Video-based Object Recognition using Novel Set-of-Sets Representations

Yang Liu[1*] Youngkyoon Jang[2*] Woontack Woo[2] Tae-Kyun Kim[1]
[1]Imperial College London, UK [2]KAIST, South Korea
{y.liu11, tk.kim}@imperial.ac.uk {y.jang, wwoo}@kaist.ac.kr

Abstract

We address the problem of object recognition in egocentric videos, where a user arbitrarily moves a mobile camera around an unknown object. Using a video that captures variation in an object's appearance owing to camera motion (more viewpoints, scales, clutter and lighting conditions), can accumulate evidence and improve object recognition accuracy. Most previous work has taken a single image as input, or tackled a video simply by a collection i.e. sum of frame-based recognition scores. In this paper, beyond frame-based recognition, we propose two novel ***set-of-sets*** *representations of a video sequence for object recognition. We combine the techniques of bag of words for a set of data spatially distributed thus heterogeneous, and manifold for a set of data temporally smooth and homogeneous, to construct the two proposed set-of-sets representations. We also propose methods to perform matching for the two representations respectively. The representations and matching techniques are evaluated on our video-based object recognition datasets, which contain 830 videos of ten objects and four environmental variations. The experiments on the challenging new datasets show that our proposed solution significantly outperforms the traditional frame-based methods.*

1. Introduction

Object recognition is one of the most important topics in computer vision and has been intensively studied in the past decades. It has found applications in many areas including human-computer interaction, intelligent surveillance, industrial inspection, robotics, medical imaging, to name a few. Although numerous recognition algorithms have been developed, most of them [3, 4, 18, 22, 39]) are image-based recognition, taking a single image as input. Although they may as well be extended into a video by straightforward accumulating frame recognition scores as in [21], exploiting a full potential of videos for object recognition still remains. Especially when nowadays videos are often available owing to widespread video cameras, video-based object recognition is becoming more and more popular [7, 13, 20, 28, 31, 32, 33, 36]. Among these methods, normally a video is treated as a collection of multiple images, and/or features are augmented by feature tracking before a single image based recognition framework is exploited (see Section 2).

In the setting we consider for video-based object recognition, a user moves a mobile camera around an unknown object of interest, while putting the object roughly in the center of images and capturing multiple viewpoints in an arbitrary manner with variations in scale, clutter, and illumination. Therefore, practically there is no assumption that a query video has the same or similar camera motion/viewpoint as those of model videos. This contrasts the fields of gesture, activity, and behaviour recognition where motion is the key discriminative information to use. We formulate the problem as representation and matching of *sets of sets*. Both queries and models are video clips, and a query video is recognised into one of objects in the dataset. Our main contributions can be summarised as:

1. We recognise a video rather than individual frames: a video captures multiple-views and smooth variations in object appearance, revealing more about object identities.
2. We explore two novel representations and matching methods for video-based object recognition, which go beyond frame-based recognition.
   (a) Video is represented as a set of (unordered) frames (ignoring temporal ordering due to random camera motion), where each frame is given a histogram vector of codewords using the BoWs technique powered by Randomised Decision Forest. Those histogram vectors along a temporal axis draw a smooth manifold, thus we perform manifold-to-manifold matching for recognition.
   (b) A set of feature vectors collected from a feature point trajectory is well constrained onto a low-dimensional manifold. Thus, a set of trajectory

*indicates equal contribution.

manifolds is obtained from a video. We propose a novel Kernel Random Forest as a codebook of manifolds, by which trajectory manifolds of a video are quantised to form a bag of manifolds.

3. The proposed methods are evaluated on new video-based object recognition datasets consisting of 830 videos. The proposed methods and their combination improve the recognition accuracies of frame-based methods by a large margin. Additionally, the combined method runs in real-time.

In the rest of the paper, literature review is done in Section 2. Section 3 formulates the video-based object recognition problem and proposes the novel video representations and matching methods. Experimental results and discussion are presented in Section 4. Conclusions are drawn in Section 5.

2. Relevant Work

There has been an increasing attention to object recognition in videos [7, 13, 20, 28, 31, 32, 33, 36]. In [28], an input video is captured by an egocentric camera, and the motion-based foreground object segmentation technique is developed to improve object recognition accuracy. In [23], A topic model is extended from still images to motion videos for unsupervised object discovery, which is limited to the type of motions in a long-range surveillance camera. A video typically contains multiple views of an object. An unsupervised multi-view feature selection algorithm is proposed to improve object recognition accuracy in [7]. Although the technique may be made useful for video-based object recognition by efficiently removing redundant views, the work tackles the problem of distributed object recognition with no temporal information considered.

A majority of algorithms [13, 20, 31, 32, 33] for object recognition in videos propose to utilize the temporal information in video and improve local video descriptors by feature tracking. In [20], the set of descriptors called Best Template Descriptor (BTD) are trained using tracked features in videos and then the vocabulary tree forms a BoWs used to recognise a single test frame. [33] proposes an efficient search space for interest points to track features, then the tracked features are exploited to recognise objects. In [32], an invariant feature is learnt by tracking image patches over time using optical flow. These methods improve feature descriptors for video, however, perform object recognition frame by frame.

The other line of works for object recognition in videos takes human faces as target objects. One example is [19] where temporal information is captured by the transition probabilities between pose manifolds, however, the strong temporal constraint harms object recognition when it moves arbitrarily. A video, therefore, has been taken as a set of unordered images and how to match a set-to-set has been intensively studied [35, 37, 14] (see below).

We briefly discuss three methods to represent a set of observations: bag of words (BoWs), probability density, and manifold methods.

BoWs is the most common representation for object recognition [31], which treats an image as a set of feature points (or descriptors), describing their semantic distributions and their structural relations [29]. Commonly a three-step framework is adopted: feature point extraction, codeword quantisation (or forming a BoWs), and classification e.g. by Nearest Neighbor, Support Vector Machine or Random Forest. Although BoWs is often used for object categorisation, it is adopted in object instance recognition as well for fast implementation [20]. It is used in our work also because we need vector representation for image and video, rather than pairwise key-point matching [24].

Probability density used for a set representation naturally provides an estimate of uncertainty. [15] employs Gaussian Processes (GPs) to place a prior probability on the spatial correlation of training data points, offering confidence estimates of new data points for probabilistic object category recognition. Gaussian, Gaussian mixtures [1], non-parametric densities and kernel methods [40] have been proposed to capture nonlinear data distributions, and various probabilistic distances e.g. KL divergence, measure the similarity between two distributions. A drawback of these methods lies in the computational complexity and they often fail when the model and query sets do not exhibit strong statistical similarity.

Another alternative to represent a set is a manifold. It has been very successful in face recognition where a set of face images draw a smooth low-dimensional manifold. The problem of classifying face image sets by Manifold-to-Manifold Distance (MMD) has been well formulated in [35], and standard techniques include Principal Angles (PA) and Kernel Principal Angles (KPA), which measure the angles between two manifolds. The similar techniques have also been successfully applied to dynamic texture recognition [6]. In [8], KPA was used to measure the similarity between two images for object recognition, where each image is a set of local image features. However, this was found poorer than other methods in [8], mainly because data variations in a set of local descriptors of an image are not smooth to be constrained on a low-dimensional manifold and forming a manifold loses discriminative information. [16] learnt a discriminative PA for object recognition with image sets, but their image description is not suitable for our work since they do not use local features.

Gesture recognition, activity recognition, and any similar problems, where motion is key discriminative information [11, 38], are different from the problem we tackle.

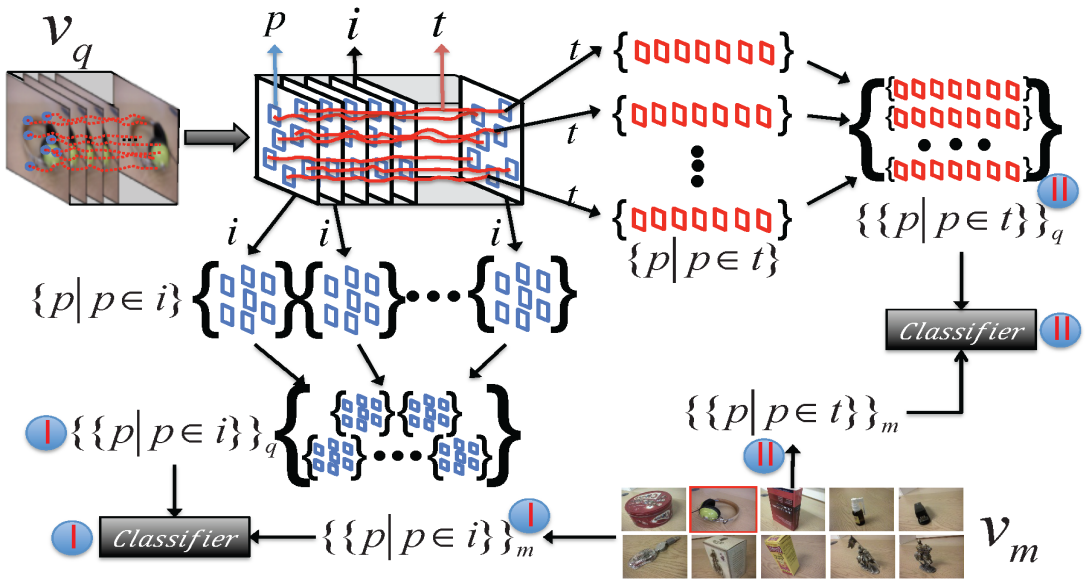


Figure 1. Set-of-sets representations for object recognition in videos. v: video, q: query, m: model, p: feature point, i: image, t: trajectory, I, II: two proposed representations. The two representations are combined at the classifier score level.

We do not assume any motion patterns for object classes, but random camera motion around objects. Structure from videos, shape recognition in point clouds [5, 12, 27] are also rather off the topic we consider. Such explicit 3D shape recovery and exploitation for object recognition can be an extension of this work in future.

3. Video-based Object Recognition

3.1. Problem formulation and overview

Video-based object recognition takes a video rather than a single image as input. The videos of the same object may contain common parts of the object at different viewpoints and scales.

As shown in Figure 1, a video v contains a couple of elements, such as images i, feature points p (represented by SURF [2] for scale and rotation invariance), and trajectories t of the feature points (see Section 4.2 for details). On one hand, a video v is, temporally, a set of images $\{i\}$ (frames), and each image i is, spatially, a set of feature points $\{p|p \in i\}$. On the other hand, a video v can be seen spatially as a set of trajectories $\{t\}$, and each trajectory t is, temporally, a set of similar feature points $\{p|p \in t\}$ along the trajectory. Therefore, a video forms the notion of a set of sets $\{\{p\}\}$, and the problem of video-based object recognition primarily concerns how to represent and match a query set of sets and a model set of sets, i.e. establishing a kernel function, $K(v_q, v_m) = K(\{\{p\}\}_q, \{\{p\}\}_m)$. We may then use any kernel classifier in a multi-class setting.

We propose two different set-of-sets representations for a video (see Figure 1): (1) The first one represents each image i in a video as a set of feature points $\{p|p \in i\}$, then represents all images of the video as a bigger set $\{\{p|p \in i\}\}$. (2) The second one first takes each trajectory in a video t as a set of feature points $\{p|p \in t\}$, then all trajectories of the video are formed as a bigger set $\{\{p|p \in t\}\}$. Two popular means of representing a set, Bag of Words (B) and Manifold (M) methods are exploited respectively for a set of data spatially (thus being multi-modally distributed) and a set of data temporally (thus being smoothly changed). Manifolds enclose BoWs in the first representation (Section 3.2), while BoWs embrace Manifolds in the second representation (Section 3.3). The two representations are combined at the classifier score level.

3.2. Manifold of BoWs

In this representation, we see a video v as a set of images $\{i\}$, where each image is a set of feature points $\{p\}$ (see Figure 2). Due to random camera motion, although a video is composed of images in a sequence, it is treated as an unordered set of images. Those images collected along a temporal axis exhibit smooth data changes well constrained on a low-dimensional manifold (see Figure 7(a)). Manifold-to-manifold matching is a well-studied area. We exploit the Kernel Principal Angle (KPA), a nonlinear extension of the Principal Angle (PA) by a kernel trick, to measure the similarity between the two manifolds by cosines of the principal angles [37].

Given a query video v_q and a model video v_m, each image in them is first represented as a set of feature points.

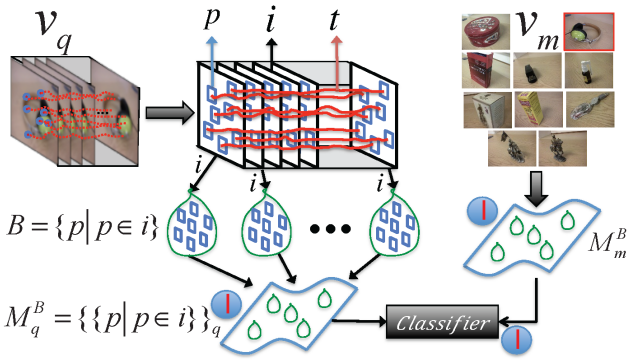


Figure 2. Manifold-of-BoWs based matching.

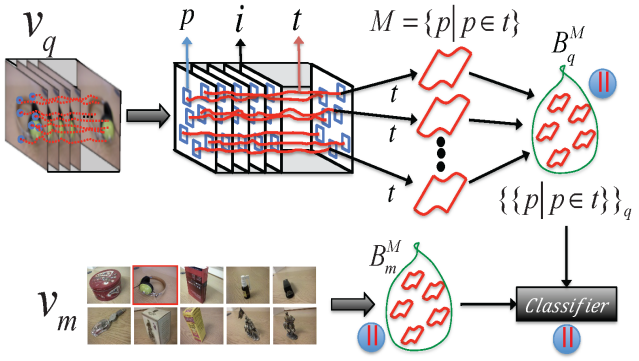


Figure 3. Bag-of-Manifolds based matching.

Those spatially distributed feature points embody a distinctive appearance. They are represented as a bag of word histogram vector B using a Randomised Decision Forest as a fast and discriminative codebook [26, 30]. Quantisation of feature points using this compact and discriminative codebook helps robust image matching. Then the query video and the model video can be represented respectively as two sets of BoWs. KPA takes the two sets as input, computes two best-matched manifolds M_q^B, M_m^B respectively and outputs the principal angles as the similarity between the two sets. The Manifold of BoWs captures appearance changes (due to e.g. viewpoint, scale or lighting changes) of an object, reflects a set-property and facilitates robust data matching by interpolation [9]. Therefore the similarity between the query video and the model video is calculated as follows:

$$\begin{aligned} K(v_q, v_m) &= K(\{i\}_q, \{i\}_m) \\ &= K(\{\{p|p \in i\}\}_q, \{\{p|p \in i\}\}_m) \\ &= K(\{B\}_q, \{B\}_m) \\ &= K(M_q^B, M_m^B) \end{aligned} \tag{1}$$

where the kernel function K is given by KPA (itself a positive semi-definite valid kernel) as

$$K(M_q^B, M_m^B) = \cos\theta = \max_{\phi(u)\in M_q^B} \max_{\phi(v)\in M_m^B} \phi(u)^T \phi(v) \tag{2}$$

where $\phi(u), \phi(v)$ are arbitrary vectors on the respective manifolds and $\phi(\cdot)$ is a nonlinear mapping from an input space onto a feature space by a kernel function $k(x, x') = \phi(x)^T \phi(x')$. In the experiment, we used the RBF kernel $k(x, x') = \exp(-|x - x'|^2/\sigma^2)$ and the minimum kernel principal angle i.e. the maximum cosine value between the two manifolds, which measures how well two manifolds intersect, to represent the similarity.

3.3. Bag of Manifolds

In this representation, we see a video v as a set of trajectories $\{t\}$, where each trajectory is a set of feature points $\{p\}$ along it (see Figure 3). Feature points in the initial frame are tracked in following frames to form trajectories (see Section 4.2 for details). Note that the feature points in a trajectory are feature vectors (SURF in our experiments) rather than 2D locations, so a trajectory is equal to a set of feature vectors. While the feature points within a trajectory are alike i.e. smoothly changed (see Figure 7(b)), those of different trajectories have distinctive appearances.

In contrast to the method in Section 3.2 where we first represent images as BoWs then a video as a manifold of BoWs, here we use the two techniques in the reversed order. The manifold method is first used to learn each trajectory as a manifold, then the BoWs technique is used to collect all trajectories i.e. manifolds, to form a so called Bag of Manifolds. The proposed method can be seen as a temporal extension of the standard Bag of Words.

It has been mentioned in the previous section that KPA can find two best-matched manifolds for two sets and output the similarity, and RF can be used as a fast and discriminative codebook. In order to learn trajectories as manifolds as well as to make a codebook of manifolds, we propose a novel method named Kernel Random Forest (KRF), which incorporates KPA into standard RF. Note that the input data for our KRF are trajectories, each of which is a set of feature vectors rather than a single vector (which is a typical input to the standard RF).

All trajectories in model videos are used as input of KRF to construct the codebook of manifolds. In each split node of KRF, a reference trajectory is chosen and compared with every input trajectory by KPA to split the node. For this process, all input trajectories are given as manifolds, so we refer to a trajectory as a manifold here (see Figure 4). At every split node, a certain amount of manifolds (10 in the experiments) are randomly selected, from which we choose

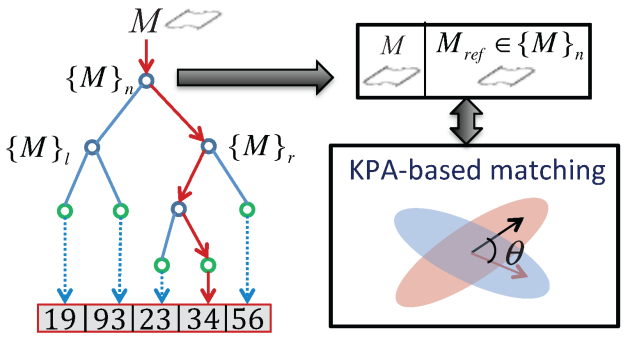


Figure 4. Kernel Random Forest Codebook.

the one (M_{ref}) that gives us the best split in terms of information gain. Each tree projects the manifolds to a higher-dimension feature space by using the kernel trick. The feature space is split into two regions by a split function f:

$$f(M, M_{ref}) = KPA(M, M_{ref}) - t = \cos\theta - t \quad (3)$$

where M is an input manifold. M_{ref} (the reference manifold) and t (the threshold) are chosen to maximise the information gain [30]. We use the RBF kernel for KPA, the maximum cosine value of the kernel principal angle to measure the similarity.

When the tree-growth is done, the leaf nodes of KRF serve as codewords of manifold codebook. The manifolds of trajectories in a video are quantised into codewords by passing them down the KRF to the leaf nodes. The histogram of the codewords then forms the Bag of Manifolds B^M for the video. Given a query video v_q and a model video v_m, the similarity between the two videos is measured by

$$\begin{aligned} K(v_q, v_m) &= K(\{t\}_q, \{t\}_m) \\ &= K(\{\{p|p \in t\}\}_q, \{\{p|p \in t\}\}_m) \\ &= K(\{M\}_q, \{M\}_m) \\ &= K(B_q^M, B_m^M) \end{aligned} \quad (4)$$

where the kernel function is obtained by Euclidean distance or cross correlation of the two histogram vectors B_q^M and B_m^M (which provides a valid kernel).

4. Experimental Results

4.1. New datasets for video-based object recognition

For evaluating the proposed method, our own video-based object dataset is presented. The dataset comprises challenging objects (less-textured, fully 3D-shaped and very similar appearance), including candybox, headset, book, throat spray, stapler, computer mouse, box1, box2, model1-flag and model2-shield (see Figure 5(a)). The last two objects were deliberatively chosen as similar objects. The dataset contains 230 videos acquired from the 10 objects. Each object has 23 videos including 3 training videos and 20 testing videos. The video is in the resolution of 640×480, 24-bit color coded, and 25 frames per sec are extracted.

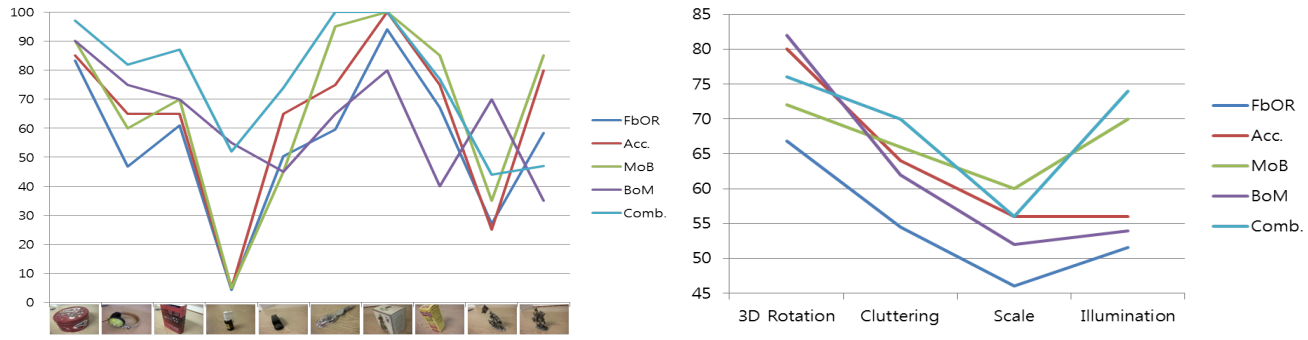
(a)

(b)

Figure 5. Dataset examples: (a) representative images for 10 object classes; (b) first row shows the 3D rotations, second row the cluttering, third row the scale changes, and last row the illumination changes.

Three turntable sequences (10-20 seconds) of an object captured in different illumination conditions are used as training videos. The testing videos (4-10 seconds) are recorded under challenging environments with four different kinds of variations: 3D rotations, clutters, severe scale changes and spot lighting movements, as shown in Figure 5(b). Each consists of 5 videos. In the case of 3D rotations, we randomly moved a camera around an object without any other effects. In the case of cluttering, the target object is in the center and the closest distance, while other registered objects serve as background. For the scale variations, significant scale changes are included, varying from taking up one tenth of view space to only one third of target objects appearing in view. For the illumination changes, we used a randomly moving spot illuminator around an object. All test videos include arbitrary 3D rotations.

We have collected the second dataset, an enlarged one, in order to train a classifier for the proposed video-based method. It contains 60 videos per object. We use 15 videos (of random 3D rotations) for training and 45 videos (15 videos per each of clutter, scale and illumination changes) for testing. Each video lasts for 4-10 seconds.



(a) (b)

Figure 6. Comparisons on DBv1 for: (a) different objects, (b) data variations (best viewed in colour).

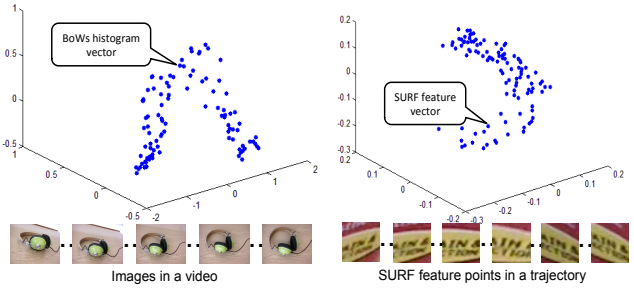
4.2. Implementations

We use FAST corner detector [34] and SURF descriptors [2] for real-time implementations. For discriminative and fast descriptor quantisation, we use Random Decision Forest as a codebook, which has been very successful in the relevant fields [26, 30]. Important parameters such as the Gaussian kernel parameters in SVM, maximum depth of tree and tree numbers in RFs, are set to report the best accuracy. In the experiments, e.g. the dimension of SURF descriptors is 64, maximum depth of a tree and tree numbers for the RF codebook are fixed to 4 and 16 in all experiments. In order to better consider the quantisation errors of unseen data, odd-numbered frames of the training videos are exploited to build the RF codebook and even-numbered frames to make model histogram vectors.

Feature tracking is performed by KLT [25] to obtain trajectories. Since we consider a moving camera, some tracked feature points disappear when the tracked area is occluded or out of camera views. We add more feature points in newly appeared parts so that a video maintains an adequate number of the tracked feature points. The number of tracked feature points dynamically change according to the feature points detected by FAST corners in the present frame, which also reflects characteristics of an object of interest. Using the tracked feature points, we form SURF descriptors [2], to support scale and rotation-invariant recognition to a certain degree. Please note that the main contributions of this work do not lie in the descriptors and trackers, but in the novel representations of a video for object recognition. Our expectation is that the proposed solution would achieve better recognition accuracies with better detectors, descriptors and trackers used: refer to [10] for comparative studies.

4.3. Evaluation results and discussion

We have compared our proposed method (two representations individually and their combined) with the frame-based object recogniser (FbOR) where the average frame recognition accuracy is reported, and the accumulating-frames recogniser (Acc) which takes results of all frames to vote for video class. Every image is represented as a BoWs histogram vector using the techniques described in Section 4.2, and is classified into one of object classes using NN, SVM in the one-vs-one setting, or RF classifier.



(a) (b)

Figure 7. Data smoothness: (a) smooth manifold of BoWs, (b) smooth trajectory manifold.

Figure 7 shows an example of manifold in each representation (thus the manifold of BoWs histogram vectors and the trajectory manifold) using the first three principal components obtained by PCA (for the visualization purpose only). Both draw a smooth data manifold in a low-dimensional space. Combining two proposed representations and matching methods is considered. As will be shown below, the tendency is that the two methods deliver quite different recognition results on different objects and data variations in the datasets, i.e. they exhibit not strongly correlated error models. Here we consider the simple SUM rule [17] at the score level, a better combining method remains as a future work.

Figure 6 shows the comparison results for different objects and data variations using the NN classifier. FbOR exhibits poor recognition accuracies on the less-textured objects (headset, throat spray and stapler) and similar-shaped objects (model1-flag and model2-shield), with the worst 4.4% on the throat spray, as shown in Figure 6(a), while the video-based methods show much higher recognition accuracies. Especially, the proposed combined method ob-

Table 1. Video-based object recognition accuracies. *:[26], **:[30], ***:Best template descriptor method [20]. The average frame recognition accuracies are reported for FbOR and BTD methods.

Method	DB v1 (23 videos/obj.) (%)			DB v2 (60 videos/obj.) (%)		
	NN	SVM	RF	NN	SVM	RF
FbOR	**54.88**	59.89*	56.27**	62.18	**68.72***	67.94**
	NN			NN	SVM	
BTD***	56.5			63.1	66.44	
Accumu. Frames	**65**			64.67	**71.55**	
Manifold of BoWs[I]	67			69.56	70.44	
Bag of Manifolds[II]	64			63.78	69.11	
Comb. set of sets[I,II]	**74.5**			71.3	**77.11**	

tains a significant accuracy gain. Figure 6(a) shows that the bag of manifolds (BoM) delivers better accuracies on the less-textured and similar objects, while the accumulating frames and manifold of BoWs (MoB) exhibit quite similar performance tendency i.e. correlated. The bag of manifold method plays a role to boost the combined accuracy by compensating the other method. In Figure 6(b), the manifold of BoWs shows the best individual accuracies on the clutter, scale, and illumination variations, while the bag of manifolds shows the best accuracy on the 3D rotations. The combined method again outperforms the best individual accuracies.

Table 1 summarises the results of the proposed solutions and their combined, the frame-based object recognition (FbOR) methods, the accumulating frame recognition and BTD [20], on the two of our datasets. The RF codebook and SVM classifier were exploited in the FbOR-SVM [26], and the RF codebook and RF classifier were exploited in the FbOR-RF [30]. In [20], the descriptors called BTD (best template descriptor) were learnt from training videos, and integrated into the vocabulary tree, which was replaced with RF in our experiment. The best accuracy obtained by FbOR on the primary dataset DB v1 is 59.89% using SVM. As expected, using SVM or RF as a classifier improves the accuracy of Nearest Neighbor (NN) classifier in the FbOR. The accuracies of [20] appear similar to those of FbOR. Note that their aim is to improve the run-time speed. Using a better feature tracker might further improve the accuracy of [20]. The proposed video-based methods (the best accuracy 67% obtained by the manifold of BoWs and NN) deliver a more than 10% accuracy gain over the frame-based methods (**54.88%** by FbOR-NN). The combined method (**74.5%**), which is about 10% better than the accumulating frames (**65%**), significantly improves the accuracies of the best individual. As shown in Figure 6, the proposed methods exhibit quite uncorrelated errors, while compensating each other to boost the combined accuracy.

In the results on the DB v2, the accuracies of NNs are improved by SVMs. The overall tendency on the DB v2 is similar to the DB v1. The frame-based recognition accuracy (**68.72%** using SVM) is improved to **77.11%** by the method combining the manifold of BoWs and bag of manifolds, which yet leads the accumulating frame method (**71.55%**). Due to the shorter and random camera moving videos (instead of the turn-table sequences in the DB v1) used for training, the accuracy gains of the proposed methods over the baselines drop a bit, however, the combined method still delivers the significant improvement.

Using a better feature detector and tracker might further improve the accuracies of the bag of manifold method and then the combined. At present, there are tracking errors involved in the trajectories, which degrade the performance.

The proposed methods and their combined were implemented in a machine of Intel Core i7 2.2 GHz processor with 8GB DRAM. The accumulating frame method (Acc) runs in real-time on every frame, and the manifold of BoWs (MoB) and bag of manifolds (BoM) are executed in a short time interval to correct recognition errors and improve the recognition confidence of the Acc.

5. Conclusions

We have tackled video-based object recognition and presented novel *set-of-sets* representations and respective matching methods. In the experiments using the new video datasets consisting of 830 videos, we showed that 1) taking a video input rather than a single image provides better object recognition accuracies, and 2) the proposed method outperforms the frame-based recognition methods and their accumulation. Additionally, the proposed method runs in real-time.

In future, we will consider an extension of our present framework to object categorisation rather than object instance recognition. To further improve the recognition accuracies, using 3D depth videos as input, better feature trackers and combining methods are to be explored. Active learning in the manner that user cooperation is maximised to recognise objects in videos is an interesting direction for future study.

Acknowledgement: This work was in part supported by the Global Frontier R&D Program on <Human-centered Interaction for Coexistence>, National Research Foundation of Korea(MSIP)(2010-0029751).

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, 2005.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 2008.

[3] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[4] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[5] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008.

[6] A. B. Chan and N. Vasconcelos. Classifying video with kernel dynamic textures. In *CVPR*, 2007.

[7] C. M. Christoudias, R. Urtasun, and T. Darrell. Unsupervised distributed feature selection for multi-view object recognition. In *CVPR*, 2008.

[8] J. Eichhorn and O. Chapelle. Object categorization with svm: Kernels for local features. *Technical report, Max Planck Inst., Tuebingen*, 2004.

[9] A. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *CVPR*, 2003.

[10] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision*, 2011.

[11] T. Glaser and L. Zelnik-Manor. Incorporating temporal context in bag-of-words models. In *ICCV*, 2011.

[12] A. Golovinskiy, V. Kim, and T. Funkhouser. Shape-based recognition of 3d point clouds in urban environments. In *ICCV*, 2009.

[13] V. Gouet-Brunet and B. Lameyre. Object recognition and segmentation in videos by connecting heterogeneous visual features. *Computer Vision and Image Understanding*, 2008.

[14] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, 2011.

[15] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian process for object categorization. In *ICCV*, 2007.

[16] T.-K. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *ECCV*, 2006.

[17] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence*, 1998.

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[19] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, 2003.

[20] T. Lee and S. Soatto. Video-based descriptors for object recognition. *Image and Vision Computing*, 2011.

[21] B. Li, R. Chellappa, Q. Zheng, and S. Z. Ser. Model-based temporal object verification using video. *Image Processing*, 2001.

[22] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[23] D. Liu and T. Chen. A topic-motion model for unsupervised video object discovery. In *CVPR*, 2007.

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[25] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.

[26] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *NIPS*, 2007.

[27] M.-T. Pham, O. J. Woodford, F. Perbet, A. Maki, B. Stenger, and R. Cipolla. A new distance for scale-invariant 3d shape recognition and registration. In *ICCV*, 2011.

[28] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010.

[29] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of copmlex human activities. In *ICCV*, 2009.

[30] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[31] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[32] D. Stavens and S. Thrun. Unsupervised learning of invariant features using video. In *CVPR*, 2010.

[33] D.-N. Ta, W.-C. Chen, N. Gelfand, and K. Pulli. Surftrac: Efficient tracking and continuous object recognition using local feature descriptors. In *CVPR*, 2009.

[34] M. Trajkovic and M. Hedley. Fast corner detection. *Image and Vision Computing*, 1998.

[35] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008.

[36] J. Winn and A. Criminisi. Object class recognition at a glance. In *CVPR, video track*, 2006.

[37] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 2003.

[38] L. Xu and P. Mordohai. Automatic facial expression recognition using bags of motion words. In *BMVC*, 2010.

[39] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[40] S. Zhou and R. Chellappa. Probabilistic distance measures in reproducing kernel hilbert space. *SCR Technical Report, University of Maryland*, 2004.