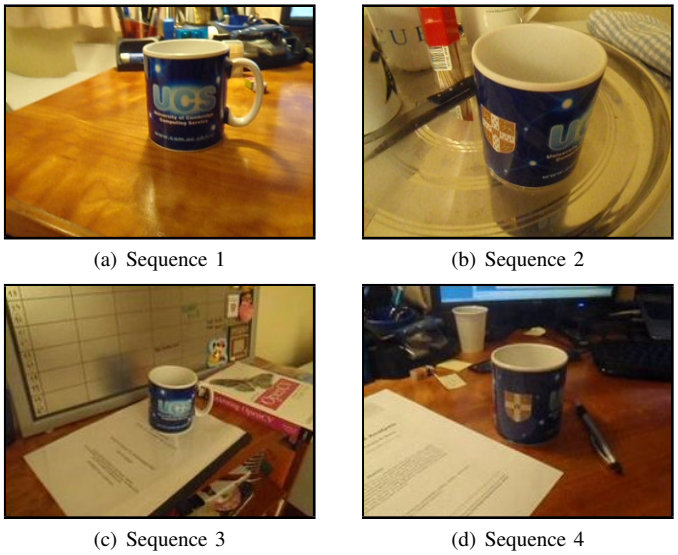
Descriptor Transition Tables for Object Retrieval using Unconstrained Cluttered Video Acquired using a Consumer Level Handheld Mobile Device

Warren Rieutort-Louis
University of Cambridge
United Kingdom

Ognjen Arandjelović
University of St Andrews
United Kingdom
ognjen.arandjelovic@gmail.com

***Abstract*—Visual recognition and vision based retrieval of objects from large databases are tasks with a wide spectrum of potential applications. In this paper we propose a novel recognition method from video sequences suitable for retrieval from databases acquired in highly unconstrained conditions e.g. using a mobile consumer-level device such as a phone. On the lowest level, we represent each sequence as a 3D mesh of densely packed local appearance descriptors. While image plane geometry is captured implicitly by a large overlap of neighbouring regions from which the descriptors are extracted, 3D information is extracted by means of a descriptor transition table, learnt from a single sequence for each known gallery object. These allow us to connect local descriptors along the 3rd dimension (which corresponds to viewpoint changes), thus resulting in a set of variable length Markov chains for each video. The matching of two sets of such chains is formulated as a statistical hypothesis test, whereby a subset of each is chosen to maximize the likelihood that the corresponding video sequences show the same object. The effectiveness of the proposed algorithm is empirically evaluated on the Amsterdam Library of Object Images and a new highly challenging video data set acquired using a mobile phone. On both data sets our method is shown to be successful in recognition in the presence of background clutter and large viewpoint changes.**



(a) Sequence 1 (b) Sequence 2

(c) Sequence 3 (d) Sequence 4

Fig. 1. Typical frames from videos of the same object, acquired at different times and in different backgrounds. Recognition of untextured ("smooth") objects across pose and illumination changes, and the presence of clutter poses a major challenge to existing methods.

I. Introduction

Owing to its pervasive application potential, computer based object recognition has been a focus of much computer vision research in the last decade. Successful proof-of-concept as well as commercial applications have been demonstrated in the context of large-scale image retrieval [1], urban scene recognition [2], augmented reality, and others. While most existing methods address the problem of object recognition using individual images, in this paper we focus on recognition from video. In other words a sequence of frames (images) of an unknown, query object is matched against a database of sequences of known, gallery objects. This problem setting is of an increasing significance considering the ease with which users can acquire and store videos (e.g. using a mobile phone camera and cloud storage), and the recognition robustness that the availability of additional data (in comparison with a single image) can provide (e.g. with respect to viewpoint).

II. Previous work

Automatic object recognition has attracted considerable research effort. Here we briefly review some of the directions taken by previously proposed approaches in the literature.

A. Holistic representations

An important source of difficulties that arise in an attempt to understand the content encoded by pixel intensities is the distributed nature of the information that can be potentially relevant to grouping decisions. The observation that the ultimate interpretation of an image fragment in the context of object recognition more often than not depends on its context, if not on the entire image, motivates the use of holistic representations.

a) Prototypes: One of the most common ways of representing objects is by a set of appearance prototypes or exemplars [3], [4]. This representation is simple, directly measurable and can thus be used irrespective of object scale or data quality in general. It also has a clear probabilistic interpretation, which means that any of a number of well understood off-the-shelf statistical methods can be applied to it. The entirety of an object's appearance is effectively described by the underlying probability density function which describes the object's possible appearance variation [5], [6].

At its core, the representation is in fact the image itself and

is thus not invariant to virtually anything at all. Background clutter poses a significant problem, just as does occlusion, as well as in the case of general, non-planar objects, changing viewpoint and illumination. Depending on the nature of the object (planarity and reflectance properties) a large number of exemplars may be needed to capture the entire corpus of appearance variation [4].

b) Model based: Model based object representation is rather different in nature from the previously discussed prototypes and the only truly view-invariant representation. Rather than describing an object in terms of how it appears in images, an object is characterized by its inherent properties such as shape and texture. As a result, this representation is not directly measurable from images. Instead, given models of objects of interest, recognition is performed by finding the model that best fits the image using back-projection: using a postulated set of viewing parameters (e.g. camera angle and illumination direction) the model is used to predict what the image should look like, which is then compared to the actual, observed appearance [7], [8], [9]. More generally, it need not be appearances that are compared but rather any measurable image features [10]. However, due to the constrained nature of this representation, it is generally suitable only for recognition within a narrow class of objects.

B. Local representations

In contrast to holistic approaches, local methods focus on describing different parts of objects first, building the representation of an entire object from bottom up.

c) Part based representations: The pictorial structures approach [11], [12] is a typical example from the group of part based representations [13]. Simultaneously using appearance and spatial information, an object is represented by a geometrically deformable configuration of different predefined (and typically manually chosen) parts. Successful examples from the literature include faces (with the eyes, the mouth and the nose as parts), the human body (with the limbs and the head as parts), motorcycles and aeroplanes. In the context of this paper, part based representations suffer from similar limitations as model based approaches.

d) Local feature based methods: In general object recognition tasks, part based approaches are largely overshadowed by the success of representations which use local descriptors [1], [14], [15], [16]. The idea is simple: at the lowest level small image patches are represented by feature vectors, which are at higher levels integrated into a consistent object description. Thus, there are three main design areas which have given rise to a variety of methods:

- which image patches are considered [17],
- how each patch is represented [15], [16], and
- how local descriptors are used to describe the entirety of an object [18], [19].

In spirit, the method proposed in this paper is local feature based with most of our contributions falling within the scope of the last of the aforementioned design issues.

III. Proposed method

Our general approach in recognizing the object in a novel, query image sequence is to compare it to training sequences of all "known" objects in the gallery and assign it to the one with the highest degree of similarity. Since the object of interest has unknown, arbitrary shape and appearance, and may be embedded in significant background clutter, extracting a model of the object's appearance from each sequence in isolation for the purpose of comparing model parameters is difficult without imposing constraints on the class, shape, or appearance of the object or the background (as was done for example by Arandjelović and Zisserman [1] who constrained their attention to sculptures which allowed them to learn and perform super-pixel-level background/foreground segmentation).

Hence in order to avoid the need for overly restrictive assumptions, we take a different approach. We merely assume that the object of interest is roughly in the centre of the video. Then when two sequences are compared with each other (one from the gallery of known objects, the other a query sequence) we seek to find the model parameters which best explain both sequences i.e. that automatically infer the common appearance elements between them. Thus, each comparison, even of sequences which correspond to different objects, produces a hypothesised model of an object. The aim is that the hypothesised model produced when correctly matching sequences are compared is that with the highest likelihood. We now explain how each of the components of our algorithm fits into the overall framework which accomplishes this. In summary, our algorithm comprises the following sequence of steps:

- Motion parallax based frame-wise scale normalization,
- Extraction of low-level spatio-termporal appearance features,
- Model parameter fitting via cross-sequence mutual likelihood maximization, and
- Quasi-volumetric foreground/background video sequence segmentation and model likelihood estimation.

A. Baseline appearance representation

At the bottom-most level, our method is based on describing small image patches i.e. local appearance [20]. This is motivated by observing that if such patches are chosen wisely, they correspond to object parts with consistent geometry and texture, and are thus less sensitive in appearance to variation in viewpoint. For such regions, representations such as the Scale Invariant Features Transform (SIFT) [21] and the related Histograms of Oriented Gradients (HOGs) [22] descriptors have been proposed and demonstrated effective in a variety of applications [23]. Being based on image intensity gradients they also show low sensitivity to illumination changes [24], [25], [26].

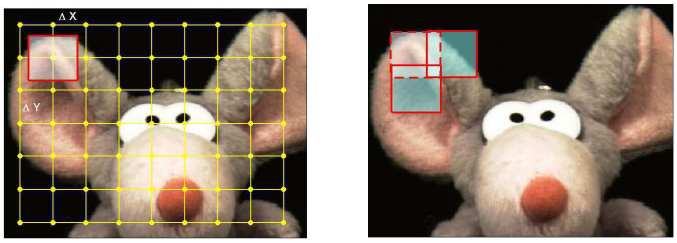
Most local descriptor based methods employ descriptors in a sparse fashion by focusing on a set of detected interest points [16]. When the number of detections is large this can achieve impressive robustness to partial occlusion and image clutter. However, a serious limitation of this approach is that it cannot handle untextured objects [1], [27]. A related problem is that of enforcing geometric constraints between local

descriptors. If no geometric constraints are used (e.g. as in the bag-of-words approach [1], [28]), the representation lacks discriminative power to distinguish between similar objects, especially two objects of the same category, or complex objects with the same basic building element [29]. For example, with this representation, both a bicycle and a metal rail fence may end up looking very similar indeed. On the other hand, devising geometric constraints suitable for general, 3D object is challenging. Lowe's use of the Hough transform effectively restricts the class of objects to nearly-planar ones or, alternatively, restricts camera viewpoint to only very small deviations.

1) Capturing view geometry through redundancy: We tackle the issue of geometry, that is, geometric constraints between different appearance features, using two complementary approaches. The first of these deals with image plane geometry i.e. the relationship between extracted local patches in a single frame. This is realized implicitly, by making the relative shift $(\Delta x, \Delta y)$ between neighbouring patches smaller than their dimensions s_x and s_y, i.e. $s_x > \Delta x$ and $s_y > \Delta y$, resulting in patch overlap. For our experiments we used 90% overlap between neighbouring pathes both in the vertical and the horizontal direction:

$$\Delta x/s_x = \Delta y/s_y = 0.1, \tag{1}$$

as illustrated conceptually in Fig 2.



(a) Dense grid (b) Neighbouring patch overlap

Fig. 2. (a) We describe the appearance of an entire video sequence, and thus both of the object of interest as well as any present clutter, by collecting local image patches collected over a dense grid. (b) Geometric relationship between patches is captured implicitly by making grid spacing smaller than the patch size. The resulting patch overlap means that the same image region contributes to multiple local descriptors.

To see how this approach captures geometric constraints in the image plane, notice that for any two arbitrary patches there is a sequence of patches, each neighbouring the previous one, that connects them. Since neighbouring patches greatly overlap and objects tend to be smooth, the difference in their appearance is small and the aforementioned patch sequence describes a manifold-like structure in the image space (for a similar idea in a different domain, that of temporal topic modelling, see the methods and analyses in [30], [31]). Extending this to the entire set of object image patches collected over our dense grid, it can be seen that this set then describes a 2D surface in the image space. At the same time, the proposed overlap solves the problem of object-grid alignment too. Because our patches are densely packed, while translating the object relative to the grid may change the appearance of any single patch, it leaves the entire set collected over the image unchanged.

a) Representing local appearance: Following previous work, we use the SIFT descriptor to describe each patch in our dense grid and then quantize it by assigning it to the nearest of the k clusters, or descriptor words, estimated by k-means clustering all descriptors extracted from all frames of the training image sequences (we used $k = 500$).

B. Scale normalization

As already mentioned in Sec II most of the existing local appearance based recognition algorithms are sparse in the sense that they focus on a relatively small number of salient, stable loci. These can be detected using one of a number of keypoint detectors [14], [32]. Considering that all modern keypoint detection algorithms explicitly consider the scale of the keypoint, local descriptors are extracted at the corresponding scale thereby achieving scale invariance. Given that in the proposed method local descriptors are collected over a dense grid, the benefit of scale invariance does not come so readily and requires a preceding normalization stage. Our approach is broadly motion parallax based.

We start by computing the optic flow field, using a variant of the well-known Lucas-Kanade algorithm. This field is modelled as comprising a translatory component (recall that our aim is to handle videos acquired in unconstrained conditions using handheld devices) and a rotational component. To correct for the former, we subtract the mean flow vector computed over a frame from the entire field. Since the remaining flow field is generated by a rotational movement of the camera with the object of interest in the centre of the view, motion parallax effected by the depth differential between the object and the background is demonstrated by a discontinuity in the magnitude of the optic flow field at the object edges. By detecting the rough object boundary based on this discontinuity the rough object size within the frame can be estimated and normalized by re-scaling the frame.

C. Descriptor transition tables

In Sec III-A1 we explained how the proposed method implicitly captures image plane geometry, that is, the relationships between different local features extracted from a single video sequence frame. We now explain how 3-dimensional geometry is learnt. The key idea revolves around the descriptor transition table representation which plays the central role in our foreground/background segmentation and the estimation of the likelihood that two sequences (gallery and query) contain the same object.

Consider an image of an object overlaid with a dense grid of overlapping image patches, such as the one previously introduced in Fig 2, and within it a particular patch at the location (x, y) with the corresponding descriptor word w_0. As the viewpoint is changed, the appearance of the patch at (x, y) changes as well, eventually sufficiently so to correspond to an entirely different descriptor word. Depending on the direction of viewpoint variation, the word observed at (x, y) may change from w_0 to any one of the words in the set of all descriptor words $\{w_i\}$.

The above allows to to define what we term the descriptor transition table (DTT) which corresponds to a particular video sequence seen in training. The value in the descriptor transition

table T at row j and column k is the probability that the observed descriptor word w_j makes the transition to the word w_k for a small viewpoint change:

$$T(j,k) = p(w_j \rightarrow w_k). \quad (2)$$

The probabilities of a transition table can be readily seen to capture the relationship between appearances of the same object from different views and thereby, implicitly, its geometry too. Broadly speaking, the spirit of the key idea here is similar to that of e.g. spatio-temporal interest points [33] or 3D LBPs [34].

b) Learning a descriptor transition table: Following our definition of a descriptor transition table, it is tempting to consider changes of descriptor words between successive frames only in the estimation of the aforementioned probabilities $p(w_j \rightarrow w_k)$ which correspond to different table entries. We do not adopt this approach as it is inherently sensitive to the actual ordering of objects views observed in a particular sequence. In other words, different video sequences can contain exactly the same views of an object but ordered differently. Thus we argue that for the purpose of DTT estimation, training video frames should be treated as a set, rather than an ordered sequence.

Our approach to populating a DTT from a training video sequence consists of considering all possible frame successions. We say that two frames I_i and I_j are "possibly successive" if their normalized distance in the image space is less than the threshold t:

$$\|I_i - I_j\| \ / \ \|I_i\| \leq t, \quad (3)$$

as illustrated in Fig 3 (in this work we used $t = 0.1$). Not only does this approach accomplish the desired independence of view sequencing [35] but it also has the advantage of using in the estimation multiple transitions per frame, rather than only a single one actually observed.

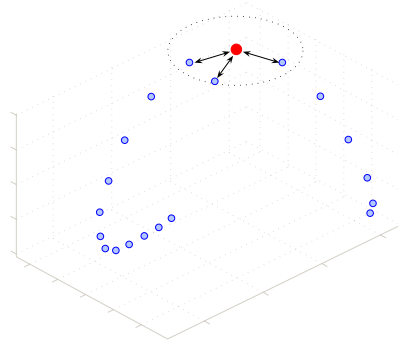


Fig. 3. The proposed learning of the descriptor transition table corresponding to an object in a training video sequence does not rely on the ordering of object views in the video. Instead we consider all descriptor word transitions between all "possibly successive" pairs of frames (after coarse background removal) as determined by their distance in the image space.

c) Applying the DTT model: We now wish to apply the learnt object appearance model in the form of a descriptor transition table, to a novel video sequence of an unknown object. We treat each track of descriptor words through a video sequence as a first order Markov chain, where the probability of observing a word w_i at "time" $n+1$ in the chain is governed by the learnt DTT:

$$p(X_{n+1} = w_i | X_n = w_j) = T(j,i). \quad (4)$$

The track starting in the first frame at the word $X_0 = w_{i(0)}$ is then produced by maximizing the likelihood:

$$p(w_{i(1)}|w_{i(0)}) \ p(w_{i(2)}|w_{i(1)}) \ldots p(w_{i(N)}|w_{i(N-1)}) \quad (5)$$

under the "bound velocity" constraint on patch correspondence:

$$\left\| \begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} - \begin{pmatrix} x_n \\ y_n \end{pmatrix} \right\| \leq d, \quad (6)$$

where (x_n, y_n) is the location of the n-th patch in the chain and $d = \sqrt{2}$ (restricting the possible transition loci to the 3×3 neighbourhood), as illustrated in Fig 4. This maximization is readily achieved using dynamic programming and the well-known Viterbi algorithm.

D. Quasi-volumetric segmentation

At this stage from a query video we have produced a set of descriptor word transitions through the sequence. Let's call one such track of transitions t_i:

$$t_i = \left\{ (w_1^{(i)}, x_1^{(i)}, y_1^{(i)}), \ldots, (w_{N_i}^{(i)}, x_{N_i}^{(i)}, y_{N_i}^{(i)}) \right\}, \quad (7)$$

where $w_j^{(i)}$ is the word at $(x_j^{(i)}, y_j^{(i)})$ that transitions to $w_{j+1}^{(i)}$ at $(x_{j+1}^{(i)}, y_{j+1}^{(i)})$. By construction, meaningful tracks should weave through the object and not trough the background. Thus, we seek to choose optimally a subset of tracks $\mathcal{T}_o$ which explains the object's appearance.

Our approach uses the Graph Cuts algorithm [36]. Motivated by the argument laid out above and in contrast to previous methods, we apply it on the descriptor track level. Unlike in the case when Graph Cuts is used on a single image, the potential of our tracks to diverge, interlace or intercept means that the underlying graph and its structure are not inherent in the basic elements that are being discriminated. Instead we construct it as follows:

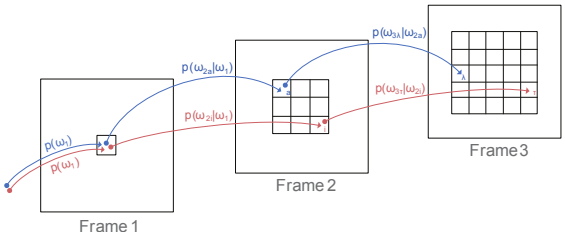
- each track corresponds to a graph node
- the cost of assigning the label "background" to the track t_j is the probability of the corresponding Markov chain in (5)
- nodes corresponding to tracks t_i and t_j are connected iff in any frame the distance between the patches they pass through is less than 2 pixels:

$$\exists k. \ (x_k^{(i)} - x_k^{(j)})^2 + (y_k^{(i)} - y_k^{(j)})^2 < 2^2 \quad (8)$$

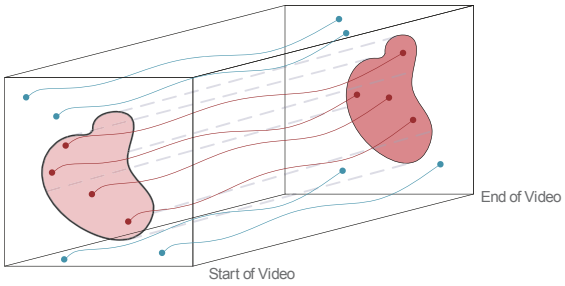
- the cost of assigning different labels to tracks t_i and t_j is:

$$e_{ij} = \sum_k \left[(x_k^{(i)} - x_k^{(j)})^2 + (y_k^{(i)} - y_k^{(j)})^2 \right]^{-1}. \quad (9)$$

Following the application of Graph Cuts, the tracks labelled as foreground define a 2D+time volume which allow the object to be segmented out, as illustrated in Fig 5. Finally, the likelihood of the same object being present in the two compared sequences can be obtained by computing the likelihood in (5)



(a) Descriptor word transitions inference



(b) Segmentation using coherence of transition paths

Fig. 4. Descriptor word transitions: (a) possible transitions are considered in frame-to-frame 9×9 neighbourhood (using the grid illustrated in Fig 2 and defined in Sec III-A1), and (b) word trajectories through the time dimension are used to perform Graph Cuts based quasi-volumetric segmentation of foreground/background in a video; see Fig 5.

using only tracks segmented out as belonging to the foreground (i.e. the best object hypothesis for the comparison).

IV. Evaluation

In this section we turn our attention to the evaluation of the proposed method. We begin by describing the data which was used to train and query different algorithms, continue with a summary of the existing methods we compared our approach with, and finish with a presentation of the results and a discussion.

A. Data sets

To evaluate the performance of the proposed algorithm and compare it to previously proposed approaches, we used two pertinent data sets. These are the publicly available Amsterdam Library of Object Images (ALOI) [37] and a highly challenging data set of video sequences acquired using a mobile phone, collected by ourselves. A comprehensive description of the ALOI can be found in the original publication. In the context of the present work it suffices to summarize it by noting that the data set is very large, comprising sets of images of 1000 objects. The set of images of a particular object corresponds to 72 different viewpoints at uniformly sampled yaw values i.e. at successive 5° rotations about the vertical axis. The ALOI contains a diverse range of objects, some of which are very much alike one another, sharing similar appearances or shapes. Examples are shown in Fig 6. The data was acquired in controlled conditions (uniform viewpoint sampling, uniform background) which allowed us to design a well-controlled evaluation protocol as a means of gaining initial insight into the strengths and weaknesses of different evaluated methods.

Unlike the ALOI, the second set we used for evaluation was acquired in highly uncontrolled conditions. In particular it

(a) Original image input (single frame from a sequence)

(b) Corresponding slice through quasi-volumetric segmentation

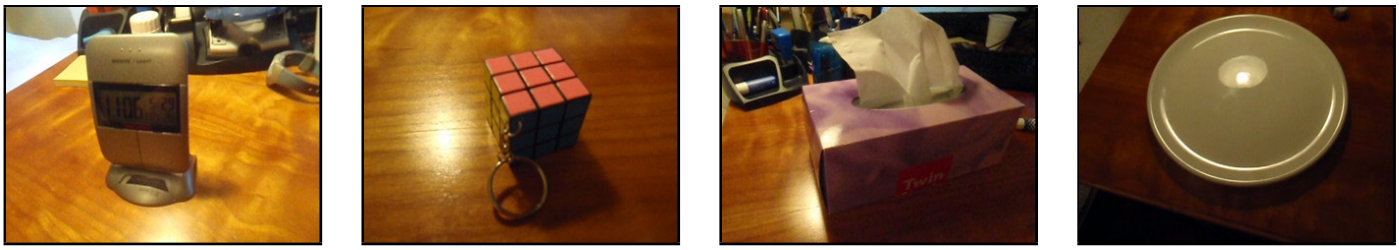
Fig. 5. (a) Typical frame from a raw video sequence, and (b) the same frame with the background removed following the proposed quasi-volumetric foreground/background segmentation of the video.

comprises 100 video sequences acquired using a mobile phone, with 2 sequences for each of the 50 objects. Objects were imaged in a room lit by artificial lighting. The placement of an object in the two sequences was different, with major changes in background clutter, illumination, pose, scale, and camera motion. Some of the challenges were already illustrated in Fig 1, while Fig 7 shows some additional examples of objects in the data set. Notice that some of the objects are untextured and some "wiry" (e.g. respectively the dining plate and the molecular model in Fig 7). In addition to general clutter, also observe the presence of shadows as well as specular reflections. We purposefully included similar objects such as, for example, a stapler and a hole-punch.

B. Baseline methods

We consider several baseline set representations which either demonstrate state-of-the-art performance in comparable recognition tasks or which have been recently described in the literature. These are: (i) sets of SIFT local descriptors, (ii) Gaussian mixture models, (iii) linear subspaces. As usual we fit Gaussian mixtures by employing probabilistic principal component mixtures and minimizing the corresponding model+data description length; following recommendations from prior work [38] for the subspaces based baseline we adopt 6-dimensional subspaces.

We adopt two baseline set similarity measures, again motivated by the reports of their good performance in the existing



(a) Digital clock (b) Rubik's cube (c) Box of tissues (d) Dinner plate

(e) Printer (f) Coffee mug (g) Box of pills (h) Molecular model

Fig. 7. Newly collected database of object video sequences: examples. Shown are representative frames from the corresponding video sequences and a succinct description of the imaged object.

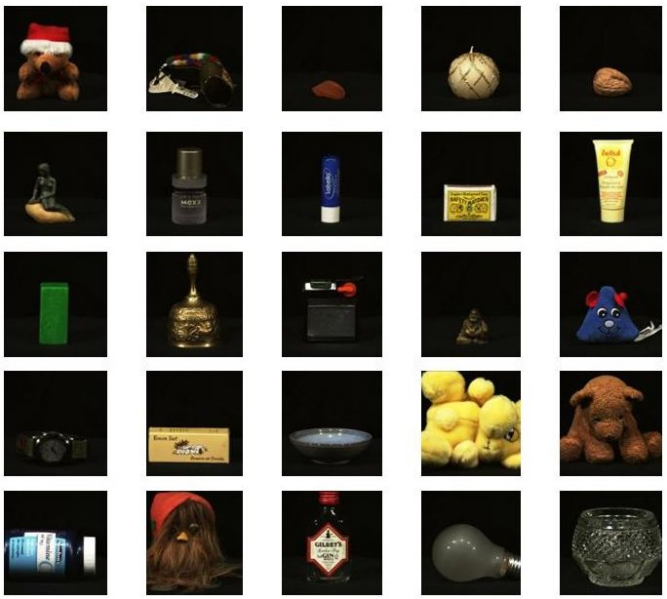


Fig. 6. Examples of objects from the Amsterdam Library of Object Images. The library includes a large number of objects (1000) with varying textural and shape properties, with many objects sharing similar appearance or shape.

literature. The first of these is the Kullback-Leibler divergence [39] applied in the context of the Gaussian mixture based representation and estimated numerically as no analytical solution exists (we shall refer to this method as Appearance+KLD). The second similarity measure we adopted and which we applied in the context of SIFT descriptor sets and linear subspaces is the algebraic method based on the maximum correlation between pairs of vectors lying in two subspaces (we shall refer to this method as SIFT+COS), which is an extension of the *maximum maximorum* ('max-max') cosine similarity between sets of exemplars $\max_{f_1 \in S_1, f_2 \in S_2} f_1^T f_2 / \|f_1\| / \|f_2\|$ [40], [41], [42]. In recent experiments [38] this method was shown to outperform a number of alternatives including by a large margin the pyramid match kernel of Grauman and Darrell [43] and the locality-constrained linear coding (LLC) of Wang *et al.* [44]. Lastly we also apply the aforementioned maximum correlation based distance on raw appearance too (Appearance+COS).

C. Results and discussion

We started our evaluation by experiments on the ALOI, designed to examine how well our algorithm copes with recognition across viewpoint changes. Generalization from a limited viewpoint range to a different, also limited viewpoint range, is a major challenge yet one that is frequently encountered in practice. We adopted the following evaluation protocol:

- for all possible viewpoint angles α (relative to an arbitrary origin of choice), the images in the viewpoint range $(\alpha, \alpha+\Delta\phi)$ of breadth $\Delta\phi$ are used as a training sequence/set,
- for all possible viewpoint range shifts $\Delta\alpha$, the images in the viewpoint range $(\alpha + \Delta\alpha, \alpha + \Delta\alpha + \Delta\phi)$ are used as the query sequence/set,
- the performance at the shift of $\Delta\alpha$ is quantified by the average recognition rate over all test cases.

Because all training and query sequences contain only a limited range of views, this protocol is much more challenging than when views are chosen as random subsets of the original 72 views. For clarity, all results reported in this paper were produced using $\Delta\phi = 40°$ – we found that the results obtained using this value are representative, qualitatively speaking, of general performance trends across different methods examined.

The key results are summarized by the plot in Fig 8 which shows the variation in the rank-1 recognition rate achieved using different methods as a function of the viewpoint change between the training and query sequences. As expected

TABLE I. THE PERFORMANCE OF DIFFERENT METHODS ON OUR NEW DATA SET OF VIDEO SEQUENCES ACQUIRED USING A MOBILE PHONE CAMERA IN THE PRESENCE OF MAJOR CLUTTER, ILLUMINATION, AND VIEWPOINT CHANGES. IN ADDITION TO THE AVERAGE RANK-1 RECOGNITION RATE THE CONFIDENCE OF CORRECT RECOGNITIONS IS QUANTIFIED BY THE RATIO OF THE SIMILARITY BETWEEN THE QUERY AND CORRECT MATCH, AND THE QUERY AND THE SECOND BEST MATCH (APPEARANCE+KLD AND APPEARANCE+COS ALGORITHMS RECOGNIZED NO OBJECT CORRECTLY SO THE CORRESPONDING QUANTITY IS UNDEFINED).

Method	SIFT+COS	Appearance+KLD	Appearance+COS	Proposed method
Rank-1 rate	0.06	0.00	0.00	1.00
Separation	1.22	N / A	N / A	1.92

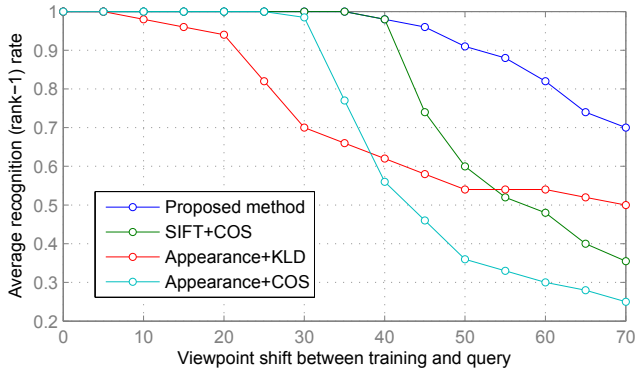


Fig. 8. The average rank-1 recognition rate achieved using different methods across viewpoint changes, using image sets constructed from the Amsterdam Library of Object Images. As expected the performance of all methods deteriorates with the increase in the viewpoint difference between training and query sequences. However the proposed method demonstrates far superior behaviour than all other methods.

both from theory and previous reports in the literature, the performance of all methods deteriorates with an increase in the viewpoint difference between training and query sequences. The most rapid deterioration is observed for the KLD based method which highlights the inherent inability of probability density based methods to generalize – if the training set does not contain representative variability, recognition performance for arbitrary novel input is likely to be poor. For viewpoint changes of moderate extent ($\Delta\alpha < 40°$) generalization is improved with subspace modelling and the use of a more invariant correlation based similarity measure (also consistent with the previous findings in the literature [38]), as witnessed by markedly better performance of the Appearance+COS algorithm. Interestingly this initial improvement is not maintained for large viewpoint changes of over $40°$. Considering the nonlinear distribution of object appearance within the corresponding image space, deterioration for linear subspace based approaches is certainly expected, yet it is unclear why it would be any greater than for the density based KLD algorithm. The use of the SIFT descriptor, which itself has been shown to show good resilience to both illumination and viewpoint, confers further benefit, with the corresponding SIFT+COS algorithm exhibiting even slower deterioration across a wide range of viewpoint changes ($\Delta\alpha < 55°$, and most significantly so for $\Delta\alpha < 40°$). However this method too is outperformed by the simple Appearance+KLD approach for changes of over $55°$. Lastly, the proposed method is readily seen to exhibit vastly superior performance in comparison with all of the other methods and across the entire range of viewpoint changes. Even for the extreme change of $70°$ it attains over 70% correct recognition rate. In comparison, the recognition rate of the Appearance+KLD approach drops to the same level already for a $30°$ viewpoint differential, for Appearance+COS for $37°$, and SIFT+COS for $46°$.

Following the highly promising findings on the ALOI in terms of the superiority of the proposed method, we next sought to evaluate how the algorithms perform on truly realistic video sequences and the newly introduced data set described in Sec IV-A. We used one of the image sequences of an object for training, and the other one (recall, in a different context, with changes in background clutter, viewpoint, camera motion, and illumination) as query. As before we initially examined the rank-1 recognition rate of different algorithms, that is, the rate at which the correct gallery sequence was found to be the best match to the query. The results are summarized in Table I (first data row). It can be immediately seen that the superiority of our algorithm over the evaluated alternatives is demonstrated best in highly challenging conditions such as those present in this data set. Our algorithm correctly identified the query object in all cases, thereby achieving perfect recognition performance. In contrast, the two appearance based approaches (Appearance+KLD and Appearance+COS) recognized none of the objects correctly, with the SIFT based algorithm coping with the challenges somewhat better but still poorly in comparison with the proposed method.

Lastly, to assess the confidence of the successful recognitions, when a successful recognition is observed we examined the ratio of the likelihoods corresponding to the top (i.e. the correct) match and the second best match (which is by implication incorrect). The results can be found in Table I (second data row). Since Appearance+KLD and Appearance+COS methods recognized no object correctly, no measurement could be taken. Comparing the results of the proposed method and that of SIFT+COS approach, we can see that not only did our algorithm exhibit vastly superior performance in terms of rank-1 recognition but also that its correct decisions were made with much greater confidence (over 50% greater class separation).

V. CONCLUSIONS

We described a novel method for object recognition that uses video sequences both as training and query input. The main novelty lies in the framework used to employ discretized local features to describe a video sequence, as well as the manner in which such features are selected in the matching process. One of the key ideas is that of the descriptor transition table which implicitly captures the 3D geometry of an object by considering the transition of a local feature from one descriptor word to another as the camera viewpoint changes. The proposed method was demonstrated as effective an in empirical evaluation on the publicly available Amsterdam Library of Object Images and a new highly challenging data set of video sequences acquired using a mobile phone.

REFERENCES

[1] R. Arandjelović and A. Zisserman, "Smooth object retrieval using a bag of boundaries." *In Proc. IEEE International Conference on Computer Vision*, pp. 375–382, 2011.

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database." *Advances in Neural Information Processing Systems*, pp. 487–495, 2014.

[3] B. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1410–1422, 2013.

[4] Z. Si and S. C. Zhu, "Learning and-or templates for object recognition and detection.." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2189–2205, 2013.

[5] O. Arandjelović, "Discriminative extended canonical correlation analysis for pattern set matching." *Machine Learning*, vol. 94, no. 3, pp. 353–370, 2014.

[6] ——, "A framework for improving the performance of verification algorithms with a low false positive rate requirement and limited training data." *In Proc. IEEE/IAPR International Joint Conference on Biometrics*, 2014, DOI: 10.1109/BTAS.2014.6996275.

[7] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis." *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.

[8] M. Hejrati and D. Ramanan, "Analysis by synthesis: 3D object recognition by object reconstruction." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2449–2456, 2014.

[9] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes." *In Proc. Asian Conference on Computer Vision*, pp. 548–562, 2012.

[10] P. Sauer, T. Cootes, and C. Taylor, "Accurate regression procedures for active appearance models." *In Proc. British Machine Vision Conference*, 2011.

[11] I. Endres, K. J. Shih, J. Jiaa, and D. Hoiem, "Learning collections of part models for object recognition." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 939–946, 2013.

[12] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1385–1392, 2011.

[13] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models." *In Proc. IEEE International Conference on Computer Vision*, pp. 1307–1314, 2011.

[14] O. Arandjelović, "Object matching using boundary descriptors." *In Proc. British Machine Vision Conference*, 2012, DOI: 10.5244/C.26.85.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF." *In Proc. IEEE International Conference on Computer Vision*, pp. 2564–2571, 2011.

[17] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches." *In Proc. European Conference on Computer Vision*, pp. 73–86, 2012.

[18] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns." *Pattern Recognition*, vol. 42, no. 3, pp. 425–436, 2009.

[19] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, 2010.

[20] Y. Biadglgne and O. Arandjelović, "Face filtering – insights from real-world data." *In Proc. International Conference on Systems, Signals and Image Processing*, pp. 65–68, 2015.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2003.

[22] N. Dalai and B. Triggs, "Histograms of oriented gradients for human detection." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[23] O. Arandjelović, "Reading ancient coins: automatically identifying denarii using obverse legend seeded retrieval." *In Proc. European Conference on Computer Vision*, vol. 4, pp. 317–330, 2012.

[24] ——, "Gradient edge map features for frontal face recognition under extreme illumination changes." *In Proc. British Machine Vision Conference*, 2012, DOI: 10.5244/C.26.12.

[25] ——, "Making the most of the self-quotient image in face recognition." *In Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, 2013, DOI: 10.1109/FG.2013.6553708.

[26] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2004.

[27] O. Arandjelović, "Matching objects across the textured–smooth continuum." *In Proc. Australasian Conference on Robotics and Automation*, pp. 354–361, 2012.

[28] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections." *In Proc. IEEE International Conference on Computer Vision*, pp. 370–377, 2005.

[29] O. Arandjelović, "Automatic attribution of ancient Roman imperial coins." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1728–1734, 2010.

[30] A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh, "Hierarchical Dirichlet process for tracking complex topical structure evolution and its application to autism research literature." *In Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, vol. 1, pp. 550–562, 2015.

[31] ——, "Discovering topic structures of a temporally evolving document corpus." 2016.

[32] T. Lindeberg, "Scale selection properties of generalized scale-space interest point detectors." *Journal of Mathematical Imaging and Vision*, vol. 46, no. 2, pp. 177–210, 2013.

[33] C. Yuan, X. Li, W. Hu, H. Ling, and S. Maybank, "3D $\mathcal{R}$ transform on spatio-temporal interest points for action recognition." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 724–730, 2013.

[34] H. Tang, B. Yin, Y. Sun, and Y. Hu, "3D face recognition using local binary patterns." *Signal Processing*, vol. 93, no. 8, pp. 2190–2198, 2013.

[35] O. Arandjelović and R. Cipolla, "Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant." *Pattern Recognition*, vol. 46, no. 1, pp. 9–23, 2013.

[36] S. J. D. Prince, *Computer Vision: Models, Learning, and Inference.*, 1st ed. Cambridge University Press, 2012.

[37] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images." *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[38] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 529–534, 2011.

[39] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection." *In Proc. IEEE International Conference on Computer Vision*, pp. 2214–2219, 2011.

[40] O. Arandjelović and R. Cipolla, "Face set classification using maximally probable mutual modes." *In Proc. IAPR International Conference on Pattern Recognition*, pp. 511–514, 2006.

[41] O. Arandjelović, "Unfolding a face: from singular to manifold." *In Proc. Asian Conference on Computer Vision*, vol. 3, pp. 203–213, 2009.

[42] ——, "Learnt quasi-transitive similarity for retrieval from large collections of faces." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[43] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features." *In Proc. IEEE International Conference on Computer Vision*, vol. 2, pp. 1458–1465, 2005.

[44] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification." *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, 2010.