

Spatio-temporal constraints for matching view-based descriptions of 3D objects

E. Delponte N. Noceti F. Odone A. Verri
DISI - Università degli Studi di Genova, Italy

{delponte, noceti, odone, verri}@disi.unige.it

Abstract

We propose a 3D object recognition method based on first extracting a compact description of image sequences, and then matching these descriptions with a two-steps strategy. The compactness of the description, made of a set of time-invariant local features, allows us to use a simple nearest neighbour matching to obtain an initial set of recognition hypotheses; spatio-temporal constraints help us to confirm or to reject these hypotheses. We report an extensive experimental analysis to assess our method against changes in illumination, abrupt scale changes, occlusions, clutter, view-point variations. Since our approach is based on the use of image sequences we also tested it with different camera motion. Our description strategy may be also applied to relatively simple poor-textured objects, since few stable features are enough for recognition.

1 Introduction

View-based 3D object recognition gained attention in the last decades as it allows to design relatively simple recognition systems without an explicit computation of the objects 3D models. View-based methods often start by finding a set of appropriate observation viewpoints for the objects of interest. Instead we believe that, in the appropriate application setting, observing an object from slightly different viewpoints, and looking for features that are distinctive in space, stable and smooth in time, can greatly help recognition. To do so, we start from an image sequence of the object and find a compressed description of it based on first extracting local keypoints and tracking them over the sequence, and then combining this information to obtain time-invariant features that will represent all the information we retain about the object. The same compressed description is extracted, at run time, from test sequences. Object localization is performed through a matching procedure that exploits the time-invariant nature of our object descriptors to emphasize matching between groups of spatially close and coeval features.

The popularity of local keypoints [1, 6, 8, 9, 3] is due to the fact that they produce compact descriptions of the image content and do not suffer from the presence of cluttered background and occlusions. The main problem with local methods is that, while observing minute details, the overall object's appearance may be lost; also, small details are more likely to be common to many different objects [10]. This may increase the number of false positives in matching and recognition. For this reason local information is usually summarized in global descriptions of the object, for instance in codebooks [1, 6]. Alternatively, closeness constraints can be used to increase the quality of matches [3]. Temporal continuity may also help to cluster related keypoints observed at different view-points during the training phase, in the case a dense sequence of the object of interest is available [4]. Extracting temporal information to the purpose of object recognition is not common, instead it is a natural choice in other recognition tasks, such as gesture recognition. Local spatio-temporal features for action recognition are described in [5], based on finding sharp changes in the temporal direction that are distinctive of a given motion. Our time-invariant features are different since they are designed to recognize appearance and not dynamics.

The state of the art on feature matching is very rich (see, for instance [2] and references therein) and it is often based on finding compromises between performance and efficiency. On this respect, the major contribution of this paper is a two-steps matching procedure that exploits the richness of our temporal features to achieve such a compromise. We first obtain a set of hypotheses on the presence of a given object on the test sequence using a simple nearest neighbour matching. Then, we refine this hypothesis using spatio-temporal constraints: time constraints allow us to focus on the more appropriate view-point range, discarding all information contained in the model that is not useful for the current match; spatial constraints expand the search area and help us to confirm or reject each hypothesis. Our modeling and matching method exploits temporal coherence *both in training and test*. For this reason it fits naturally in the video analysis framework. In spite of the redundancy of the data that we use, the descriptions we ob-

tain are compact since they only keep information which is very distinctive across time. We report very good results on test sequences of increasing difficulty, in the presence of clutter, illumination and scene changes, occlusions, similar objects.

The paper is organized as follows. Section 2 describes the time-invariant local features that are at the basis of our modeling; Section 3 reports the procedure we follow to build a model for an image sequence and to match to image sequences; Section 4 shows experimental results, while Section 5 is left to final remarks.

2 Time-invariant features

Feature extraction: Given an image sequence, we detect corners in scale-space [7] and estimate their principal direction [8]. We then track the corners over the sequence with a Kalman filter — the unknown state of the system is $x_k = \{p_k, s_k, d_k\}$, where p_k is the keypoint position at time k , s_k its scale, d_k its principal direction.

Cleaning procedure: At the end of this extraction procedure, short trajectories are discarded. Keypoints of the trajectories are first represented as SIFT [8]. We then apply a cleaning procedure that eliminates noisy trajectories: first, for each trajectory we compute the scale and the principal direction variance. Then, trajectories with high variances are further analysed in order to check whether they contain abrupt changes that could be caused by tracking errors. To do so, we perform a SSD correlation test between the first gray-level patch of the trajectory and the subsequent ones. In the presence of an abrupt change, the trajectory is discarded. After the cleaning procedure, a further processing is performed: while on the feature extraction phase, features robust to viewpoint variations are preferred, on the description phase, long trajectories may contain too much information since they are the result of observing a feature on a long period of time (and possibly a high range of views). Descriptions generated from these long trajectories tend to oversmooth the appearance information, and may lead to a high number of false positives. To this purpose we apply a *cutting phase* that cuts a trajectory into many sub-trajectories of length N . The choice of N is not crucial, and common sense rules may be applied. In our experiments, unless otherwise stated, N is set to 10.

Description: Finally, a time-invariant feature is described by (a) a spatial appearance descriptor, that is the average of all SIFT vectors of its trajectory, and (b) a temporal descriptor, that contains information on when the feature first appeared in the sequence and on when it was last observed. To increase the stability of the appearance descriptor we discard the keypoints lying outside an hyperball of fixed radius centered at the keypoints centroid.

3 Building and matching models

Building a model for a sequence: The content of an image sequence is redundant both in space and time. We obtain compressed descriptions for the purpose of recognition extracting a collection of time-invariant features over the sequence, that we call a *model* of the sequence, and discarding all the other information. We do not keep any information on the relative motion between the camera and the object, as it is not informative for the recognition purpose. In the training phase the sequence model is also a *model for the object*.

The space occupancy of our model is limited: considering that the keypoints detected in each frame are approximately 300. Thus, if every keypoint of the sequence participates to build the model, in the average case (a sequence of length 250) the model is made of approximately 75000 features. Our approach grants to obtain a compact representation of typically 1500 time-invariant features.

Matching two sequences: Let us consider a test sequence represented by a model T , and a training sequence, represented by an object model M . The problem we address is to see whether the test model T contains the object M . Our two-steps matching strategy is performed as follows: on the first stage a nearest-neighbour between each training model and the test model gives us initial hypotheses for the presence of known objects in the sequence. The second stage, based on enhancing spatio-temporally coherent matches, helps us to confirm or to reject each hypothesis. On the first stage, for each feature in M , we use histogram intersection to check whether T contains a similar feature. We set a minimum similarity threshold (usually equal to 0.6) to obtain a collection of robust matches (f_M, f_T) .

On the second stage, we use spatio-temporal constraints and perform a reverse matching procedure from the test to the training model. First, we detect the subsets of training and test models containing most matches: I_M and I_T . In this way we get a raw temporal localization of the object on the test, but also hints about its appearance, since I_M marks a particular field of view. In order to refine this detection, we reject matches which are spatially isolated.

Second, we try to increase the number of matches around the matches obtained previously. Let us consider a match (f_M, f_T) : we can specify two sets, F_M and F_T , containing features (of M and T respectively) appearing together with f_M and f_T for a chosen period of time (in our experiments 5 frames). A new matching step is then performed between F_M and F_T considering a lower similarity threshold: a pair $(\tilde{f}_M, \tilde{f}_T)$ is a new match if the features are spatially close to f_M and f_T , respectively. This new stage is repeated for each pair (f_M, f_T) belonging to I_M and I_T intervals.

The incorrect recognition hypotheses are rejected while the correct ones are enriched by further matches in the sur-



Figure 1. The 4 objects modeled in our experiments. From left: dewey, book, winnie and goofy.

rounding areas. This procedure increases the number of high scores when groups of features appear both in training and test. This is more likely to happen if training and test models contain features coming from the same objects. Experiments will show that this two-stages strategy is especially useful when the acquisition setting is very different between training and test. Thanks to the compact representation of visual information the matching phase is efficient even when the sequence is long and there are several detected keypoints.

4 Experiments

In this section we report some experiments carried out on different video sequences acquired under various conditions. We trained the system to model 4 objects: 2 planar objects (a book and a ring binder), 2 complex 3D objects with many concavities (Figure 1). Each object is represented by a training sequence of 250 frames acquired by a still camera observing the object on a turntable. Each training model has about 1000 entries.

Preliminary tests confirmed that our modeling and matching techniques do not suffer from scale and illumination changes and clutter (see Fig. 2). Table 1 shows the results of comparing training models with test models acquired in a different room, with a different illumination. Also, training and test sequences are acquired in different days and daytime. The table shows how, using our matching strategy, noticeably increases the matches: columns M1 report nearest neighbour matches, while columns M2 are obtained with our matching strategy. Table 2 presents results obtained while checking the method’s robustness when the camera moves of a more complex motion, in this case it is hand-held by the user: since the ego-motion is shaky, features are more difficult to track, and, as a result, trajectories tend to be shorter and more instable. Again, columns M1 show the hits obtained by simple nearest neighbour appearance matching, while columns M2 show how the results have been increased with our matching strategy. The advantage of this approach is apparent, our matching increases the number of positive matches, while limiting the false positive matches. Finally, we apply our method to the

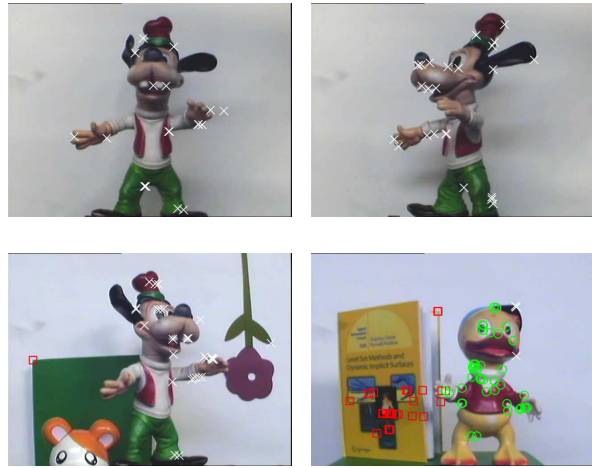


Figure 2. Matching results in the presence of illumination and scale changes, clutter, multiple objects. Circles: dewey’s features; squares: book, crosses: winnie; Xs: goofy.

	Book		Goofy		Dewey		Winnie	
	M1	M2	M1	M2	M1	M2	M1	M2
Book	85	97	4	0	11	7	8	1
Goofy	5	1	80	97	23	0	3	0
Dewey	3	1	11	2	63	93	2	0
Winnie	7	1	5	1	3	0	87	99

Table 1. Hit percentages between training and test videos acquired in different conditions with and without the use of spatio-temporal constraints (M2 and M1 respectively).

	Book		Goofy		Dewey		Winnie	
	M1	M2	M1	M2	M1	M2	M1	M2
Book	42	69	14	19	11	0	9	3
Goofy	28	7	52	81	29	0	21	17
Dewey	19	17	17	0	41	100	12	5
Winnie	11	7	17	0	19	0	58	75

Table 2. Hit percentages between training and test videos acquired with a hand-held camcorder, with and without spatio-temporal constraints (M2 and M1 respectively).

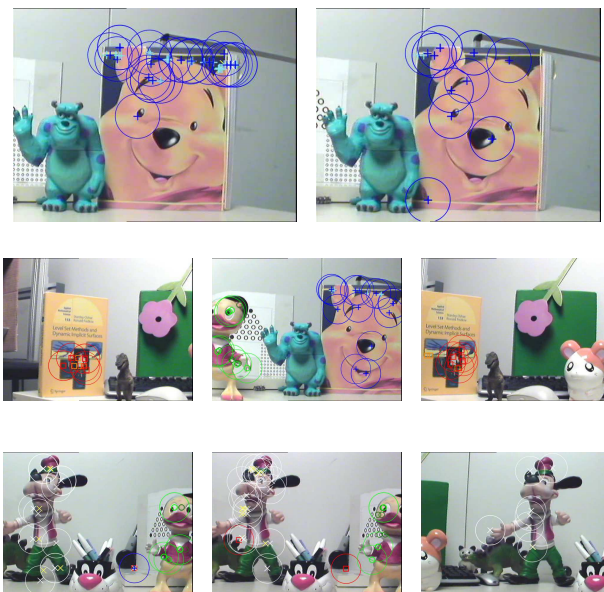


Figure 3. Matches on sample frames. Green circles: dewey. Red squares: book. Blue crosses:winnie. White Xs: goofy.

localization of objects placed in very complex scenes, containing many other objects of the same sort. The motion of the camera is almost rectilinear, displaying the objects as in a *tracking shot*. We collected 4 long videos about 500 frames each. The videos have been acquired at a variable speed, slowing the camera down in the presence of possible objects of interest. When the speed is higher, the video quality decreases and very few time-invariant features are detected. The objective of these experiments is to test the videos against the presence of our 4 objects of interest. Figure 3 shows matching examples on sample frames from our video shots. Matches are kept when the similarity value is higher than 0.6. White Xs (goofy), blue crosses (winnie), red squares (book) and green circles (dewey) represent the features matching the objects models with temporal constraints. The big circles indicate the regions where the spatial constraints are explored: yellow Xs, light blue crosses, orange squares, and light green circles are the matches obtained expanding the search on the above mentioned regions.

5 Conclusions

This paper proposed a 3D object recognition method that uses image sequences to model the objects appearance and suggested a matching strategy that exploits spatio-temporal constraints to improve the quality of matches and increase

localization performances. The time-invariant local features that we proposed are based on tracking keypoints over the sequence, keeping only stable trajectories. Thanks to our approach the overall appearance of 3D objects can be represented with a relatively small set of features. The compactness of the description allows us to use a simple nearest neighbour matching to obtain an initial set of hypothesis; spatio-temporal constraints help us to confirm or to reject these hypotheses. We performed an extensive experimental analysis to assess our method against changes in illumination, abrupt scale changes, occlusions, clutter, view-point variations. Since our approach is based on the use of image sequences we also tested it with different camera motion. The results we obtained confirm the appropriateness of the descriptions and the effectiveness of our matching strategy to perform recognition. We are currently working on an online version of the method to perform live object recognition.

References

- [1] G. Csurka, C. Dance, L. Fan, and C. Bray. Visual categorization with bag of keypoints. In *The 8th European Conference on Computer Vision - ECCV*, Prague, 2004.
- [2] E. Delponte, F. Isgrò, F. Odone, and A. Verri. Svd-matching using sift features. *Graphical models*, 68(5):415–531, 2006.
- [3] V. Ferrari, T. Tuytelaars, and L. V. Gool. Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, 67(2), 2006.
- [4] M. Grabner and H. Bischof. Object recognition based on local feature trajectories. In *I Cogn. Vision Work.*, 2005.
- [5] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.
- [6] B. Leibe, K. Mikolajczyk, and B. Schiele. Efficient clustering and matching for object class recognition. In *British Machine Vision Conference*, 2006.
- [7] T. Lindeberg. Feature detection with automatic scale selection. Technical report, CVAP, Department of numerical analysis and computing science, 1998.
- [8] D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, Corfú, Greece, 1999.
- [9] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66(3), 2006.
- [10] A. Torralba, K. Murphy, and W. Freeman. Sharing visual features for multiclass and multiview object detection. *Trans. on PAMI*. in press.