only the necessary details; in (Day,1995) the authors proposed a graphical data model for specifying spatio-temporal semantics of video data for object detection and recognition. The most important information used in (Chen, 2002) are the relative spatial relationships of the objects in function of time evolution. The model is based on capturing the video content in terms of video objects. The authors differentiate foreground video objects and background video objects. The method includes the detection of background video objects, foreground video objects, static video objects, moving video objects, motion vectors. In (Sivic, 2006) Sivic et al. developed an approach to object retrieval which localizes all the occurrences of an object in a video. Given a query image of the object, this is represented by a set of viewpoint invariant region descriptors.

2.2 Object Recognition

Object recognition is one of the most important issue in computer vision community. Some works use video to detect moving objects by motion. In (Kavitha, 2007), for example, the authors use two consecutive frames to first estimate motion vectors and then they perform edge detection using canny detector. Estimated moving objects are updated with a watershed based transformation and finally merged to prevent over-segmentation.

In geometric based approaches (Mundy, 2006) the main idea is that the geometric description of a 3D object allows the projected shape to be accurately analyzed in a 2D image under projective projection, thereby facilitating recognition process using edge or boundary information.

The most notable appearance-based algorithm is the eigenface method (Turk, 1991) applied in face recognition. The underlying idea of this algorithm is to compute eigenvectors from a set of vectors where each one represents one face image as a raster scan vector of gray-scale pixel values. The central idea of feature-based object recognition algorithms lies in finding interesting points, often occurring at intensity discontinuity, that are invariant to change due to scale, illumination and affine transformation.

Object recognition algorithms based on views or appearances, are still a hot research topic (Zhao, 2004) (Wang, 2007). In (Pontil,1998)) Pontil et al. proposed a method that recognize the objects also if the objects are overlapped. In recognition systems based on view, the dimensions of the extracted features may be of several hundreds. After obtaining the features of 3D object from 2D images, the 3D object recognition is reduced to a classification problem and features can be considered from the perspective of pattern recognition. In (Murase, 1995) the recognition problem is formulated as one of appearance matching rather than shape matching.

The appearance of an object depends on its

shape, reflectance properties, pose in the scene and the illumination conditions. Shape and reflectance are intrinsic properties of the object, on the contrary pose and illumination vary from scene to scene. In (Murase, 1995) the authors developed a compact representation of objects, parameterized by object pose and illumination (parametric eigenspace, constructed by computing the most prominent eigenvectors of the set) and the object is represented as a manifold. The exact position of the projection on the manifold determines the object's pose in the image. The authors suppose that the objects in the image are not occluded by others objects and therefore can be segmented from the remaining scene.

In (Lowe, 1999) the author developed an object recognition system based on SIFT descriptors (Lowe, 2004), more particularly, the author used SIFT keypoints and descriptors as input to a nearest-neighbor indexing method that identifies candidate object matches. The features of SIFT descriptors are invariant to image scaling, translation and rotation, partially invariant to illumination changes and affine or 3D projection. The SIFT keypoints are used as input to a nearest-neighbor indexing method, this identifies candidate object matches.

In (Wu, 2011) the authors analyzed the features which characterize the difference of similar views to recognize 3D objects. Principal Component Analysis (PCA) and Kernel PCA (KPCA) are used to extract features and then classify the 3D objects with Support Vector Machine (SVM). The performances of SVM, tested on Columbia Object Image Library (COIL-100) have been compared. The best performance is achieved by SVM with KPCA. KPCA is used for feature extraction in view-based 3D object recognition.

In (Wu, 2011) different algorithms are shown by comparing the performances only for four angles of rotation (10° 20° 45° 90°). Furthermore, the experimental results are based only on images with dimensions 128 x 128.

Chang et al. (Chang, 1999) used the color co-occurrence histogram (that adds geometric information to the usual color histogram) for recognizing objects in images. The authors computed model of color co-occurrence histogram based on images of known objects taken from different points of view. The models are then matched to sub-regions in test images to find the object. Moreover they developed a mathematical probabilistic model for adjusting the number of colors in color co-occurrence histogram.

In (Jinda-Apiraksa, 2013) the focus is on the problem of near-duplicates (ND), that are similar images that can be divided in identical (IND) and non-identical (NIND). IND is formed by transformed versions of an initial image (i.e. blurred, cropped, filtered), NIND by pictures containing the same scene or objects. In this case, the subjectivity of "how much" two image are similar is a hard problem to face off. They present a NIND ground truth derived by asking directly to ten subjects and they make it available on the web.

A high-speed and high-performance ND retrieval system is presented in the work of (Dong, 2012). They use an entropy-based filtering to eliminate points that can lead to false positive, like those associated to near-empty regions, and a sketch representation for filtered descriptors. Then they use a query expansion method based on graph cut.

Recognizing in video includes the problem of detection and in same cases tracking of the object. The paper of (Chau, 2013) is an overview on tracking algorithms classification where the authors divide the different approaches in point, appearance and silhouette tracking.

In our method we use SIFT for obtaining the object model from multiple views (multiple frames) of the object in the video. In our method the recognition of the object is performed by matching the keypoints of the sampled frames from the video with the keypoints of the objects models. Similarly to the method of Peng Chang et al. (Chang, 1999) we used object modeling for object recognition but we preferred to extract local features (SIFT) rather than global features such as the color co-occurrence histogram.

3 DATASET CREATION AND OBJECT MODELING

The recognition algorithm is based on a collection of models built from videos of known objects. To test the performance of the proposed method we first constructed a dataset of videos representing several objects. Then the modeling method is proposed.

3.1 Dataset

3.1.1 Video Description of the Object

For each object of the dataset the related video contain a 360 degree view of the object starting from a frontal position. This is done using a turntable, a fixed camera and a uniform background. Video resolution is 1280 x 720p (HD) at 30 fps and the

lenght is approximately 15 seconds.

3.1.2 Relation with Real Applications

This type of dataset try to simulate a simple video-acquisition that can be done with a mobile device (i.e. a smartphone) circumnavigating an object that have to be added to the known object database. In real application the resulting video have to be re-elaborated, for example trying to estimate motion velocity and jitter. If a video contain a partial view of the object (i.e. less than 360 degree) recognition task can be still performed but only for the visible part of the object.

3.1.3 Image Dataset

The constructed dataset is formed by videos of 25 different objects. As the angular velocity of the turntable is constant, a subset of 36 frame is sampled uniformly for each object so extracting views that differ by 10 degrees of rotation (see fig. 1). So, starting from the video dataset, an image dataset is also constructed with these samples containing 900 views of the 25 objects. Although original background is uniform, shadows, light changes or camera noise can produce a slightly changing resulting color. In the extracted views the original background is segmented and replaced with a real uniform background (i.e. white) that not produce SIFT keypoint, so storing only the visual information about the object (fig. 2).

3.2 Object Modeling

Starting from the image dataset of 900 images a reduced version is extacted to have, for each object, only a subset of the initial 36 images representing the visual model to be used for recognition.

3.2.1 Overview

For each object, the model is extracted as follow:

1. SIFT descriptors and keypoints are calculated for all views;
2. for each view, only SIFT points that match with points in previous or next view are used as view descriptors;
3. the number of point of each view is used as discrete function and local maxima and minima are extracted;
4. object model is obtained taking images corresponding to maxima and minima.

Figure 1: Complete 360 degree view of the video object.

Figure 2: On the right, the video object frame, on the left the video object, without background.

3.2.2 Maxima and Minima Extraction

Rotating an object by few degrees, most part of the object that is visible starting the rotation generally is still visible at the end. This is related to the object geometry (shape, occluding parts, symmetrics) and the pattern features (color change, edges). Calculating SIFT descriptors of two consecutive views (views that differ by 10 degrees of rotation), it is expected that a large part of the descriptors will match.

For each view, if only the keypoints matching with the previous and the next view are considered and the others are discarded, the remaining keypoints are representative of the shared visual informations in a three images range. Only repeated and visible points in at least two views are present in the resulting subset. The number of remaining points is used like a discrete similarity function and local maxima and minima are extracted. Taking local minima of this function, the related images are the most visually different in their neighborhood, so these represent views that contain a visual change of the object. Local maxima, on the other hand, correspond to pictures that contain common details in their neighborhood, so being representative of this. Only views corresponding to local maxima and minima are used to model the object, so taking the images that contain “typical” views (maxima) and visual breaking views (minima) such as in fig. 3. In Fig. 4 and 6 we plot a curve that shows, for a given view (x-axis), the number of SIFT points that match (y-axis) with points in previous or next view.

The curves shown in fig. 4 and 6 can be characterized by a lot of local maxima and minima that could correspond to views that are very close each other. This would go in the opposite direction from the objective of our method, that, on the contrary, aims to represent the object with the fewest possible views. This is the reason why we also apply a 'smooth' interpolation function to the curves shown in fig. 4 and 6. The results of 'smooth' interpolations are depicted in fig. 5 and 7, showing curves very close to the original ones (fig, 4 and 6). Furthermore the curves in fig.5 and 7 have a number of local maxima and minima lower than the curves in fig. 4 and 6. Since now on we call 'dataset model' the model of the object that consists of 36 images/views (that differ by 10 degree of rotation). We call 'full model' the model that consists of the views that correspond to local maxima and minima in not smoothed curves (such as in fig. 4 and 6). We call 'smoothed model' the model that consists of the views corresponding to local maxima and minima in smoothed curves (as in fig. 5 and 7). In tab. 1 we show, for each object, the size of full and smoothed model and the model compression. The latter is the ratio between the number of views composing the current model (i.e. 'full model' or 'smoothed model') and the number of views composing the 'dataset model'.

Figure 3: On the left side, Panda Object View corresponds to a local maxima (0 degree view) of the curve in fig. 4, on the right side Panda Object View corresponds to a local minima (110 degrees view) of the curve in fig.4.

4 PROPOSED RECOGNITION METHOD

Given the dataset and the extracted object models, we propose a method that performs recognition using a video as query input. Input query video may contain or not one of the known objects, the only hypothesis on the video is that if it contains an object of the database then the object is almost always visible in the related video even if subject to changes on scale and orientation.

4.1 Proposed Method

The proposed recognition follows this steps:

1. extract N frames from video query;
2. match every frame with all components of all models;
3. counting the number of matching points for all the views of the models and all frames of the video, take the maximum value. The object related to this match is the recognized object, if the number of matches exceeds a fixed threshold (10 in our experiments).

4.1.1 Refining Matches

If the models give a complete representation of the appearance of the object, step two is crucial for recognition task. Experimental results shows that results can be corrupted in real-word query because cluttered background can lead to incorrect or multiple matches.results can be corrupted in real-word query because cluttered background can lead to incorrect or multiple matches. To make a more robust matching phase, it is important to exclude these noisy points. This can be done using RANSAC

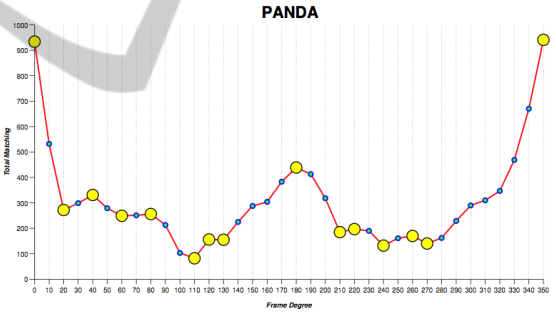

Figure 4: The chart of matching keypoints for all views of Panda Object. Yellow circles are local maxima and minima.

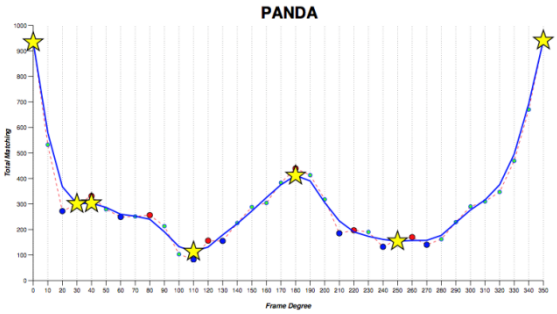

Figure 5: The smoothed chart for Panda Object (blue line). Yellow stars are local maxima and minima. Red dash line is the original chart.

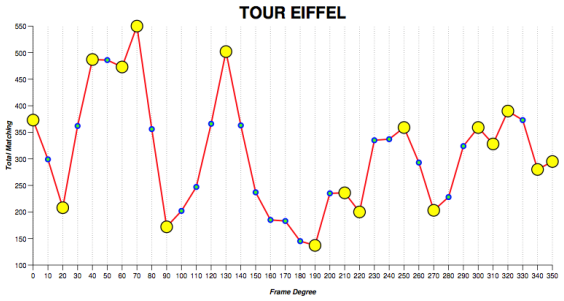

Figure 6: The chart of matching keypoints for all views of Tour Eiffel Object. Yellow circles are local maxima and minima.

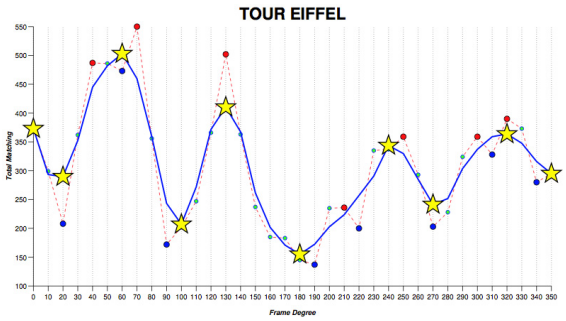

Figure 7: The smoothed chart for Tour Eiffel Object. Yellow stars are local maxima and minima. Red dash line is the original chart.

(Fischler, 1981) in the matching operation, to exclude points that don't fit an homography transformation (fig. 8). Furthermore, considering that a single keypoint of an image can have more than one match with keypoints of the second image, we consider multiple matches of the same keypoint as a single match.

5 RESULTS

Object recognition using video and image dataset was done using the MATLAB implementation of SIFT present in (Vedaldi, 2010) and RANSAC implementation present in (Kovesi, 2003) following the process described in section 4.1. To achieve matches with less but more robust points the

Table 1: In this table, experimental results and statistical values about the video object modeling are shown: object id, object name, the number of the views composing the object 'full model' and the object 'smoothed model', the compression factor (i.e the ratio between the number of object model views and the number of all the object views in the dataset).

obj. ID	name	full model	compression	smoothed model	compression
1	Dancer	14	38.89%	10	27.78%
2	Bible	15	41.67%	9	25.00%
3	Beer	7	19.44%	5	13.89%
4	Cipster	12	33.33%	5	13.89%
5	Tour Eiffel	17	47.22%	10	27.78%
6	Energy Drink	17	47.22%	7	19.44%
7	Paper tissue	13	36.11%	13	36.11%
8	Digital camera	13	36.11%	7	19.44%
9	iPhone	13	36.11%	9	25.00%
10	Statue of Liberty	17	47.22%	11	30.56%
11	Motorcycle	9	25.00%	7	19.44%
12	Nutella	19	52.78%	9	25.00%
13	Sunglasses	23	63.89%	15	41.67%
14	Watch	16	44.44%	9	25.00%
15	Panda	15	41.67%	7	19.44%
16	Cactus	17	47.22%	11	30.56%
17	Plastic plant	19	52.78%	9	25.00%
18	Bottle of perfume	13	36.11%	5	13.89%
19	Shaving foam	10	27.78%	8	22.22%
20	Canned meat	20	55.56%	9	25.00%
21	Alarm clock (black)	15	41.67%	11	30.56%
22	Alarm clock (red)	15	41.67%	8	22.22%
23	Coffee cup	20	55.56%	11	30.56%
24	Cordless phone	15	41.67%	7	19.44%
25	Tuna can	17	47.22%	7	19.44%
Tot.		381		219	
Mean Value		15.24	42.33%	8.76	24.33%

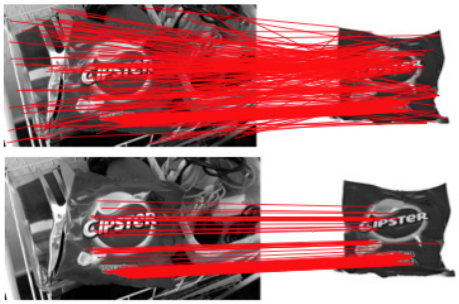

Figure 8: The images show the matches with (lower) and without (upper) RANSAC.

threshold of the match function used was 2 instead of the default 1.5 value. The difference of the resulting number of points can be seen in fig. 9. The proposed method was tested with 30 different videos. Each video contains one of the known object except five videos that contain unknown objects. Query videos have an average length of 4 seconds and the first step of the method is performed with a uniform frame sampling rate fixing N (the number of the selected frames per video) at 4 (so approximately one frame for second). In fig. 10 best match number is shown with relationship to the number of experiments (step 3). In step 3 the selection of an appropriate threshold (10) is performed by statistical analysis of the correct match. The chart in fig. 10 shows that the best matches, for each object, are distributed into two major groups. In tab.2 recognition correctness results are shown for each test video query, including the original id and name for the present object (or NO OBJ# for unknown object). Total recognition performance is shown in tab. 3, with an average precision of the system of 83%. The number of matches performed is 291, so only 24% of the full dataset dimension of 900. In fig. 8 an example of correct recognition is shown. Fig. 11 shows the matches for an unrecognized object (dancer) and for a correct not recognition of unknown object.

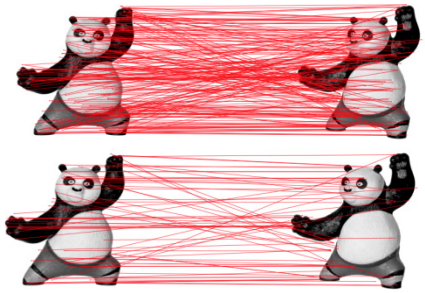

Figure 9: Matching results with different thresholds: 2 (lower) and default value, 1.5 (upper).

Table 2: Video object recognition correctness results.

obj. ID	name	result
1	Dancer	uncorrect
2	Bible	correct
3	Beer	correct
4	Cipster	correct
5	Tour Eiffel	correct
6	Energy Drink	correct
7	Paper tissue	correct
8	Digital camera	correct
9	iPhone	correct
10	Statue of Liberty	correct
11	Motorcycle	correct
12	Nutella	correct
13	Sunglasses	uncorrect
14	Watch	correct
15	Panda	correct
16	Cactus	uncorrect
17	Plastic plant	uncorrect
18	Bottle of perfume	correct
19	Shaving foam	correct
20	Canned meat	correct
21	Alarm clock (black)	correct
22	Alarm clock (red)	correct
23	Coffee cup	uncorrect
24	Cordless phone	correct
25	Tuna can	correct
	NO OBJ1	correct
	NO OBJ2	correct
	NO OBJ3	correct
	NO OBJ4	correct
	NO OBJ5	correct

Table 3: The precision of video object recognition system.

Testset size	correct	uncorrect	precision
30	25	5	83.33%

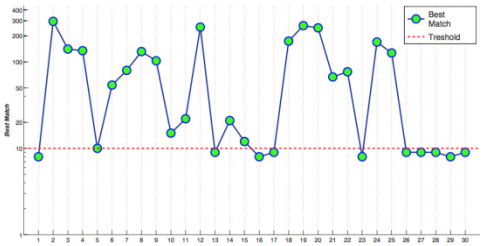

Figure 10: The best matches results for each object and for unknow objects (NO OBJ).

ACKNOWLEDGEMENTS

The authors wish to acknowledge Christian Caruso for helping us in the implementation and experimental phases.

6 CONCLUSIONS AND FUTURE WORKS

In this paper we proposed a new method for video object recognition based on video object models. The results of video object recognition, in terms of accuracy are very encouraging (83%). We created a video dataset of 25 video object, it consists of 360 degree-views of the objects. From the video dataset an image dataset is also constructed by sampling the video frames. It contains 900 views of the 25 objects. Our method for object modeling gives, as result, a compact and complete representation of the objects, it achieves almost 76% data compression of the models. With regard to object recognition method, one of the possible improvement is to refine the selection of the frames for the query in the objects models database. Given a video, the camera motion could be estimated and the frame samples extracted according to motion, for example trying to get a frame every fixed angular displacing. Best results should be reached using a sampling rate that approximate the rate used in the dataset creation. If the video is long enough to have a high number of selected frames, the same modeling process could be used in the query to increase time performance of the recognition, preserving the accuracy taking only the most relevant views.

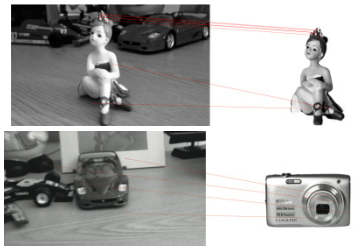

Figure 11: Two examples of results: a false negative (the dancer) and a true negative (unknown object).

REFERENCES

Li, Z. N., Zaiane, O. R., Tauber, Z., 1999. Illumination Invariance and Object Model Content-Based Image and Video Retrieval. *In Journal of Visual Communication and Image Representation*, vol 10, pp 219-224.

Z. Li and B. Yan., 1996 Recognition Kernel for content-based search. *In Proc. IEEE Conf. on Systems, Man, and Cybernetics*, pages 472-477.

Day, Y. F., Dagtas, S., Iino, M., Khokhar, A., Ghafoor, A., 1995. Object-oriented conceptual modeling of video data. *In Proceedings of the Eleventh International Conference on Data Engineering*.

Chen, L., Ozsu, M. T., 2002. Modeling of video objects in a video databases. *In Proceedings of IEEE International Conference on Multimedia and Expo*.

Sivic, J., Zisserman, A., 2006. Video Google: Efficient visual search of videos. In Toward Category-Level Object Recognition, pp. 127-144, Springer.

Vedaldi, A., Fulkerson, B., 2010. VLFeat: An open and portable library of computer vision algorithms. *In Proceedings of the International Conference on Multimedia*.

Kavitha, G., Chandra, M. D., Shanmugan, J., 2007. Video Object Extraction Using Model Matching Technique: A Novel Approach. In *14th IWSSIP, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services,* pp. 118-121.

Mundy, Joseph L. 2006. Object recognition in the geometric era: A retrospective. Toward category-level object recognition. pp.3-28.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints, *In International Journal of Computer Vision n. 60 vol.2 pp. 91-110*, Springer.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. In *Journal of cognitive neuroscience* vol.3, n.1, pp. 71-86, MIT press.

Zhao, L. W., Luo, S. W., Liao, L. Z., 2004. 3D object recognition and pose estimation using kernel PCA. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics.*

Wang, X. Z., Zhang, S. F., Li, J., 2007. View-based 3D object recognition using wavelet multiscale singular-value decomposition and support vector machine. In *ICWAPR.*

Pontil, M., Verri, A., 1998. Support vector machines for 3D object recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.*20 n.*6, pp. 637-646.

Murase, H., Nayar, S. K., 1995. Visual learning and recognition of 3-D objects from appearance. In *International journal of computer vision*, vol.*14 n.*1, pp. 5-24. Springer.

Lowe, D. G., 1999. Object recognition from local scale-invariant features. In . *The proceedings of the seventh IEEE international conference on Computer vision.*

Chang, P., Krumm, J., 1999. Object recognition with color cooccurrence histograms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Wu, Y. J., Wang, X. M., Shang, F. H., 2011. Study on 3D Object Recognition Based on KPCA-SVM. In *International Conference on Information and Intelligent Computing,* vol.18 pp. 55-60. IACSIT Press, Singapore.

Fischler, Martin A and Bolles, Robert C.,1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography *In Communications of the ACM*, vol. 24, num.6, pp. 381–395.

Kovesi, P., 2003. MATLAB and Octave Functions for Computer Vision and Image Processing. [online] Available at: <http://www.csse.uwa.edu.au/~pk> [Accessed September 2013]

Jinda-Apiraksa, A., Vonikakis, V., Winkler, S., 2013. California-ND: An annotated dataset for near-duplicate detection in personal photo collections. *In Proceedings of 5th International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt, Austria.

CVIPLab, 2013. Computer Vision & Image Processing Lab, Università degli studi di Palermo Available at: <https://www.dropbox.com/sh/sqkq03tsembdu4m/N1mCVCFxGQ>

Dong, W., Wang, Z., Charikar, M., Li, K., 2012. High-confidence near-duplicate image detection. *In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*.

Chau, D. P., Bremond, F., Thonnat, M., 2013. Object Tracking in Videos: Approaches and Issues. arXiv preprint arXiv:1304.5212.