Neurocomputing 171 (2016) 982–990

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A comparative study of video-based object recognition from an egocentric viewpoint

Mang Shao*, Danhang Tang, Yang Liu, Tae-Kyun Kim

Department of Electrical and Electronic Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form
2 July 2015
Accepted 13 July 2015
Communicated by Liang Lin.
Available online 31 July 2015

Keywords:
Object instance recognition
Egocentric video
Comparative study

ABSTRACT

Videos tend to yield a more complete description of their content than individual images. And egocentric vision often provides a more controllable and practical perspective for capturing useful information. In this study, we presented new insights into different object recognition methods for video-based rigid object instance recognition. In order to better exploit egocentric videos as training and query sources, diverse state-of-the-art techniques were categorised, extended and evaluated empirically using a newly collected video dataset, which consists of complex sculptures in clutter scenes. In particular, we investigated how to utilise the geometric and temporal cues provided by egocentric video sequences to improve the performance of object recognition. Based on the experimental results, we analysed the pros and cons of these methods and reached the following conclusions. For geometric cues, the 3D object structure learnt from a training video dataset improves the average video classification performance dramatically. By contrast, for temporal cues, tracking visual fixation among video sequences has little impact on the accuracy, but significantly reduces the memory consumption by obtaining a better signal-to-noise ratio for the feature points detected in the query frames. Furthermore, we proposed a method that integrated these two important cues to exploit the advantages of both.

1. Introduction

Video-based object recognition (VbOR) methods have emerged during the last decade, but attracted less attention than image-based methods. Its motivation, however, has been further established by a recent research [1], by revealing that how human brain can effortlessly interpret a multitude of objects with different identity-preserving transformations. After exposing a monkey's visual system to an artificial visual world without temporal contiguity, neuroscientists observed that inferior temporal cortex neurons began to lose their capacity for being transformation invariant. This strongly encourages the exploitation of temporal information in object recognition tasks. For instance, some recent studies attempted this by using learned trajectory descriptors [2,3] or viewpoint invariant features [4,5] during visual fixation in video clips from different aspects.

On the other hand, the exploitation of spatial cues, either in 2D image layouts [6] or 3D object structures [7], is a flourishing branch of object recognition. The viewpoint-invariant theorem [8] states that the essential component of object recognition, regardless of viewing conditions, is structural information. Encoding object structural information requires only a small amount of memory, yet it is capable of producing a multitude of object representations via their interrelations and mental rotations. In the field of computer vision, stereo vision is often utilised to obtain precise depth perception, and hence 3D structure. On top of that, some recent studies have obtained impressive performance by using multi-view images to reconstruct 3D information to support object recognition [9], semantic segmentation [10] and pose estimation tasks [11,7]. Recently, due to the growing use of wearable vision devices, e.g., *Google Glass*, research into egocentric videos has attracted more and more attention. As one of the useful source of spatial information, egocentric vision has the advantages of being controllable during capturing informative viewpoints and being more practical than turntable settings.

In this comparative study, our goal is to explore the potential usage of egocentric videos for training and as query sources for the recognition of rigid 3D objects in realistic scenes. In particular, we aim to exploit the temporal and spatial cues provided by egocentric videos and to answer the following questions. Are they helpful? If so, are they helpful in terms of accuracy or efficiency? Can they be combined? It is worth noting that there have been

* Corresponding author. Tel.: +447703729830.
E-mail address: ms2308@ic.ac.uk (Shao).

http://dx.doi.org/10.1016/j.neucom.2015.07.023
0925-2312/

recent advances in object *category* recognition [12–14], but only a small number of studies have investigated the problems of *instance* object recognition [15], particularly in egocentric videos [16,17]. Therefore we highlight our contributions as below:

- We captured a *Sculptures in Victoria and Albert (V&A) Museum* dataset from an egocentric viewpoint.
- We categorised and compared diverse state-of-the-art object recognition frameworks and their video-based extensions.
- We proposed a hybrid solution that combines the advantages of both temporal and spatial cues.

2. Methods

Given exemplar videos of target objects, the purpose of VbOR is to identify them in query videos. Due to the egocentric setting in our study, each video captured multiple views of only one target object that appeared roughly in the centre. Therefore, the whole video was assigned and recognised with one label. In this comparative study, we focused on the methods represented by the taxonomy shown in Fig. 1. In terms of utilising spatial information, these methods can be categorised mainly into 2D and 3D approaches. Among the 2D approaches, there are three different ways to represent videos: image-based, set-based and video-based. In image-based methods, each video is treated as independent images, where a straightforward combination of individual results is applied to obtain the final output of the video. In set-based methods, each video is treated as a set of unordered images with underlying mathematical structure, such as a manifold. In video-based approaches, each video is represented as a set of ordered images, i.e., with temporal information. By contrast, 3D-based VbOR utilises reconstructed 3D information from multi-view images. This is a relatively new area with only a small set of methods. Thus we consider these methods as a separate category. In the following subsections, we analyse the pros and cons of each framework. Comparative evaluation can be found in Section 5.

2.1. Image-based methods

To select representative image-based methods, we adopted three baselines from state-of-the-art object recognition frameworks based on their image classification techniques, i.e., (a) point-to-point (P2P), (b) image-to-image (I2I) and (c) point-to-class (P2C), as illustrated in Fig. 2. The image classification results are combined later via voting.

Point-to-point methods measure the similarity between two images based on their corresponding local image appearance, which is usually encoded by a feature descriptor.

In the seminal paper by Lowe [18], image classification was performed by matching a set of keypoints detected in image regions. Using robust fitting algorithms, e.g., *RANSAC* [19], the correspondences can be constrained further by dominant transformation between the matched pairs. This technique can improve the recognition precision significantly. But it may fail when there is no similar viewpoint in the database to a query image. Recent advances in graph matching [20,21] have relaxed the geometric constraint between point correspondences for articulated or deformable object recognition. However, these methods are generally computationally expensive and infeasible for large-scale problems.

Image-to-image methods compute the vector of visual word frequencies in images to facilitate similarity measurement. In general, I2I methods are efficient and suitable for large-scale problems because of the compactness of their image representations. The Euclidean distance in a feature space reflects the similarity between features. Thus we can also apply learning-based classifiers, e.g., linear support vector machine (SVM) and Random Forests, to facilitate a better generalisability and efficient recognition. I2I methods have been applied widely to various image classification tasks, e.g., scene recognition [6], image categorisation [22,23], object recognition [24] and video image retrieval [25]. These methods have achieved state-of-the-art performance on most publicly available benchmark datasets of image classification. [12,13]. However, despite the success of these methods, the vector quantisation process may degrade the discriminatory power of individual image features, which is crucial for instance recognition problems.

Point-to-class (P2C) methods have also achieved impressive results on several benchmark datasets in recent years. The concept was emphasised in [26] to sidestep the negative effects of vector quantisation in I2I methods, and later improved and extended in [27,28]. The basic idea is to directly measure the similarity between query features and training features in every object class without vector quantisation. Compared with P2P methods, P2C has better generalisability, because images are decomposed into image features that can be matched simultaneously across all training images. This approach is also suitable for large-scale problems because the feature-matching procedure can be accelerated to real-time using approximated nearest neighbour algorithms. The main drawback is that P2C methods are based on non-parametric classifiers and consequently consume more memory because all the features are retained.

2.2. Set-based methods

Set-based methods aim to capture the inherent characteristics of a set based on the assumption that the members of the set follow a particular statistical distribution, as shown in Fig. 3. For

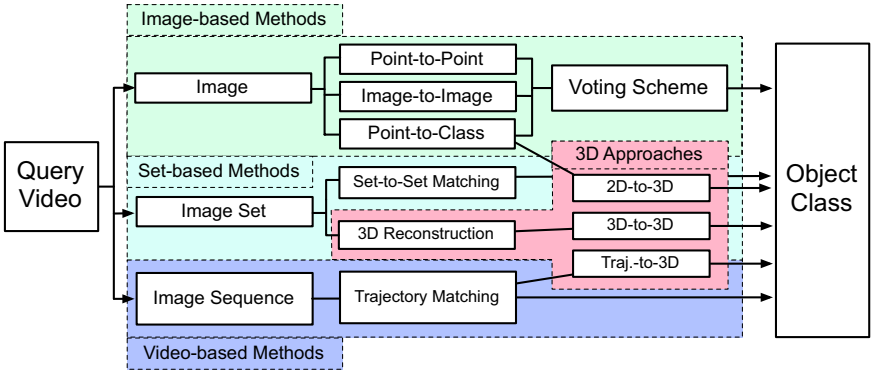


Fig. 1. Method categorisation and experimental setup.

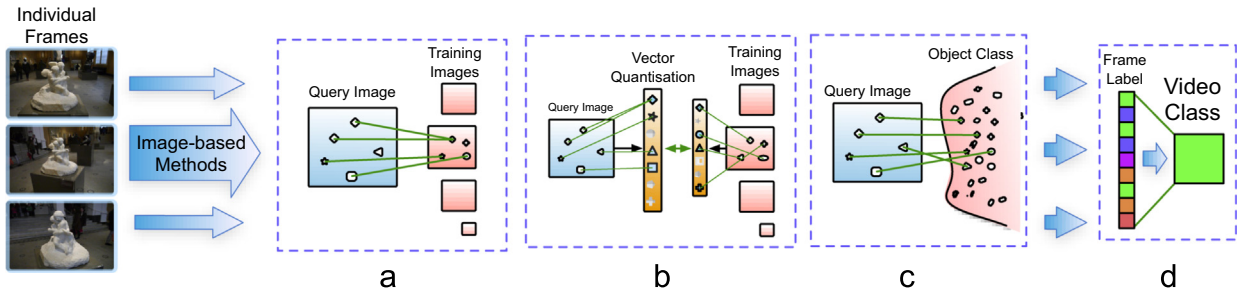


Fig. 2. Image-based methods selected from three state-of-the-art object recognition frameworks.

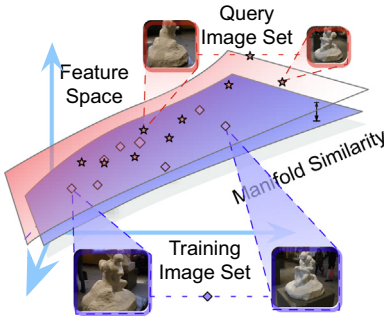


Fig. 3. Toy example of set-based methods.

the VbOR problem, the appearance of an object in each frame is constrained by identity-preserving image variations, i.e., viewpoint, scale or illumination changes. If we consider that an image is a data point in a high-dimensional space, the manifold that is spanned by the image variations can be learned using subspace or manifold techniques [29,30]. The video-to-video similarity can then be estimated based on their manifold intersection, e.g., their largest principal angle. In previous studies, these techniques have achieved superior performance in different tasks, such as face recognition [31], head pose estimation [32] and object pose estimation [29].

The motivation for applying set-based techniques to object recognition problems is that the unseen views of an object can be interpolated from existing images, which leads to significant improvements in generalisability. However, in dynamic real-world scenarios, estimating the subspace or manifold from an image set is always challenging as the distribution of images in a set is often highly non-linear due to the existence of complex background noises and object variations.

2.3. *Video-based methods*

Video-based methods exploit the temporal coherence between adjacent frames in the video. For the VbOR problem, temporal coherence can be used to learn better representations from videos based on feature tracking [3], or to remove unstable local features [25]. In addition, applying video-based techniques may facilitate learning the variation among object parts, improving the signal-to-noise ratio, or compressing the representation of video data. An example is shown in Fig. 4 where the trajectories can be extracted via tracking and later used for matching.

Extracting spatio-temporal coherence is important in many tasks, such as action recognition, video surveillance and object tracking. However, it is not trivial to do so in an egocentric setting, since the camera can move in an arbitrary manner and the time-ordering does not reflect any characteristics of the object's identity. Greater computational power is also consumed due to the additional tracking process.

2.4. *3D-based methods*

A different approach is to utilise 3D geometric cues, earlier works such as [33] requires hand-crafted 3D CAD models as input, whilst in this study we focus on more recent methods that reconstruct models from multi-view images, as shown in Fig. 5. 3D-based object recognition is a relatively new yet attractive research field, which has been popularised by the emergence of low-cost depth cameras. In the case of rigid objects, 3D geometry can be treated as one of the most nuisance-invariant cues that can be obtained from video. In the literature, [34] provides a comprehensive survey of how 3D CAD models can be used in content-based retrieval systems. Several recent studies, including object recognition [35], landmark recognition [36] and camera pose estimation [37], have exploited 3D models reconstructed by photogrammetric methods, such as stereo matching [38] and structure-from-motion [39].

However, the use of 3D object models for object recognition has several limitations. Photogrammetric methods require camera calibration to retrieve the absolute scale and location of an object, and the 3D point cloud generated is generally sparse, which requires more computation. Moreover, the object is often required to be static in the scene.

3. Implementation

As illustrated in Fig. 1, we implemented baseline methods and extended them by adapting geometric and temporal validation techniques. Although there exist frameworks for invariances of image features and receptive field responses under more general classes of visual transformations [40,41], in this paper we restrict ourselves to scale invariance as implemented in standard SIFT [18] in all experiments.

Image-based methods: We adopted the framework proposed in [18] as the baseline for the P2P approach, the standard bag-of-words approach [42] for I2I, and *NBNN* [26] for P2C. In P2P and P2C, geometric validation was achieved by strictly applying *RANSAC* [19] to correspondences based on a perspective transformation. In I2I, we employed Spatial Pyramid Matching (*SPM*) with a uniformed grid (as in [6]), and spatial consistency using *Video Google* [25]. Additionally, extensions for each image-based method are implemented. In I2I, we replaced the clustering method (k-means) by sparse coding with max pooling, as described in [22], to reduce the error from vector quantisation. Apart from that, better distance functions were employed from [25]. In P2C, we implemented *Local-NBNN* [43] as a state-of-the-art version of *NBNN*.

Set-based methods: A kernel approach, *Kernel Principle Angles* (*KPA*) [30], is implemented to compute the principal angles in the feature space as the basis of the manifold-based method. We collected bag-of-words representations of video frames as the feature set based on the assumption that images in a sequence are

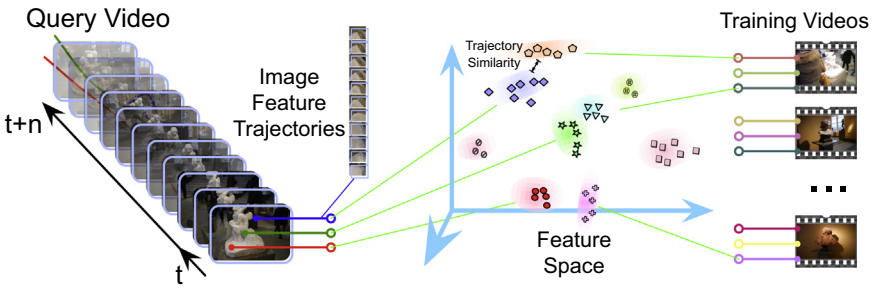


Fig. 4. A toy example of trajectory matching methods based on feature tracking.

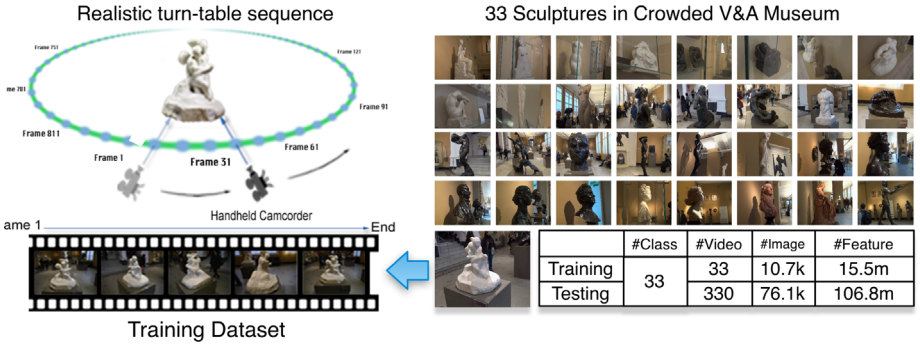


Fig. 5. Toy example of 3D-based methods where each video is treated as an unordered image set.

highly correlated so that they lie on a low-dimensional manifold, which spans the variations in the object.

Video-based methods: For video-based methods, we tracked interesting points bidirectionally [44] with *Kanade-Lucas-Tomasi feature tracker* (*KLT*) [45]. Three video-based methods were evaluated in the *Local-NBNN* framework: unstable feature removal (*Filtering*), averaged-trajectory matching (*TM*) and trajectory matching by KPA (*TM+KPA*). Additionally, we added a recent method [17] into comparison. In *Filtering*, only unstable feature points are rejected; in *TM*, each trajectory is encoded into a single feature vector by averaging its feature points; and in *TM+KPA* and Liu et al.'s method [17], *KPA* is applied to obtain trajectory similarity measurements.

3D-based methods: According to the general framework of 2D-to-3D image classification systems described in the literature [35,46,7], we first reconstructed 3D object point cloud models from the training videos using the *VisualSfM* toolkit [47], which is a structure-from-motion based photogrammetric modelling program, where foreground segmentation is used to cleanse the noisy 3D point clouds. As shown in Fig. 5, each 3D point corresponds to a set of image features from video sequences during appearance-based feature matching and 2D-to-3D geometric validation can then be applied to constrain the correspondences to a rigid transformation. We applied *ePnP* [48] to facilitate efficient 2D-to-3D transformation estimation and 3D *RANSAC* for 3D-to-3D estimation, according to [49].

Hybrid methods: Furthermore, we propose a hybrid method combining a video-based and a 3D-based method to incorporate benefits from both. To avoid unnecessary experiments, we chose to combine the best method out of each category, i.e., *Local-NBNN+TM* and 2D-to-3D+ePnP, by empirical results (see Section 5). In practice, each object 3D point cloud and video trajectories corresponds to a set of similar features, which can be encoded by simply averaging the set of feature vectors into a single representation for better *Local-NBNN* performance. After matching video trajectories to 3D object points, *ePnP* validation is performed between the trajectories' coordinates of each frame and object 3D point coordinates. Similar to the 3D-based methods, only the correspondences that pass geometric validation are taken into account for the later voting procedure.

4. Dataset

Several egocentric datasets have been made publicly available [50,51]. However, to the best of our knowledge, there is no suitable benchmark dataset for evaluating complex rigid object recognition methods with egocentric videos. Thus, we constructed a new video dataset with 33 less textured sculptures in cluttered museum scenes, as shown in Fig. 6.

In total, 363 videos (30 fps, 720 × 576 pixels) were captured by amateur users with a hand-held camcorder in a crowded museum.[1] We have collected 33 different sculptures served as object instances, each of which has a training video and 10 testing videos. The training videos were captured at 180 or 360 degrees from azimuth around the sculptures, depending on their positions. For testing videos, we deliberately added different nuisance including extreme views, large scale changes, occlusions, light reflection and temporal object disappearance.

Our video-based dataset has many unique properties compared with standard image-based datasets: (i) high correlation within each video sequence, (ii) high inter-class correlation between classes, and (iii) various types of additional nuisances. The high correlation between images within the video sequence poses several challenges: how to extract useful cues from the correlated video frames (e.g., 3D geometry and temporal coherence), and how to remove redundant information from large video datasets to improve efficiency while maintaining the discriminatory power. In addition, the sculptures were similar in their appearance and physical structure. Therefore the dataset has a high inter-class correlation, which causes difficulties for

[1] The dataset and feature will be available at http://www.mangshao.net/vadataset.

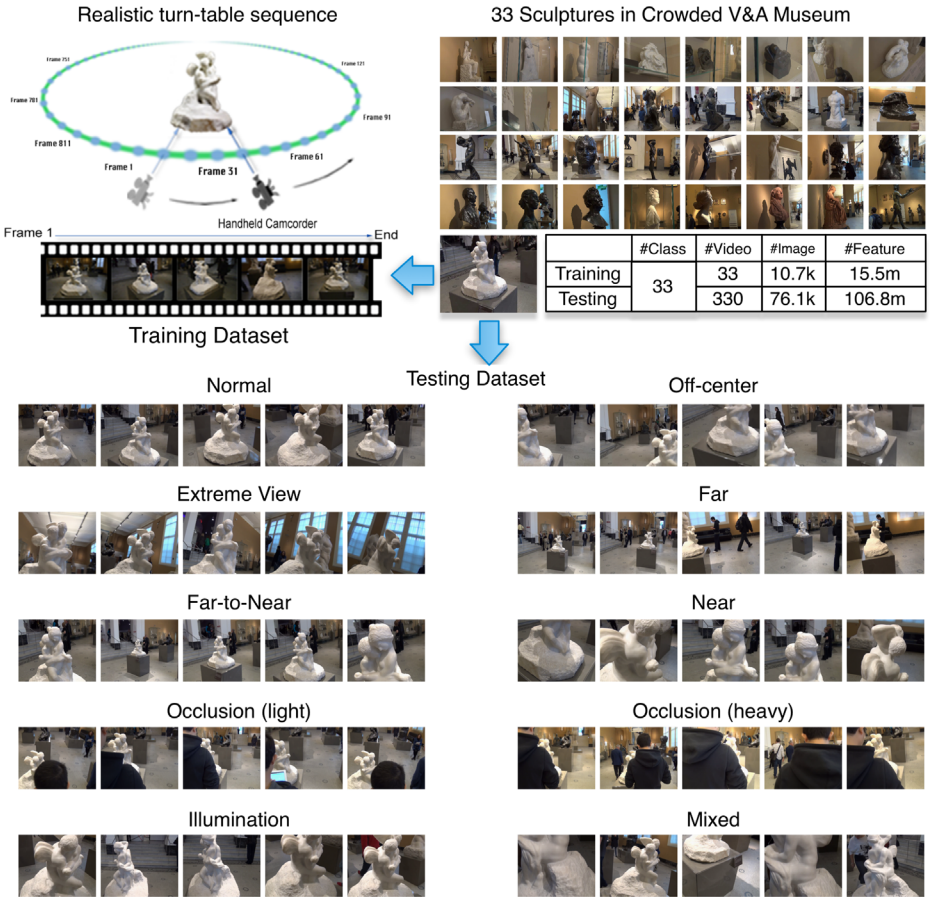


Fig. 6. Illustration of the collected dataset.

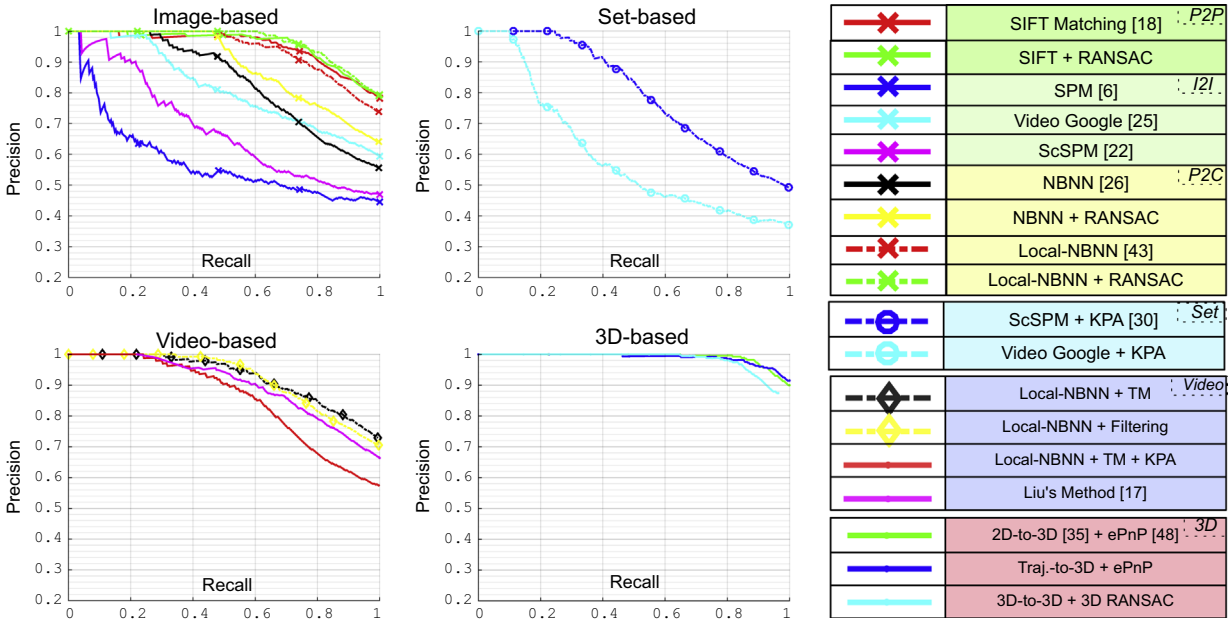


Fig. 7. Full evaluation of the video classification results based on the precision–recall curves (all figures are best viewed in colour), including image-based methods (with voting), set-based methods, video-based methods and 3D-model based methods. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

many linear classifiers, e.g., linear SVM. Furthermore, the most common technique used for object recognition, vector quantisation, would greatly reduce the discriminatory power of features and degrade the recognition accuracy.

5. Evaluation

In Figs. 7 and 8, an overview of all the experiments performed with our dataset is provided. The category of each experiment was

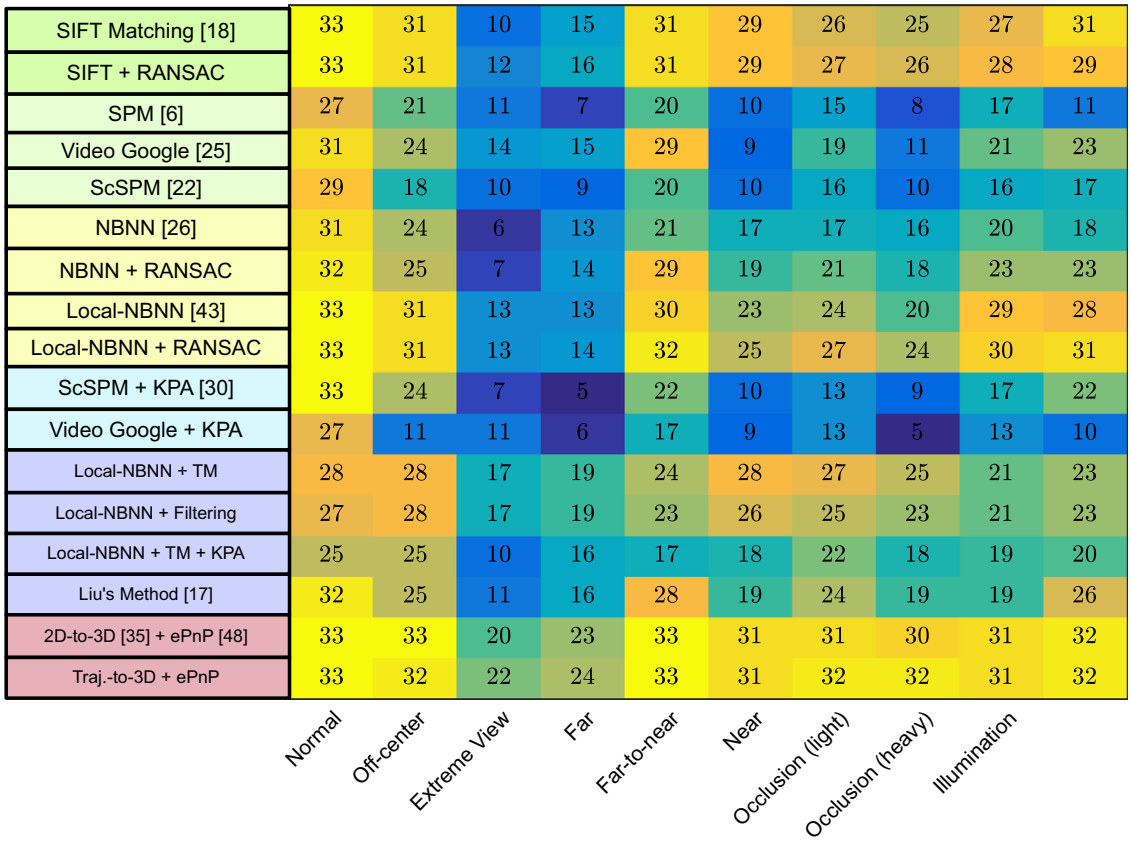


Fig. 8. We divided our *V&A* dataset into 10 subsets by different types of nuisance. Each column represents one subset which contains 33 videos (one per class), whilst each row gives the results of a method. The numbers indicate the amount of successfully classified videos.

determined by the representation of query testing videos. Traditionally, object recognition evaluations report the accuracy as percentage. To better demonstrate the impact of each method, we measured their performance in the form of a precision–recall (PR) curve, which was inspired by the ratio test in [18], since it can be easily generalised to evaluate the performance of image classification. For each query video, the assigned object class was deemed acceptable only if the ratio between the highest and second-highest class probability was above a certain threshold, otherwise it was considered as a false negative. If the highest class was the same as the ground truth, it was considered as a true positive. By testing all possible thresholds, a full PR-curve is obtained. It is worth noting that some of the results obtained with image-based methods contrast with their performance using public datasets. We consider this is mainly due to the aforementioned uniqueness of our dataset.

Image-based methods: Only the final voting accuracy for each video is shown for image-based methods. In general, I2I methods had poor performance due to the quantisation of image descriptors, as described in the literature [26]. The larger (approximately $10k$) visual vocabulary with a better distance function (Bhattacharyya distance) in *Video Google* or *SPM based on sparse coding (ScSPM)* improved the results obtained with *bag-of-words* to some extent, but it was still not as good as other methods.

The P2P and P2C methods achieved similar accuracy and outperformed I2I methods, owing to the following reasons. Firstly, there was no feature quantisation in P2P and P2C and thus no loss of discriminatory power. Secondly, the geometric relationship among features is partially lost in I2I methods, whereas the robust estimation method *RANSAC* in P2P and P2C methods constrains the spatial distribution of image features, which is favourable for rigid object recognition.

Set-based methods: Overall, set-based manifold methods did not have significant impact on the performances of I2I methods. The main difficulty was related to the complex nuisance effects in scenes such as under extreme view or occlusions, since manifold-based methods are generally prone to set complexity and outliers. In addition, *KPA* [30] improves *ScSPM* [22] but it degrades *Video Google* [25]. The results showed that the advantage of applying *KPA* was reduced when the vocabulary size increased. Since the high heterogeneity of image representations caused the failure of KPA when determining dependencies within the set, thereby leading to inaccurate estimates of the subspace from the image set.

Video-based methods: The results of the comparisons between four methods, *Filtering*, *TM*, *TM+KPA*, Liu et al.'s method [17] and their baseline *Local-NBNN*, have shown that: (i) the formation of a trajectory did not improve the recognition accuracy, (ii) averaging the trajectory obtained a similar performance, and (iii) *KPA* was computationally expensive and not suitable for application to trajectory matching.

These contradictory results can be explained as follows: (i) To reduce unstable features and prevent long-term drifting, a bidirectional validation was applied to the trajectories. As shown in Fig. 9(c), approximately 90% of the features used in training and 67% in the testing dataset were filtered. However, the filtering process did not generate extra inliers and the inherent advantage of *Local-NBNN* is its robustness to noisy feature points, which explains their similar accuracy. (ii) Fig. 9(a) shows that the trajectories were short due to the strict spatio-temporal constraint and unstable features increased due to the use of a hand-held camera. This feature is actually favourable to averaged trajectory matching, since the features in most of the trajectories are nearly identical, hence their mean is representative. (iii) However, this also explains the poor performance of the *KPA* approach because short trajectories were far from sufficient to

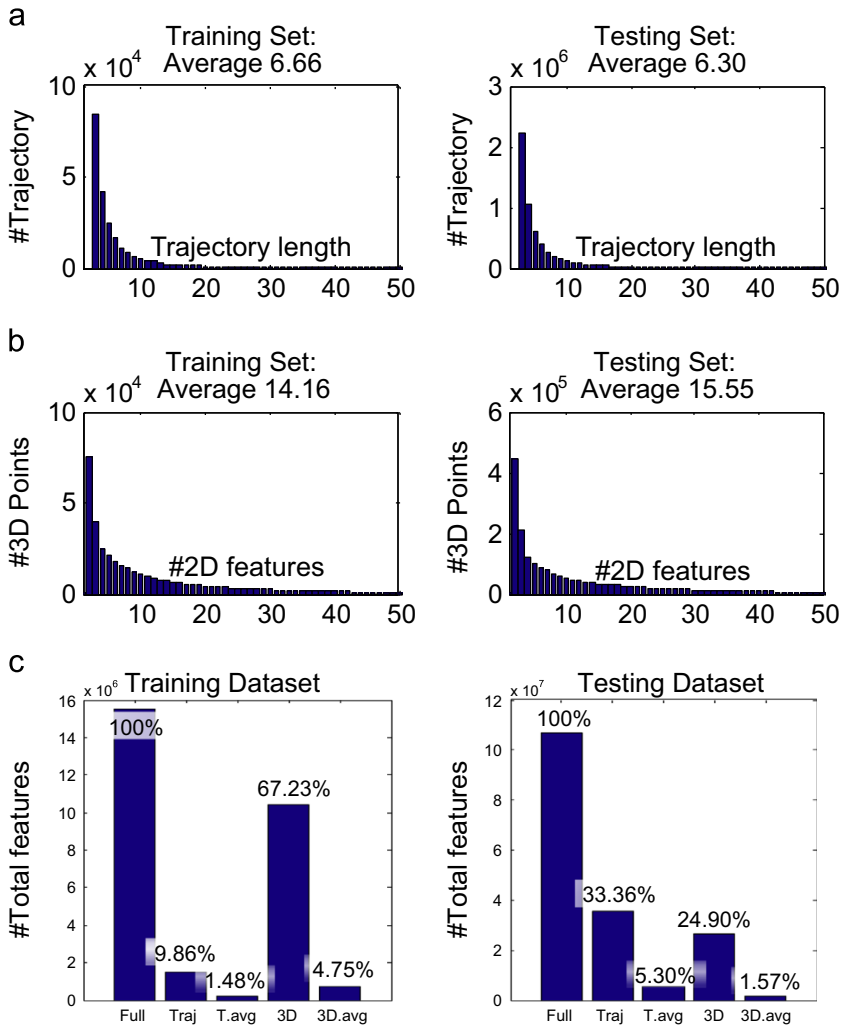


Fig. 9. Histogram showing (a) the number of frames crossed per trajectory, (b) the number of SIFT features allocated per 3D point and (c) the percentage compression of the dataset after forming the trajectory and 3D points.

span the variations in the object parts, thus they lacked discriminatory power. In addition, it should be noted that the application of *KPA* is computationally intense with massive trajectories due to its high complexity.

However, by using trajectory averaging, the dataset was compressed to 1.48% for training and to 5.30% for testing compared with their original sizes, as shown in Fig. 9(c). This reduced the computational power and memory requirements, especially when using non-parametric classifiers.

3D-based methods: Retrieving 3D mesh model of objects from the training video and performing 2D-to-3D geometric validation has increased the recognition performance dramatically for two reasons. (i) In training videos, the reconstruction process rejected features that were not consistent with the object geometry, such as pedestrians or specularities from light reflection, thereby resulting in a large increase in the signal-to-noise ratio of the database. (ii) Fig. 9(b) shows the long-tail distribution of a number of features allocated to 3D points, which indicates that a considerable amount of 3D points contained features that covered a large range of view. The 3D geometry is viewpoint-invariant according to the assumption of the objects rigidity, thus the 2D-to-3D geometric validation strictly constrained the correspondences, which resulted in a higher confidence in the final voting of the

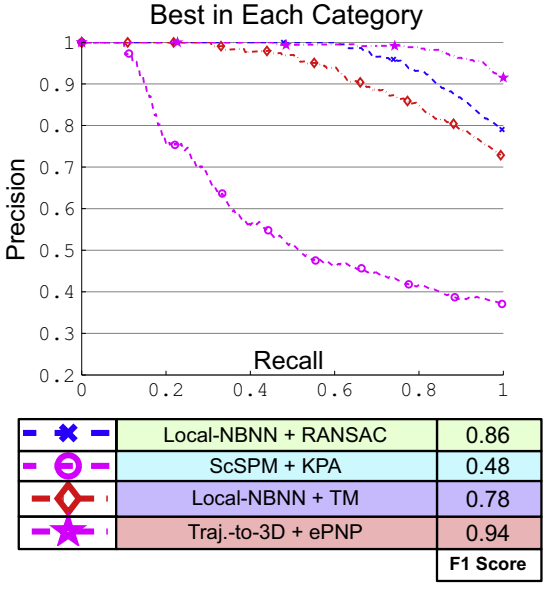
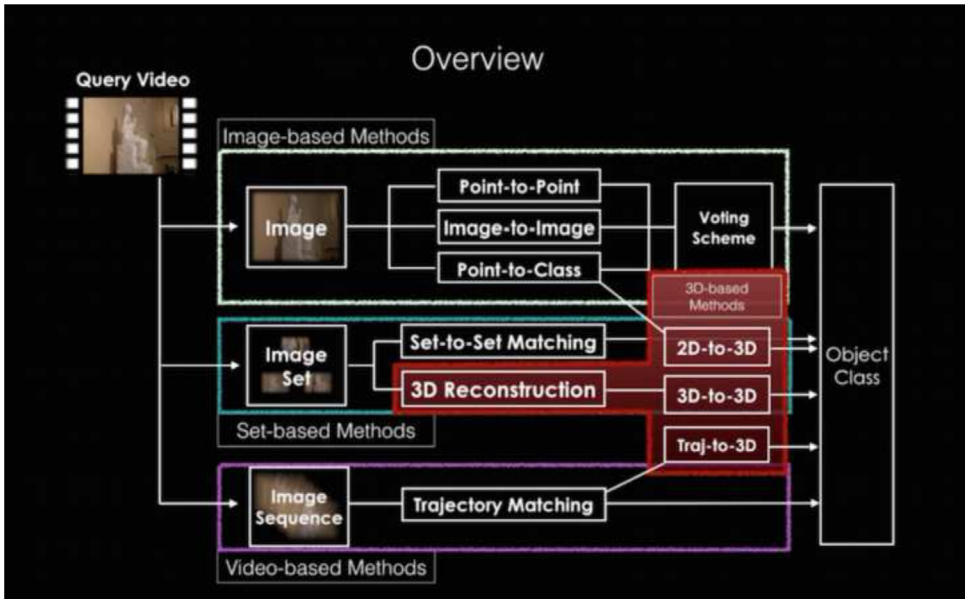


Fig. 10. Precision–recall curves with the hybrid method compared with the best method from each method category.



Video S1. Visualisation of different video-based object recognition methods and dataset examples.A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.neucom.2015.07.023.

object class. This is especially helpful for VbOR because the object class can usually be determined from a few confident frames within the video. Furthermore, the training dataset was compressed to 67.23% after 3D reconstruction and to 4.75% after it was averaged further into a single feature vector, according to Fig. 9(c). 3D-to-3D also achieved good recognition accuracy, but the reconstruction process during the testing stage required too much memory and computational power, which made it inefficient compared with 2D-to-3D methods.

Hybrid methods: The combined method exploited the advantages of 3D-based and video-based methods, and achieved one of the best P–R rates, as shown in Fig. 10. It also compressed the training dataset to 4.75% by averaging the features in each 3D point and it compressed the testing dataset to 5.30% by averaging the trajectories, leading to significant reduction in memory consumption whilst maintaining the discriminatory property within the target objects.

6. Conclusion

We have performed a comparative study of various object recognition methods and their video-based extensions for the rigid object instance recognition problem in egocentric videos. Based on the empirical evaluation results, we conclude that video- and 3D-based methods not only outperformed image-based methods, but are also capable of compressing the dataset into a more compact form. In particular, utilising 3D geometric constraints greatly improves the video classification accuracy, whilst tracking significantly reduces the query data size by rejecting unstable image regions and by forming trajectories. We have found that the most promising method for achieving the best object recognition performance with egocentric videos is to train the object classes with geometric constraints and to classify the query videos with spatio-temporal constraints. Thus we have developed and validated a hybrid method combining advantages from both constraints.

Appendix A. Supplementary material

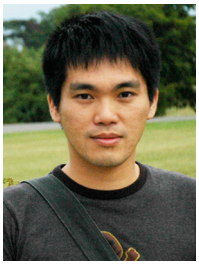
The following are the supplementary data to this paper:
Video S1.

References

[1] N. Li, J.J. DiCarlo, Unsupervised natural visual experience rapidly reshapes size invariant object representation in inferior temporal cortex, Neuron 67 (6) (2010) 1062–1075.
[2] T. Lee, S. Soatto, Video-based descriptors for object recognition, Image Vis. Comput. (2012) 639–652.
[3] N. Noceti, E. Delponte, F. Odone, Spatio-temporal constraints for on-line 3d object recognition in videos, Comput. Vis. Image Underst. 113 (12) (2009) 1198–1209.
[4] D. Stavens, S. Thrun, Unsupervised learning of invariant features using video, in: Computer Vision and Pattern Recognition (CVPR), IEEE, San Francisco, 2010, pp. 1649–1656.
[5] W. Zou, A. Ng, S. Zhu, K. Yu, Deep learning of invariant features via simulated fixations in video, in: Advances in Neural Information Processing Systems (NIPS), 2012, pp. 3212–3220.
[6] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition (CVPR), vol. 2, IEEE, New York, 2006, pp. 2169–2178.
[7] I. Gordon, D.G. Lowe, What and where: 3D object recognition with accurate pose, Toward Category-level Object Recognition, Springer, Siracusa (2006) 67–82.
[8] M.A. Peterson, G. Rhodes, Perception of Faces, Objects, and Scenes, Oxford University Press, New York, 2003.
[9] J. Knopp, M. Prasad, G. Willems, R. Timofte, L. Van Gool, Hough transform and 3d surf for robust three dimensional classification, in: European Conference on Computer Vision (ECCV), Springer, Heraklion, 2010, pp. 589–602.
[10] S.Y. Bao, M. Bagra, Y.-W. Chao, S. Savarese, Semantic structure from motion with points, regions, and objects, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, 2012, pp. 2703–2710.
[11] S. Savarese, L. Fei-Fei, 3d generic object categorization, localization and pose estimation, in: International Conference on Computer Vision (ICCV), IEEE, Minneapolis, 2007, pp. 1–8.
[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Miami, 2009, pp. 248–255.
[13] M. Everingham, L. van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[14] L. Lin, P. Luo, X. Chen, K. Zeng, Representing and recognizing objects with massive local image patches, Pattern Recognit. 45 (1) (2012) 231–240.
[15] F. Rothganger, S. Lazebnik, C. Schmid, J. Ponce, 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints, Int. J. Comput. Vis. 66 (3) (2006) 231–259.
[16] X. Ren, C. Gu, Figure-ground segmentation improves handled object recognition in egocentric video, in: Computer Vision and Pattern Recognition (CVPR), IEEE, San Francisco, 2010, pp. 3137–3144.
[17] Y. Liu, Y. Jang, W. Woo, T.-K. Kim, Video-based object recognition using novel set-of-sets representations, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Columbus, 2014, pp. 533–540.
[18] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[19] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[20] M. Cho, K.M. Lee, Progressive graph matching: making a move of graphs via probabilistic voting, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, 2012, pp. 398–405.
[21] O. Duchenne, A. Joulin, J. Ponce, A graph-matching kernel for object categorization, in: International Conference on Computer Vision (ICCV), IEEE, Barcelona, 2011, pp. 1792–1799.
[22] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Miami, 2009, pp. 1794–1801.
[23] S.S. Bucak, R. Jin, A.K. Jain, Multiple kernel learning for visual object recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1354–1369.
[24] X. Liu, L. Lin, S. Yan, H. Jin, W. Tao, Integrating spatio-temporal context with multiview representation for object recognition in visual surveillance, IEEE Trans. Circuits Syst. Video Technol. 21 (4) (2011) 393–407.
[25] J. Sivic, A. Zisserman, Efficient visual search of videos cast as text retrieval, Pattern Anal. Mach. Intell. 31 (4) (2009) 591–606.
[26] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Anchorage, 2008, pp. 1–8.
[27] T. Tuytelaars, M. Fritz, K. Saenko, T. Darrell, The nbnn kernel, in: International Conference on Computer Vision (ICCV), IEEE, Barcelona, 2011, pp. 1824–1831.
[28] R. Behmo, P. Marcombes, A. Dalalyan, V. Prinet, Towards optimal naive bayes nearest neighbor, in: European Conference on Computer Vision (ECCV), Springer, Heraklion, 2010, pp. 171–184.
[29] L. Mei, J. Liu, A. Hero, S. Savarese, Robust object pose estimation via statistical manifold modeling, in: International Conference on Computer Vision (ICCV), 2011.
[30] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, J. Mach. Learn. Res. 4 (2003) 913–931.
[31] R. Wang, S. Shan, X. Chen, Q. Dai, W. Gao, Manifold-manifold distance and its application to face recognition with image sets, Image Process. 21 (10) (2012) 4466–4479.
[32] J. Wu, M.M. Trivedi, A two-stage head pose estimation framework and evaluation, Pattern Recognit. 41 (3) (2008) 1138–1158.
[33] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, D. Jacobs, A search engine for 3d models, ACM Trans. Graph. 22 (1) (2003) 83–105.
[34] J.W. Tangelder, R.C. Veltkamp, A survey of content based 3d shape retrieval methods, Multimed. Tools Appl. 39 (3) (2008) 441–471.
[35] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, F. Wu, Y. Rui, Efficient 2d-to-3d correspondence filtering for scalable 3d object recognition, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Portland, 2013, pp. 899–906.
[36] A. Irschara, C. Zach, J.-M. Frahm, H. Bischof, From structure-from-motion point clouds to fast location recognition, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Miami, 2009, pp. 2599–2606.
[37] T. Sattler, B. Leibe, L. Kobbelt, Improving image-based localization by active correspondence search, in: European Conference on Computer Vision (ECCV), Springer, Florence, 2012, pp. 752–765.
[38] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, Pattern Anal. Mach. Intell. 30 (2) (2008) 328–341.
[39] B. Ummenhofer, T. Brox, Dense 3d reconstruction with a hand-held camera, In: Pattern Recognition: Joint 34th DAGM and 36th OAGM Symposium, Graz, Austria, August 28-31, 2012, Proceedings, 2012, pp. 103–112.
[40] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: a survey, Found. Trends® Comput. Graph. Vis. 3 (3) (2008) 177–280.
[41] T. Lindeberg, Invariance of visual operations at the level of receptive fields, PLoS ONE 8 (2013) 66990. http://dx.doi.org/10.1371/journal.pone.0066990 http://arxiv:hepth/1210.0754.
[42] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition (CVPR), vol. 2, IEEE, 2005, pp. 524–531.
[43] S. McCann, D.G. Lowe, Local naive bayes nearest neighbor for image classification, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, 2012, pp. 3650–3656.
[44] Z. Kalal, K. Mikolajczyk, J. Matas, Forward–backward error: automatic detection of tracking failures, in: International Conference on Pattern Recognition (ICPR), IEEE, Istanbul, 2010, pp. 2756–2759.
[45] B.D. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision, in: International Joint Conferences on Artificial Intelligence (IJCAI), vol. 81, 1981, pp. 674–679.
[46] A. Collet, M. Martinez, S.S. Srinivasa, The moped framework: object recognition and pose estimation for manipulation, Int. J. Robot. Res. 30 (10) (2011) 1284–1306.
[47] C. Wu, Visualsfm: a visual structure from motion system, 2011.
[48] V. Lepetit, F. Moreno-Noguer, P. Fua, Epnp: an accurate o (n) solution to the pnp problem, Int. J. Comput. Vis. 81 (2) (2009) 155–166.
[49] D.A. Forsyth, J. Ponce, Computer Vision: A Modern Approach, Prentice Hall Professional Technical Reference, 2002.
[50] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, P. Beltran, Guide to the Carnegie Mellon University Multimodal Activity (cmu-mmac) Database, 2009.
[51] X. Ren, M. Philipose, Egocentric recognition of handled objects: benchmark and analysis, in: Computer Vision and Pattern Recognition Workshops (CVPR-WS), IEEE, Miami, 2009, pp. 1–8.

Mang Shao received the BSc and MEng degrees in electrical and electrical engineering from Imperial College London, United Kingdom, in 2012. In 2013, he is working on the PhD degree in Imperial Computer Vision and Learning Lab at Imperial College London. His research interests include object recognition and 3D object pose estimation.



Danhang Tang is a PhD candidate in the Imperial Computer Vision and Learning Lab, which belongs to the Electrical and Electronic Engineering Department of Imperial College London. Prior to his PhD, he received a 1st honor MSc degree from University College London and a BSc degree from Sun Yat-sen University. From 2007 to 2009, he worked as a system architect for Evryx Technologies Ltd., in support for SnapNow, one of the first image recognition apps in the world. During this time he also participated in drafting the visual search specification for China Telecom. His research topic is articulated object detection and pose estimation.

Yang Liu received MSc degree from Chinese Academy of Sciences (National Laboratory of Pattern Recognition, Institute of Automation) in 2011, and BSc degree from Beihang University (School of Automation Science and Electrical Engineering) in 2008. His current research topic is object recognition in videos.

Tae-Kyun Kim received the BSc and MSc degrees from the Korea Advanced Institute of Science and Technology in 1998 and 2000, respectively, and the PhD degree from the University of Cambridge in 2007. He is a research fellow in the Sidney Sussex College at the University of Cambridge. He was a research staff member at the Samsung Advanced Institute of Technology during 2000–2004. His research interests include computer vision, statistical pattern classification, and machine learning. The joint proposal of Samsung and NEC for face image descriptor, for which he developed main algorithms, is the international standard of ISO/IEC JTC1/SC29/WG11.